Minor Applied Data Science: *Energy in transition*

Het voorspellen van het energievebruik per dag

De Haagse Hogeschool, Den Haag

24 januari 2023

Groep 3

Nour Albrzawi (19015046)

Job van der Laken (20170637)

Senna Sjambar (17118050)

Bjorn Verspoor (15126382)

Sjoerd de Witte (20157843)

Begeleiding: Jeroen Vuurens, Tony Andrioli, Edward van Noort en Factory Zero

Abstract

Het doel van dit onderzoek is het voorspellen van het energieverbruik per dag van huishoudens in Zoetermeer die gebruik maken van de integrated Climate Energy Module van Factory Zero.

Dit word gedaan door toepassing van Machine Learning modellen en neurale netwerken. Er is onderzoek gedaan naar de toepassing van de volgende modellen: Linear Regression, Support Vector Regression, XGBoost en Long Short Term Memory.

Uit het onderzoek is gebleken dat het voorspellen van het energieverbruik met de aangeleverde dataset complex is. De meeste modellen laten een slecht gemiddelde zien van de evaluatiemetrieken die zijn toegepast. Deze hebben namelijk een R2 score van onder de 0. Het LSTM model presteerde het beste met een RMSE score van 4.614 en een R2 score van -0.048.

Introductie

Opdrachtgever

De opdrachtgever voor dit project, Factory Zero, en gelijkgestemde partners, zijn ervan overtuigd dat het mogelijk is om de behoefte aan hogere kwaliteit en betere prestaties rondom energievoorziening tegen een lagere prijs, te halen doormiddel van innovatie-gestuurde industriële bouwpraktijk. Om dit mogelijk te maken is door Factory zero de integrated Climate Energy Module(iCEM) ontwikkeld en geproduceerd. De gecentraliseerde energiemodule is een volledig elektrisch oplossing voor ventilatie, warm water, verwarmen en koelen. Optioneel kun je de iCEM uitbreiden met zonnepanelen. De prestaties van de module kun je in real-time monitoren via een display in de woonkamer (Factory Zero, 2022).

Opdracht

De opdrachtgever heeft de mogelijkheid geboden om onderzoek te doen naar mogelijke verbeteringen ten opzichte van het in kaart brengen van de in en uit-stroom van energie naar huizen, van zowel het netwerk als geïnstalleerde zonnepanelen. Hiervoor heeft de opdrachtgever een dataset verstrekt.

De relevante punten uit de dataset bestaan uit energieverbruik per huishouden, energie opgewekt van de zonnepanelen per huishouden, energie van het net geleverd per huishouden en energie terug geleverd aan het net in kalenderjaar 2019 van een wijk rijtjeshuizen in Zoetermeer, Nederland.

Het uiteindelijke doel van het onderzoek is het voorspellen van het energieverbruik per huis, per dag. Deze informatie is nuttig voor Factory Zero, omdat met deze informatie vervolgens te bepalen is wat de optimale financiële rendabiliteit van een eventuele batterijinstallatie zou zijn.

Uitvoering

Om het onderzoek uit te voeren is literatuuronderzoek gedaan naar energie, zonnepanelen, batterijen, Factory Zero, de integrated Climate Energy Module(iCEM) van Factory Zero en concepten rondom energievoorspelling. Ook zijn meerdere gesprekken gevoerd met een werknemer van Factory Zero (Meneer Baldiri). Om voorspellingen te maken is gebruik gemaakt van diverse Machine Learning modellen en neurale netwerken. Deze zijn opgesteld en gerund in Jupyter Notebook. Data is verzameld en gemanipuleerd om bruikbaar te maken in diverse modellen. Vervolgens is de data op verschillende modellen getraind en getest en vervolgens gevalideerd.

Methodologie

Data

De data data was aangeleverd door Factory Zero bestond uit metingen die zijn gedaan voor 120 energie neutrale rijtjeshuizen in Zoetermeer met een iCEM uitgebreid met zonnepanelen. De metingen liepen uiteen van warmtepomp, energieverbruik tot CO_2-sensor. Deze metingen zijn allemaal in 2019 gedaan.

