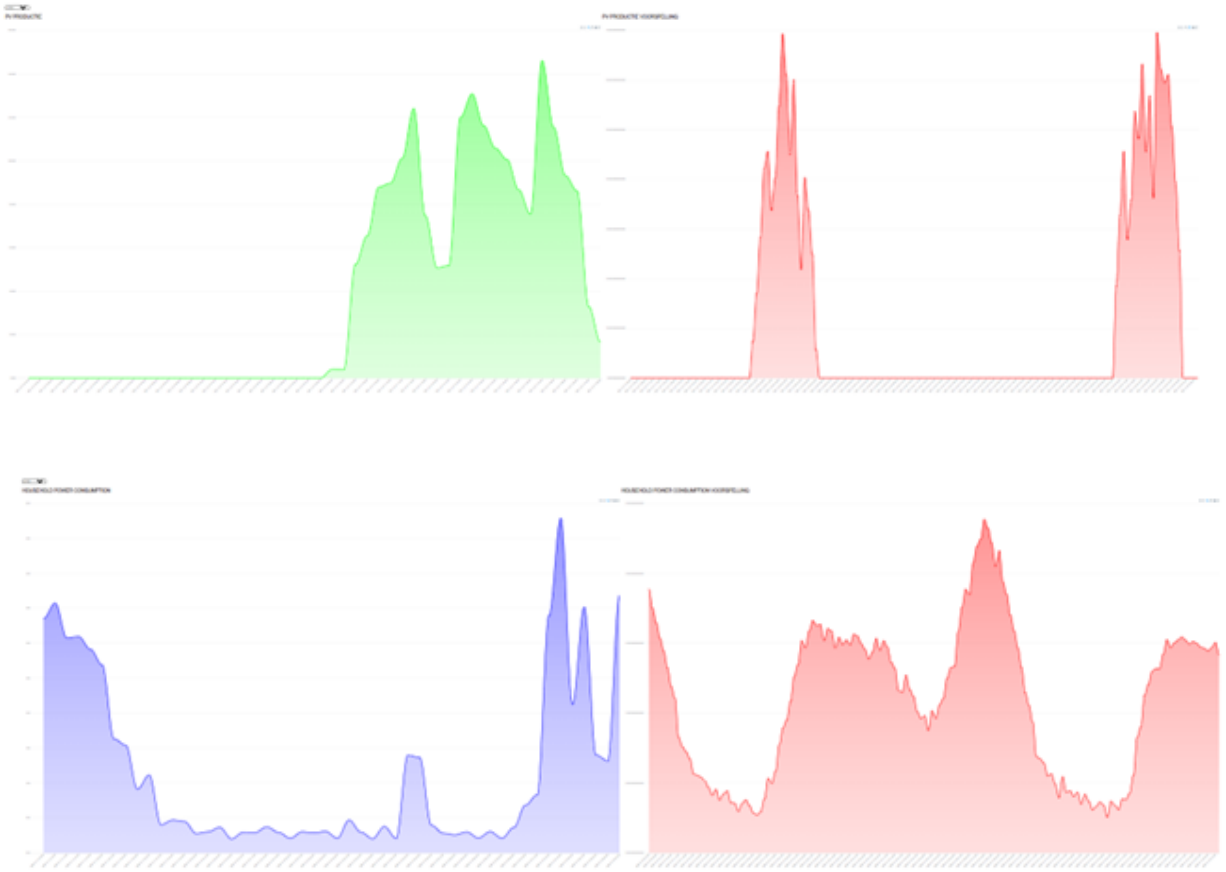
Dynamische Energiescan



by

Axel Frederiks (2202630)
Oscar Theelen (2204154)
Jonah Siemers (2202907)
Menno Rompelberg (2204333)

08/07/2024

Supervisors
Shannen Dolls
Jan Baljan
Reinder Bernhard

Inhoud

Introductie

Aanleiding

De aanleiding voor dit project is dat Woonwijzerwinkel, een bedrijf gespecialiseerd in verduurzaming en energiebesparing, een pitch wil voorbereiden voor TenneT. Het doel van de pitch is om TenneT te overtuigen van de waarde van P1-sensoren, die momenteel ontwikkeld worden, en de omliggende systemen. Met zo een P1-sensor kan men namelijk live het energieverbruik monitoren. Dit project zal dienen als bewijsvoering voor Woonwijzerwinkel om toestemming te krijgen om hun sensoren te installeren mogen op een lokaal industrieterrein.

Doelstelling

Het project richt zich op de ontwikkeling van een innovatief platform dat de visualisatie van energieverbruik data mogelijk maakt en tegelijkertijd toekomstig energieverbruik kan voorspellen door middel van modellen. Dit platform heeft als doel om bedrijven te helpen hun energieverbruik beter te begrijpen en te optimaliseren, wat voor autonomie zorgt bij bedrijven die duurzame energie opwekken en bijdraagt aan een kostenefficiënter energiebeheer. Dit omvat het verkennen van het onderzoekslandschap en het onderzoeken van openbare datasets op beschikbaarheid en bruikbaarheid voor dit onderwerp.

De doelstelling van dit project is om een systeem te ontwerpen dat toekomstige situaties op het gebied van weersomstandigheden, energieopwekking en energieverbruik kan voorspellen en visueel kan weergeven

Theoretisch kader

The Smart Grid is een geavanceerd elektriciteitsnet dat digitale technologie omvat voor tweerichtingscommunicatie tussen nutsbedrijven en consumenten, waardoor de efficiëntie en betrouwbaarheid van de elektriciteitstransmissie wordt verbeterd. “The Smart Grid is not just about utilities and technologies; it is about giving you the information and tools you need to make choices about your energy use.” *(U.S. Department of Energy, z.d.)*

In de tekst zal vaker de afkorting PV voor Photovoltaic gebruikt worden; “Een fotovoltaïsche cel, ook wel PV-cel genoemd, is een zonnecel die licht omzet in elektriciteit.” *(Wikipedia-bijdragers, 2023)*

Modellen:

LSTM, Long Short-Term Memory is een verbeterde versie van het recurrent neural network, ontworpen door Hochreiter & Schmidhuber. LSTM blinkt uit in sequentie voorspelling en legt lange termijn afhankelijkheden vast. Ideaal voor tijdreeksen, machinale vertaling en spraakherkenning vanwege de afhankelijkheid van de volgorde.
(GeeksforGeeks, 2024)

GRU, Gated Recurrent Unit is een type recurrent neural network (RNN) dat in 2014 werd geïntroduceerd door Cho et al. als een eenvoudigere alternatieve versie van Long Short-Term Memory (LSTM) netwerken. Net als LSTM kan GRU sequentiële data verwerken zoals tekst, spraak en tijdreeksdata.Het basisidee achter GRU is het gebruik van gating-mechanismen om de verborgen toestand van het netwerk bij elke tijdstap selectief bij te werken.
(GeeksforGeeks, 2023)

ARIMA, een AutoRegressive Integrated Moving Average, of ARIMA, is een statistisch analysemodel dat tijdreeksgegevens gebruikt om ofwel de dataset beter te begrijpen of toekomstige trends te voorspellen. Een statistisch model is autoregressief als het toekomstige waarden voorspelt op basis van eerdere waarden. Bijvoorbeeld, een ARIMA-model kan proberen om de toekomstige prijzen van een aandeel te voorspellen op basis van zijn eerdere prestaties, of de winst van een bedrijf te voorspellen op basis van eerdere periodes.
(Hayes, 2024)

SVM, support vector machine is een machine learning- algoritme onder toezicht dat wordt gebruikt voor zowel classificatie als regressie. Het hoofddoel van het SVM-algoritme is het vinden van het optimale hypervlak in een N-dimensionale ruimte die de datapunten in verschillende klassen in de kenmerkruimte kan scheiden.
(GeeksforGeeks, 2023)

Transformer, Een transformer model is een neuraal netwerk dat context en daarmee betekenis leert door relaties in sequentiële data, zoals de woorden in deze zin, te volgen. Transformer Modellen passen een evoluerende set wiskundige technieken toe, genaamd attention of self attention, om subtiele manners te detecteren waarop zelfs verre data-elementen in een reeks elkaar beïnvloeden en van elkaar afhankelijk zijn.
(Merritt, 2024)

EDA, Exploratory Data Analysis verwijst naar de methode van het bestuderen en verkennen van datasets om hun overheersende eigenschappen te begrijpen, patronen te ontdekken, uitschieters te lokaliseren en relaties tussen variabelen te identificeren. EDA wordt normaal gesproken uitgevoerd als een voorlopige stap voordat meer formele statistische analyses of modellering worden ondernomen.
(GeeksforGeeks, 2024a)

GHI, Globale Horizontale Irradiantie (GHI) is de hoeveelheid terrestrische irradiantie die valt op een horizontaal oppervlak ten opzichte van het aardoppervlak. Als GHI niet direct kan worden gemeten, kan het worden berekend uit de directe normale irradiantie (DNI) en de diffuse horizontale irradiantie (DHI) met behulp van de volgende vergelijking:

$$GHI = DNI \cdot \cos(\theta) + DHI$$

waarbij: GHI de globale horizontale irradiantie is, DNI de directe normale irradiantie is, DHI de diffuse horizontale irradiantie is en θ de zonnehoogte hoek is (de hoek tussen de zon en het verticale vlak). Deze formule combineert de directe component van de zonnestraling die door de zonneschijf wordt geleverd (aangepast aan de horizontale oriëntatie) met de diffuse component die van de hemel komt.
(Global Horizontal Irradiance, n.d.)

Methoden

Onderzoeksontwerp

Het hoofddoel van het onderzoek was om een gepast AI model te vinden om voorspellingen te doen op een serie van tijd. Voor dit onderzoek is er dus eerst gezocht naar bestaande AI modellen die tijdseries voorspellen en zijn deze hierna vergeleken met elkaar op accuratie en snelheid.

Dataverzameling

In dit hoofdstuk wordt de methodologie beschreven die is gebruikt voor het verzamelen van de bronnen en andere data die ten grondslag liggen aan dit verslag. Het doel van deze sectie is om transparantie te bieden over de zoekstrategie, de gebruikte databases, datasets en zoekmachines, en de criteria voor de selectie van de bronnen.

Zoekstrategie

Een belangrijk startpunt voor onderzoek was het eerste gesprek met de opdrachtgever. Hieruit zijn de meeste requirements opgesteld, maar ook een hoop belangrijke zoektermen en sleutelwoorden achterhaald. Hierna is het onderzoek aangevuld met relevante zoektermen die het onderzoek naar voren bracht, maar ook bracht deze meer vragen, en/of aanbevelingen naar voren tijdens gesprekken met de klant en docent/onderzoekers van het lectoraat.
Uit de bronnen die voortkomen uit de dalijk genoemde zoektermen werd vooral gekeken naar de titels en daarna de abstracts van de papers. Er waren veel bronnen te vinden over dit onderwerp en daarom is alleen naar de eerste pagina van resultaten gekeken steeds. Bij bronnen die vrijwel hetzelfde onderzoek hebben uitgevoerd is gekozen voor de bron die het meest divers was in gebruikte methodes.

Definiëren van Zoektermen en Sleutelwoorden

De eerste zoektermen uit het klantgesprek waren: “PV opbrengst, stroomverbruik, bedrijven, industrieterrein, voorspellen, data visualisatie, python” en vanuit het lectoraat: "tijdreeksdata, machine learning, neurale netwerken, weersvoorspellingen”. Later in het project waren de zoektermen meer model gericht en gebruikte zoektermen als: “ARIMA, GRU, Transformers, SVM” in combinatie met bovenstaande termen. Deze modellen komen ook voort uit reeds gevonden papers en wederom aanbevelingen van docent/onderzoekers.

Gebruik van Zoekmachines

Voor het vinden van papers om kennis op te doen en vergelijkbare onderzoeken te vinden is Google Scholar gebruikt. De zoektermen die eerder zijn benoemd zijn hiervoor naar het Engels vertaald om meer mogelijke resultaten te krijgen. Zo is bijvoorbeeld de combinatie: “machine learning, pv prediction, weather prediction” geprobeerd.

Voor algemene uitleg over termen die nog onduidelijk waren is over het algemeen Google gebruikt. Dit leidde meestal naar bekende sites zoals GeeksforGeeks om verduidelijking te krijgen. Verder is Google ook gebruikt om datasets te zoeken, met Kaggle in gedachten als een goede databank op aanraden van onderzoekers van het lectoraat.

Raadplegen van Gespecialiseerde Databases

Met de term “pv datasets” was het eerste resultaat de website van TU Delft die in eigen woorden “the most extended list of open-source photovoltaic (PV) power databases" had. Deze uitspraak werd versterkt door het feit dat de opvolgende 5 zoekresultaten allemaal ook in de lijst van de TU stonden. De datasets van de TU zijn afgewogen en daaruit gekozen, maar dit wordt later in het verslag extra toegelicht.

Verdere databanken zoals Sciencedirect en mdpi hebben we niet specifiek bezocht, omdat deze papers ook te vinden zijn in de data van Google Scholar.

Referentie Management

In dit document is gebruikgemaakt van de bronvermelding methode APA om bronvermeldingen te organiseren en om de gebruikte bronnen op een herkenbare manier te presenteren. De APA-stijl (American Psychological Association) biedt richtlijnen voor het citeren van verschillende soorten bronnen, zoals boeken, artikelen, websites en meer, wat helpt bij het waarborgen van consistentie en duidelijkheid.

Keuze van machine learning modellen

Om het doel van een accurate voorspelling te behalen moet er gekozen worden voor een voorspellingsmodel. Er blijken meerdere veelbelovend te zijn vanuit het paper *van Markovics en Mayer (2020)* dus zullen er ook meerdere worden getest. Bijlage A laat zien dat lineaire modellen slecht scoren voor een vergelijkbaar probleem. Uit hun onderzoek blijken een Support Vector Machine (SVM), neuraal netwerk (MLP) en verschillende ensemble methoden goed te werken.

