We realise a paper does not use an index, however it could be nice to have an overview

1 Introduction 4
Introduction to the topic 4
Relate to current knowledge 4
Introduce our work, give purpose and the main objective 4
2 Methodology 5
2.1 Data collection & cleaning 5
2.1.1 Data collection 5
2.1.2 Data cleaning 6
2.2 Featured models 7
2.2.1 MVLR 7
2.2.2 SVR 7
2.2.3 MLP 7
2.2.4 LSTM 7
2.3 Evaluation Metrics 9
2.4 Forecasting methods 10
2.4.1 Energy Production 11
2.4.2 Energy Consumption 12
3 Results ..... **Error! Bookmark not defined.**
Production ..... **Error! Bookmark not defined.**
Main discoveries ..... **Error! Bookmark not defined.**
Consumption ..... **Error! Bookmark not defined.**
Main discoveries ..... **Error! Bookmark not defined.**
Summary comparison ..... **Error! Bookmark not defined.**
4 Conclusions and discussion 16
Answers 16
Explanation 16
Further implications 16
Factory Zero (decision not modelled) 16
5 Recommendations 17
5.1 The booster and consistent intervals ..... **Error! Bookmark not defined.**
5.2 More hours between predictions ..... **Error! Bookmark not defined.**
5.3 Identifying peaks in the data ..... **Error! Bookmark not defined.**
5.4 Get more data ..... **Error! Bookmark not defined.**

6 References 18

Χομμεντεδ [ϑB1]: 3 Results, we moeten het subkopje niet daadwerkelijk main discoveries noemen

1 Introduction [feedback of mr Rahola not yet processed]

1.1 M1

The European Commission has set the ambitious goal of reducing the impact of global warming with the Paris agreement (United Nations Climate Change). The Netherlands has released their NDC in November 2019, in which the plans for the coming 30 years are presented. The Netherlands has set a goal of reducing the CO2 emission, relative to 1990, with 49% by 2030 and with 80-95% by 2050 (Rijksoverheid). Electricity makes up around 26% of the energy-wise CO2 emissions. In 2015 the portion of renewable energy of the electricity consumption was circa 12%. As a result of the reduction of CO2 emissions this percentage will increase to circa 41% by 2023. One of the ways to achieve that is to create more net zero emission buildings (NZEB). These buildings are isolated better and have as main power source one that is low on CO_2 emissions. They are however still connected to the main grid to sell and buy electricity in case the supply is more or less than required. At this point in time NZEBs sell their surplus of energy to the net at a profit. In the near future this will not be as straightforward. As the energy demand is increasing, the load of the grid is as well. Meaning that the electricity providers will deal out fines for overloading the net by selling electricity at points of high electricity traffic. This overloading of the electricity network can be prevented by developing a way in which these NZEBs can store their energy efficiently, will it be through batteries or for example the hot water reservoir. In addition to adding a storage system there has to be developed a way in which the amount of energy consumed and produced can accurately be predicted of the NZEBs. According to these predictions electricity providers can decide whether to raise or lower the prices and through that stabilize the electricity output of the NZEB households.

Χομμεντεδ [ϑB2]: Gekke zin

Χομμεντεδ [ΔΧϖ(3P2]: Vergad het woordje "is" zo beter?

Χομμεντεδ [ϑB4P2]: And have a s main power source klinkt niet alsof het oké engels is

Χομμεντεδ [ϑB5]: 'A way has to be developed' ofzo

Χομμεντεδ [ΔΧϖ(6P5]: Die snap ik niet

Χομμεντεδ [ϑB7P5]: In addition to adding this storage system, a way in which.. has to be developed
Dat zou ik van die zin maken

Χομμεντεδ [ϑB8]: Welke predictions?

Χομμεντεδ [ΔΧϖ(9P8]: De vorige zin maakt dat toch wel duidleijk?

Χομμεντεδ [ϑB10]: Provider toch?

Former research indicates that such predictions of energy production and/or consumption can very well be done through the use of machine learning and neural network prediction models. Studies such as that from (Liu, et al., 2019) state that the machine learning approach is promising but the difficulty lies in determining what model to apply in order to gain the best result. Several studies, among which (Jana, Ghosh, & Sanyal), (Kim & Cho, 2019), (Khovalyg & Heidari, 2020), (al A. H., 2020), have shown LSTM to hold great promise in the effort of predicting energy consumption of residential facilities. The complication to neural network models such as LSTM is however the determination of the hyperparameters in order to gain optimal performance. In order to gain insight into the capabilities of both machine learning as well as neural network models when it comes to the prediction of energy production and consumption of a NZEB, this research conducts the performances of 2 machine learning models and 2 neural network models for both the prediction of energy production and consumption. To be more specific, the performances of SVR (support vector regression), MVLR (multi variate linear regression), MLP (multi-layer perceptron) and LSTM (long short term memory). The data used in this research is provided by the company Factory Zero (over ons, sd), a company that revolves around the vision of making NZEB housing affordable for everyone. The company offered the data of 120 identical NZEB households located in Zoetermeer (Zuid Holland, the Netherlands) during the year 2019. The goal of this research is answering the question:

Χομμεντεδ [ϑB11]: Waar is de referentie?

Χομμεντεδ [ΔΧϖ(12P11]: Dat wordt verderop aangegeven, dit is meer een introductie zin

Χομμεντεδ [ϑB13P11]: Hmm ja moet wel kunnen dan denk ik! Maar ik laat deze comment staan omdat ik het er niet 100% mee eens ben

'What is a suitable machine learning or neural networks model to predict the energy production and consumption of a NZEB 1 day in advance on hourly resolution?'