Voor dit onderzoek is voornamelijk data van de SmartMeter en de Solar gebruikt, die bijhoudt hoeveel energie er aan het huis wordt geleverd en hoeveel energie het huis terug levert aan het net en hoeveel energie er opgenomen is van het zonnepaneel. De metingen worden om de 5 minuten gedaan, maar doordat de apparatuur niet altijd goed werkt was dit niet altijd het geval. De huizen zijn daarom allemaal gecheckt op foute metingen, door naar het plot van het energieverbruik van een jaar te kijken. Soms ontbraken er maar een aantal metingen en dit kon worden opgelost door interpolatie van de data, maar bij andere huizen ontbraken soms maanden aan data. Uit de dataset zijn daarom de volgende huizen gehaald [8, 13, 18, 21, 25, 27, 29, 31, 32, 33, 34, 35, 36, 45, 49, 53, 59, 62, 65, 68, 78, 82, 85, 86, 87, 89, 90, 96, 97, 101, 103, 107, 108, 109,111, 118,119], deze huizen hadden een dusdanig fout in de metingen dat ze niet bruikbaar waren voor dit onderzoek.

Verder is er ook gekeken naar data van het KNMI over de zonnesterkte. Dit is in een enkel model meegenomen.

Voor dit onderzoek wordt er getracht het energieverbruik van de volgende dag te voorspellen, daarom is er gekozen om de metingen lineair te interpoleren over een dag. Het energieverbruik van de voorgaande dag(en) wordt in elk model gebruikt.

Modellen

In dit onderzoek zijn verschillende methoden gebruikt om het energieverbruik te voorspellen. Statistische modellen zoals ARIMA hadden volgens Factory Zero tot nu toe tot niks geleid, waarschijnlijk omdat de data erg onregelmatig is. Daarom is de focus gelegd op Machine Learning modellen en neurale netwerken, gezien deze ook het best om kunnen gaan met onvoorspelbare waardes. Als eerste is er gekeken naar Linear Regression. Daarna ook naar Long Short Term Memory (LSTM), XGBoost en Support Vector Machine (SVM). Hieronder wordt per model verteld waarom ervoor gekozen is en hoe de data is voorbereid om het desbetreffende model te trainen.

Voor alle modellen die hieronder staan beschreven is de tijdreeksdata eerst stationair gemaakt. Dit is gedaan omdat op deze manier het gemiddelde en de standaarddeviatie van de data niet veranderd over tijd, waardoor een model hier beter op kan leren. Dus eerst wordt de stationaire data voorspeld waarna elke voorspelling wordt opgeteld bij het energieverbruik van de vorig dag om het zo weer terug te zetten naar een voorspelling voor de tijdreeksdata.

Linear Regression

Linear Regression is het simpelste en meest gebruikte algoritme in Machine Learning modellen. Bij een linear regression model worden verbanden ontdekt tussen de features en de doelvariabelen. Met linear regression kunnen alleen numerieke waarden voorspeld worden (Linear Regression in Python met sklearn, 2020).

Voor dit onderzoek is eerst gekeken naar linear regression. Daarvoor zijn de vorige twee weken aan energievebruik en de dag van de week meegenomen als feature en is de dag van de week omgezet naar dummyvariabelen.

Support Vector Regression

Support Vector Regression, SVR, is een type van Support Vector Machine, SVM, echter wordt dit type gebruikt om regressieproblemen op te lossen. Het idee achter SVR is om het hypervlak te vinden dat de marge (e) maximaliseert, wat de afstand is tussen het hypervlak en de dichtstbijzijnde datapunten, in een hoogdimensionale ruimte. De datapunten die het dichtst bij het hypervlak liggen zijn de support vectors. Het doel is om het hypervlak te vinden dat de marge maximaliseert terwijl het nog steeds nauwkeurige voorspellingen doet (Awad & Khanna, 2015).

Een SVR-model kan goed omgaan met onregelmatigheden in de data. Dit maakt het dan ook geschikt om toe te passen op het voorspellen van het energievebruik. Daarnaast biedt SVR een betere balans tussen de voorspellingsnauwkeurigheid en de berekensnelheid vergeleken met Multiple Linear Regression en Neurale Netwerken (Shao, Wang, Bu, Chen, & Wang, 2020).

Voor het trainen van het model is het energieverbruik van de vorige twee weken meegenomen als feature samen met welke dag van de week het is. Dit wordt toegepast met een sliding window, dus de dagen 1 t/m 14 voorspellen het energievebruik voor dag 15. De dagen 2 t/m 15 voorspellen het energievebruik voor dag 16 enzovoorts. De dag van de week is omgezet naar dummyvariabele ŧ. Als feature wordt dus de stationaire waarde meegenomen, hieruit komt ook een voorspelde stationaire waarde. Voor het valideren is dit eerst weer omgezet naar het daadwerkelijke energieverbruik. Tot slot wordt er nog door middel van een GridSearch onderzocht wat de beste hyperparameters zijn.

