Research Project – portfolio

Nour Albrzawi (19015046)

Task definition

Introduction

Het project Energy Transition is door Factory Zero gestart en samen met de docenten van de minor gewerkt om de mogelijkheid aan de groepjes te bieden het project uit te voeren. De opdrachtgever, meneer Baldiri heeft een grote dataset naar ons toegestuurd om dit project uit te kunnen voeren. De dataset bevat 121 excelbestanden van 120 huishoudens in Zoetermeer in 2019. De laatste dataset "nummer 121" is een unit list waarin de units per soort meting worden aangegeven. Bijvoorbeeld kWh voor het meten van energieverbruik.

Aanleiding en doelstelling

Deze 120 huisjes hebben allemaal zonnepanelen geïnstalleerd. Zij gebruiken energie van het net, maar zij leveren ook energie opgewekt door hun zonnepanelen aan het net. De prijs van saldering gaat volgende jaren stijgen waardoor het steeds duurder wordt. Dit project is gestart om het energieverbruik per dag per huishouden te voorspellen. Eerst was het doel om ook te kijken naar de optimale batterijgrootte. Daar is een begin aan gemaakt maar niet tot een resultaat toegekomen (de literatuur onderzoek staat in de portfolio onder Domain Knowledge". Nu is het doel om het energieverbruik per huis per dag te voorspellen om te kijken wanneer het beste is om energie van het net te kopen/ aan het net te verkopen.

Gebruikte data

De datasets bevatten meerdere sheets. Voor het berekenen van het energieverbruik is dit formule gebruikt. Energieverbruik = smart_in + solar_out – smart_out. Het formule is van een onderzoekvorige studenten vorig jaar opgehaald. Om het energieverbruik te berekenen zijn de sheets 'SmartMeter' en 'Solar' gebruikt. Daarnaast is ook gekeken naar de zonsterkte om als feature in de voorspelmodellen te gebruiken. De zonsterkte is door een ander groepslid opgezocht en van KNMI opgehaald.

Onderzoeksvraag

De onderzoeksvraag is nu als volgt:

- *Wanneer is het beste om energie van het net te kopen/ aan het net te verkopen?*

Er is dus gekeken naar het energieverbruik per huis. Elk huis moet dan op basis van zijn verbruik beslissen wanneer het beste moment is om energie van het net te kopen en wanneer juist niet.

Evaluation

Het energieverbruik is voorspeld met machine learning modellen en neurale netwerk. De gebruikte machine learning modellen zijn: Linear Regression, Polynomial Features, XGboost en Support Vector Machine. Naast deze modellen is ook gekeken naar neurale netwerk Long Short Term Memory (LSTM) op advies van de docent. Deze modellen zijn allemaal gekozen omdat dit probleem geen

classificatie maar regressie probleem is. Tijdens deze minor is geleerd welke modellen bij welk soort probleem van toepassing zijn.

Future work

Tijdens dit onderzoek zijn er voorspellingen gedaan zowel per dag als per uur. Echter is toen gebleken dat het voorspellen per uur veel tijd kost en minder goede resultaten geeft. Dat komt door weinig/ geen informatie over het huishouden per huisje. De eerste aanbeveling zou dus kunnen zijn om te kijken naar energieverbruik per uur per huisje. Daarvoor zijn informatie van de huizen nodig m.b.t. bijvoorbeeld hoeveel personen in elk huisje zitten, hoe laat welke machine zij aandoen etc.

Verder zou ook gekeken kunnen worden naar de optimale batterijgrootte. Het bepalen van een rendabel batterijgrootte kan het beste met statistiek gedaan worden. Dat blijkt uit gesprekken met meneer van Noort. Met behulp van een batterij kan energieverbruik bespaard worden en kunnen mensen daarnaast ook bepalen wanneer zij energie aan het net verkopen en wanneer van het net inkopen. 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 warmtepomp te maken heeft met het energieverbruik. Er is dus gekeken naar wanneer de warmtepomp aan/ uit is, om dit mee te nemen in de voorspelmodellen. Echter is gebleken dat de warmtepomp in de dataset altijd 0 is, wat betekent dat hij altijd uit staat. Daarnaast kan de “holiday mode” ook invloed hebben op het energieverbruik in de zin van als iemand op vakantie is, dan gebruikt hij minder energie. In de dataset staat ”holiday mode” altijd uit. Er zou dus gekeken kunnen worden naar de betrouwbaarheid van de dataverzameling.