1.2 M2

We did research about predicting the energy consumption of commercial buildings. However we concluded that individual housing has not gotten the same amount of research put into it yet. This could be caused by the complications resulting from the

Χομμεντεδ [ϑB14]: Geen We

privacy sensitivity of gaining data from residential houses [4]. Recently more research is being conducted on the prediction of residential housing energy consumption and production [5] [6]. They do this because

What is a suitable model to predict energy consumption and production of an nZEB, one day in advance with hourly resolution?

1.3 M3

Multiple machine learning models (SVR, MVLR, MLP & LSTM) have been tested to evaluate their performance in forecasting the energy production and consumption of nZEBs. These models have been using data from 120 nZEB row-houses located in South-Holland. Data will be explained in a new chapter called Data.

Forecasting energy use and energy production of NZEB row houses

*Jefry Yehia el Bhwash[1], Levy Duivenvoorden[1], Cornelis Gerrit van Drunen[1], Amin Mansouri[1], Niels van Schaik[1] and Tadeo Baldiri Salcedo Rahola[2]**

[1] Faculty of Technology, Innovation and Society, The Hague University of Applied Sciences, Rotterdamseweg 137, 2628 CN,Delft, The Netherlands
[2] Escuela Politécnica Superior, University Francisco de Vitoria, Carretera Pozuelo a Majadahonda,Pozuelo de Alarcón, Madrid, Spain

Χομμεντεδ [Β(15]: verkeerde vestiging

Χομμεντεδ [Β(16]: misschien onze universiteit

Abstract - eeeeeeeeeeeeeeee

Eeeeeeeeee

Eeeeeeeeeeeeee

Eeeeeeeeeee

Χομμεντεδ [ΔΛ(17]: Abstract moet nog worden geschreven.

1 Introduction

Introduction from the previous page will be put here after reviewing.

Introduction to the topic

The European Commission has set the ambitious goal of reducing the impact of global warming with the Paris agreement. The Netherlands has released their NDC in November 2019 in which the plans for the coming 30 years are presented. The Netherlands has set a goal of reducing the CO2 emission, relative to 1990, with 49% by 2030 and with 80-95% by 2050 [3]. Electricity makes up around 26% of the energy-wise CO2 emissions. In 2015 the portion of renewable energy of the electricity consumption was circa 12%. As a result of the reduction of CO2 emissions this percentage will increase to circa 41% by 2023. One of the ways to achieve that is to create more Nearly Zero-Energy Buildings (nZEBs). These buildings are isolated better and use a low in CO2 energy source as their main power source. They are still connected to the main grid to sell and buy energy in case the supply is more or less than required. At the this point in time nZEBs sell their surplus of energy to the net at a profit. In the near future this will not be as straightforward. As the energy demand is increasing, the load of the grid is as well. Meaning that the energy providers will deal out fines for overloading the net by selling energy at points of high energy traffic[1]. To be able to keep nZEBs profitable they need to make smart decision about it. To achieve this, the prediction of energy consumption and energy production is essential.

Relate to current knowledge

S

Introduce our work, give purpose and the main objective

S

Χομμεντεδ [ΔΛ(18]: (Amin:)
Als het commentaar van Baldiri is verwerkt kan je deze hier gelijk verwerken.

2 Methodology

2.1 Data collection & cleaning

2.1.1 Data collection

The data used consists of the historical data from the Smart Meter and Solar Panel Inverter of 120 Net-zero energy buildings (NZEBs) located in Zoetermeer (Zuid-Holland, Netherlands). The data from one such NZEB is transmitted to data servers via the communication module of the iCEMs (integrated climate energy module). These iCEMs are used for energy storing and heating up the houses. The component that are looked futher into is the inverter and the smartmeter. The data consisting of several worksheets containing all the meter data collected from the NZEB over the year 2019 with a resolution of 5 min.

Besides the received data, weather data is also used. The weather data is used for forecasting. This data was received from nearby weather station located in Voorschoten (within 15km radius from Zoetermeer) gathered by the royal Dutch meteorological institute; KNMI. From this data the feature Global Sun Irradiance was extracted, which defines the intensity of sunlight [J/cm^2] on 1 hour resolution. This feature is used I a 24hr shifted version for the energy production models since in a real-time setting this data is only available the next day. In order to obtain the energy consumption of the NZEB the following calculation was made.

$$Energy\ consumption = smart_{in} + solar_{out} - smart_{out} \tag{1}$$

This calculation was needed since the smart meter that registers the amount of electricity that enters the NZEB also registers the electricity produced by the solar panels. The exact cause of this phenomenon has not been found yet, but due to the insignificant amount of data it takes up this negative data was flattened to 0 in order for the models to not get confused.

Besides the data received from FZ a form of weather data was also acquired, since more sunlight means more solar energy, thus making data collected by weather stations or possible weather forecasts most valuable in this effort. It was possible to receive data from a nearby weather station located in Voorschoten (within 15km radius from Zoetermeer) gathered by the royal Dutch meteorological institute; KNMI. From this data the feature Global Irradiance was extracted, which defines the intensity of sunlight [J/cm^2] on 1 hour resolution. This feature is used in a 24hr shifted version for the energy production models since in a real-time setting this data is only available the next day. In order to obtain the energy consumption of the NZEB the following calculation was made:

This calculation was needed since the smart meter that registers the amount of electricity that enters the NZEB also registers the electricity produced by the solar panels. Although the equation would theoretically not permit a negative outcome, a negative energy consumption was still found in >1% of the data. The exact cause of this phenomenon has not been found yet, but due to the insignificant amount of data it takes up this negative data was flattened to 0 in order for the models to not get confused.