XGBoost

XGBoost is een Machine Learning-techniek die vaak wordt gebruikt voor verschillende soorten problemen. Het is een ensemble-techniek die gebruik maakt van beslissingsbomen. Een belangrijk kenmerk van beslissingsbomen is dat ze goed zijn in het afhandelen van non-lineaire relaties tussen variabelen. Dit maakt ze geschikt voor het voorspellen van tijdreeksen, aangezien veel tijdreeksen niet-lineaire patronen bevatten.

Een studie uit 2019 toonde aan dat XGBoost geschikt kan zijn voor tijdreeks voorspellingen (Abbasi, R. A., Javaid, N., Ghuman, M. N. J., Khan, Z. A., & Ur Rehman, S, 2019).

De features voor dit model zijn het energieverbruik van de vorige dag, de dag van de week en de maand van het jaar. De data is gesplitst in een traindataset van 10 maanden (jan t/m okt) en een test dataset van 2 maanden (nov, dec). De beste hyperparameters zijn gevonden met behulp van GridSearch.

LSTM

LSTM staat voor Long Short Term Memory en is een uitbreiding van de RNN's, ofwel Recurrent Neural Networks. Deze neurale netwerken worden onder andere gebruikt om voorspellingen te maken met sequentiële data. LSTM's en RNN's werken goed op dit soort data, omdat er tijdens het maken van een voorspelling rekening wordt gehouden met voorgaande inputs, i.p.v. het individueel behandelen van elke input. Omdat het energieverbruik per dag een vorm is van sequentiële data, is er in dit onderzoek gebruikt gemaakt van een LSTM.

Uit de literatuur blijkt dat LSTM's gebruikt kunnen worden voor het voorspellen van tijdreeksdata. Een studie uit 2020 "LSTM based long term energy consumption prediction with periodicity" (Wang, 2020) beschreef dat LSTM een grote potentie heeft om energieverbruik te voorspellen. In het onderzoek "Predicting residential energy consumption using CNN-LSTM neural networks", worden

LSTM's ook gebruikt om het energieverbruik te voorspellen (Kim, 2019). Naast LSTM's worden hier ook Convelutional neural networks (CNN's) toegepast. In dit onderzoek is alleen ingegaan op de LSTM's zelf.

In dit onderzoek is de sequentiële data te lang voor een RNN, waardoor er een exploding gradient problem of vanishing gradient problem kan ontstaan. Dit houdt in dat bij het toepassen van gradient descent binnen het neural netwerk, over de optimale gewichten heen wordt gesprongen of juist amper naar de optimale oplossing toe wordt bewogen.

Voor het model is als feature het energieverbruik van elke dag in één hele week (7 dagen) gebruikt, waarbij de volgende dag na deze week dan de doelwaarde is. De data is gesplitst in twee delen, de eerste 10 maanden als traindata en de laatste 2 maanden als testdata. Om het model te trainen, wordt er gebruik gemaakt van 100 epochs, 35 hidden layers en de Mean Squared Error (MSE) als loss functie.

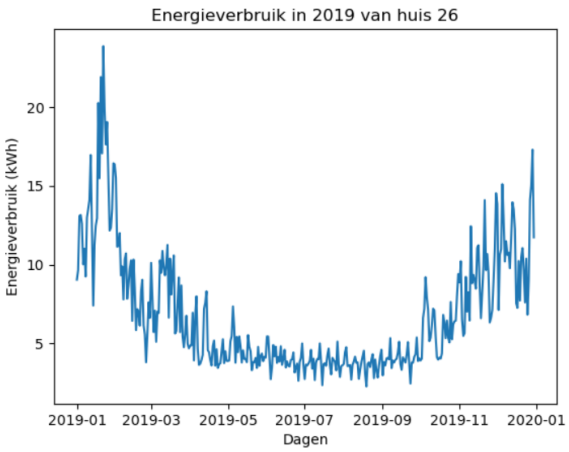
Evalueren

De modellen worden geëvalueerd met de coefficient of determination (R2) en root mean squared error (RMSE).