Op aanraden van een docent/onderzoeker van het lectoraat Data Intelligence worden de volgende modellen ook onderzocht: ARIMA, GRU, LSTM en Transformers. Deze hebben allemaal gemeen dat ze goed werken met temporele data en ook voorspellingen over langere periode kunnen geven.

Proces

Week 1 begon zonder dataset, waardoor er geen direct werk kon worden verricht. In week 2 was de dataset nog steeds niet beschikbaar, maar er werd wel een API beloofd en een gesprek gepland met Martien (De ontwerper van de P1 sensoren) om de verdere stappen te bespreken. In week 3 was er nog steeds geen dataset of API ontvangen, waarna een verkenning werd uitgevoerd op Martiens device om te kijken of het gebruikt kon worden, echter is gebleken dat dit niet de richting op ging waar de opdracht voor bedoeld was.

In week 4 werd voor het eerst data ontvangen, echter waren dit Emap-bestanden, met deze data is niks ondernomen verder doordat het niet uit te lezen was. Uiteindelijk is er contact opgenomen met Tim van HBI en werd er een alternatieve dataset beloofd. De levering van de dataset van Tim werd in week 5 vertraagd. Er is daarentegen niet stilgezeten, er is een start gemaakt aan het werken met open datasets als alternatief voor de ontbrekende gegevens.

Week 6 stond in het teken van actieve analyse van de beschikbare data. Er werd een Exploratory Data Analysis (EDA) gemaakt op de open dataset om de kwaliteit en bruikbaarheid van de gevonden openbare data te beoordelen en mogelijke inzichten te identificeren. Alle modellen werden uitgewerkt, behalve de transformer aangezien deze niet gerealiseerd kon worden door de complexiteit van deze modellen. In week 7 werd de echte data ontvangen, waarna hier ook een EDA op werd uitgevoerd om de kwaliteit en bruikbaarheid van de data te beoordelen voor mogelijke toekomstige groepen. Dit leidde tot een aanbevelingsrapport, echter is deze dataset niet verwerkt in het uiteindelijke product aangezien hier helaas geen tijd meer voor was doordat de data pas laat geleverd werd.

In week 8 en 9 is alles afgerond en zijn de laatste modellen getraind om zo een volledig product af te leveren met de visualisatie inbegrepen.

De tabel die het proces van elke week bevat is te vinden in bijlage I.

Product

Requirements

De requirements zijn verdeeld in verschillende categorieën en prioriteitsniveaus (Must, Should, Could)

Voorspelling:

- *F11 (Must)***:** Het systeem kan toekomstige energieverbruik voorspellen. Aan de hand van input data uit een relevante dataset. Dit is een essentiële functionaliteit die aan de kern van het programma zal liggen..
- *F12 (Must)*: Het systeem kan toekomstige energieproductie voorspellen. Aan de hand van input data uit een relevante dataset. Net als requirement F11 is dit een cruciale eis van de applicatie, echter heeft dit betrekking tot een andere dataset.

Visualisatie:

- *F31 (Must)*: Het systeem weergeeft huidige data en visualiseert de voorspelde data in een webapplicatie. Dit zorgt ervoor dat gebruikers eenvoudig inzicht krijgen in de originele data en voorspelde data..

Functionaliteit:

- *F41 (Could)*: Het systeem geeft opties om energie te handelen met bedrijven. Op basis van het voorspelde energieverbruik en de voorspelde energieopbrengst. Hoewel dit een nuttige toevoeging is, is het niet essentieel binnen de scope van dit project.
- *F42 (Could)*: Het systeem zorgt voor de aankoop van extra energie wanneer energieverbruik boven energieproductie dreigt te stijgen. Ook voor deze requirement geldt dat het een goede toevoeging is, maar buiten de essentiële scope valt.

Opslag:

- *F51 (Should)*: Het systeem slaat alleen voor een bepaalde tijd data op om zo te veel opslagverbruik te voorkomen. Het is belangrijk om efficiënt om te gaan met opslag aangezien er gewerkt wordt met zeer grote hoeveelheden.

Model

De applicatie zal gebruik maken van een voorspellingsmodel om toekomstig energieverbruik te kunnen voorspellen. Dit model zal verschillende factoren in overweging nemen, zoals historische verbruiksgegevens, weersomstandigheden, energieopwekking, bedrijfsactiviteiten en seizoensgebonden variaties. Door deze data te analyseren, kan het model anticiperen op pieken en dalen in het energieverbruik, waardoor efficiënter energiebeheer mogelijk wordt.

Naast het voorspellen van energieverbruik, zal het model ook een belangrijke rol spelen bij het optimaliseren van de energieopslag en -distributie. Op basis van de voorspellingen van hoeveel eigen energie wordt opgewekt (bijvoorbeeld door zonnepanelen of windturbines), kan het model adviseren over de meest kosteneffectieve strategieën. Dit omvat beslissingen over het opslaan van overtollige energie in batterijsystemen voor later gebruik of het verhandelen van energie met andere bedrijven op het bedrijventerrein die op dat moment meer energie nodig hebben.

Datasets

Openbare Datasets

Vanwege het gebrek aan toegang tot live data, wordt de beschikbare data zodanig gemanipuleerd en gepresenteerd dat deze een simulatie van real-time data biedt. Door deze benadering kan alsnog de dynamiek en actuele relevantie van real-time gegevens worden nagebootst, ondanks de beperkingen in directe data-acquisitie. Hierdoor blijft de functionaliteit behouden en kunnen analyses en beslissingen worden gebaseerd op gegevens die de illusie van actuele tijdelijkheid geven.

Voor het onderzoek naar datasets zijn diverse openbare gegevensbronnen onderzocht. Op aanbeveling van de docent-onderzoeker van het lectoraat Data Intelligence zijn de datasets gezocht via de volgende websites: Kaggle en TU Delft. Uiteindelijk is besloten om de volgende datasets van deze platforms te gebruiken:

Gebruikte datasets

Household Electric Power Consumption
Metingen van het elektriciteitsverbruik in één huishouden met een bemonsteringsnelheid van één minuut over een periode van bijna vier jaar. Er zijn verschillende elektrische grootheden en enkele deelmeterwaarden beschikbaar. Dit archief bevat 2075259 metingen verzameld tussen december 2006 en november 2010 (47 maanden). Deze data is afkomstig uit de Verenigde Staten. *(Kaggle, 2016)*

2022_15min_data Amstelveen
Metingen van de zonnepaneel opbrengst op een onbekende plek in Amstelveen. Deze metingen hebben om de 15 minuten plaatsgevonden over een periode van een jaar. Deze dataset bevat 33601 metingen van 2021-12-26 tot 2022-12-10 (ongeveer 12 maanden). Doordat er een jaar aan data is en dus alle seizoenen voorkomen is deze dataset goed bruikbaar. De metingen hebben plaatsgevonden in Amstelveen, Nederland *(TU Delft, n.d.)*. In samenwerking met deze dataset wordt de PVLib *(Anderson. K, 2023)* library gebruikt om zo de zonne-data op te halen van de benodigde tijdstippen op de benodigde plek.

Beschrijving van de Kenmerken

Er zijn een aantal bruikbare datasets online gevonden die gebruikt kunnen worden om te gebruiken binnen een proof of concept, aangezien er geen data direct is geleverd vanuit de opdrachtgever. Twee van deze datasets zijn uiteindelijk gebruikt: 2022_15min_data (voor de energieopwekking) en household_power_consumption (voor het energieverbruik). Bijlage K

Niet-gebruikte datasets

Data Platform - Open Power System Data
Deze bevat verschillende soorten tijdreeksgegevens die relevant zijn voor de modellering van energiesystemen, namelijk elektriciteitsprijzen, elektriciteitsverbruik (belasting) en de opwekking en capaciteit van wind- en zonne-energie. De gegevens worden verzameld per land, controlegebied of biedzone. De geografische dekking omvat de EU en enkele buurlanden. Alle variabelen worden weergegeven in uren resolutie. Waar originele gegevens in hogere resolutie (half uur of kwartier) beschikbaar zijn (*Data Platform,* 2020).

Energy Usage From DOE Buildings
Energiegegevens uit een selecte portefeuille van gebouwen die eigendom zijn van de stad (DOE). Deze gegevens zijn afkomstig uit gebouwen die behoren tot het Department of Education (DOE). Deze data is afkomstig uit de Verenigde Staten *(data.world, 2024)*.

PVDAQ (*PV Data Acquisition*)
Biedt toegang tot fotovoltaïsche prestatiegegevens verzameld door NREL voor systemen in het hele land (Verenigde Staten). NREL is een nationaal laboratorium van het Amerikaanse ministerie van Energie, Office of Energy Efficiency en hernieuwbare energie, beheerd door de alliantie voor duurzame energie (*NREL: Developer Network*, z.d.).

Onderzoek naar modellen

Uit onderzoek is gebleken dat de volgende modellen als kandidaten kunnen worden beschouwd. Deze methoden zijn aangeraden door de docent onderzoeker van het lectoraat Data Intelligence. Daarnaast worden deze methoden bekrachtigd door verschillende studies *(Hyndman, R. J. & Athanasopoulos, G.)* die onderzoek hebben gedaan naar de onderstaande methodes waardoor ze relevant zijn in deze situatie.

ARIMA (*AutoRegressive Integrated Moving Average*)
Beschrijving: ARIMA is een model voor tijdreeks voorspellingen dat zowel autoregressieve als bewegende gemiddelde componenten combineert, en ook integratie om stationariteit te bereiken.
Toepassing: Geschikt voor het voorspellen van tijdreeksen zoals verkoopcijfers, economische indicatoren en temperatuurgegevens.*(Hyndman, R. J. & Athanasopoulos, G.)*

GRU (*Gated Recurrent Unit*)
Beschrijving: GRU is een type recurrent neural network (RNN) dat gates gebruikt om het doorgeven van informatie te regelen, wat helpt om het probleem van de verdwijnende gradient tegen te gaan.
Toepassing: Voor sequentiële gegevens zoals tekst, spraakherkenning en tijdreeks voorspellingen. *(Cho. K., 2014)*

LSTM (*Long Short-Term Memory*)
Beschrijving: LSTM is een ander type RNN dat speciale cellen gebruikt om lange-termijn afhankelijkheden in sequentiële gegevens te behouden.
Toepassing: Gebruikt voor taken zoals taal modellering, machinevertaling en tijdreeksanalyse. *(Hochreiter, S. & Schmidhuber, J.,1997)*.

SVM (*Support Vector Machine*)
Beschrijving: SVM is een classificatie- en regressie-algoritme dat probeert een optimale hypervlak te vinden dat de verschillende klassen in de gegevens scheidt.
Toepassing: Geschikt voor beeldherkenning, tekstclassificatie en bio-informatica
(Cortes, C., & Vapnik, V.,1995)

Lineaire Regressie
Beschrijving: Lineaire regressie is een eenvoudig regressiemodel dat de relatie tussen een afhankelijke variabele en een of meer onafhankelijke variabelen modelleert door een rechte lijn te passen.
Toepassing: Gebruikt voor het voorspellen van continue uitkomsten zoals huizenprijzen, verkoopprognoses en risico-inschattingen.
(Weisberg, S.,2005).

Logistische Regressie
Beschrijving: Logistische regressie is een model voor binaire classificatie dat de kans modelleert dat een bepaalde gebeurtenis optreedt.
Toepassing: Veel gebruikt voor credit scoring, ziekte voorspelling en spamdetectie (Alleen classificatie).
(Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X., 2013).

Decision Tree
Beschrijving: Decision trees zijn boomstructuren waarin elke interne knoop een "test" op een attribuut vertegenwoordigt, elke tak een uitkomst van de test, en elk blad een klasse of waarde.
Toepassing: Zowel classificatie- als regressieve taken, zoals klantsegmentatie en foutdiagnose *(Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. , 1984)*.

Random Forest
Beschrijving: Random Forest is een ensemble learning methode die meerdere decision trees traint en de gemiddelde voorspelling van de individuele bomen gebruikt om te verbeteren.
Toepassing: Bekend om zijn nauwkeurigheid en robuustheid in taken zoals beeldherkenning, fraude-detectie en voorspellend modelleren.
(Breiman, L. , 2001)

Transformers

Beschrijving: Transformer Architecture is een model dat gebruikmaakt van zelf aandacht en een hele zin omzet in één enkele zin. Dit is een grote verschuiving ten opzichte van de manier waarop oudere modellen stap voor stap werken, en het helpt de uitdagingen te overwinnen die te zien zijn in modellen als RNN's en LSTM's.

Toepassing: Transformers zijn slechts beperkt tot NLP en worden ook gebruikt voor computervisie- taken zoals beeldclassificatie, objectdetectie en het genereren van afbeeldingen.