Χομμεντεδ [ϑΒ19]: Ik dacht dat dit voor Nearly Zero Energy Building stond

Χομμεντεδ [ΔΛ(20Ρ19]: De benaming klopt hier.
In het nederlands heet het nul-op-de-meter huizen.

Χομμεντεδ [ϑΒ21]: Dit is in de intro al volledig uitgeschreven

Χομμεντεδ [ΔΛ(22]: Bron

Χομμεντεδ [ϑΒ23]: /cm2 is wat wij gebruiken

Χομμεντεδ [ΔΛ(24]: Wat heeft dit met de naam van de paragraaf “Data Collection” te maken?
Eerder heeft het te maken met data preprocessing.

Χομμεντεδ [ΔΛ(25]: Schrijf in de tegenwoordige tijd.

Χομμεντεδ [ϑΒ26]: Ik vind deze zin heel raar lopen, maar ik ben net wakker

Χομμεντεδ [ΔΛ(27]: Je legt hier aan de lezer uit waarom je iets doet.
Dit hoort niet, je moet alleen schrijven hoe wij het hebben gedaan.
Anders komt het over alsof we er niet zeker van zijn.

Χομμεντεδ [ΔΛ(28]: Bron

Χομμεντεδ [ΔΛ(29]: Dit kan korter, van de KNMI is de zonirradiantie verkregen.

Χομμεντεδ [ϑΒ30]: /cm2 is wat wij gebruiken

Χομμεντεδ [ΔΛ(31]: Wellicht een plotje/schets invoegen om het duidelijker te maken.

Χομμεντεδ [ϑΒ32]: Theoretisch gezien kan het wel, maar praktisch gezien niet, want praktisch is het onmogelijk voor de smart meter om meer te verkopen dan dat er geproduceerd wordt. Wij weten dit maar we geven dit nergens aan

Χομμεντεδ [ΔΛ(33]: Dit mag niet in de methodologie, dit hoort in de discussie.
Het is iets wat is opgevallen tijdens het onderzoek wat niet hoort te gebeuren.

Χομμεντεδ [ΔΛ(34]: Wat heeft dit met de naam van de paragraaf “Data Collection” te maken?
Eerder heeft het te maken met data preprocessing.

2.1.2 Data cleaning

Due to Data Freezes and possible internet outages (a period of time where the sensors did not indicate its findings) that were causing large spikes in the data, it was not consistent enough to use at the minimal resolution of 5 min. Instead the data was resampled to 1 hour, since the price is also calculated at 1 hour. However some NZEB datasets still contained large timestamp differences which made it unusable to make predictions on. To indicate which NZEB datasets were best to make a heatmap. After looking in to the data, a heatmap is made. In the heatmap is shown that 10% of the NZEBs were found to be the optimal choice, and used to learn models on. For both the prediction of energy production and consumption.

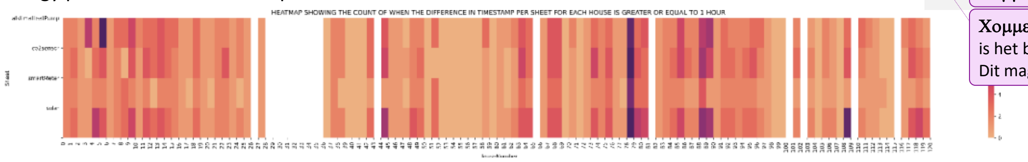

Figure 1: Heatmap indicating the amount of timestamp differences greater or equal to 1 hour per NZEB. The white spaces indicate NZEBs that were not usable due to extreme amounts of inconsistencies (1000+).

Further examination of the required shape of the input data for a time series based machine learning model proved that the timestamp indicating the data points was not optimal to use. Therefore these timestamps were converted to Date-Time objects (a variable used to indicate a date and time of a datapoint). ~~Another inconsistency that occurred was at the calculation of the energy consumption of a NZEB . Although the equation would theoretically not permit a negative outcome, a negative energy consumption was still found in >1% of the data. The exact cause of this phenomenon has not been found yet, but due to the insignificant amount of data it takes up this negative data was flattened to 0.~~

Due to Data Freezes and possible internet outages (a period of time where the sensors did not indicate its findings) that were causing large spikes in the data, it was not consistent enough to use at the minimal resolution of 5 min. Instead the data was resampled to 1 hour, since this is the maximum interval FZ would prefer. However some NZEB datasets still contained large timestamp differences which made it unusable to make predictions on. To indicate which NZEB datasets were best to use the heatmap shown below was made.

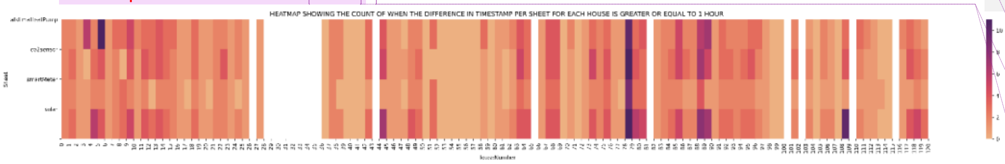

Figure 2: Heatmap indicating the amount of timestamp differences greater or equal to 1 hour per NZEB. The white spaces indicate NZEBs that were not usable due to extreme amounts of inconsistencies (1000+).