De R2 score zegt iets over de hoeveelheid variantie in de afhankelijke variabele (het energieverbruik) die wordt verklaard door de features. Deze waarde ligt tussen de 0 en 1, hierbij geeft 1 het beste model aan. Alle variantie in het energieverbruik wordt dan volledig verklaart door de features. Waardes lager dan een 0 betekend dat het model slechter werkt dan dat het gemiddelde wordt voorspeld (van Heijst, 2022). In dit onderzoek worden regressieanalyses gedaan met verschillende modellen en de R2 score is een evaluatiemetriek die van toepassing is op dit soort regressieanalyses. Op deze manier is het mogelijk om de modellen te testen en te vergelijken.

De RMSE komt in de literatuur ook vaak naar voren bij het evalueren van tijdreeksdata. Het geeft aan wat de gemiddelde afwijking is van elke gemaakte voorspellingen, dus hoe lager de RMSE des te beter.

Naast de scores zijn er verschillende plots gemaakt van het huishouden nummer 26. Dit is gedaan om een beter beeld te krijgen van de voorspelling op de test data voor elk Machine Learning model. Er is gekozen voor huis 26, omdat hier geen grote onverwachte pieken of dalen in te zien is over het hele jaar. Dit is te zien in Figuur 1, aan het begin en eind van het jaar (koude periode) werd veel energie verbruikt en in het midden (warme periode) werd minder energie verbruikt.

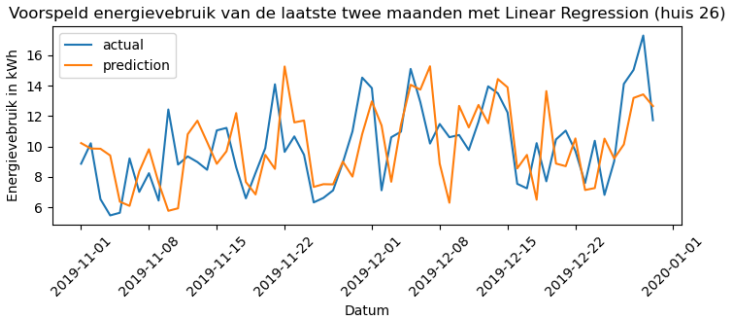


Figuur 1: Het energieverbruik in kWh uit 2019 van huishouden 26.

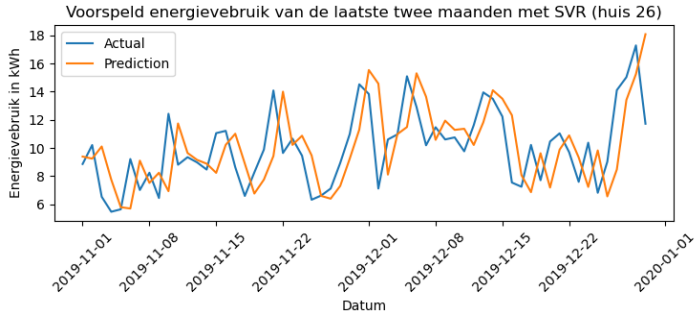
Resultaten

Tabel 1: De gemiddelde RMSE en R2 scores alle gebruikte huishoudens van de toegepaste voorspelmodellen.

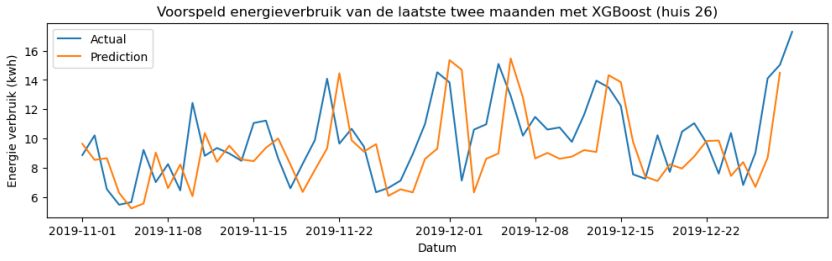
	Multiple Linear Regression	Support Vector Regression	XGBoost	LSTM
RMSE	5.964	5.284	6.574	4.614
R2	-0.745	-0.382	-0.910	-0.048



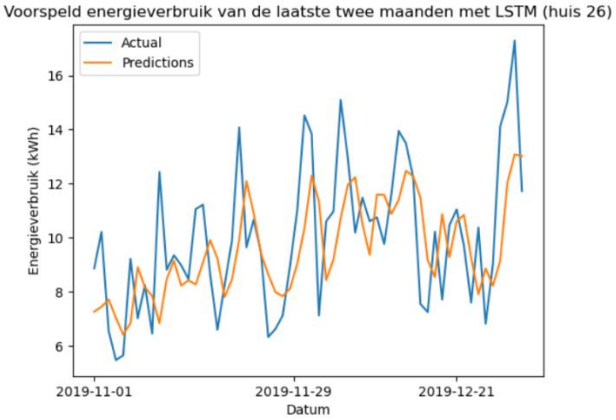
Figuur 2: Het voorspeld en daadwerkelijk energieverbruik in kWh met Linear Regression van de laatste twee maanden van huis 26 met een R2 score van -0.073



Figuur 3: Het voorspeld en daadwerkelijk energieverbruik in kWh met SVR van de laatste twee maanden van huis 26 met een R2 score van -0.068.