(GeeksforGeeks, 2023b)

Resultaten

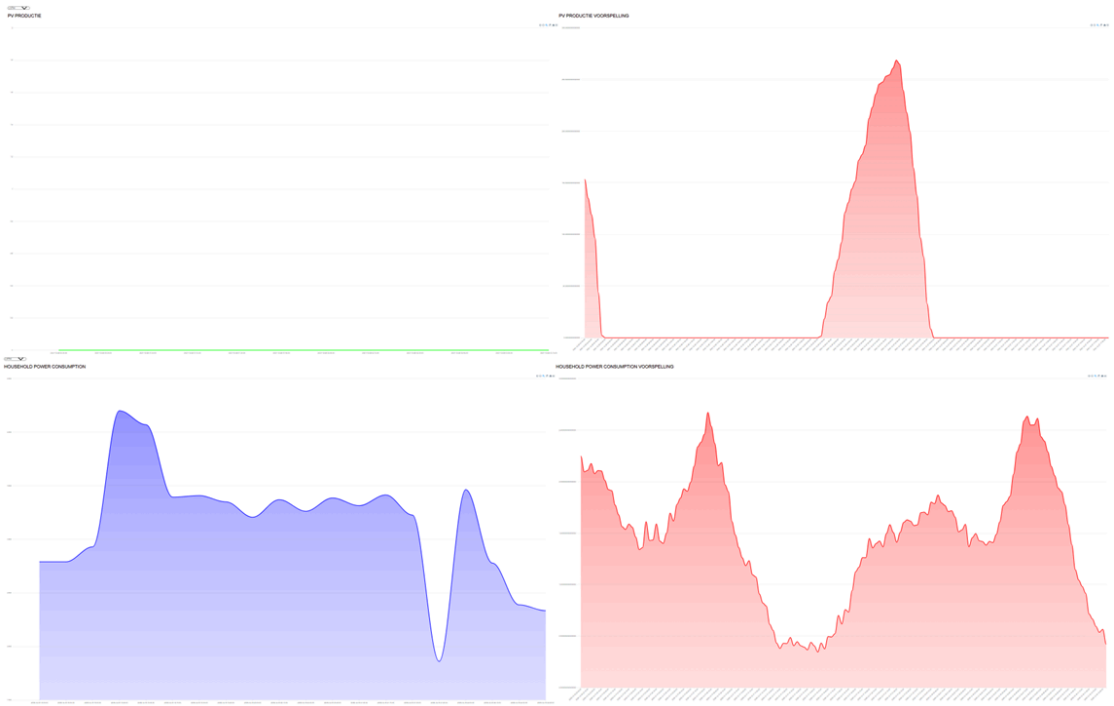
Applicatie

Requirements

De belangrijkste requirements zijn F11, F12 en F31. Deze drie zijn uiteindelijk ook allemaal gerealiseerd in het eindproduct. F51 was ook een belangrijke vereiste op het gebied van performance van het product. Aangezien er veel data wordt verzameld en opgeslagen. Er is aandacht gestoken in het optimaliseren van de opslag, echter is dit niet de hoofdfocus geweest. De twee functionaliteit requirements, F41 en F42 zijn uiteindelijk buiten scope gelaten doordat hier geen ruimte meer voor was.

Data visualisatie

- Het platform zal verschillende soorten grafieken en diagrammen bieden om historische, actuele en toekomstige energieverbruik data te visualiseren. Denk hierbij aan lijngrafieken, staafdiagrammen en histogrammen
- Gebruikers kunnen verschillende tijdsperioden selecteren om te bekijken.
- De visualisatie wordt geregeld door een Flask web applicatie, deze werkt met python en is dus makkelijk te gebruiken met machine learning.



Een voorbeeld van de werking van de applicatie. De gebruiker komt op een pagina die alle belangrijke informatie toont in een oogopslag.

Household Electric Power Consumption

Onderzoek aan dataset: Household Electric Power Consumption

1. Inleiding

Deze dataset bevat metingen van het elektriciteitsverbruik in een huishouden met een frequentie van één minuut over een periode van bijna 4 jaar (van december 2006 tot november 2010). De dataset bevat 2.075.259 metingen met negen kenmerken en enkele ontbrekende waarden (ongeveer 1,25% van de rijen).

2. Beschrijving van de Kenmerken

De dataset bevat de volgende kenmerken:

ID	Formaat	Beschrijving
Datum	dd/mm/yyyy	De datum van de meting
Tijd	hh:mm	De tijd van de meting
Global_active_power	Kilowatt	Globaal huishoudelijk gemiddelde actieve vermogen per minuut
Global_reactive_power	Kilowatt	Globaal huishoudelijk gemiddelde reactieve vermogen per minuut
Voltage	Volt	Gemiddelde spanning per minuut
Global_intensity	Ampère	Gemiddelde spanning per minuut
Sub_metering_1	Wattuur actieve energie	Energie submetering nr. 1. Dit komt overeen met de keuken, die voornamelijk een vaatwasser, een oven en een magnetron bevat.
Sub_metering_2	Wattuur actieve energie	Energie submetering nr. 2. Dit komt overeen met de wasruimte, die een wasmachine, een droger, een koelkast en een lamp bevat.
Sub_metering_3	Wattuur actieve energie	Energie submetering nr. 3. Dit komt overeen met een elektrische boiler en een airconditioner.

Analyse van Ontbrekende Waarden

De dataset bevat enkele ontbrekende waarden (ongeveer 1,25% van de rijen). Dit komt neer op ongeveer 181.853 ontbrekende waarden. Deze worden aangevuld met het gemiddelde van de andere waardes.

Heatmap van Correlatiematrix

Op de heatmap in bijlage C is te zien dat er geen zeer sterke correlaties zijn tussen tijd en andere data. Dit kan betekenen dat het moeilijker zal zijn om accurate voorspellingen te doen, omdat er geen duidelijke patronen zijn om op te baseren.

Resultaten van standaard modellen Household Electric Power Consumption (Kaggle, 2016)

Ondanks dat te doen was met een sequentieel probleem met deze dataset is er gekeken naar wat lineaire regressie, decision tree regressie, random forest en SVM toepassingen, hier zijn een aantal resultaten uit voortgekomen die in te zien zijn in bijlage L.
In het kort, alle eerder genoemde modellen in dit kopje zijn niet geschikt voor de doelstelling van dit project.

ARIMA getraind met volledige set

Naast de standaardmodellen zijn er ook een aantal sequentiële modellen getest op deze dataset. ARIMA is hier een van. Hoewel ARIMA een lineair model is, is er dan ook snel ondervonden dat dit model niet werkt op deze dataset. ARIMA scoort dan wel met een Mean Squared Error van 0.09288927338340065. Maar zoals te zien is in bijlage E komt de voorspelling totaal niet overeen met wat er verwacht wordt

LSTM vs GRU:

Uit tests is gebleken dat zowel LSTM als GRU twee geschikte kandidaten waren voor het voorspellen van toekomstige verbruikswaarden uit de dataset. Deze modellen zijn getraind op 20 epochs. De gekozen hoeveelheid epochs is voortgekomen uit hoeveel rekenkracht de machines hadden waarop ze zijn getraind. In een tabel in bijlage F is te zien dat de loss voor LSTM en GRU ongeveer gelijk blijven. Dit ondersteund dan ook het standpunt dat de twee modellen ongeveer hetzelfde presteren.

Zoals te zien is in bijlage F overschat LSTM en onderschat GRU gemiddeld. LSTM scoort met een Mean Absolute Error van 0.08773131750727241 (kilowatt) en GRU scoort een Mean Absolute Error van 0.10309738475432642 (kilowatt). Deze resultaten liggen vrij dicht op elkaar Om deze reden is ervoor gekozen om een ensemble te vormen tussen LSTM en GRU als een extra optie, aangezien deze dan het beste van beide werelden bevat.

Ook is te zien dat voorspellingen volledig buiten de originele dataset ongeveer aansluiten op de originele data.

PV dataset met zonnestraling

Onderzoek aan dataset 2022_15min_data_GHI.csv

1. Inleiding

De dataset is een samenvoegsel van de originele dataset afkomstig van Residential PV power plant en de uit PVLIB verkregen zonnestraling data. Op deze manier is er een koppeling te leggen tussen de opbrengst van PV cellen (Photo Voltaic, oftewel zonnepanelen) en het weer.

2. Beschrijving van de Kenmerken

De dataset bevat de volgende kenmerken:

ID	Formaat	Beschrijving
PV Productie (W)	W	Het vermogen in Joule/seconde
DateTime	YYYY/MM/DD hh:mm:ss	Een object van tijd
GHI (W/m^2)	W/m^2	Global horizontal irradiance
Year	YYYY	Het jaar van de meting
Month	MM	De maand van de meting
Day	DD	De dag van de meting
Hour	hh	Het uur van de meting
Weekday	0 - 6	De dag van de week waarin maandag = 0 en zondag = 6
Minute	mm	De minuut van de meting
Second	ss	De seconde van de meting

Analyse van Ontbrekende Waarden

Er zijn 95 ontbrekende GHI-waarden. Voor de rest zijn alle benodigde datapunten gevuld. Deze punten worden aangevuld met de functie fillna() van pandas. Hierbij wordt de methode "*ffill()*" gebruikt, deze vult ontbrekende waarden in door de laatste geldige waarneming door te geven aan de volgende geldige.

Heatmap van Correlatiematrix

In de heatmap in bijlage D zijn de correlaties tussen verschillende waarden weergegeven. Hieruit blijkt een correlatie tussen de PV-data en de GHI-data, wat aangeeft dat de PVLIB-data representatief is en dus als databron kan worden gebruikt.

Resultaten van standaard modellen 2022_15min_with_GHI (Kaggle, 2016)

Ondanks dat dit een sequentieel probleem is, is er gekeken naar wat lineaire regressie, decision tree regressie, random forest en SVM toepassingen, hier zijn een aantal resultaten uit voortgekomen die in te zien zijn in bijlage M. In het kort, alle eerder genoemde modellen in dit kopje zijn niet geschikt voor de doelstelling van dit project.

ARIMA getraind met volledige set

Naast de standaardmodellen zijn er ook op deze dataset een aantal sequentiële modellen getest. ARIMA is hier een van. Hoewel ARIMA een lineair model is, is er dan ook snel ondervonden dat dit model niet werkt op deze dataset. Net zoals bij de andere dataset is te zien in bijlage G dat voorspellingen totaal niet overeenkomen met een gewenst resultaat.

LSTM vs GRU:

Uit tests is gebleken dat zowel LSTM als GRU twee geschikte kandidaten waren voor het voorspellen van toekomstige verbruikswaarden uit de dataset. Deze modellen zijn getraind op 20 epochs. De gekozen hoeveelheid epochs is voortgekomen uit hoeveel rekenkracht de machines hadden waarop ze zijn getraind. In een tabel in bijlage H is te zien dat de loss voor LSTM en GRU ongeveer gelijk blijft. Dit ondersteund dan ook het standpunt dat de twee modellen ongeveer hetzelfde presteren.

Zoals te zien is in bijlage H overschat LSTM en onderschat GRU gemiddeld. LSTM scoort met een Mean Absolute Error van 137.9038284167211 (watt) en GRU scoort een Mean Absolute Error van 110.92523288719134 (watt). Deze resultaten liggen vrij nauw op elkaar.
Om deze reden is ervoor gekozen om een ensemble te vormen tussen LSTM en GRU als een extra optie, aangezien deze dan het beste van beide werelden bevat.

Ook is te zien dat voorspellingen volledig buiten de originele dataset ongeveer aansluiten op de originele data.

Keuze in modellen

Als keuzes zijn de GRU en LSTM modellen gekozen, deze zijn te selecteren in de flask app omgeving. Er is voor beide gekozen aangezien GRU statistisch gezien beter scoort, maar LSTM na evaluatie een realistischere voorspelling biedt. Later is er ook besloten een ensemble optie te bieden. Aangezien een van de modellen overschat en de andere onderschat wordt er het gemiddelde genomen tussen de twee om een eventueel beter resultaat te bereiken.

Aanbevelingen en bevindingen voor P1 sensor data

Allereerst bevat de gestuurde data meerdere kolommen zonder data. Deze zijn weggehaald voordat de set geleverd is. In het kader van gegevensbescherming is het essentieel om bepaalde kolommen te verwijderen die gevoelige en persoonlijk identificeerbare informatie bevatten. Een aantal van deze kolommen, zoals **owner_name** en **organisation_name**, bevatten direct persoonlijk identificeerbare gegevens. Het opnemen van namen van eigenaren of organisaties kan leiden tot het identificeren van individuen en hun associaties.8

Daarnaast zijn er kolommen zoals **device_name**, **app_region**, en **device_type** die indirect persoonlijke gegevens kunnen onthullen. Specifieke apparaatnamen, regio's waar de app wordt gebruikt, en typen apparaten kunnen bijdragen aan het profileren van gebruikers, vooral als deze informatie uniek of zeldzaam is.

Kolommen die te maken hebben met locatie, zoals **building_name**, **room_id**, en **floor**, zijn eveneens risicovol. Deze kolommen kunnen details onthullen over de fysieke locatie van gebruikers, wat kan leiden tot ongewenste blootstelling van persoonlijke locaties en bewegingen binnen een gebouw.

Verder zijn er kolommen die verband houden met tijd en registratie, zoals **device_registration_date** en **premium**. De registratie datum van een apparaat kan worden gebruikt om de activiteiten en gedragingen van gebruikers te traceren, terwijl de status van een premium abonnement inzicht kan geven in de financiële status en voorkeuren van een gebruiker.

Ten slotte zijn er de kolommen **ui_x_axis** en **ui_y_axis** waarvan niet zeker is waar ze voor dienden, maar de term ‘ui’ suggereert dat het met een interface te maken had en dat is niet relevant voor het maken van voorspellingen.

Door deze kolommen te verwijderen, kan ervoor gezorgd worden voor de bescherming van de privacy van gebruikers en voorkomen dat gevoelige en persoonlijk identificeerbare informatie wordt blootgesteld.

De eerste bruikbare kolom is de **time** kolom. Deze is voor tijdseries data belangrijk en heeft bij deze dataset een probleem. De P1 sensoren maken gebruik van LORA voor datatransmissie en hebben daarom inconsistentie in tijden. Meerdere studies benoemen dat “dirty timestamps” een negatieve invloed kunnen hebben op het maken van tijdsgebonden voorspellingen met deze data. (Wang & Wang, 2020)(Song et al., 2021)

Discussie

Prestaties

De uiteindelijk gekozen modellen waren LSTM en GRU en een ensemble van de twee modellen. Hoewel de resultaten goed lijken te zijn wijken ze statistisch gezien aardig af van de werkelijkheid. De hele reden om een ensemble te vormen was, omdat een van de modellen te hoog voorspelde en de ander te laag. Ook al heeft het een beter resultaat opgeleverd, is het nog niet perfect en is er dus ruimte voor verbetering.