For both the prediction of energy production & consumption 10% of the NZEBs were found to be the optimal choice, and used to learn models on. Further examination of the required shape of the input data for a time series based machine learning model proved that the timestamp indicating the data points was not optimal to use. Therefore these timestamps were converted to Date-Time objects (a

Χομμεντεδ [ΝϖΣ35]: To do:
- heatpump booster peaks hebben we weggehaald en staat er nog niet tussen in dit stukje
- Define booster

Χομμεντεδ [ΔΛ(36]: Wellicht is de naam Data preprocessing toepasselijker.

Χομμεντεδ [ΔΛ(37]: Je mag alleen zeggen dat dit gevonden is als er een bron of een resultaat is die het onderbouwd.

Χομμεντεδ [ΔΛ(38]: .

Χομμεντεδ [ΔΛ(39]: Waar komt dit nummer vandaan? Hoe is het berekend?
Dit mag je niet neerzetten.

Χομμεντεδ [ΔΛ(40]: Dit moet in de discussie, zoals eerder gezegt.

Χομμεντεδ [ϑΒ41]: Dit is net ook gezegd

Χομμεντεδ [ΔΛ(42]: Ik zou hier een tweede paragraaf maken:
1. Kiezen van de beste huizen.
2. Data inconsistenties in de data.

Χομμεντεδ [ΔΛ(43]: Het komt over alsof we er niet zeker van zijn wat we hebben gedaan.

Χομμεντεδ [ΔΛ(44]: Je mag alleen zeggen dat dit gevonden is als er een bron of een resultaat is die het onderbouwd.

Χομμεντεδ [ϑΒ45]: Ik denk dat dit een minimum resolutie is/ laagste resolutie

Χομμεντεδ [ϑΒ46]: Dit willen ze omdat de resolutie van de energie prijs ook in uren is

Χομμεντεδ [ΔΛ(47]: Dit hoort in het abstract, waarin je onderzoeksvraag is verwerkt.

Χομμεντεδ [ΔΛ(48]: In de heatmap hieronder moeten alleen de datasets worden weergegeven die ook echt zijn gebruikt.

Χομμεντεδ [ΔΛ(49]: Dit loopt niet lekker.

Χομμεντεδ [ΔΛ(50]: Geen & teken.
Gebruik and.

Χομμεντεδ [ΔΛ(51]: Waar komt dit nummer vandaan? Hoe is het berekend?
Dit mag je niet neerzetten.

variable used to indicate a date and time of a datapoint). Another inconsistency that occurred was at the calculation of the energy consumption of a NZEB . Although the equation would theoretically not permit a negative outcome, a negative energy consumption was still found in >1% of the data. The exact cause of this phenomenon has not been found yet, but due to the insignificant amount of data it takes up this negative data was flattened to 0.

Welke huizen zijn er gekozen uit het plotje hierboven?

Een correlatiematrix van de overwogen features?

2.2 Featured models

2.2.1 MVLR

The multivariate linear regression (MVLR) is the most simple machine learning model. This model looks at the linear relations between multiple input features. Then when all the coefficients are known a prediction can be made.

Because of the relation between the input variables and the output, the output is in an higher dimension. So the output is not an (linear) line.

2.2.2 SVR

According to the literature the SVR, Support Vector Regression, is a simple machine learning model with good results. The SVR is based on the SVM (Support Vector Machine) that is mostly used for classification.

The SVR looks in the given data at hyperplanes to predict the next value. The SVR is insensitive to outliers due to its nature. It tries to find an clear distinction between the input and the desired output. Due to the insensitivity to outliers the SVR chooses its own dataset. So the SVR might find the general pattern, but high variance in the data is a problem for it.

The SVR can be tuned to make more distinction between the outliers and output. The parameter for this is C, the most common value for this is 1.

2.2.3 MLP

A multilayer perceptron (MLP) is the most basic form of an neural network. The basic concept is as follows, an input layer consist of all the input features for the model. Then the hidden nodes will process this data in such an manner that the output will be the target. This process is repeated many times, every time is called an epoch. After each epoch the backpropagation will be calculated, afterwards this backpropagation is used to update the hidden nodes.

During the literature study it was noted that an MLP of 2 hidden layers works optimally. An MSE-loss (mean squared error) was used as the metric to optimize for.

2.2.4 LSTM

A long short term memory (LSTM) is an more complex neural network. The LSTM is an form of recurrent neural network (RNN). This means that is learns, mostly, from sequences of data. The LSTM combines the RNN with an kind of memory called the cell- and hidden state. Thus this is an self-learning and remembering neural network.

Χομμεντεδ [ΔΛ(52]: Dit moet in de discussie, zoals eerder gezegt.

Χομμεντεδ [ϑB53]: Dit is net ook gezegd

Χομμεντεδ [ΔΛ(54]: Ik zou hier een tweede paragraaf maken:
1. Kiezen van de beste huizen.
2. Data inconsistenties in de data.

Χομμεντεδ [ΔΛ(55]: Het komt over alsof we er niet zeker van zijn wat we hebben gedaan.

Χομμεντεδ [ΔΛ(56]: ?

Χομμεντεδ [ϑB57]: [28, 37, 40, 42, 51, 56, 58, 70, 99, 100, 105, 115]

Χομμεντεδ [ΔΛ(58]: ?