Figuur 4: Het voorspeld en daadwerkelijk energieverbruik in kWh met XGBoost van de laatste twee maanden van huis 26 met een R2 score van -0.117.



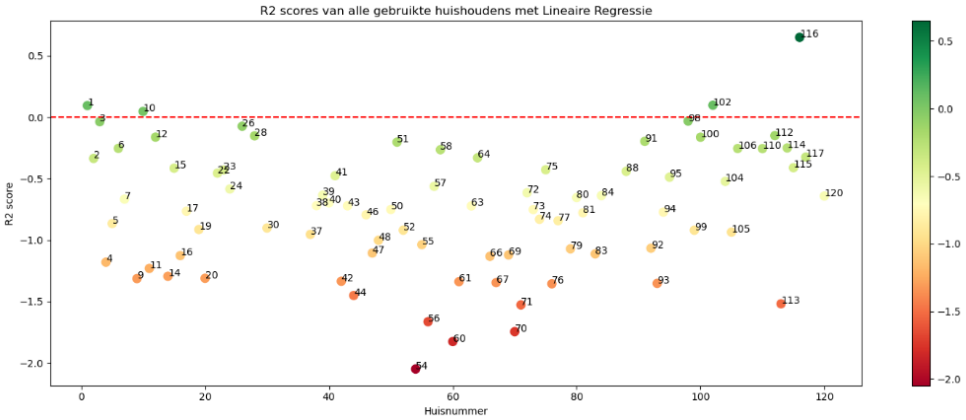
Figuur 5: Het voorspeld en daadwerkelijk energieverbruik in kWh met LSTM van de laatste twee maanden van huis 26 met een R2 score van 0.269.

Uit de resulaten van Tabel 1 is van de R2 score af te lezen dat de voorspellingen allemaal slechter presteren dan het gemiddelde energieverbruik voorspellen. Uit de RMSE score is af te leiden dat de voorspelling er gemiddeld tussen de 4.5 en 6.5 energieverbruik in kWh per dag naast ligt. Gemiddeld verbruikt een koelkast met losse vriezer 520 kWh per jaar (Rijksoverheid, 2020). Dit betekend dat de voorspellingen er ongeveer 3 tot 4.5 koelkasten naast ligt.

Uit de resultaten is verder af te leiden dat LSTM de hoogste R2 score heeft en de laagste RMSE score.

Analyse van resultaten

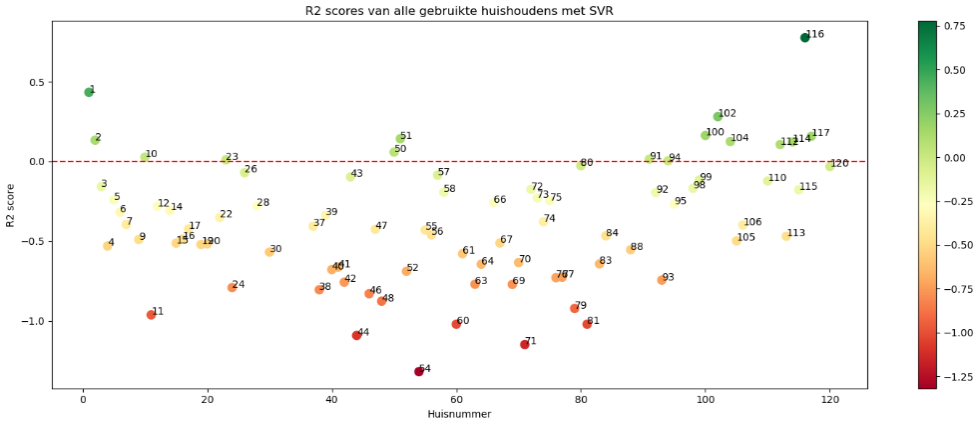
Bij de afbeeldingen: Figuur 2, Figuur , Figuur 4 en Figuur 5, van huis 26 is te zien dat de voorspelde waarde soms op de goede plek zit, maar over het algemeen is er steeds een dag lag te zien. Deze lag is ook aanwezig bij de andere huishoudens. Een verklaring hiervan kan zijn dat het meenemen van de vorige dag in de features erg zwaar weegt en daardoor een lag veroorzaakt.