Struikelblokken

Het verkrijgen van data vanuit bedrijven bleek nogal een uitdaging te zijn door AVG wetten en andere hindernissen op dit gebied. Uiteindelijk is er wel data geleverd, echter is hier niet mee doorgewerkt, omdat er niet genoeg tijd was hiervoor.

Rekenvermogen trainen

Het trainen van de Machine Learning modellen vereist veel rekenkracht van de computer, daarom is het voordelig als er gebruikgemaakt kan worden van de GPU. Op het moment van het project is het alleen mogelijk om tensorflow te draaien op een Nvidia GPU. Als men de modellen wil trainen op een niet-NVIDIA systeem zal het op de CPU moeten draaien. Dit duurt een aanzienlijk stuk langer.

Conclusie

Het project heeft aangetoond dat het mogelijk is om met behulp van machine learning-modellen nauwkeurige voorspellingen te doen van energieverbruik en -opwekking op basis van eerder verzamelde data. De keuze tussen een combinatie van LSTM en GRU modellen biedt een robuuste basis voor verdere ontwikkeling en implementatie. Deze resultaten vormen een solide basis voor Woonwijzerwinkel om hun P1 sensoren te promoten en te testen op een lokaal industrieterrein, zoals beoogd in samenwerking met TenneT.

Voor de toekomst lijken Transformers een veelbelovende onderzoeksrichting. Voorspellingen op de klant geleverde data, zoals de P1 dataset, zullen ook verder inzicht bieden in de toepasbaarheid van de bevindingen van dit onderzoek op de specifieke situatie.

Literatuurlijst

Anderson, K., Hansen, C., Holmgren, W., Jensen, A., Mikofski, M., and Driesse, A. “pvlib python: 2023 project update.” Journal of Open Source Software, 8(92), 5994, (2023). DOI: 10.21105/joss.05994.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. Routledge. https://www.routledge.com/Classification-and-Regression-Trees/Breiman-Friedman-Stone-Olshen/p/book/9780412048418

Cho. K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. https://arxiv.org/abs/1406.1078

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

Data Platform – Open Power System data. (2020, 6 oktober). https://data.open-power-system-data.org/time_series/2020-10-06

Energy Usage From DOE Buildings - dataset by city-of-ny. (2024, 27 april). data.world. https://data.world/city-of-ny/mq6n-s45c

GeeksforGeeks. (2023, March 2). Gated recurrent unit networks. GeeksforGeeks. https://www.geeksforgeeks.org/gated-recurrent-unit-networks/

GeeksforGeeks. (2023a, June 10). Support Vector Machine (SVM) algorithm. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/

GeeksforGeeks. (2023b, December 10). Transformers in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/getting-started-with-transformers/

GeeksforGeeks. (2024, June 10). What is LSTM Long Short Term Memory? GeeksforGeeks. https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory

GeeksforGeeks. (2024a, May 16). What is Exploratory Data Analysis? GeeksforGeeks. https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

Global Horizontal Irradiance. (n.d.). PV Performance Modeling Collaborative (PVPMC). https://pvpmc.sandia.gov/modeling-guide/1-weather-design-inputs/irradiance-insolation/global-horizontal-irradiance/

Hayes, A . (2024, April 6). Autoregressive Integrated Moving Average (ARIMA) Prediction Model. Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). John Wiley & Sons. https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473

Household electric power consumption. (2016, 23 augustus). Kaggle. https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. https://otexts.com/fpp3/arima.html

Kapp, S., Choi, J., & Hong, T. (2023). Predicting industrial building energy consumption with statistical and machine-learning models informed by physical system parameters. Renewable & Sustainable Energy Reviews, 172, 113045. https://doi.org/10.1016/j.rser.2022.113045

Markovics, D., & Mayer, M. J. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. Renewable & Sustainable Energy Reviews, 161, 112364. https://doi.org/10.1016/j.rser.2022.112364

Merritt, R. (2024, June 26). What is a transformer model? | NVIDIA Blogs. NVIDIA Blog. https://blogs.nvidia.com/blog/what-is-a-transformer-model/

PVDAQ (PV Data Acquisition) API | NREL: Developer Network. (z.d.). https://developer.nrel.gov/docs/solar/pvdaq-v3/

PV power databases. (n.d.). TU Delft. https://www.tudelft.nl/en/ewi/over-de-faculteit/afdelingen/electrical-sustainable-energy/photovoltaic-materials-and-devices/dutch-pv-portal/pv-power-databases (download: https://pvportal-3.ewi.tudelft.nl/PVP3.1/Open_Databases/Systeem_Amstelveen.zip)

U.S. Department of Energy. (z.d.). Smart Grid: The Smart Grid | SmartGrid.gov. https://www.smartgrid.gov/the_smart_grid/smart_grid.html

Visser, L., AlSkaif, T., & Van Sark, W. (2022). Operational day-ahead solar power forecasting for aggregated PV systems with a varying spatial distribution. Renewable Energy, 183, 267–282. https://doi.org/10.1016/j.renene.2021.10.102

Weisberg, S. (2005). Applied linear regression (3rd ed.). John Wiley & Sons. https://www.wiley.com/en-us/Applied+Linear+Regression%2C+4th+Edition-p-9781118386088

Wikipedia-bijdragers. (2023, 11 juli). Fotovoltaïsche cel. Wikipedia. https://nl.wikipedia.org/wiki/Fotovolta%C3%AFsche_cel

Song, S., Huang, R., Cao, Y., & Wang, J. (2021). Cleaning timestamps with temporal constraints. ˜The œVLDB Journal, 30(3), 425–446. https://doi.org/10.1007/s00778-020-00641-6

Wang, X., & Wang, C. (2020). Time Series Data Cleaning with Regular and Irregular Time Intervals. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2004.08284

Github repository
https://github.com/Ozziehman/Energiescan

Bijlage A

Metrieken voor machine learning modellen

Samenvatting van de vijf statistieken en de looptijd, gemiddeld voor de 16 geteste PV-installaties, voor alle 24 machine learning-modellen en de zes invoer- en afstemming gevallen. Groene en rode kleuren staan voor betere en slechtere waarden dan de mediaan, terwijl vetgedrukte cijfers de beste resultaten per rij aangeven. Deze tabel komt uit de paper van Markovics en Mayer (2022)

	Input	Hyperp.	Linear models												Kernel Ridge	Support Vector Machines	Neural Network	Neighbors	Decision Tree	Ensembles						
			LR	Lasso	Ridge	EN	Lars	OMP	BR	ARD	PAR	RAN-SAC	TR	Huber	KR	SVM	MLP	KNN	DT	RF	ET	ABR	GBR	XG-Boost	LGBM	Cat-Boost
nRMSE	Basic	default	52.6	94.0	52.6	78.1	52.6	53.3	52.6	52.6	81.2	54.0	52.8	53.0	52.6	53.2	52.1	59.6	72.9	56.2	56.3	56.5	**52.1**	55.6	53.6	54.2
		tuned	52.6	52.6	52.6	53.4	52.6	52.5	52.6	52.5	52.7	**53.8**	52.8	52.6	**51.5**	52.5	51.5	52.8	52.7	52.0	51.7	53.2	51.7	51.8	52.4	51.7
	Complex	default	49.4	94.0	49.4	72.6	58.4	52.8	49.4	49.4	77.1	56.7	49.9	50.1	49.4	46.7	46.7	53.1	63.9	48.7	47.7	51.8	**46.7**	50.6	48.4	48.8
		tuned	49.4	49.7	49.4	49.8	49.5	49.4	49.4	49.4	49.4	50.3	49.9	49.4	**45.3**	46.6	45.4	46.4	48.2	46.1	45.9	49.7	46.5	46.4	46.7	46.1
	Low resolution	default	52.1	94.0	52.1	84.8	52.1	65.7	52.1	52.1	79.9	56.1	51.2	52.2	52.1	47.2	**46.2**	54.2	64.6	52.7	51.4	55.5	48.0	52.2	49.9	50.4
		tuned	52.1	52.6	52.1	55.1	52.6	63.2	52.1	52.1	51.5	55.3	51.4	52.1	**46.0**	46.6	46.2	47.1	51.3	49.1	47.8	57.3	48.2	48.3	48.8	47.4
nMAE	Basic	default	37.4	82.7	37.4	68.8	37.4	38.2	37.4	37.4	59.9	36.3	36.2	35.8	37.4	**35.1**	36.5	40.9	48.7	39.0	39.1	44.9	36.6	38.5	37.2	37.6
		tuned	37.4	37.7	37.5	40.4	37.4	37.4	37.4	37.4	37.5	36.4	36.2	36.8	36.2	37.1	36.3	38.0	36.9	36.4	36.5	38.9	**36.0**	36.8	36.7	36.6
	Complex	default	35.3	82.7	35.3	63.7	40.9	37.8	35.3	35.3	56.0	37.9	34.3	34.3	35.3	**27.9**	30.8	32.7	38.8	31.0	30.7	39.5	30.9	32.6	30.9	31.5
		tuned	35.3	35.7	35.3	35.9	35.4	35.3	35.3	35.3	35.0	34.2	34.3	34.9	**29.6**	30.8	29.6	30.4	31.8	30.2	30.2	36.4	30.8	30.8	30.9	30.5
	Low resolution	default	39.8	82.7	39.8	74.8	39.8	49.8	39.8	39.8	58.8	39.5	38.2	38.5	39.8	**29.2**	30.4	34.3	39.9	33.5	33.1	43.2	32.8	34.0	32.4	32.9
		tuned	39.8	40.3	39.8	43.4	40.1	48.0	39.8	39.8	38.7	39.6	38.3	39.6	31.0	**30.3**	31.0	32.2	34.9	33.4	33.0	42.5	33.1	33.7	34.8	32.5
Corr	Basic	default	82.9	-1.1	82.9	82.1	82.9	82.3	82.9	82.9	61.3	82.5	82.9	82.9	82.9	83.0	**83.3**	78.1	69.6	80.3	80.2	81.1	83.2	80.8	82.2	81.8
		tuned	82.9	82.9	82.9	82.6	82.9	82.9	82.9	82.9	82.9	82.4	82.9	82.9	**83.7**	83.0	83.6	82.8	82.8	83.3	83.5	82.5	83.5	83.4	83.0	83.5
	Complex	default	85.1	-1.1	85.1	82.6	78.7	82.7	85.1	85.1	65.3	81.0	84.9	84.7	85.1	**87.2**	86.8	83.2	76.7	85.6	86.2	84.5	86.8	84.5	85.8	85.5
		tuned	85.1	84.9	85.1	84.8	85.0	85.1	85.1	85.1	85.3	84.6	84.9	85.0	**87.6**	86.9	87.6	87.0	85.9	87.2	87.3	85.2	86.9	86.9	86.8	87.1
	Low resolution	default	83.6	-1.1	83.6	70.9	83.6	71.5	83.6	83.6	66.6	81.0	84.0	83.2	83.6	86.8	**87.1**	82.3	76.1	83.1	83.9	81.7	86.0	83.5	84.9	84.6
		tuned	83.6	83.3	83.6	82.1	83.2	73.9	83.6	83.6	84.1	81.1	83.9	83.5	**87.2**	86.9	87.1	86.6	83.8	85.3	86.2	79.2	85.8	85.8	85.5	86.4
Skill	Basic	default	35.7	-15.0	35.7	4.4	35.7	34.8	35.7	35.7	0.6	34.0	35.4	35.2	35.7	34.9	36.3	27.1	10.8	31.3	31.1	30.8	**36.3**	32.0	34.5	33.7
		tuned	35.7	35.7	35.7	34.7	35.7	35.7	35.7	35.7	35.6	34.2	35.4	35.7	**37.1**	35.7	37.0	35.5	35.6	36.5	36.8	34.9	36.7	36.7	35.9	36.8
	Complex	default	39.7	-15.0	39.7	11.1	28.7	35.4	39.7	39.7	5.8	30.7	39.0	38.7	39.7	42.9	42.9	35.1	21.8	40.4	41.6	36.7	**42.9**	38.1	40.8	40.3
		tuned	39.7	39.2	39.7	39.1	39.5	39.7	39.7	39.7	39.6	38.5	39.0	39.6	**44.6**	43.0	44.5	43.3	41.1	43.7	43.9	39.2	43.1	43.2	42.9	43.6
	Low resolution	default	36.3	-15.0	36.3	-3.8	36.3	19.7	36.3	36.3	2.3	31.4	37.4	36.1	36.3	42.3	**43.5**	33.7	21.0	35.5	37.1	32.1	41.3	36.1	39.0	38.4
		tuned	36.3	35.7	36.3	32.7	35.7	22.7	36.3	36.3	37.0	32.4	37.2	36.4	**43.8**	43.0	43.5	42.5	37.2	39.9	41.6	29.9	41.1	41.0	40.3	42.1
nMBE	Basic	default	0.16	**0.00**	0.16	0.00	0.16	-0.01	0.15	0.15	-1.16	-3.97	-1.78	-2.98	0.16	-0.73	-0.20	-0.98	-1.04	-0.60	-0.53	4.96	0.03	-0.40	-0.25	-0.28
		tuned	0.16	0.12	0.15	0.09	0.14	0.16	0.15	0.16	3.14	-3.63	-2.03	-0.18	0.14	2.71	0.03	-1.22	-0.03	0.00	-0.10	0.68	-0.39	0.26	**0.00**	0.10
	Complex	default	1.14	**0.00**	1.14	0.00	2.61	-0.02	1.13	1.13	11.38	-0.46	-0.88	-0.90	1.14	2.65	0.42	0.58	-0.02	-0.04	-0.11	-1.34	0.08	-0.01	0.05	0.03
		tuned	1.14	0.49	1.13	0.42	0.97	1.14	1.13	1.12	5.89	-0.50	-1.04	1.25	0.37	3.04	0.09	0.06	0.14	**-0.02**	0.03	-0.30	-0.15	0.29	0.07	0.09
	Low resolution	default	2.90	**0.00**	2.90	0.00	2.90	0.19	2.90	2.90	17.27	3.48	4.24	1.97	2.90	2.19	0.67	0.47	-0.58	-0.46	-0.47	-0.18	0.76	0.50	0.47	0.66
		tuned	2.90	2.47	2.89	1.09	2.68	0.71	2.90	2.90	8.44	-3.42	4.04	3.17	0.81	1.64	0.31	1.61	0.19	0.21	**0.11**	0.73	0.75	0.76	0.79	0.39
Time	[s]	default	0.2	0.1	0.1	0.2	0.2	0.1	0.2	0.5	0.4	2.0	130.5	4.4	1091.8	700.9	321.8	6.8	6.0	24.9	11.6	20.3	132.	18.7	3.6	120.0
		tuned	0.2	1.8	0.1	0.5	0.2	0.1	0.2	0.3	0.5	0.8	98.1	3.9	1044.6	597.1	73.5	23.9	2.5	15.4	8.7	278.5	83.7	9.4	1.6	99.0