Χομμεντεδ [ϑB59]: Hier gaan we geen vrolijke resultaten uit halen denk ik, kan me vaag herinneren dat consumptie niet heel lekker ging met de correlatie matrix

Χομμεντεδ [ΔΛ(60P59]: Dat kan, zorg er dan wel voor dat het duidelijk is waarom we deze features hebben gekozen. Ik weet dat er een paper was die een correlatiematrix had geprobeert en daarna eervoor koos om (ongeveer) dezelfde features als ons te gebruiken.
(Je kan me appen voor de bron als je hem niet kan vinden)

Χομμεντεδ [ϑB61]: De wiskundige formules toevoegen & literatuur over de modellen, want nu leggen wij iets uit over modellen terwijl we gewoon een paar studenten zijn

Χομμεντεδ [ϑB62]: Je kan dit denk ik niet zeggen zonder bron

Χομμεντεδ [ϑB63]: Zou opnieuw aangeven welke paper dit zegt

The input data is restructured in such an way that it is compatible with the LSTM (a three dimensional tensor). Then the data is fed into the LSTM and the output is of size 100. The model tries to predict only one value, so an linear layer after the LSTM does this.

Every epoch the LSTM is being trained and after each epoch the backpropagation was done and the model was updated according to the parameters.

Χομμεντεδ [ΔΛ(64]: Plaatjes en formules?

2.3 Evaluation Metrics

In order to evaluate the performance of the models and be able to compare them with each other there has to be selected a range of loss functions which are well-applicable to each of the models. Below is described the 4 loss functions used to evaluate each of the models performance.

Χομμεντεδ [9B65]: Meestal schrijf je getallen uit en deze zin loopt niet lekker

The R^2 (pronounce R-squared) error represents the linearity between the predicted value (typically called y_{hat}) and the actual value (y). Usually the value that comes out of the equation is between zero and one. A score of one is the best score possible.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2}$$

The root mean squared error (RMSE) is the error that shows the root of the mean squared error. The RMSE is more sensitive to peaks or outliers in the data due to the root. The lower the RMSE value, the better.

$$\text{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

The mean absolute percentage error (MAPE) is the mean absolute error represented as an percentage. Due to the rather small numbers in the data the outcome can become rather large. The lower the percentage the better.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i}$$

The mean absolute error (MAE) is the most basic form of errors. This measures the average between the absolute difference between the target and the model output. The lower the value the better.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

The MAE is the most relatable to the data, due to the same dimensions of input and error.

Χομμεντεδ [ΔΛ(66]: Dit is de loss function (de loss die het NN en de LSTM proberen te minimaliseren).
Wat doet dit in evaluation metrics?

2.4 Forecasting methods

To predict energy consumption of a NZEB requires a completely different approach as to predicting energy production of one such house, due to the distinct difference in data. Energy production is more periodic compared to the more hectic data of energy consumption. The initial approach was to create two models, one for energy production and one for energy consumption. First we will discuss energy production.

2.4a Data scaling

All of the considered models will work better if the input data has been scaled. The scaled used for this is the standardscaler. This is due of the Gaussian like distribution of the data. The formula for this scaler is:

$$x_{scaled}(i) = \frac{x(i) - m}{u} \qquad [?]$$

In the formula ? the x(i) is the i'th datapoint, m is the mean value over the entire dataset, u is the standarddeviation and $x_{scaled}(i)$ is the scaled i'th datapoint.

After applying this formula almost all of the input data falls between -1 and 1.

2.4b Loss function

The LSTM and the MLP need an function to optimize for so it can evaluate and optimize how well it performs. The Huber loss is used for this. This loss function uses two different functions. One for the points with high variance and one for the points with an low variance. Both formulas are shown below:

$$loss(x,y) = \frac{1}{n}\sum_i z_i \qquad [?a]$$

$$z_i = \begin{cases} 0{,}5(x_i - y_i)^2 & if\ |x_i - y_i| < 1 \\ |x_i - y_i| - 0{,}5 & otherwise \end{cases} \qquad [?b]$$

Where $loss(x,y)$ is the loss for the output, x_i, from the model, y_i the target is and n the amount of datapoints in the batch.

Commented [ΔΛ(67]: Dit is ook een mening, je kan wel zeggen dat energie consumptie een lagere periodiciteit heeft.

Commented [9B68]: ...this, the initial...

Commented [ΔΛ(69]: 1 tot 4 schrijf je volledig uit in een paper.

Commented [ΔΛ(70]: Nog even een juiste plaats hiervoor vinden.

Commented [ΔΛ(71]: Nog even een juiste plaats hiervoor vinden.

2.4.1a Energy Production

When it comes to the prediction of the hourly amount of solar energy produced of the next day by the NZEB, each of the models described earlier were implemented separately. Generally machine learning models, such as SVR, MVLR and MLP, cannot look back into their timeseries dataset like the LSTM can by remembering. Therefore these need a different way to look back into their data for comparable results. Therefore two datasets were created, one for each model type, displayed in Table 1. The main difference being that the one for SVR, MVLR and MLP contains the energy production of 24 hours ago till 168 hours ago. representing a comparable way in which the window of the LSTM looks over the data.

Χομμεντεδ [ΔΛ(72]: 1 tot 4 volledig uitschrijven.

Χομμεντεδ [ΔΛ(73]: uitschrijven

Table 1 Contents and specifications of the used datasets.