Figuur 6: R2 scores van alle gebruikte huishoudens met Linear Regression.

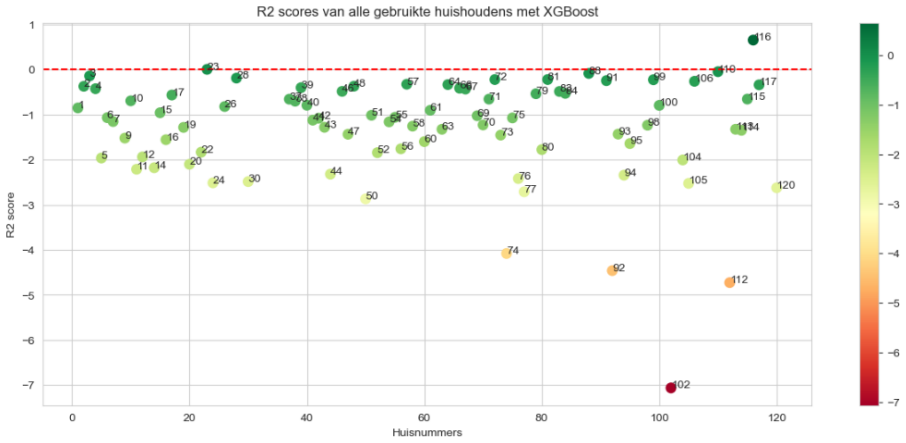
Bij de geplotte R2 scores uit Figuur 6 van Lineair Regression is duidelijk te zien dat op vier huishoudens na de rest allemaal onder de 0 liggen. Dit laat zien dat Linear Regression, op de manier

dat het tijdens dit onderzoek is gebruikt, niet geschikt is om het energieverbruik van de volgende dag te voorspellen.



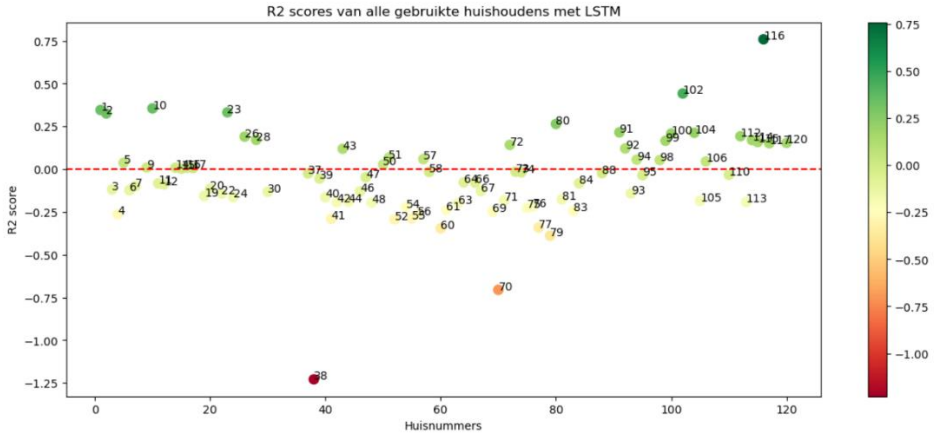
Figuur 7: R2 scores van alle gebruikte huishoudens met Support Vector Regression.

Uit de geplotte R2 scores van Figuur 7 is van de verschillende huishoudens duidelijk af te leiden dat er enkele huizen goed te voorspellen zijn, maar het overgrote gedeelte niet. De meeste liggen namelijk onder een R2 score van 0.



Figuur 8: R2 scores van alle gebruikte huishoudens met XGBoost.

In Figuur 8 is te zien dat de R2 scores bij het gebruik van XGBoost niet goed zijn. Op een paar uitzonderingen na liggen de meeste huisjes tussen de 0 en -2. In vergelijking met de SVR en LSTM-modellen presteert XGBoost aanzienlijk slechter.