Bijlage B

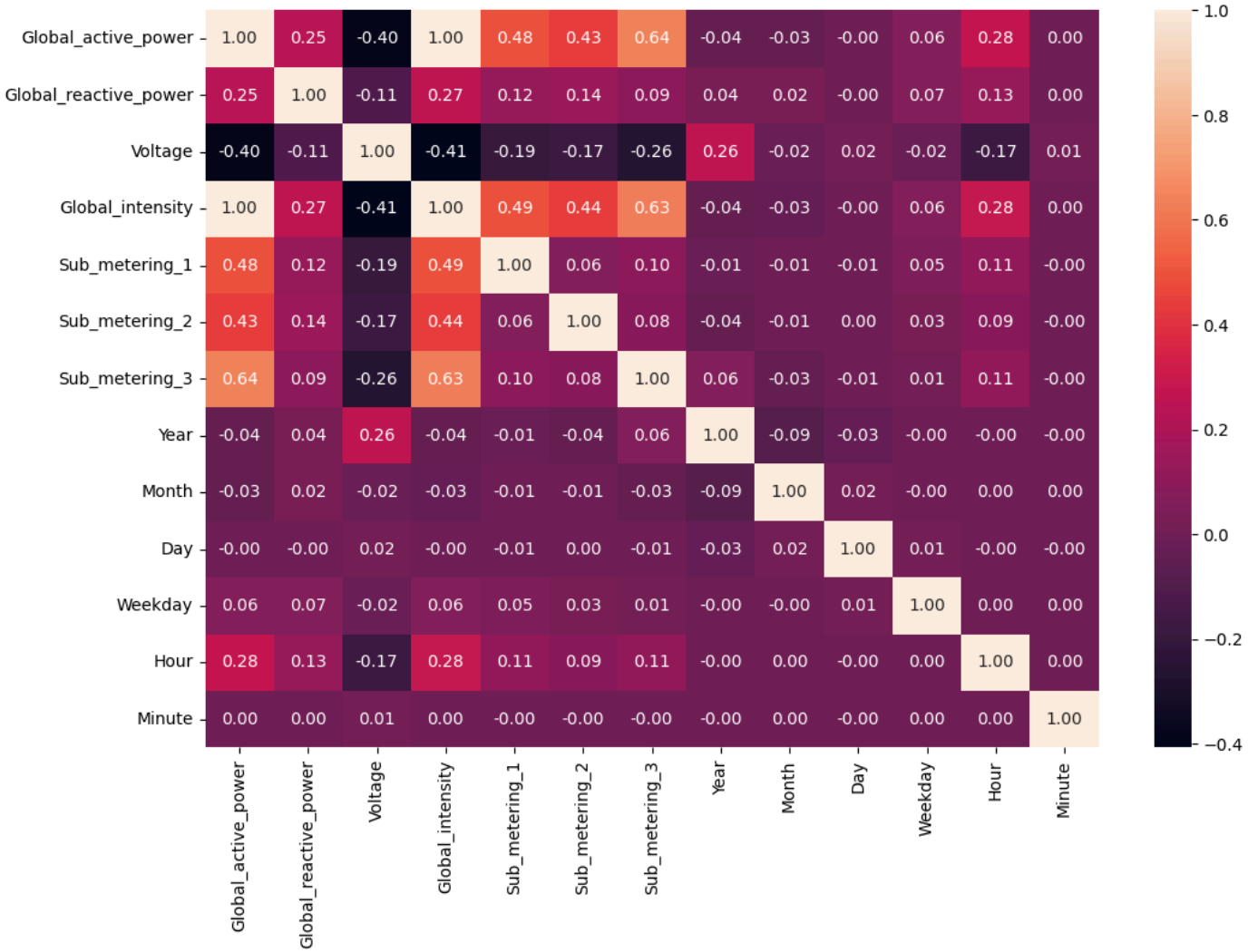
ARIMA Model Selectie

Optimalisatie van AIC voor Tijdreeksen Voorspellingen

ARIMA Model	AIC	Tijd (seconden)
(1,1,0)(0,0,0)[0] intercept	42.181.571	3.32
(0,1,1)(0,0,0)[0] intercept	42.181.269	5.22
(0,1,0)(0,0,0)[0] intercept	42.180.024	5.64
(0,1,0)(0,0,0)[0]	42.178.025	2.69
(1,1,2)(0,0,0)[0] intercept	31.414.052	29.78
(2,1,1)(0,0,0)[0] intercept	30.929.892	25.56
(3,1,1)(0,0,0)[0] intercept	30.751.928	29.28
(2,1,2)(0,0,0)[0] intercept	30.493.939	42.61
(3,1,2)(0,0,0)[0] intercept	30.488.280	60.70
(4,1,1)(0,0,0)[0] intercept	29.758.890	43.33
(4,1,2)(0,0,0)[0] intercept	29.660.460	58.09
(4,1,2)(0,0,0)[0]	29.658.463	28.72
(5,1,1)(0,0,0)[0] intercept	29.411.174	67.73
(5,1,2)(0,0,0)[0] intercept	28.952.685	113.46
(5,1,2)(0,0,0)[0]	28.950.688	57.91
(4,1,3)(0,0,0)[0] intercept	28.843.151	104.87
(4,1,3)(0,0,0)[0]	28.841.004	58.69
(4,1,4)(0,0,0)[0] intercept	28.682.827	153.79
(4,1,4)(0,0,0)[0]	28.680.831	67.94
(5,1,4)(0,0,0)[0] intercept	28.637.094	132.97
(5,1,3)(0,0,0)[0] intercept	28.636.026	113.90
(5,1,4)(0,0,0)[0]	28.635.012	69.04
(5,1,3)(0,0,0)[0]	28.633.997	66.04

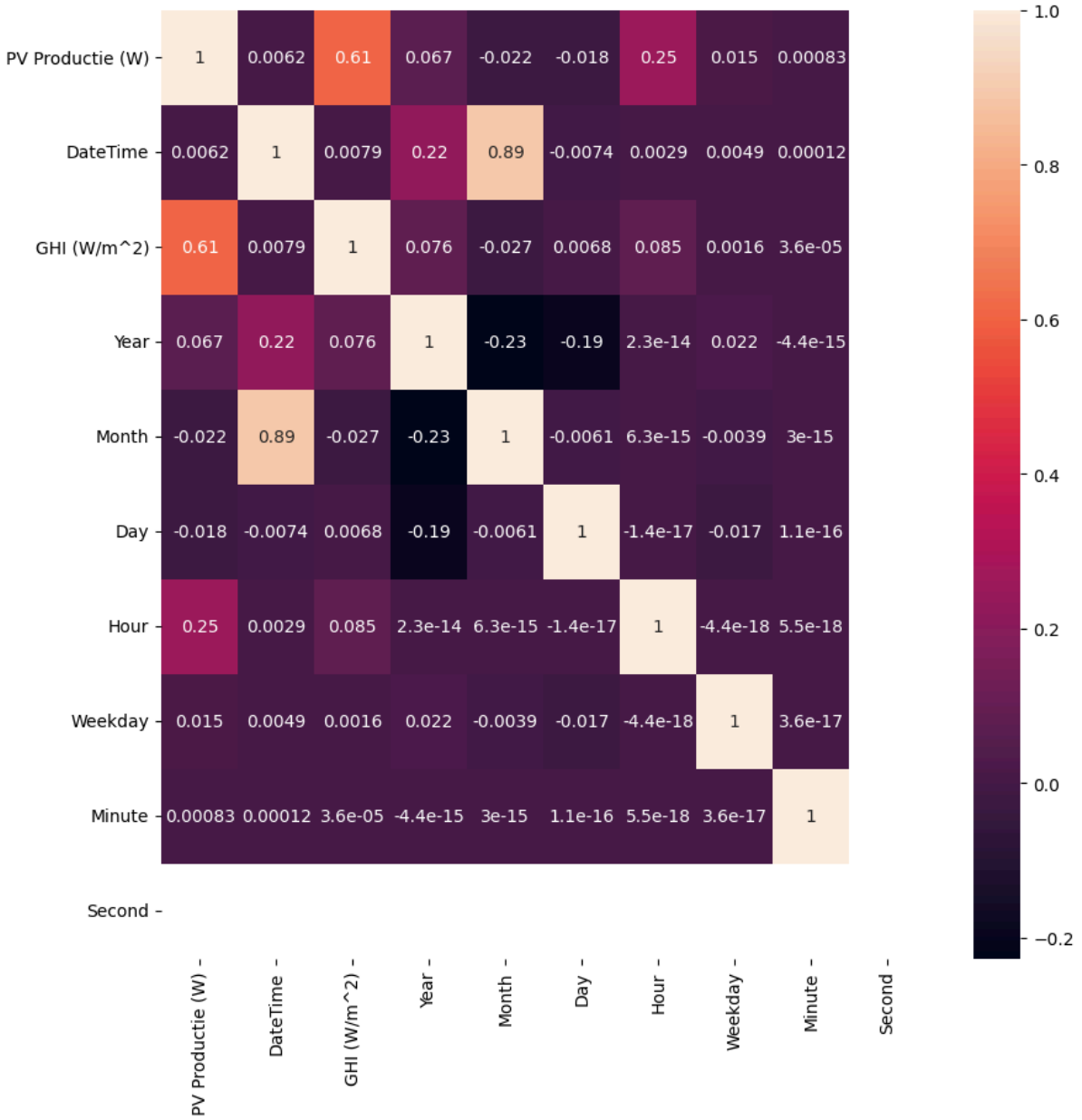
Bijlage C

Household Power Consumption | Heatmap



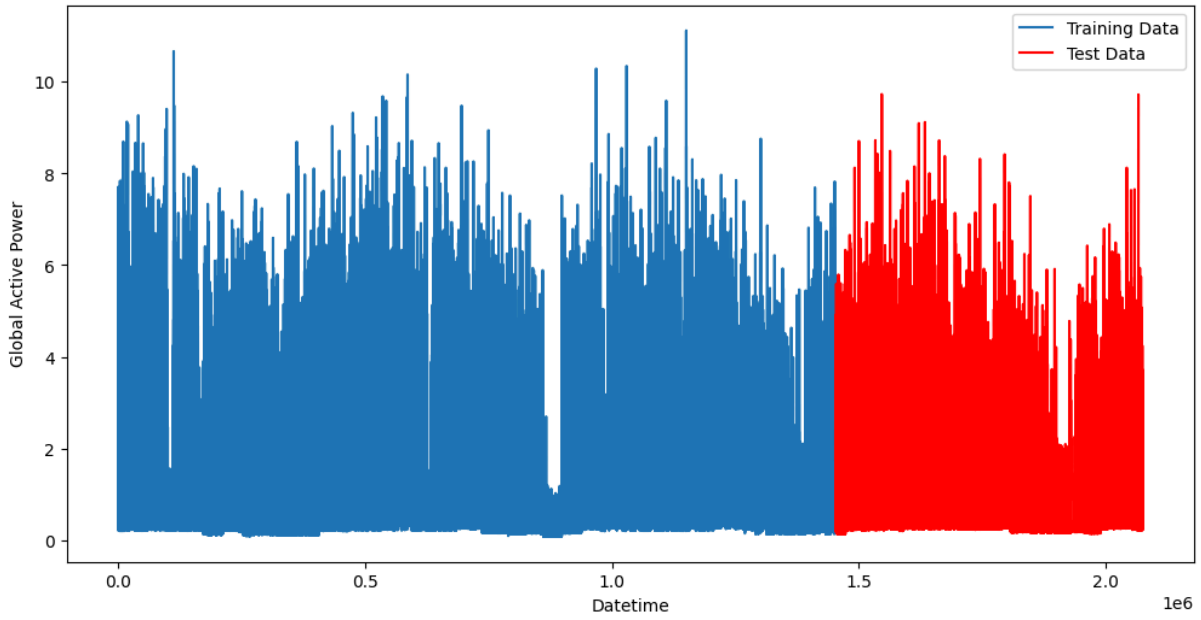
Bijlage D

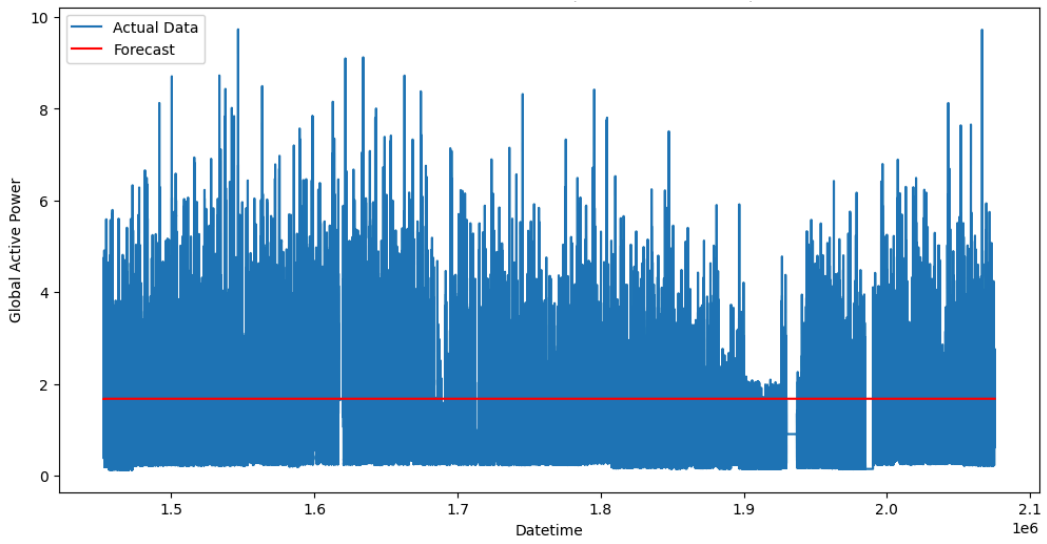
2022_15min_data_with_GHI | Heatmap



Bijlage E

ARIMA | Household power consumption





GRU getraind op 10% van de dataset (duurt zeer lang om volledig te doen)