DATASETS		
	SVR, MVLR and MLP:	**LSTM:**
FEATURES	• History Energy production 24 hours till 168 hours in the past • The mean energy production of the previous day and the previous week • Hour of the prediction, one-hot encoded • Global irradiance of 24 hours in the past	• History Energy production 24 hours till 168 hours in the past • Hour of the prediction • Global irradiance of 24 hours in the past
SPLITTING METHOD	Train: 2019-01 : 2019-08 Validate: 2019-09 : 2019-11 Test: 2019-12	

What distinguishes energy production greatly from energy consumption is that energy production at first glance looks significantly more periodic (Figure 1) than energy consumption, thus raising the suspicion that simple machine learning models such as MVLR or SVR might already perform well. However the possibility that more complicated neural networks such as MLP or LSTM might be able to compete still remains, therefore two of such models were created and will be discussed next.

Χομμεντεδ [9Β74]: Dit is al eerder gezegd

Χομμεντεδ [ΔΛ(75]: Het lijkt hier net alsof we niet zeker zijn of ons onderzoek resultaten gaat opleveren. 😉

Χομμεντεδ [9Β76]: two

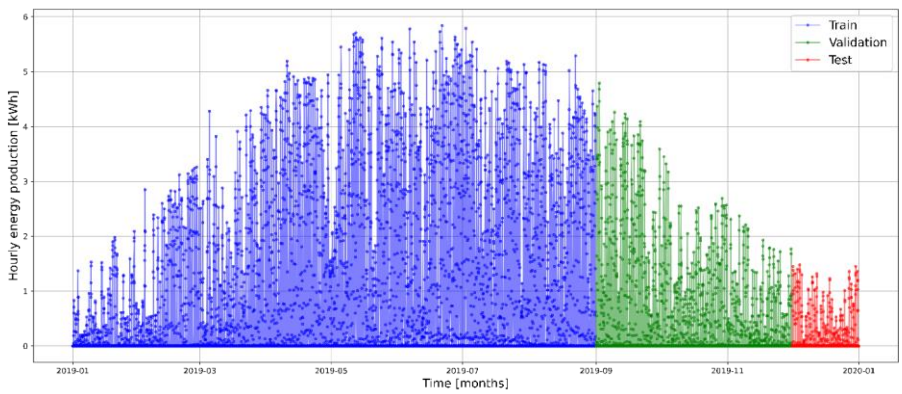

Figure 1: The solar energy production over the course of a whole year (2019) from one of the NZEBs split up in 3 parts; blue: training data (8 months), green: validation data (3 months), red: testing data (1 month).

2.4.2b Energy Consumption

Compared to energy production, predicting the hourly amount of energy consumed by the NZEB requires a more complicated design. The same models as Energy production were used; MLP, SVR, MVLR and LSTM.

The dataset being used is almost the same as production, except for a few slight changes. In Table 2 the dataset is shown.

Table 2 Contents and specifications of the used datasets.

DATASETS		
	SVR, MVLR and MLP:	**LSTM:**
FEATURES	- History Energy consumption 24 hours till 168 hours in the past - The mean energy production of the previous day and the previous week - Hour and weekday of the prediction, one-hot encoded	- History Energy consumption 24 hours till 168 hours in the past - Hour of the prediction - Hour and weekday of the prediction
SPLITTING METHOD	Train: 2019-01 : 2019-08 Validate: 2019-09 : 2019-11 Test: 2019-12	

The consumption data is less periodic than the production data. This makes the problem more complex.

Plot of consumption data.

Χομμεντεδ [ΔΛ(77]: Graag een plotje.

2.4.Models

LSTM

In order for the LSTM to be able to improve itself it uses the Adam optimizer. To grant the LSTM a greater capability of finding the true minimum within the loss landscape, a learning rate scheduler was added to the model, this varies the learning rate of the Adam optimizer in a sinusoidal fashion. In addition, the scheduler was reset after 10 epochs during the process of running the model. This way the scheduler will be able to complete multiple cycles of increasing and decreasing the learning rate whilst training the model.

Χομμεντεδ [ΔΛ(78]: Dit is vakjargon.

A LSTM is an form of RNN, this means that the data has to be recurrent. Therefore the LSTM is trained with an transfer window. The window moves over the dataset in which it takes the 7 hours prior to the prediction. These will be fed into the LSTM.

An dataloader was used to get the best performance. The theory is that the LSTM is able to more easily pick up daily patterns due to the smaller batches. The size of these batches was chosen to be 64, this means that the LSTM trains in groups of 64 hours.

Twelve houses were used for the LSTM. It was chosen to use the same model parameters for every house. Before training the LSTM is initialized with random weights. After which the LSTM trains for one hundred epochs on the training dataset, the most optimal for house number 28.

After training, the LSTM is validated on the validation and testing dataset.

MLP

The multilayer perceptron is an feed forward network. In other words, it tries to predict the target by considering the input variables, so it only forwards the output.

The optimal parameters was a layersize (in the hidden layer) of 128 nodes. When using two (linear-) layers and an ReLU activation function between the two layers.

After considering the amount of epochs to train for the optimal amount is one hundred epochs. There is however no learning rate scheduler.

The optimal parameters were decided on one house. Where the same parameters were used on the other 11 houses. The network was assigned random weights at the beginning of training of each house.

3 Results

The results were determined by running the 4 models on 12 randomly chosen houses from a list of houses that have the lowest number of one-hour gaps. After each model was run on these houses the average of the metrics MAE, MSE, MAPE and R^2 was calculated and shown the corresponding table.

Χομμεντεδ [ΔΛ(79]: Waarom staan de resultaten hier?