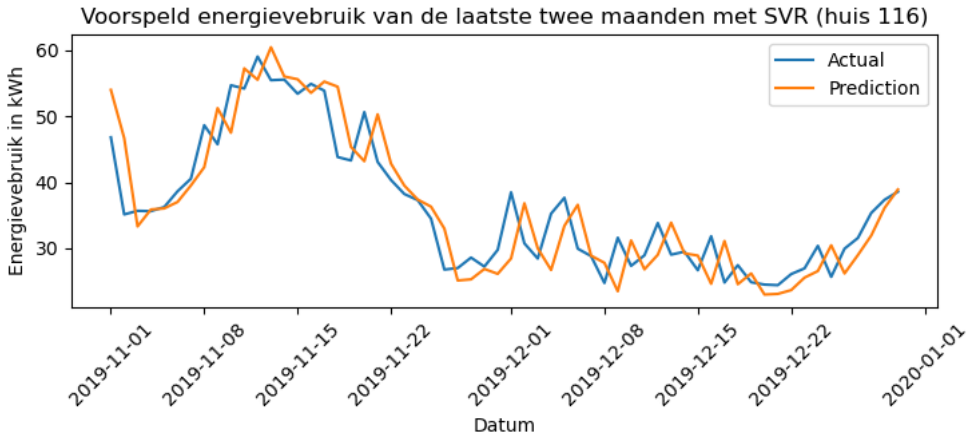
Figuur 9: R2 scores van alle gebruikte huishoudens met LSTM.

Uit Figuur 9met alle R2 scores van het LSTM-model is te zien dat er voor sommige huishoudens een goede voorspelling gemaakt kan worden, deze hebben dan een R2 boven de 0.25. Voor veel huishoudens is het nog lastig om een goede voorspelling te maken, waardoor de gemiddelde R2 score erg tegenvalt.

Discussie

Een wens van de opdrachtgever was om het energieverbruik per uur te voorspellen gezien de prijzen voor het inkopen en verkopen van energie aan het net per uur fluctueert. Hier is een begin aan gemaakt door de data te interpoleren naar uur, maar gezien de onregelmatigheid in de data was het niet mogelijk hier nuttige voorspellingen op te doen. Hierdoor is uiteindelijk de focus komen te liggen op het energieverbruik voorspellen per dag.

Uit de resultaten is gebleken dat geen enkel model, op de manier dat het is toegepast tijdens dit onderzoek, goed presteert. Dit komt omdat de gemiddelde R2 score rond de 0 ligt of lager. Dit betekent dus dat even goed het gemiddelde energieverbruik voorspeld kon worden in plaats van een Machine Learning model een voorspelling te laten doen.



Figuur 10: Het voorspeld en daadwerkelijk energieverbruik in kWh met LSTM van de laatste twee maanden van huis 116.

Opvallend aan Figuur 10 is dat huis 116 bij alle modellen het best presteert terwijl hier nog steeds een lag van een dag inzit. Dit valt te verklaren doordat het verloop van het energieverbruik minder fluctueert dan de rest van de huizen. Hier is namelijk te zien dat in de eerste maand het verbruik hoog is en in de laatste maand laag. Dit komt omdat bij het berekenen van de R2 de variantie in het hoge energieverbruik en het lage energieverbruik tegen elkaar worden weggespeeld.

Hoewel de modellen wel zijn te vergelijken op de uitkomst die ze nu hebben genereerd door middel van de verschillende evaluatiemetrieken, is het lastig om een definitief antwoordt te geven welk model het beste werkt bij dit probleem. Dit komt doordat elk model verschillende features heeft gebruikt, omdat in sommige modellen bepaalde features makkelijker mee zijn te nemen dan in andere.

Conclusie

Uit dit onderzoek is gebleken dat het energieverbruik van de meeste huishoudens uit de dataset van Factory Zero lastig te voorspellen is. Bij elk model dat is toegepast, is te zien dat er een dag lag zit in de voorspelling. De modellen laten dan ook stuk voor stuk een slecht gemiddelde zien van de evaluatiemetrieken die zijn toegepast. Toch blijkt dat het energieverbruik voor sommige huishoudens wel redelijk te voorspellen is met een R2 van boven de 0.2. Dit komt dan nog het vaakst voor bij LSTM en SVR, deze presteren over het algemeen nog het best van alle modellen. Daarna komen lineaire regressie en XGBoost, waarbij het af en toe mogelijk is om een aannemelijke voorspelling te maken.

Aanbevelingen

Tijdens dit onderzoek zijn zowel per dag als per uur voorspellingen gedaan. Uiteindelijk is gebleken dat het voorspellen per uur veel minder goede resultaten gaf bij onze modellen. Dat is te verklaren doordat er meer informatie aan een Machine Learning model moet worden gegeven als het op uur basis een voorspelling wilt geven.

De eerste aanbeveling is om tijd te investeren in het kijken naar het energieverbruik per uur per huisje. Daarvoor zou informatie van de huizen met betrekking tot hoeveel personen in elk huisje wonen, wat voor apparaten aanwezig zijn die veel verbruiken, wat het verbruik in kWh van elk veel verbruikend apparaat is en rond hoe laat de machines meestal worden aangezet zeer waardevol zijn.

Verder kan er gekeken worden naar de optimale batterijgrootte. Er is tijdens dit onderzoek een begin gemaakt op dit gebied, maar niet genoeg om aantoonbare resultaten te presenteren. Het bepalen van een rendabel batterijgrootte kan het beste met statistiek berekend worden. Met behulp van een batterij kan energieverbruik bespaard worden en is er meer ruimte om te bepalen wanneer aan het net verkopen rendabel is tegenover eigen verbruik. De energieprijzen verschillen per uur, daarom is het handig om eerst het energieverbruik per uur te voorspellen en daarna te kijken naar de optimale batterijgrootte.

Tijdens een gesprek met meneer Baldiri is gebleken dat de warmtepomp te maken heeft met het energieverbruik. Er is dus gekeken naar wanneer de warmtepomp aan/ uit is, om dit eventueel mee te kunnen nemen in de voorspelmodellen. Echter is gebleken dat de warmtepomp in de dataset altijd de variabel 0 aangeeft, wat zou moeten betekenen dat hij altijd uit staat. Daarnaast kan de “holiday mode” ook invloed hebben op het energieverbruik. Als iemand op vakantie is, dan word er dankzij die instelling veel minder energie verbruikt door de warmtepomp. In de dataset staat

”holiday mode” altijd uit. Er zou dus gekeken kunnen worden naar de betrouwbaarheid van de dataverzameling.

Een interessante toevoeging in een veel later stadium zou zijn om gebruikers bijvoorbeeld via een app de mogelijkheid te geven om op elk moment hun inwoner aantallen, apparaten (met kWh), en langdurige afwezigheid(vakantie) in te stellen of aan te passen om de voorspelling meer accuraat te maken. Dit zou gestimuleerd worden door het feit dat betere voorspellingen tot minder energiekosten zou leiden.

Literatuur

Abbasi, R., Javaid, N., Ghuman, M., Khan, Z., & Ur Rehman, S. (2019, maart). Short term load forecasting using XGBoost. *Springer*, pp. 1120-1131. Opgeroepen op januari 19, 2023

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and.* New York: Apress. Opgeroepen op januari 10, 2023, van https://link.springer.com/book/10.1007/978-1-4302-5990-9

Factory Zero. (2022). *Product*. Opgeroepen op januari 3, 2023, van Factory Zero: https://www.factoryzero.nl/product/

Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. *Energy*, *182*, 72-81. https://doi.org/10.1016/j.energy.2019.05.230

Linear Regression in Python met sklearn. (2020, augustus 25). Opgehaald van data science partners: https://pythoncursus.nl/linear-regression-python/#wat-is-linear-regression-python

Rijksoverheid. (2020). *Hoe kan ik het stroomverbruik van elektrische apparaten verminderen?* Opgeroepen op januari 23, 2023, van Iedereen doet wat: https://www.iedereendoetwat.nl/mogelijkheden/elektrische-apparaten#:~:text=Televisie%20en%20stereo%20met%20randapparatuur,%E2%82%AC60%2C%2D%20per%20jaar.

Shao, M., Wang, X., Bu, Z., Chen, X., & Wang, Y. (2020, maart). Prediction of energy consumption in hotel buildings via support vector machines. *Elsevier*, 1-9. doi:https://doi.org/10.1016/j.scs.2020.102128

van Heijst, L. (2022, juli 5). *Regressieanalyse uitvoeren, interpreteren en rapporteren*. Opgeroepen op januari 12, 2023, van Scribbr: https://www.scribbr.nl/statistiek/regressieanalyse/#:~:text=De%20'R%20Squared'%20geeft%20aan,de%20afhankelijke%20variabele%20verklaard%20wordt

Wang, J., Du, Y., & Wang, J. (2020). LSTM based long-term energy consumption prediction with periodicity. *Energy, 197*. doi:https://doi.org/10.1016/j.energy.2020.117197