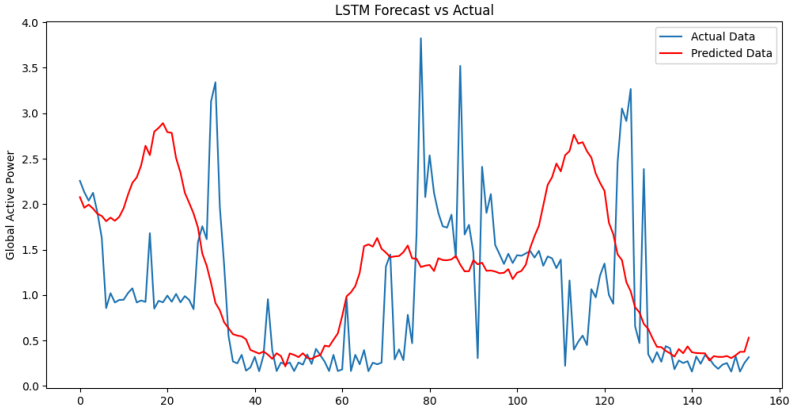
Mean Squared Error: 0.09288927338340065

Bijlage F

LSTM en GRU | Household power consumption

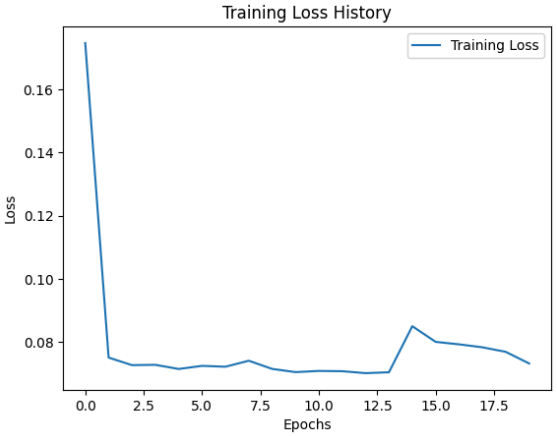
Epoch	*LSTM loss*	*GRU loss*
1	*0,1748*	*0.1797*
2	*0.0750*	*0.0787*
3	*0.0726*	*0.0753*
4	*0.0727*	*0.0735*
5	*0.0714*	*0.0724*
6	*0.0724*	*0.0719*
7	*0.0721*	*0.0715*
8	*0.0740*	*0.0715*
9	*0.0714*	*0.0709*
10	*0.0704*	*0.0708*
11	*0.0708*	*0.0712*
12	*0.0707*	*0.0702*
13	*0.0701*	*0.0709*
14	*0.0703*	*0.0695*
15	*0.0850*	*0.0691*
16	*0.0800*	*0.0688*
17	*0.0792*	*0.0769*
18	*0.0783*	*0.0766*
19	*0.0768*	*0.0713*
20	*0.0731*	*0.0698*

LSTM Forecast op testset

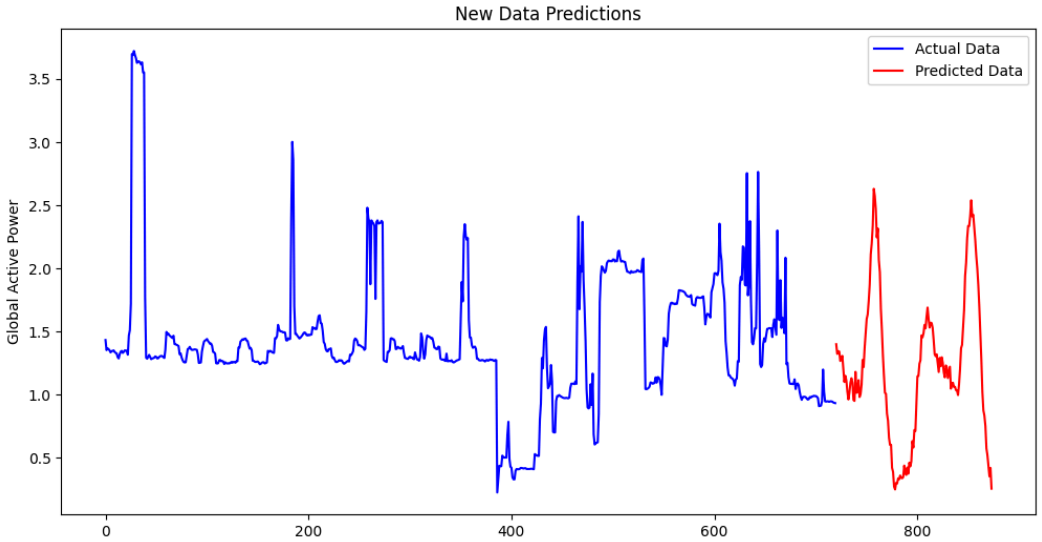


Mean Absolute Error: 0.08773131750727241 (kilowatt)

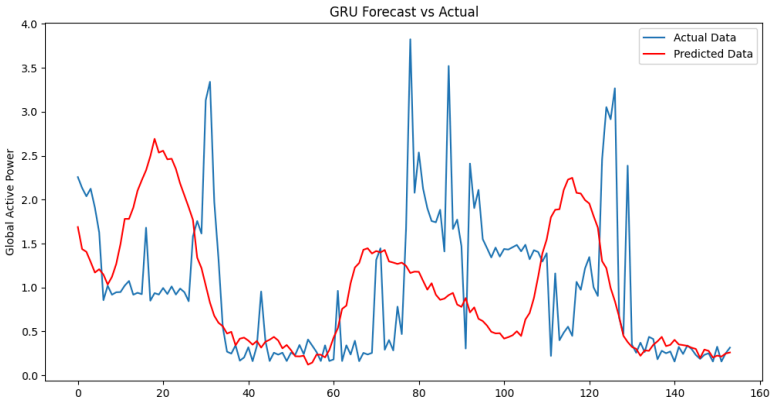
Loss LSTM



Forecast buiten data uit dataset (echte voorspelling LSTM)

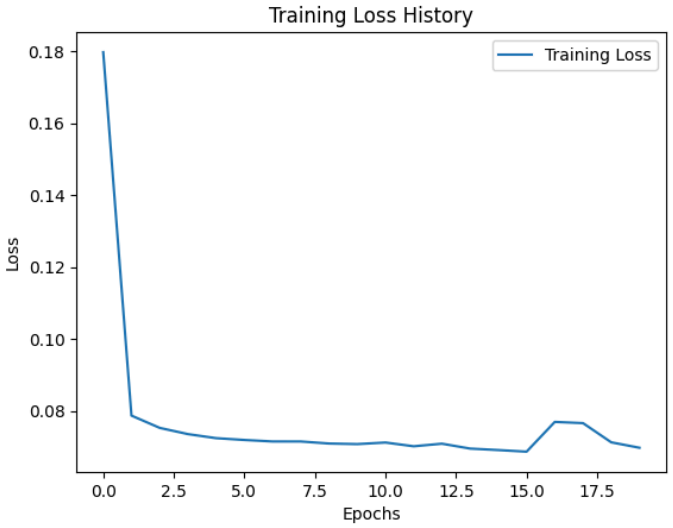


GRU Forecast op testset

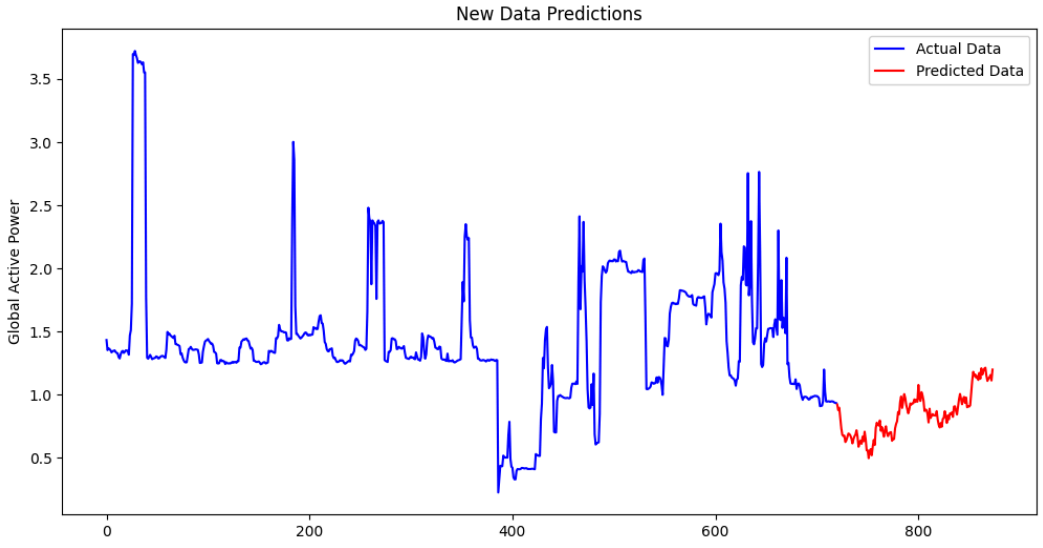


Mean Absolute Error: 0.10309738475432642 (kilowatt)

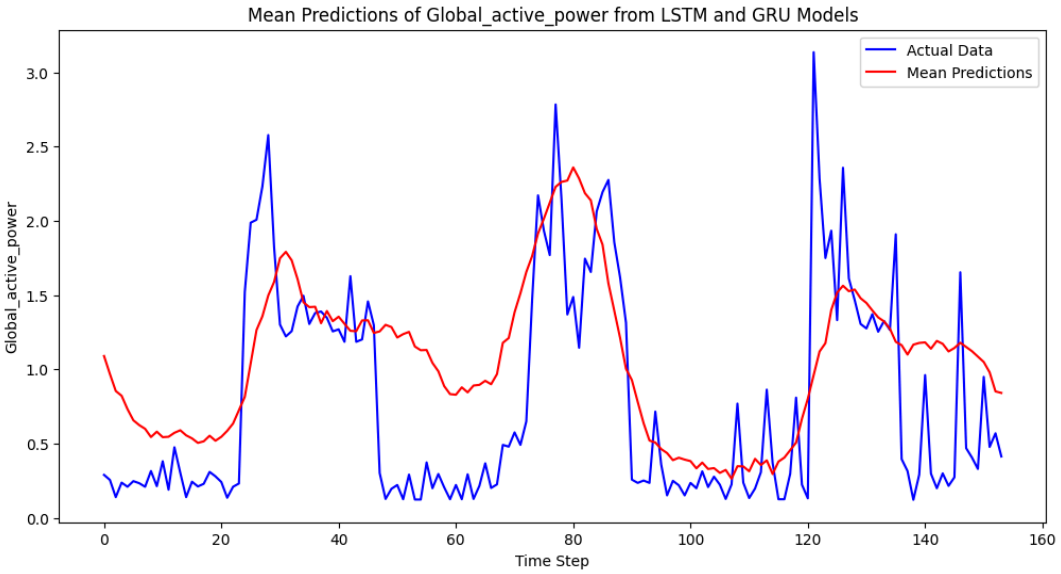
Loss GRU



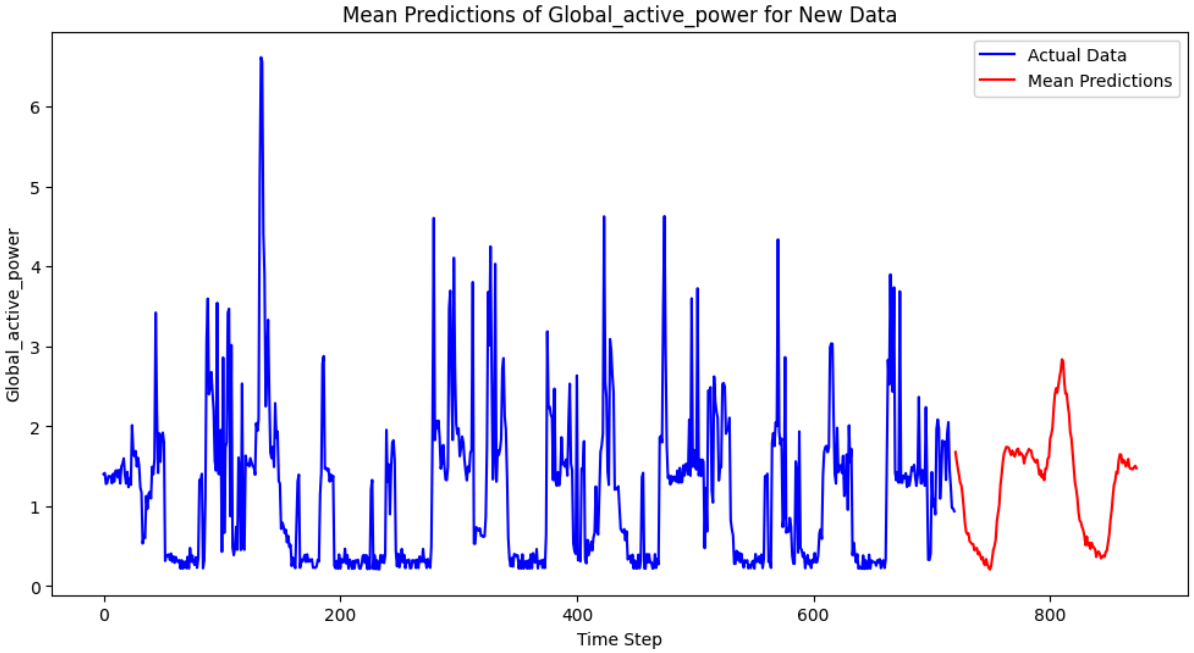
Forecast buiten data uit dataset (echte voorspelling GRU)