Conclusions

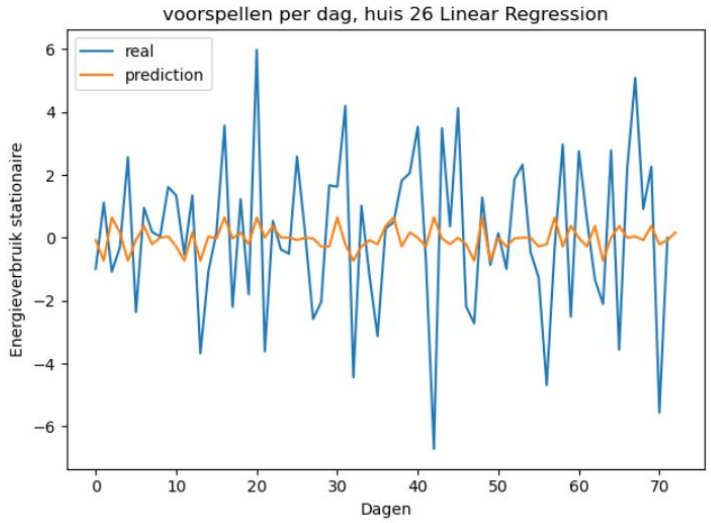
Zelf heb ik alle modellen behalve XGboost toegepast, maar ik heb dieper gekeken naar LinearRegression en LSTM. De modellen zijn geëvalueerd met de coefficient of determination (R2) en root mean squared error (RMSE). De R2 score zegt iets over de hoeveelheid variantie in de afhankelijke variabele, die wordt verklaard door de features. Deze waarde ligt tussen de 0 en 1, hierbij geeft 1 het beste model aan. 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 hoe beter. Hieronder zijn meer details over de modellen beschreven.

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. In figuur 1 is huis 26 als voorbeeld toegevoegd en zijn de laatste 70 dagen van het jaar (de testdata) geplot.

```
R2_score: -0.140
MAE: 1.601
MSE: 4.581
```

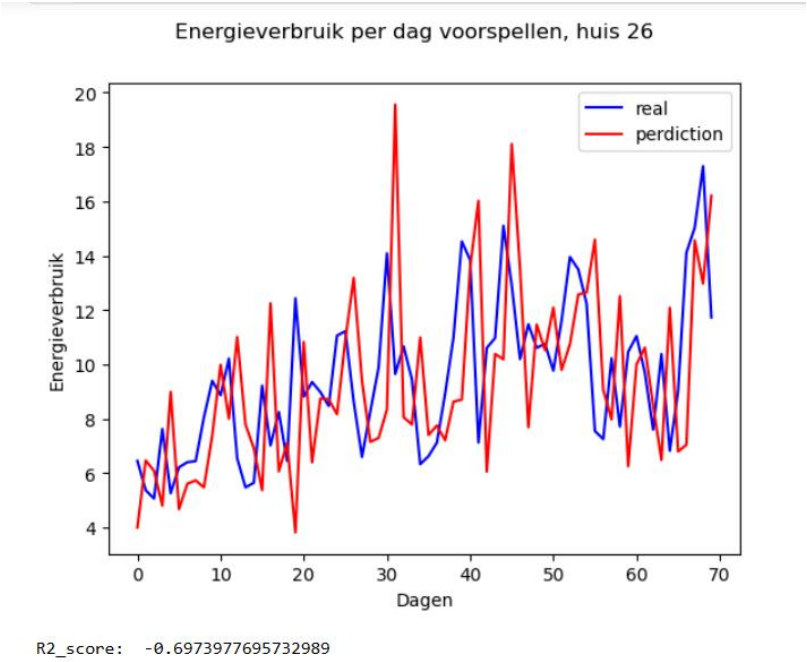


Figuur 1: voorspellen per dag met Linear Regression

LSTM

Long Short-Term Memory (LSTM) is een type recurrent neuraal netwerk (RNN) dat specifiek is ontworpen om sequentiële gegevens te verwerken, zoals tijdreeksen, spraak en tekst. Omdat dit project met tijdreeksen te maken heeft, lijkt het ons verstandig naar LSTM te kijken.

Voor dit model is het energieverbruik als feature en target gebruikt. Er is gebruikgemaakt van een sliding window van 14 dagen aan data. Een voorbeeld: als X de eerste 14 dagen van het jaar is, dan is y dag 15. Dat betekent dat LSTM model het energieverbruik van de 15^{e} dag van het jaar voorspelt. Dan wordt X van dag 2 t/m dag 15 en y dag 16. Het model voorspelt dus de 16^{e} dag enzovoort. In figuur 2 is huis 26 als voorbeeld toegevoegd en zijn de laatste 70 dagen van het jaar (de testdata) geplot.



Figuur 2: voorspellen per dag met LSTM

Analyses en conclusie

Uit de resultaten van alle huisjes is een gemiddeld van alle scores berekend door de groepsleden. Het is echter lastig te zeggen welk model het beste voorspelt. Maar over het algemeen werkt LSTM nu het beste met een gemiddeld R2 score van - 0.048 en RMSE van 4.614.

Planning

Tijdens dit project is gewerkt volgens scrum. Elke week wordt verteld wat elke groepslid heeft gedaan, waar hij/zij vastloopt en wat zijn/haar planning voor de komende week is. Meestal wordt dat op de vrijdagen gedaan.

Voor het paper zijn de hoofdstukken verdeeld over alle groepsleden en is een deadline gesteld. Iedereen heeft zijn stuk(ken) op tijd geschreven en samen hebben wij uiteindelijk gekeken naar de hele paper.

Bronnen

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