Consumption

The results in **Error! Reference source not found.** show that the LSTM overall outperforms the other models at predicting consumption. The same table concludes that SVR performs the worst overall.

Table 3 Consumption Model Results

Model	MAE	MSE	MAPE	R2
LSTM	0.44	0.43	69.54	-3.19
MLP	0.57	0.65	99.78	-1.77
MVLR	0.41	0.34	93.29	-4.09
SVR	0.65	0.90	398.45	-5.08

In **Error! Reference source not found.** the four models for consumption are shown in combination with the ground truth. The LSTM again is closest to predicting the timing and height of the peaks, as well as the mean of the data.

The MLP seems to perform the worst out of these models when it was the second worst according to **Error! Reference source not found.**.

Χομμεντεδ [ΝϖΣ80]: Maak hier 4 losse afbeeldingen van. Nus is de tekst onleesbaar en met 4 losse afbeeldingen kan je nog de scale veranderen zonder mega veel ruimte in te nemen.

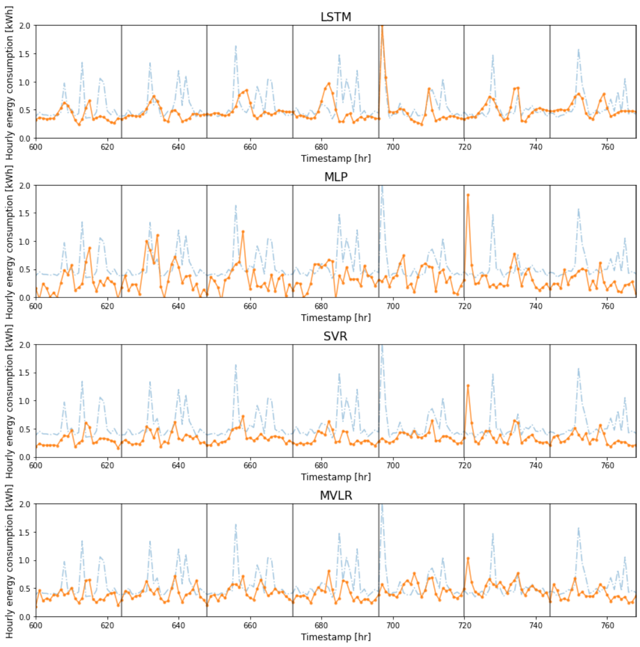

Figure 1 Consumption Model Graphs

Production

The results in **Error! Reference source not found.** show that the SVR is performing the best overall at predicting the production. The MLP performs the worst overall.

Table 4 Production Model Results

Model	MAE	MSE	MAPE	R2
LSTM	0.17	0.07	1105848556.51	0.15
NN	0.50	0.37	21149948633.33	-1.78
MVLR	0.19	0.08	59336116708.33	0.53
SVR	0.14	0.05	42122716831.25	0.50

In **Error! Reference source not found.** the four models for production are shown in combination with the ground truth. The SVR shows the lowest percentage of negative production predictions which results in the SVR having better predictions at night than the other models. The LSTM seems to perform better at predicting the shape of the production during the day, however the LSTM predicts negative production during the night.

The MLP seems to perform the worst out of these models which corresponds to the results of **Error! Reference source not found.**.

Χομμεντεδ [ΝωΣ81]: Zie figure 1 (weer 4 losse afbeeldingen)

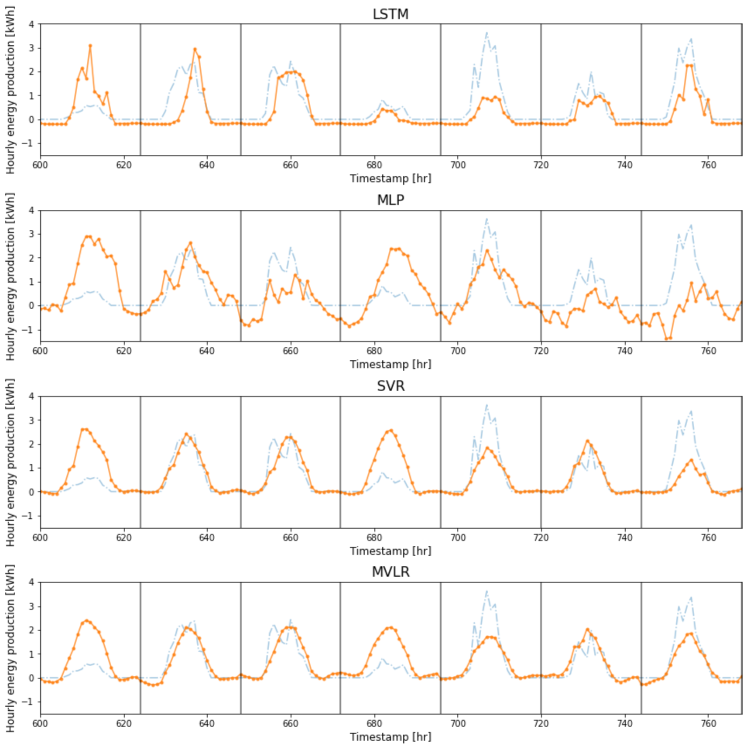

Figure 2 Production Model Graphs

4 Conclusions and discussion

Answers

Consumption

For consumption the LSTM performs best and shows the most potential.

Production

For production the SVR performs the best, however the LSTM does show a lot of potential as well.