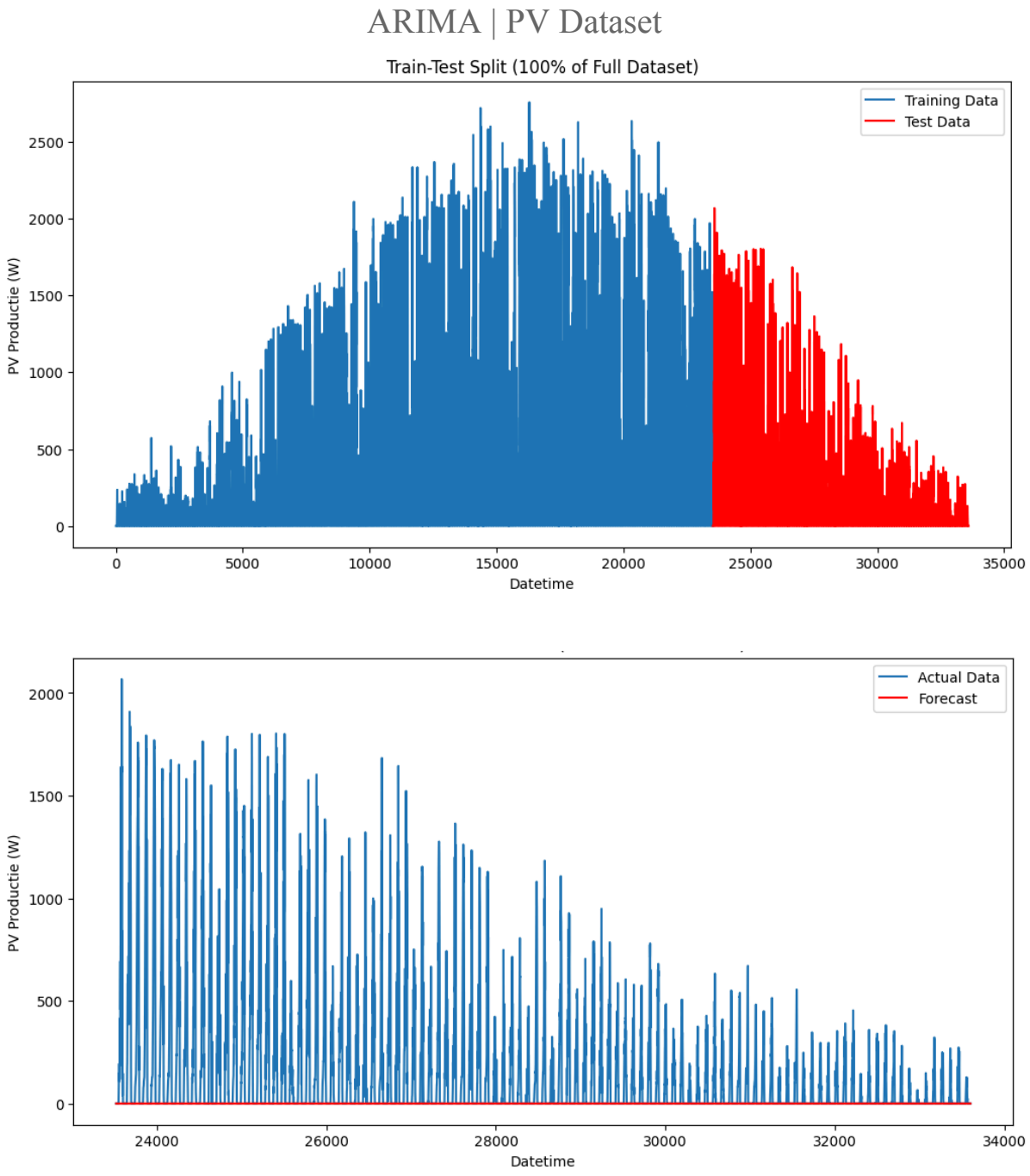
GRU en LSTM samen



Mean Absolute Error: 0.47356576407870116 (kilowatt)



Bijlage G



Er is te zien dat het ARIMA model een horizontale rechte lijn voorspelt, dit duidt erop dat het onbruikbaar is aangezien het geen voorspellingen maakt die in de buurt komen van de realiteit.

Bijlage H

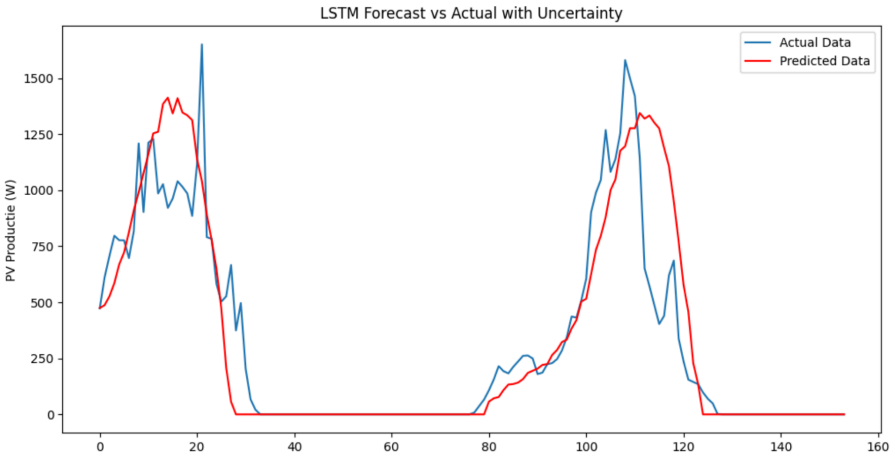
LSTM en GRU | PV Dataset

Loss per epoch

Epoch	*LSTM loss*	*GRU loss*
1	*0.2862*	*0.2945*
2	*0.0598*	*0.0631*
3	*0.0574*	*0.0589*
4	*0.0560*	*0.0569*
5	*0.0555*	*0.0555*
6	*0.0547*	*0.0551*
7	*0.0541*	*0.0544*
8	*0.0537*	*0.0538*
9	*0.0531*	*0.0539*
10	*0.0528*	*0.0533*
11	*0.0524*	*0.0531*
12	*0.0520*	*0.0528*
13	*0.0519*	*0.0526*
14	*0.0512*	*0.0522*
15	*0.0515*	*0.0520*
16	*0.0514*	*0.0521*
17	*0.0534*	*0.0517*
18	*0.0520*	*0.0514*
19	*0.0507*	*0.0512*
20	*0.0504*	*0.0511*

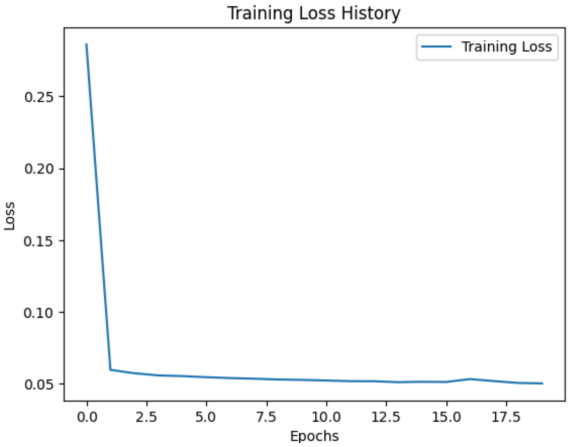
LSTM Forecast

Forecast op testset

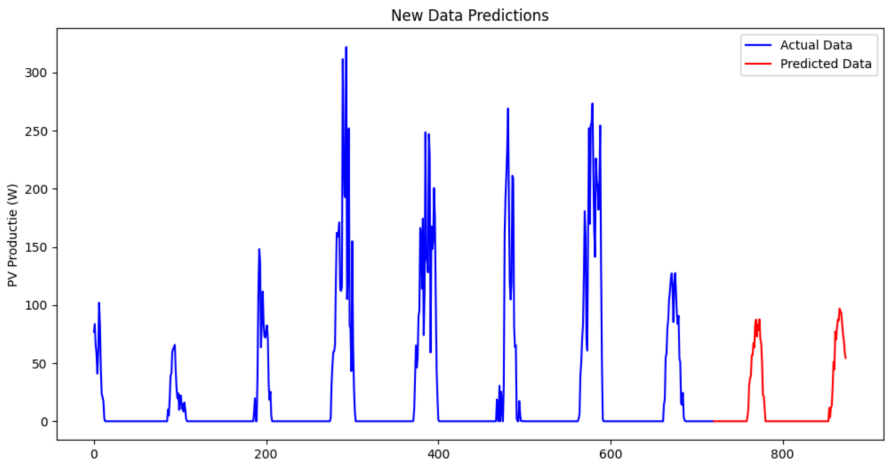


Mean Absolute Error: 137.9038284167211 (watt)

Loss LSTM

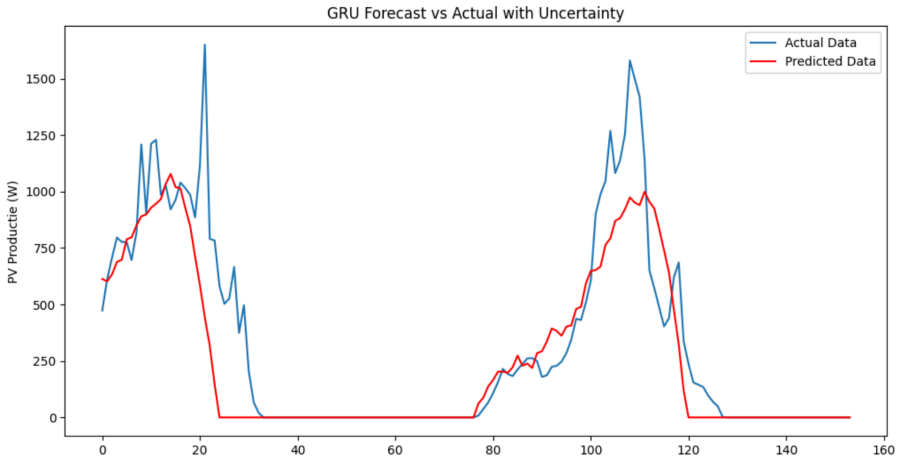


Forecast buiten data uit dataset (echte voorspelling)



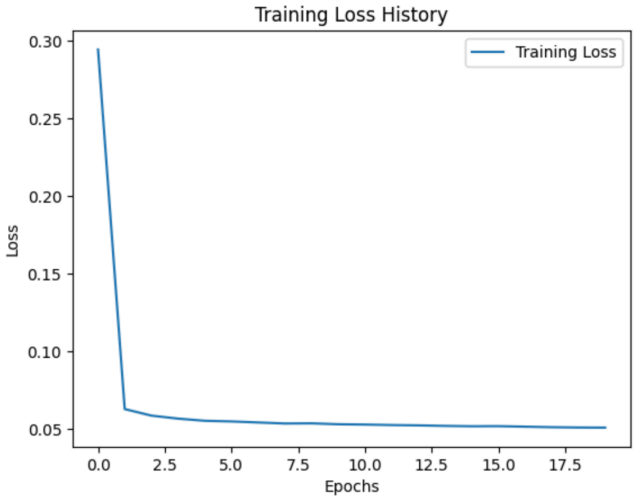
GRU Forecast

Forecast op testset

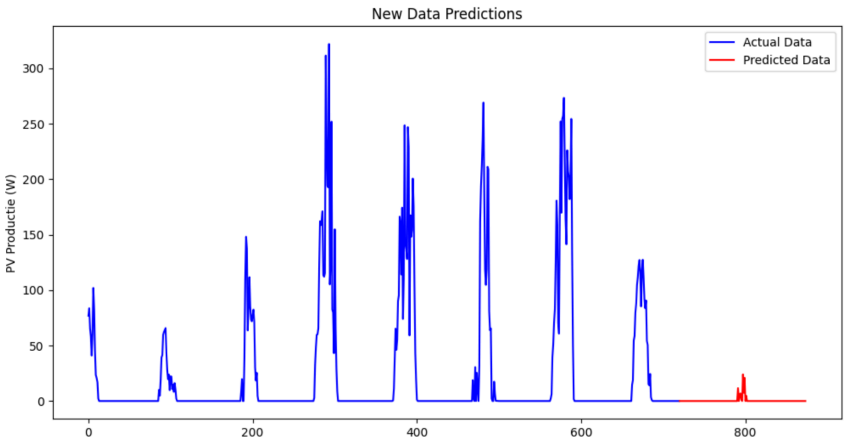


Mean Absolute Error: 110.92523288719134 (watt)

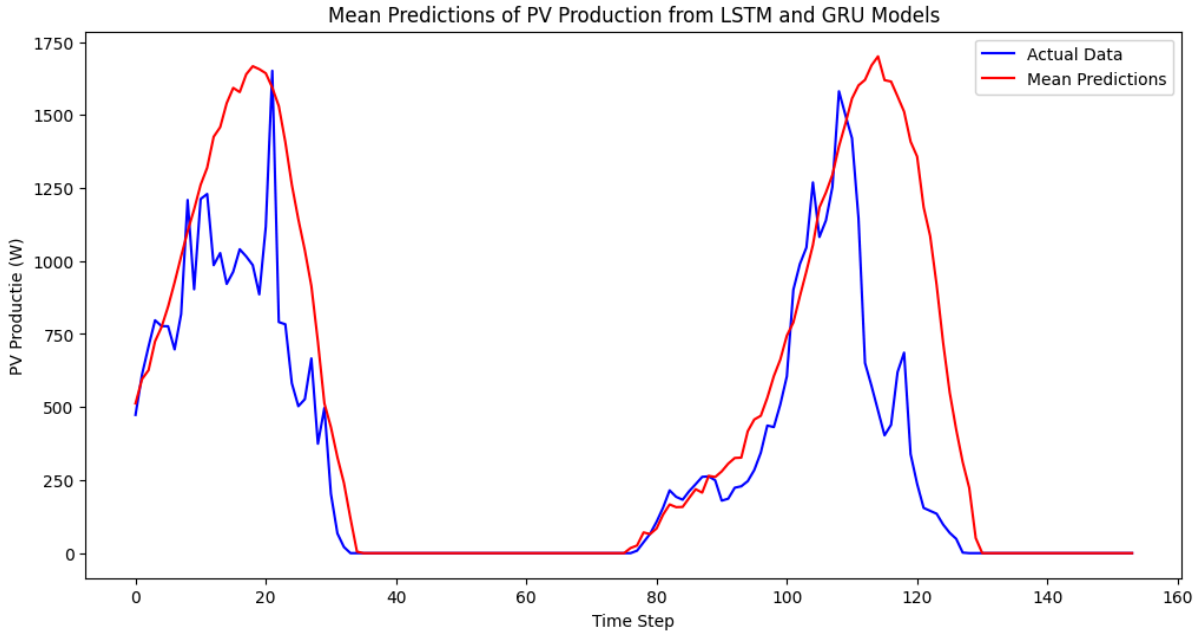
Loss GRU



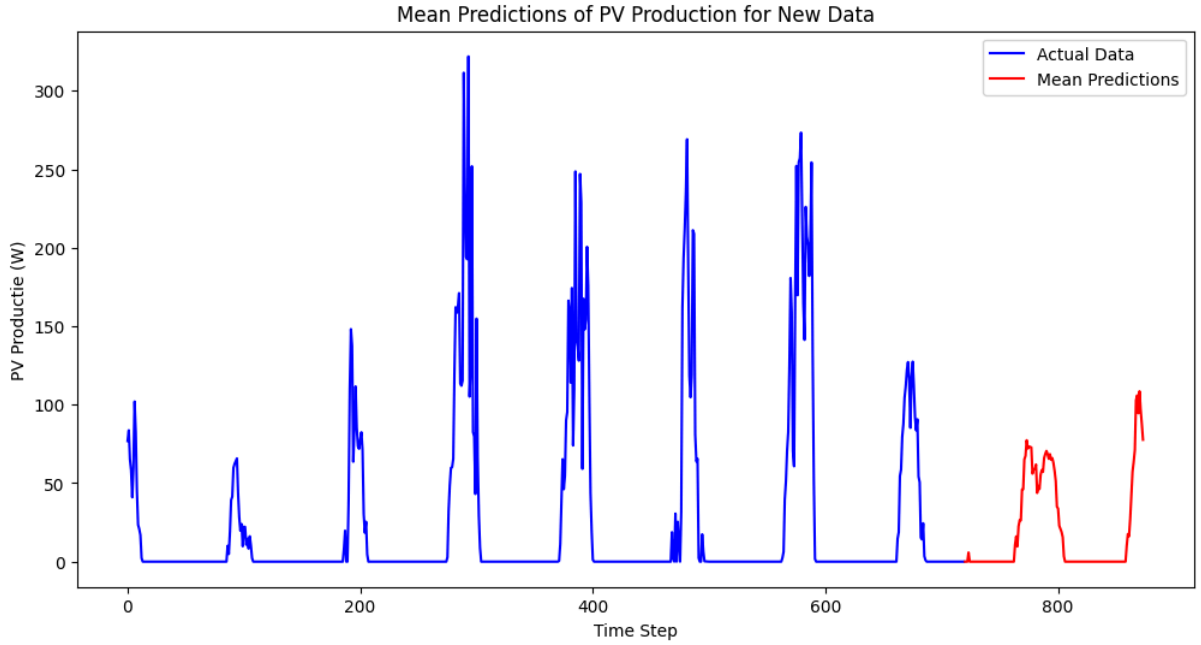
Forecast buiten data uit dataset (echte voorspelling)



GRU en LSTM samen



Mean Absolute Error: 188.10287973719653 (watt)



Bijlage I

Process tabel

Week	Status	Gebeurtenis
1	Geen dataset beschikbaar.	Geen specifieke acties ondernomen vanwege het ontbreken van data.
2	Nog steeds geen dataset beschikbaar.	Er is een API beloofd. Een gesprek is gepland met Martien om verdere stappen te bespreken.
3	Nog steeds geen dataset of API ontvangen.	Verkenning uitgevoerd op Martiens device om te kijken of het gebruikt kan worden.
4	Eerste data ontvangen.	Emap-bestanden ontvangen. De opdracht en verwachtingen zijn duidelijk gecommuniceerd naar Tim en Martien. Tim heeft beloofd een CSV-bestand te leveren. Een onvolledig CSV-bestand is ontvangen van Reinder.
5	Vertraagde levering van de dataset van Tim.	Start gemaakt met werken aan een open dataset als alternatief voor de ontbrekende gegevens.
6	Actieve analyse van beschikbare data.	Exploratory Data Analysis (EDA) uitgevoerd op de open dataset om de kwaliteit en bruikbaarheid van de data te beoordelen en om mogelijke inzichten te identificeren. Alle modellen uitgewerkt behalve de transformer
7	Werkelijke data ontvangen	EDA uitgevoerd op de werkelijke dataset geleverd vanuit Tim om de kwaliteit en bruikbaarheid van de data te beoordelen voor mogelijke toekomstige groepjes (leidt tot aanbevelingsrapport).
8/9	Afronding product	De visualisatie is in deze weken afgerond en de modellen zijn voor de laatste keer volledig getraind.

Bijlage J

Requirements tabel

ID	Categorie	Beschrijving	Prioriteit	Voldaan?
F11	Voorspelling	Het systeem kan toekomstig energieverbruik voorspellen aan de hand van input data uit een relevante dataset.	Must	Ja
F12	Voorspelling	Het systeem kan toekomstige energieproductie voorspellen aan de hand van input data uit een relevante dataset.	Must	Ja
F31	Visualisatie	Het systeem weergeeft huidige data en visualiseert de voorspelde data in een webapplicatie.	Must	Ja
F41	Functionaliteit	Het systeem geeft opties om energie te handelen met bedrijven. Op basis van het voorspelde energieverbruik en de voorspelde energieopbrengst.	Could	Nee
F42	Functionaliteit	Het systeem zorgt voor de aankoop van extra energie wanneer energieverbruik boven energieproductie dreigt te stijgen.	Could	Nee
F51	Opslag	Het systeem slaat alleen voor een bepaalde tijd data op om zo te veel opslagverbruik te voorkomen.	Should	Ja

Bijlage K

Dataset kenmerken

Naam	Interval	Records	Features	Tijd	Formaat
Dataset Enerlynk*	N/A	N/A	N/A	N/A	.emap, .a, .b
Dataset opdrachtgever	1 dag	154	8	1 januari 2024 30 mei 2024	.csv
Household Electric Power Consumption	1 minuut	2.075.261	8	16 december 2006 26 november 2010	.csv
2022_15min_data , Amstelveel dataset	15 minuten	33601	2	26 december 2021 10 december 2022	.csv
Energy Usage From DOE Buildings	-	-	-	-	.csv
PVDAQ	-	-	-	-	.csv
Data Platform	-	-	-	-	.csv
Dataset Sqippa - Eastron SDM230	15 minuten	-	24	31 januari 2022 3 juni 2024	.csv
Dataset Sqippa - Eastron SDM630	1 uur	-	24	29 januari 2024 3 juni 2024	.csv
Dataset Sqippa - Milesight WS522	20 minuten	-	24	13 juni 2022 3 juni 2024	.csv
Dataset Sqippa - P1 Sensor	variabel**	43.519	40	-	.csv

Notitie:

* Het was niet mogelijk om uit de dataset van Enerlynk data te extraheren. De formaten zijn niet te lezen door open-source software en is alleen uit te lezen door Huawei zelf. Daarom wordt deze niet gebruikt als dataset en zijn de kenmerken N/A .

** De hardware voor de P1 metingen gebruikt LORA en heeft daarom inconsistente tijdsintervallen.

Bijlage L

Standaard modellen | Household Electric Power Consumption

Linear Regression

Mean Squared Error	1.0161300098782997
Root Mean Squared Error	1.0080327424634081
r^2 score	0.08437353750557064

Een gemiddelde afwijking van ongeveer 1 kilowatt is niet acceptabel, dit model zal niet werken. Overigens is de r2 score ook erg laag en is het model dus niet bruikbaar.

Decision Tree Regression

Mean Squared Error	0.32019260886867207
Root Mean Squared Error	0.5658556431358374
r^2 score	0.7114770522224829

De mean squared errors zijn net acceptabel en de r2 score is matig tot goed. Dit model zou bruikbaar kunnen zijn, maar bevat niet de mogelijkheid om een een sequentie voorspellingen uit te voeren.

Random Forest

Mean Squared Error	0.3202778076427239
Root Mean Squared Error	0.5659309212640037
r^2 score	0.7114002803022211

Voor random forest geldt dezelfde conclusie als voor decision tree.

Een aantal methoden van Support Vector Machines (SVR, NuSVR, LinearSVR)

SVR

Mean Squared Error	1.6083436903968689
Root Mean Squared Error	1.268204908678747
r^2 score	0.005018400567125436

Uit het Support Vector Regressiemodel is te concluderen dat de mean squared errors onacceptabel groot zijn en de r2 score onacceptabel laag is. Het support vector regressiemodel is onbruikbaar, ook deze zou geen sequenties kunnen voorspellen.

NuSVR

Mean Squared Error	1.5727541439064692
Root Mean Squared Error	1.2540949501160066
r^2 score	0.027035425971298044

Uit de Nu Support Vector Regression methode is hetzelfde te concluderen als uit het standaard Support Regression Model.

LinearSVR

Mean Squared Error	3.1129607611038046
Root Mean Squared Error	1.7643584559561032
r^2 score	-0.9257940299380629

Het Lineaire Support Vector Regressiemodel is met uitstek het slechtste van de eerder genoemde modellen, met een Root Mean Squared Error van 1.76 kilowatt en een r2 score van maarliefst -0,92, dit betekent dat het nemen van het gemiddelde van de dataset een beter resultaat zal opleveren dan de voorspelling uitvoeren.

Bijlage M

Standaard modellen | 2022_15min_with_GHI (PV)

Linear Regression

Mean Squared Error	133828.17177795636
Root Mean Squared Error	365.82532960137735
r^2 score	0.4067401055388381

Het lineair regression model is niet bruikbaar bij deze dataset aangezien deze dermate slecht presteert dat het onbruikbaar is.

Decision Tree Regression

Mean Squared Error	55874.36257745038
Root Mean Squared Error	236.37758476101405
r^2 score	0.7523091139526206

Het Decision Tree regressie model zou bruikbaar zijn voor 1 enkele voorspelling doordat de scores wel acceptabel zijn. Echter is het nodig dat er in een sequentie data voorspeld kan worden.

Random Forest

Mean Squared Error	33901.193844403555
Root Mean Squared Error	184.122768403051
r^2 score	0.8497161067431465

Het Random forest regressie model presteert zichtbaar beter dan de enkele decision tree, maar hiervoor geldt dezelfde reden voor het niet overwegen van dit model.

Een aantal methoden van Support Vector Machines (SVR, NuSVR, LinearSVR)

SVR

Mean Squared Error	162026.193583936
Root Mean Squared Error	402.52477387601374
r^2 score	0.2817383572643054

Het support vector regression model is niet geschikt voor gebruik doordat de scores aanduiden dat de voorspellingen te ver en te onverklaarbaar van de werkelijke waarden af liggen.

NuSVR

Mean Squared Error	161106.29797306386
Root Mean Squared Error	401.38049027458203
r^2 score	0.28581625181947246

Voor het geteste Nu Support Vector Regressie model geldt dezelfde conclusie als voor de standaard Support Vector Regressie toepassing.

LinearSVR

Mean Squared Error	151778.80220903436
Root Mean Squared Error	389.5879903295716
r^2 score	0.327165013287546

Ook voor deze instantie geldt het feit dat SVR niet toepasbaar is op de vorm van de dataset.