Explanation

Further implications

Factory Zero (decision not modelled)

5 Recommendations

While looking through the data we encountered large spikes in energy usage in somewhat consistent intervals. After looking in to these spikes we concluded that the booster was causing these spikes (figure 2). This was to kill of potential Legionella bacteria. The interval between these spike was inconsistent enough that we noted an improvement in the prediction models after we removed the peaks. This is why we suggest to set the booster on a consistent interval. This might help the models to identify the peaks in energy that the booster causes and thus could improve the model.

Χομμεντεδ [ΔΛ(82]: BRON/figuur?

Χομμεντεδ [ϑΒ83]: Je ziet hier dat er 2 spikes van de booster komen een paar van de heatpump en de rest onbekend, dus ik denk dat je dit anders zou kunnen verwoorden

Χομμεντεδ [ΔΛ(84]: Het is niet zeker dat dit zal helpen, totdat bewezen is dat dit zou helpen.

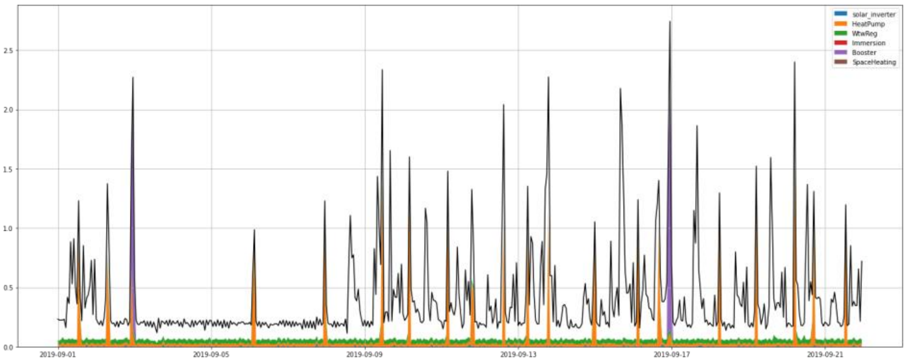

Figure 2: Energy use of house #37 categorised by appliances that use energy over a period of 3 weeks

After some testing it also appeared that the interval in which the model predicts also effects the accuracy of the model (figure 3). It might be worth looking more deeply in this interval to achieve better results than the 1 hour interval that was used in this paper.

Χομμεντεδ [ΝϖΣ85]: Komt een figuur dat de resultaten van verschillende intervallen weergeeft. (als je weet welke notebook dat is msg me ff)

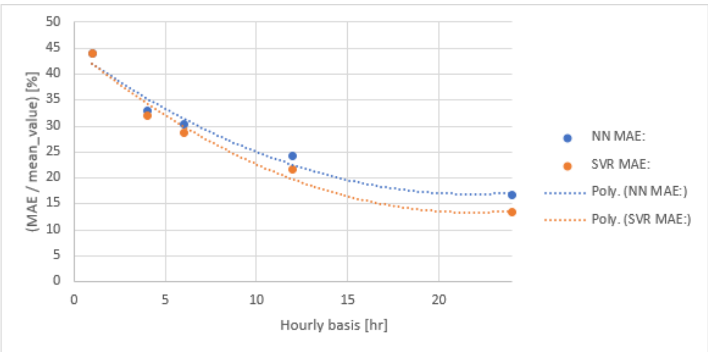

Figure 3: Hourly basis and its MAE

One of the problems that occurred while creating the models was inconsistent peaks and predicting these peaks. It may be worth it to see if models benefit from a separate model that predicts peaks or other means to detect these.

6 References

al, A. H. (2020). Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series. *Solar Energy 207*, 626-639.

al, K. X. (2019). Short-term Time Series Data Prediction of Power Consumption Based. *IOP Conference Series: Materials Science and Engineering*, 646.

Azhar Mat Daut, M., Yusri Hassan, M., Abdullah, H., Yusri Hassan, M., Abdul Rahman, H., Pauzi Abdullah, M., & Hussin, F. (n.d.). Buildingelectrical energy consumption forecasting analysis using conventional andartificial intelligence methods: A review. *Elsevier*. Retrieved from ScienceDirect.

Fan, C., Xiao, F., & Wang, S. (n.d.). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Elsevier*. Retrieved from ScienceDirect.

Jana, R., Ghosh, I., & Sanyal, M. (n.d.). A granular deep learning approach for predicting energy consumption. *Elsevier*. Retrieved from ScienceDirect.

Khan, P. W., Byun, Y.-C., Lee, S.-J., Kang, D.-H., Kang, J.-Y., & Park, H.-S. (2020). Machine Learning-Based Approach to Predict Energy Consumption of Renewable and Nonrenewable Power Sources. *Energies*. Retrieved from MDPI.

Khovalyg, D., & Heidari, A. (2020). Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series decomposition. *Elsevier*. Retrieved from ScienceDirect.

Kim, T.-Y., & Cho, S.-B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. *Elsevier*. Retrieved from ScienceDirect.

Liu, Z., Wu, D., Han, Y., Lun, L., Gao, J., Jin, G., & Cao, G. (2019). Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction. Energy Exploration & Exploitation. *Elsevier*. Retrieved from ScienceDirect.

over ons. (n.d.). Retrieved from factoryzero: https://factoryzero.nl/over-ons/

Rijksoverheid. (n.d.). *Energieagenda: naar een CO_2-arme energievoorziening.* Retrieved from rijksoverheid.nl: https://www.rijksoverheid.nl/documenten/rapporten/2016/12/07/ea

United Nations Climate Change. (n.d.). *The Paris Agreement.* Retrieved from UNFCCC: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement