Model Experimenten

Digital Twin 3.0

Handpicked Agencies

Breda

Door: Koen Pijnenburg

Introductie

Voor de uitbreiding van het Twindle project zullen machine learning modellen gebruikt worden om de luchtkwaliteit te voorspellen. Dit zullen modellen zijn voor luchtvochtigheid, temperatuur, CO2 en TVOC. In dit document zal voor ieder model toegelicht.

Inhoudsopgave

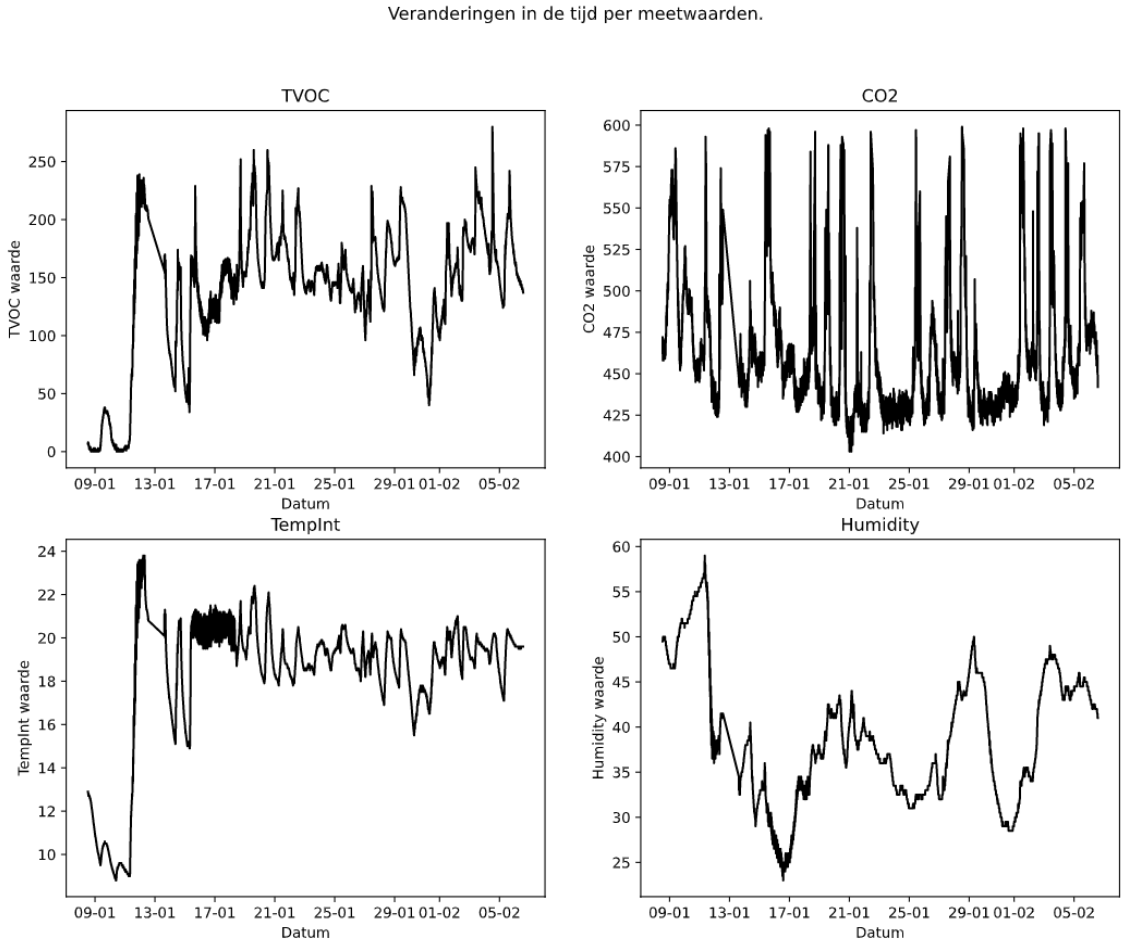
Gegevensoverzicht

Om te bepalen welke doelen, evaluatiemethoden en modellen toegepast kunnen worden zullen de patronen in de luchtvochtigheid, temperatuur, CO2 en TVOC verder onderzocht worden.

Aangezien deze waarden tijdreeksen zijn zullen de patronen onderzocht worden door middel van, onder andere, tijdreeksgrafieken, scatterplots en correlelogrammen.

Tijdreeksen

De onderstaande tijdreeksen geven het verloop van de verschillende meetwaarden in de tijd aan. Dit geeft een eerste inzicht in eventuele patronen, correlaties en seizoensgebondenheid.

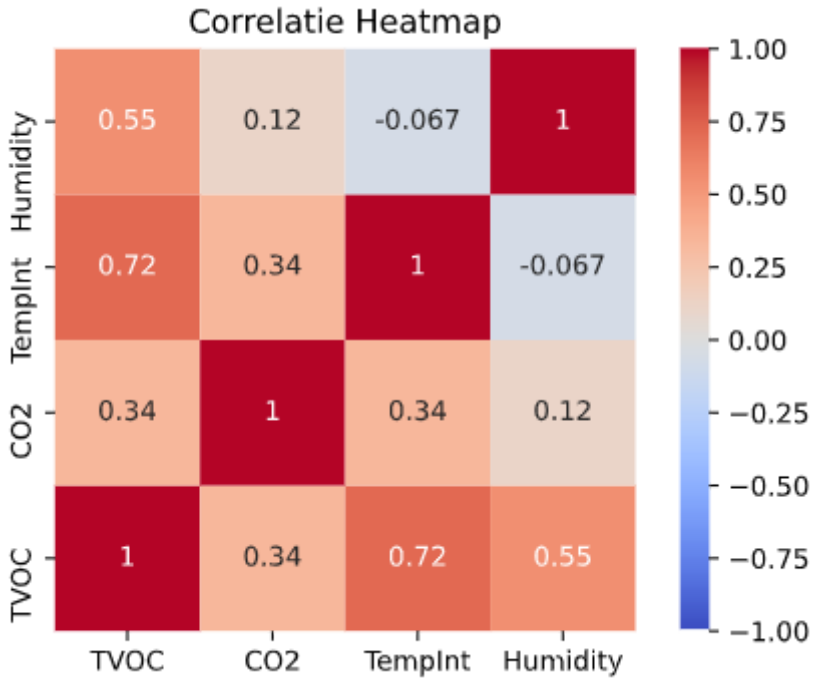


Afbeelding 1: Veranderingen in tijd.

Hierin valt te zien dat alle meetwaarden waarschijnlijk seizoensgebonden zijn maar geen trend bevatten, dit zal in de komende hoofdstukken verder onderzocht worden. Opvallend is dat er veel anomalieën plaats hebben gevonden tussen 09-01 en 13-01. In deze periode is er gewisseld van message queue systeem waardoor sommige metingen niet kloppen. Tijdens het trainen van het model zullen deze data waarschijnlijk buiten beschouwing gelaten moeten worden om ruis te voorkomen.

Correlaties

Het kan zijn dat een of meerdere meetwaarden aan elkaar of zichzelf gerelateerd zijn. Dit zal invloed hebben op welke modellen toepasbaar zijn en welke data gebruikt kan worden.

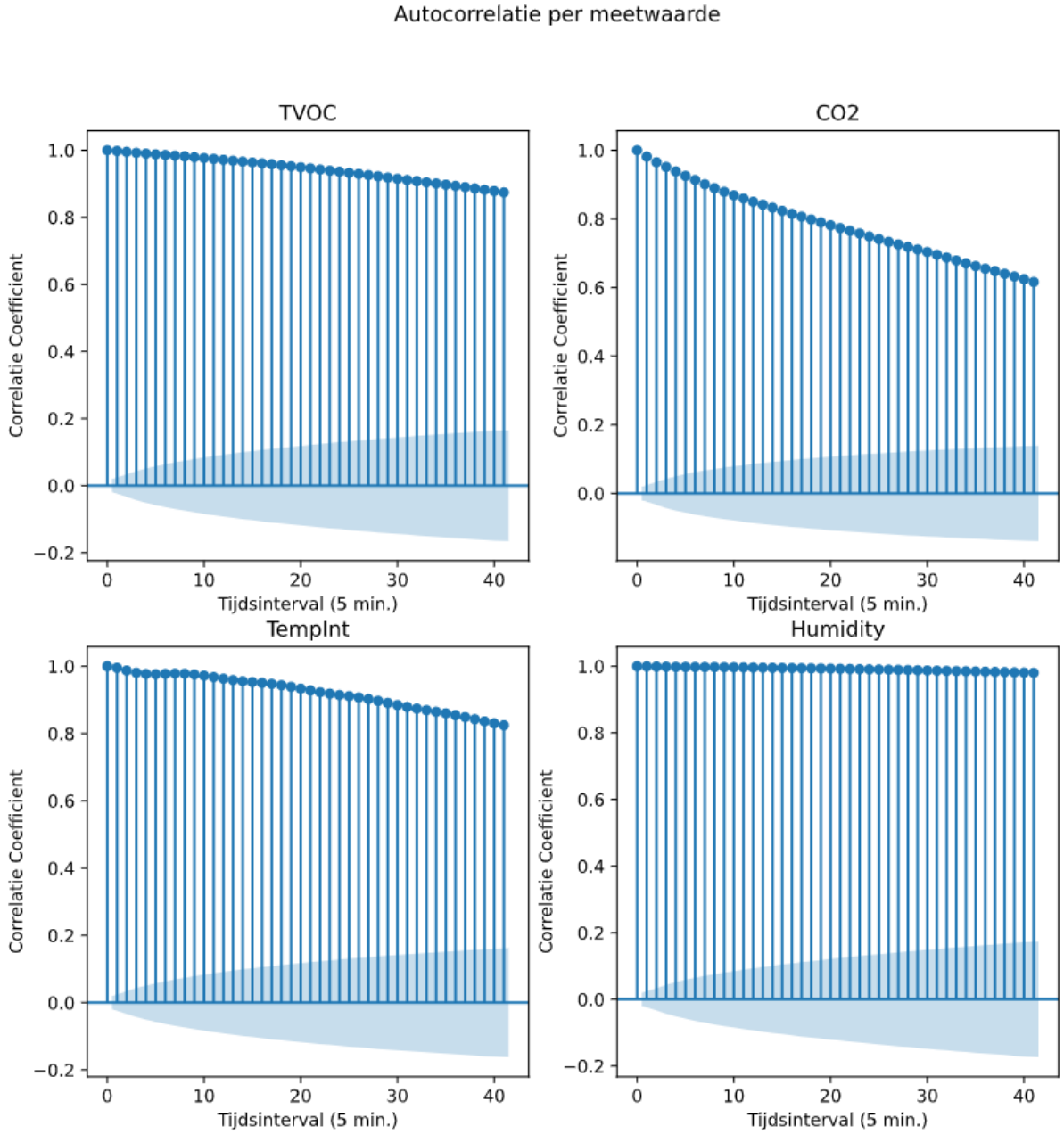


Afbeelding 2: Onderlinge correlaties

Uit de bovenstaande heatmap kan geconcludeerd worden dat er, positieve, onderlinge correlaties aanwezig zijn. In de onderstaande tabel kan een opsomming van deze correlaties gevonden worden.

Meetwaarde	**Erg sterk (> 0.7)**	**Sterk (> 0.5)**	**Zwak (< 0.5)**
Luchtvochtigheid		TVOC	CO2, Temperatuur
Temperatuur	TVOC		CO2, Luchtvochtighied
CO2			TVOC, Temperatuur, Luchtvochtigheid
TVOC	Temperatuur	Luchtvochtigheid	CO2

Bij tijdreeksen moet er rekening gehouden worden met autocorrelatie. Dit betekend dat er een verband is tussen een meting op een bepaalde tijd en de metingen die er na gemaakt zijn. In de onderstaande correlelogrammen valt de relatie tussen een punt in de tijd en de veertig nakomende te zien.

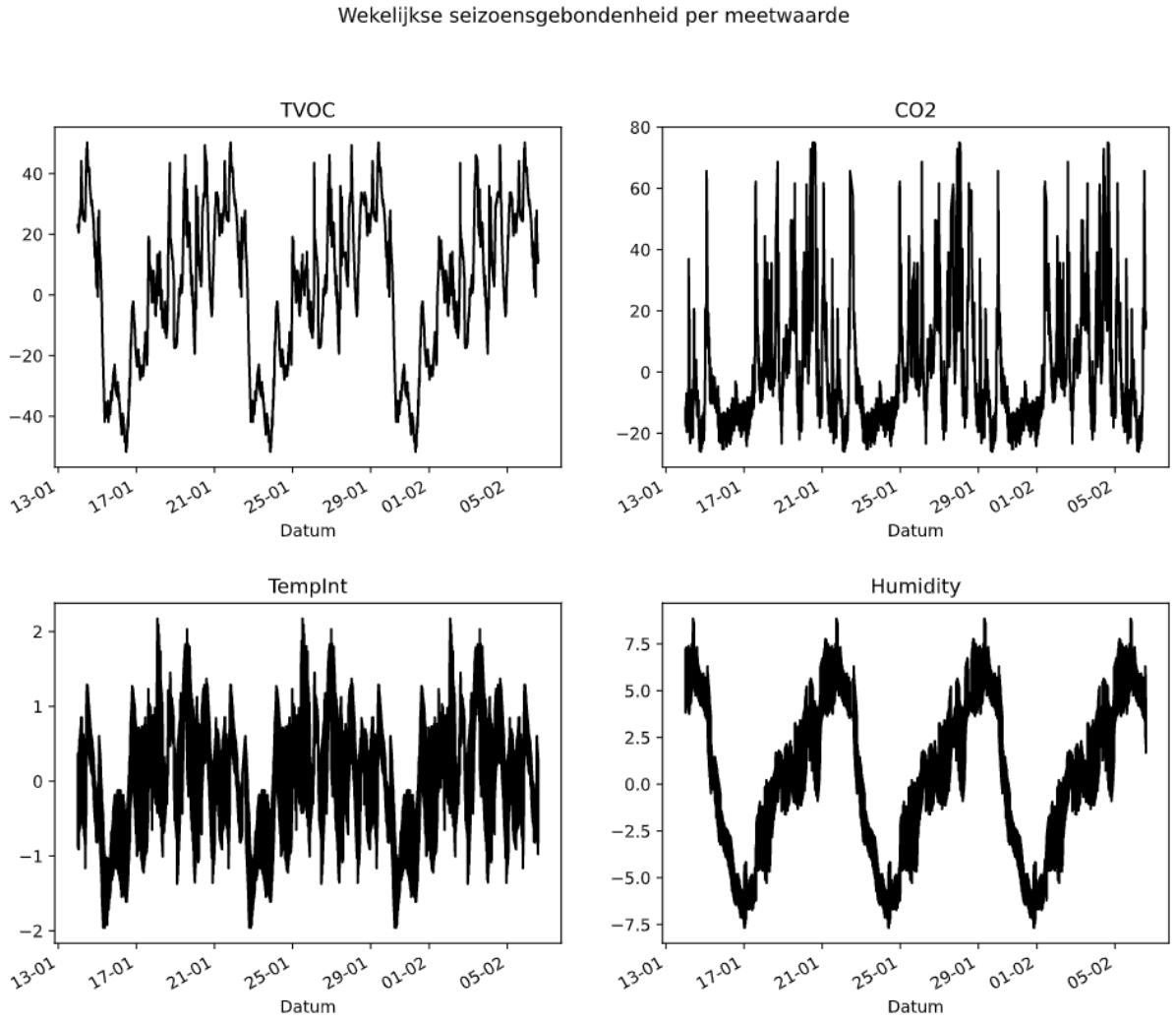


Afbeelding 3: Autocorrelatie

Er kan geconcludeerd worden dat alle meetwaarden erg sterke autocorrelaties bevatten. Dit kan voor problemen zorgen tijdens normale least-squares regressie technieken. Er zal dus gezocht moeten worden naar technieken om deze auto correlatie te verwijderen of algoritmes die hier juist gebruik van kunnen maken.

Seizoensgebondenheid

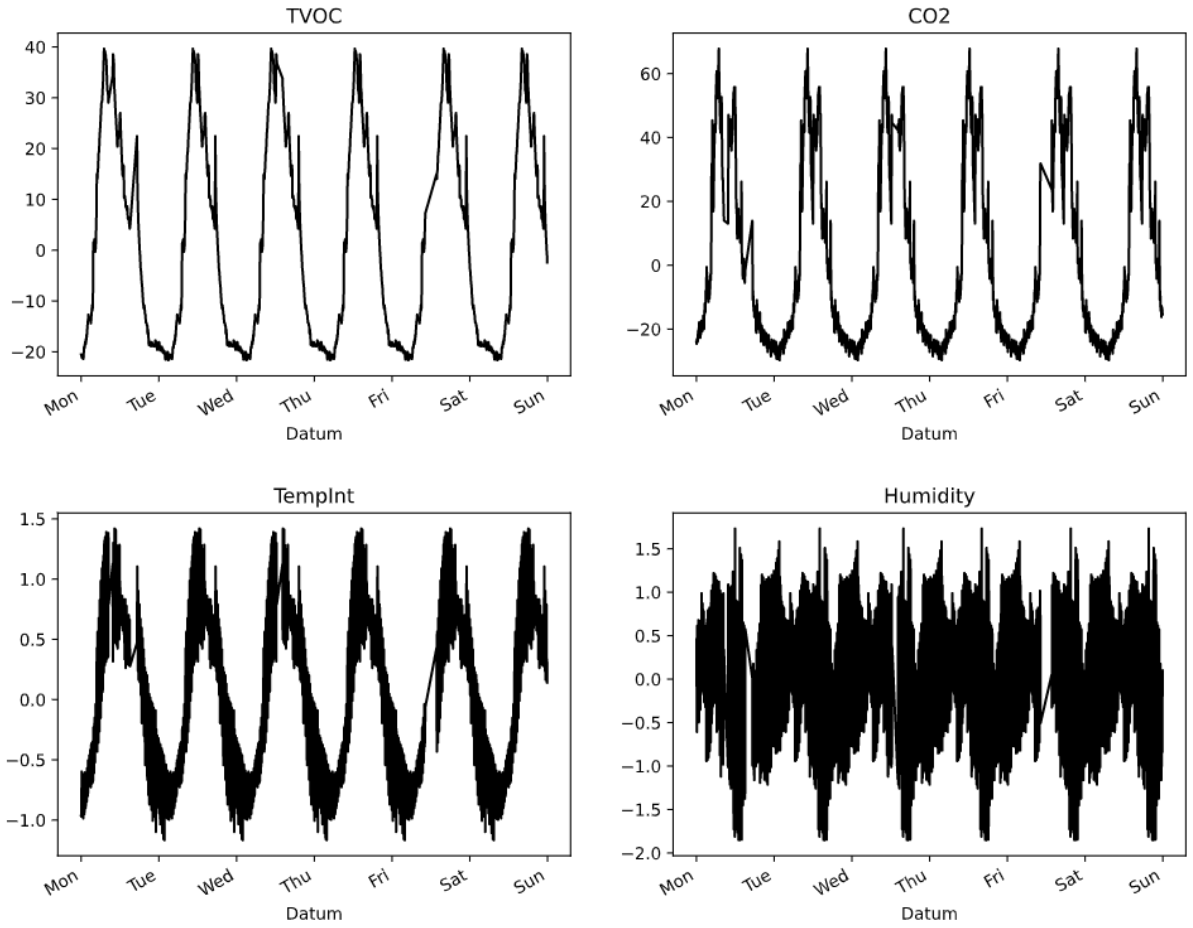
Net zoals bij autocorrelatie verwachten traditionele regressie technieken dat de data niet seizoensgebonden is of trends bevat, ook wel stationair genoemd. Door gebruik te maken van de ‘seasonal_decompose’ methode uit het statsmodels package kunnen seizoensgebondenheid en trends gevisualiseerd worden.



Afbeelding 4: Wekelijkse Seizoensgebondenheid

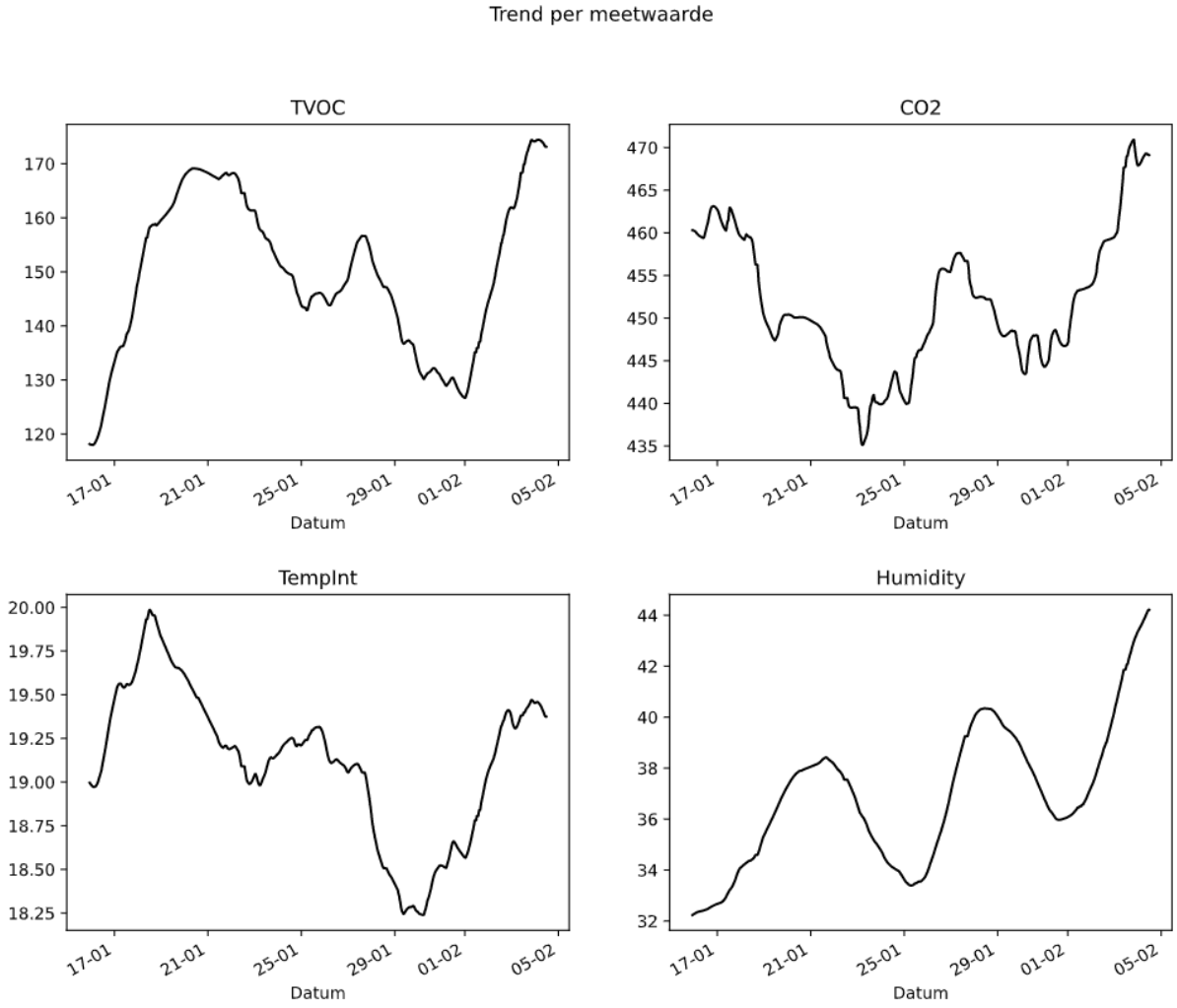
Aan de hand van de bovenstaande grafieken kan geconcludeerd worden dat alle meetwaarden wekelijks of dagelijkse cyclusen vertonen. Zie onderstaande afbeelding voor de dagelijkse cycli.

Dagelijkse seizoensgebondenheid per meetwaarde.



Deze patronen kunnen verklaart worden door het feit dat er doordeweeks meer personen aanwezig zijn dan in het weekend en de temperatuur en airconditioning centraal geregeld worden.

Zoals in de onderstaande afbeelding gezien kan worden zijn er geen sterke trends in de TVOC, CO2 en temperatuur waarden aanwezig. De luchtvochtigheid lijkt een kleine positieve trend te volgen.



De stationariteit van een tijdreeks kan verder verduidelijkt worden door de Dickey-Fuller test uit te voeren. Deze methode resulteert een een waarde, p-waarde, die nul is bij stationariteit. De meeste machine learning algoritmes verwachten dat deze waarde niet boven de 0.05 uitkomt.

Meetwaarde	**p-waarde**	**Stationair**
TVOC	0.000212	Ja
CO2	0.000000	Ja
Temperatuur	0.000000	Ja
Luchtvochtigheid	0.270564	Nee

Deze test bevestigt de eerdere aanname dat alle waarden behalve luchtvochtigheid stationair zijn.

Bevindingen

De gegevens die verwerkt moeten worden zijn tijdreeksen. Dit soort tijdgebonden data moet aan specifieke eisen voldoen om gebruikt te kunnen worden in traditionele regressie en classificatie algoritmes.

Hiervoor moet de data niet seizoensgebonden zijn, geen trends bevatten en stationair zijn. De meeste meetwaarde voldoen niet aan deze eisen. Dit betekend dat de data getransformeerd dient te worden voor deze gebruikt wordt. Een voorbeeld van zo'n transformatie is de verschillen tussen twee meetpunten berekenen en zo de trends en seizoensgebondenheid te verminderen.

Een andere optie is om algoritmen te gebruiken die juist gebruik kunnen maken van deze informatie.

Eisen

In dit hoofdstuk zal worden toegelicht aan welke eisen de modellen moeten voldoen. Dit zal gedaan worden door de evaluatie methoden te beschrijven en doelen te stellen waar de modellen aan moeten voldoen.

Doelen

Vanuit de product owner is het doel gesteld dat de modellen tenminste 90% accuraat moeten zijn. De meetwaarden die gemodelleerd dienen te worden zijn continu en niet uit te drukken in procent accuracy. R2 Score kan gebruikt worden voor dit soort modellen.

Het geeft een nummer tussen 0.0 en 1.0 aan wat geïnterpreteerd kan worden als een percentage. Wanneer een R2 Score van 0.9 wordt behaald zal het model aan de eisen van de product owner voldoen.

Wel moet bij deze methode goed gecontroleerd worden op overfitting. De methode die hiervoor gebruikt zal worden wordt verder besproken in het volgende hoofdstuk.

Evaluatiemethoden

Om te kunnen bepalen hoe goed een model presteert zullen meerdere evaluatiemethoden worden toegepast. Deze kunnen onderverdeeld worden in datavoorbereiding en scoring methodieken.

Datavoorbereiding

Om overfitting te voorkomen is het belangrijk dat de data die gebruikt wordt tijdens de training van het model niet wordt gebruikt tijdens het validatieproces. Aangezien dat de modellen voorspellingen maken over de toekomst zijn standaard cross validation en splitting technieken niet toepasbaar (Brownlee, 2019).

In plaats hiervan zal de eerste 70% van de data gebruikt worden om te modellen te trainen. Dit laat 30% van de data over die gebruikt zal worden om het model te valideren.

Scoring

De modellen zullen beoordeeld worden op twee scores; R2 Score, Root Mean Square Error (RMSE).

R2 Score

R2 is een score voor lineaire regressiemodellen. Deze waarde geeft het percentage van de variantie in de afhankelijke variabele aan dat de onafhankelijke variabelen gezamenlijk verklaren. R2 meet de sterkte van de relatie tussen uw model en de afhankelijke variabele op een handige schaal van 0,0 - 1,0.

RMSE

Dit is een cijfer dat in dezelfde eenheid is als het target. Geeft de gemiddelde foutmarge weer. Wordt gebruikt om duidelijk te maken in hoeverre de gemiddelde voorspelling kan afwijken.

Model Beschrijvingen

In dit hoofdstuk wordt een overzicht van de modellen gegeven. Dit wordt gedaan door de benodigde data, transformaties en outputs te benoemen.

Er zijn meerdere opties beschikbaar om de data te modelleren. Deze kunnen opgedeeld worden in de volgende categorieën; lineaire regressie, simpele forecasting methoden en complexe forecasting methode.

Lineaire Regressie

In een voorgaande versie van dit rapport was een eerste versie van de lineaire regressie modellen beschreven, zie *Bijlage 1: Lineaire Regressie (versie 1)*. Tijdens dit hoofdstuk zal hier naar gerefereerd worden.

Voorbereiding

Wanneer lineaire regressie algoritmen worden toegepast op tijdreeksen wordt er verwacht dat de dataset stationair is en er geen correlatie aanwezig is. Zoals eerder onderzocht is is dit niet het geval. De onderstaande technieken kunnen gebruikt worden om de data stationair te maken en autocorrelaties te verminderen (Hyndman & Athanasopoulos, 2018, p. 1) .

- Differencing; Het verschil tussen een meting en de voorgaande.
- Seasonal differencing; Het verschil tussen een meting vandaag en een meting van gisteren op dezelfde tijd.

Daarnaast zal de sequential data omgezet moeten worden naar paren van inputs en outputs (Brownlee, 2019b). Hiervoor zal de Pandas `shift()` methode toegepast worden. Onderstaand valt een voorbeeld te vinden van deze methode

Modellen

Standaard ‘least-squares’ passen het best bij tijdreeks data. De twee onderstaande technieken zullen worden toegepast.

- Linear regression
- Multiple linear regression

Naast de evaluatiemethoden zoals beschreven in het voorgaande hoofdstuk zal ook residual analyse worden toegepast. Tijdens deze analyse worden de residuals onderzocht op:

1. Autocorrelatie wat eventueel overgebleven seizoensgebondenheid aan duid.
2. Patronen die non-lineariteit aanduiden.

Exponential Smoothing

Voorspelling die gemaakt worden met exponential smoothing methoden zijn gemiddelden van eerdere waarnemingen, waarbij de eerdere waarnemingen zwaarder meewegen dan de oudere (Hyndman & Athanasopoulos, 2018, pp. 237 – 289).

De Holt-Winters seasonal methode zal gebruikt worden om te kunnen profiteren van de trends en seizoensgebondenheid die aanwezig is in de data.

Voorbereiding
Er zal gebruik gemaakt worden van het `statsmodels` package. Veel modellen en vanuit dit package verwachten dat de data geïndexeerd is met timestamps en voorzien is van een frequentie.

Modellen
- `ExponentialSmoothing` uit het `statsmodels.tsa.holtwinters` package

Autoregressive Integrated Moving Average

Autoregressive integrated moving average (ARIMA) modellen maken gebruik van een combinatie van de volgende technieken.
- Autoregression (AR)
- Moving average (MA)

Het kan gebruik maken van de trends, autocorrelatie en seizoensgebondenheid van de data.

Voorbereiding
Aangezien voor deze techniek ook het `statsmodel` package zal worden toegepast zal dezelfde voorbereiding als bij het Exponential Smoothing model zal gebruikt worden.

Modellen
- `ARIMA` uit het `statsmodels.tsa.arima.model` package
- `auto_arima` van het `PMDARIMA` package voor model selection.

Model Beoordelingen

De in het voorgaande hoofdstuk genoemde modellen zijn ontwikkeld en geëvalueerd. In dit hoofdstuk zal worden toegelicht wat de resultaten waren en wat er eventueel verbeterd kan worden in komende iteraties.

Lineaire Regressie

Voor de lineaire regressie modellen is het Sci-Kit learn package toegepast. De onderstaande algoritmes zijn uitgetest.

- LinearRegression (LR)
- KNearestNeigbors (KNN)
- RandomForestRegressor (RF)

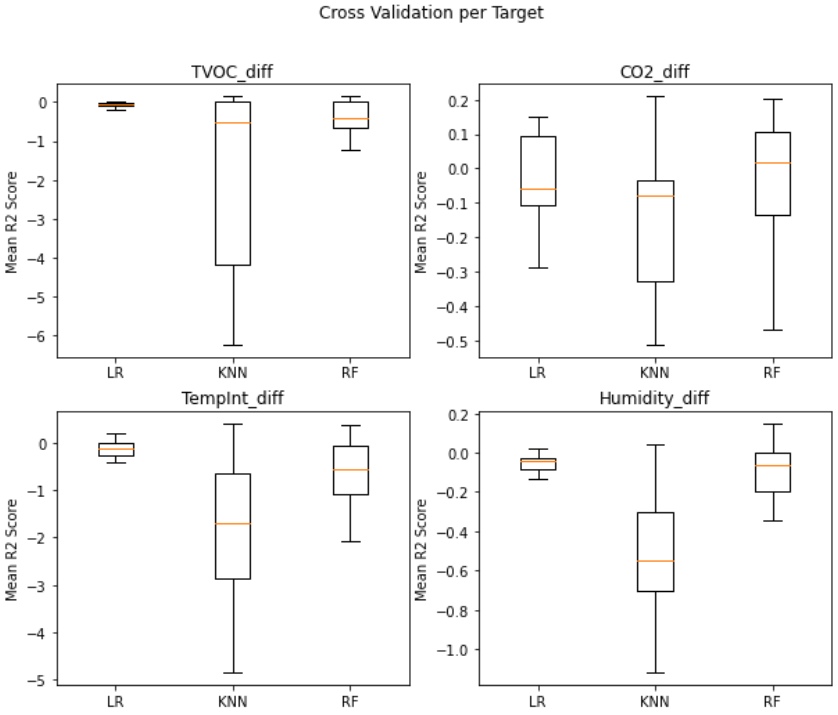
Targets & features

In de onderstaande tabel is een overzicht te vinden van welke targets voorspelt wordt door middel van een aantal features. Deze features zijn bepaald door de onderlinge correlaties zoals te zien zijn in *Afbeelding 2: Onderlinge correlaties.*

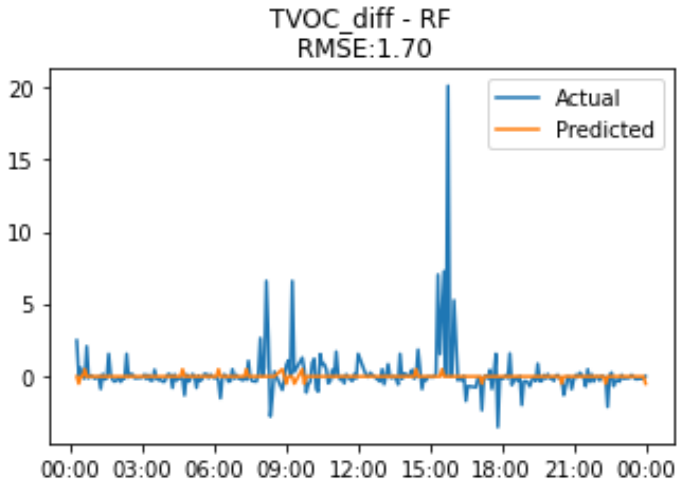
Target	**Features**
TVOC	CO2_difference_t-1, Humidity_difference_t-1, TVOC_difference_t-1
CO2	CO2_difference_t-1
TempInt	TVOC_difference_t-1, TempInt_difference_t-1
Humidity	TVOC_difference_t-1, Humidity_difference_t-1

Resultaten

Om de modellen te verifiëren is cross validation met als scoring metric R2 toegepast. De resultaten hiervan zijn in de onderstaande afbeeldingen te zien. Aan deze afbeeldingen is af te leiden dat geen enkel model een patroon in de features heeft kunnen ontdekken waaruit het target kan worden afgeleid.



Afbeelding ?: R2 cross validation

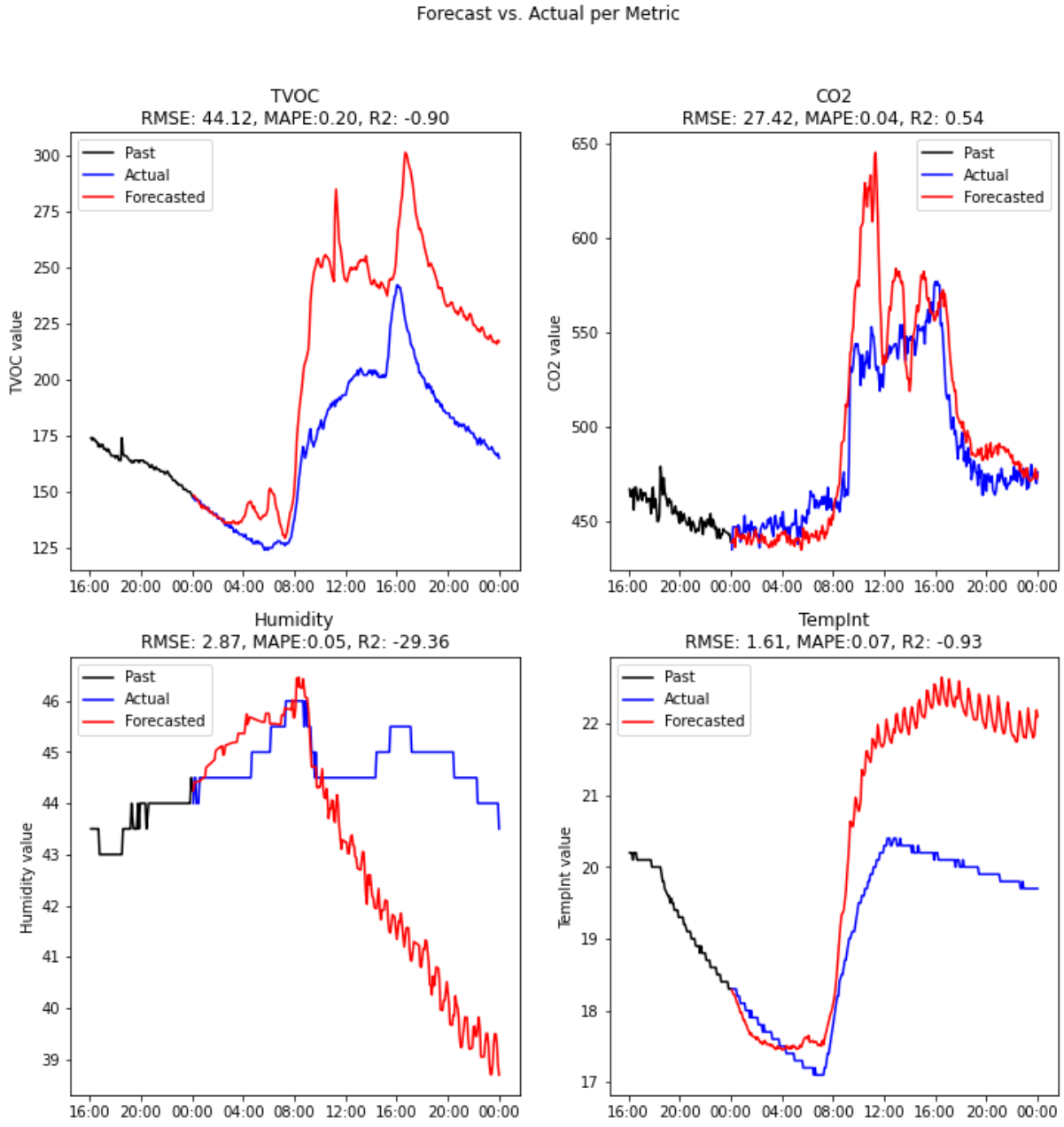


Afbeelding ?: TVOC Voorspelling voorbeeld

Waarschijnlijk missen er nog cruciale features bij deze modellen om een accurate voorspelling te kunnen maken. In een volgende iteratie zal verder onderzocht moeten worden wat invloed heeft op de verandering in de targets.

Exponential Smoothing

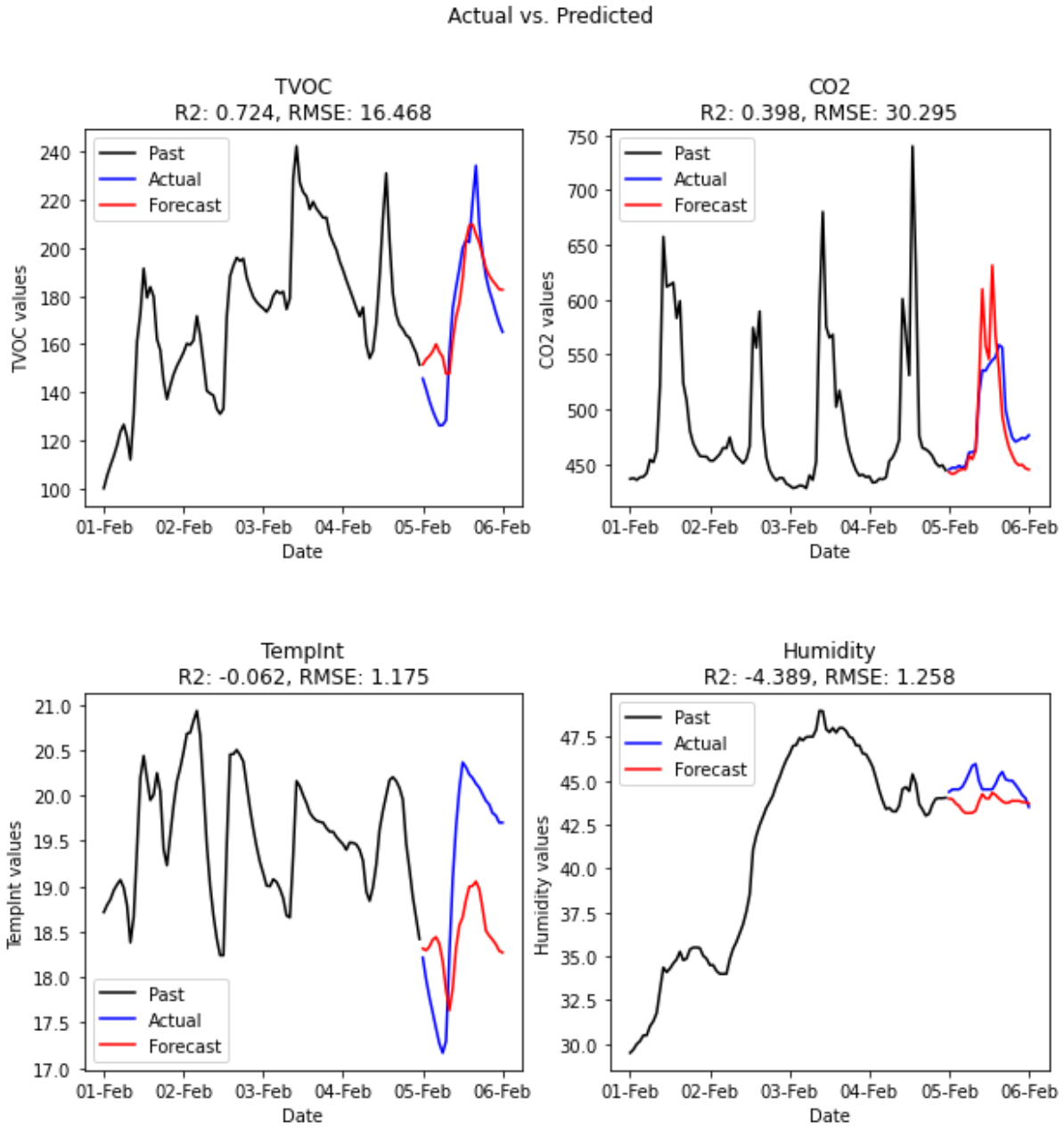
Zoals aangegeven in het modelbeschrijving hoofdstuk is het statsmodel package toegepast om de Exponential Smoothing modellen te ontwikkelen. In de onderstaande afbeelding valt een voorbeeld van de voorspellingen en de scores te zien per target.



Voor de meetwaarden die een sterker dagelijks patroon volgen zoals TVOC en CO2 presteren deze modellen redelijk goed voor voorspellingen binnen 1 a 2 uur. Buiten deze tijdsintervallen worden ze steeds minder accuraat. Voor meetwaarden die afhankelijk zijn van andere factoren, zoals luchtvochtigheid en temperatuur, zijn de resultaten minder accuraat.

ARIMA

Het proces om de ARIMA modellen te ontwikkelen is geautomatiseerd door middel van het pmdarima package. De auto_arima functie probeert meerdere configuraties en kiest hieruit de het best presterende model. De resultaten hiervan vallen in de onderstaande afbeelding te zien.

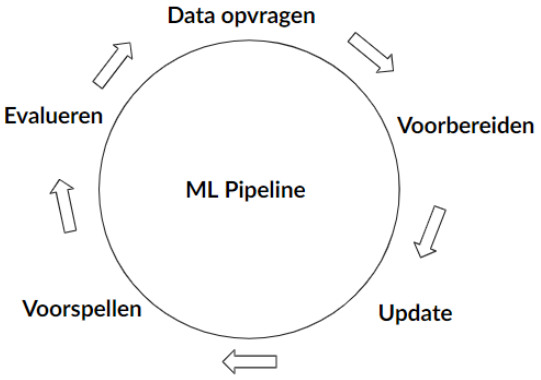


Net als bij de Exponential Smoothing modellen presteert deze techniek beter op de TVOC en CO2 meetwaarden. Deze waarden hebben ook een wekelijkse seizoensgebondenheid die momenteel nog niet gemodelleerd is. Wanneer er meer data beschikbaar is zal dit meegenomen worden in het model wat waarschijnlijk resulteert in hogere scores.

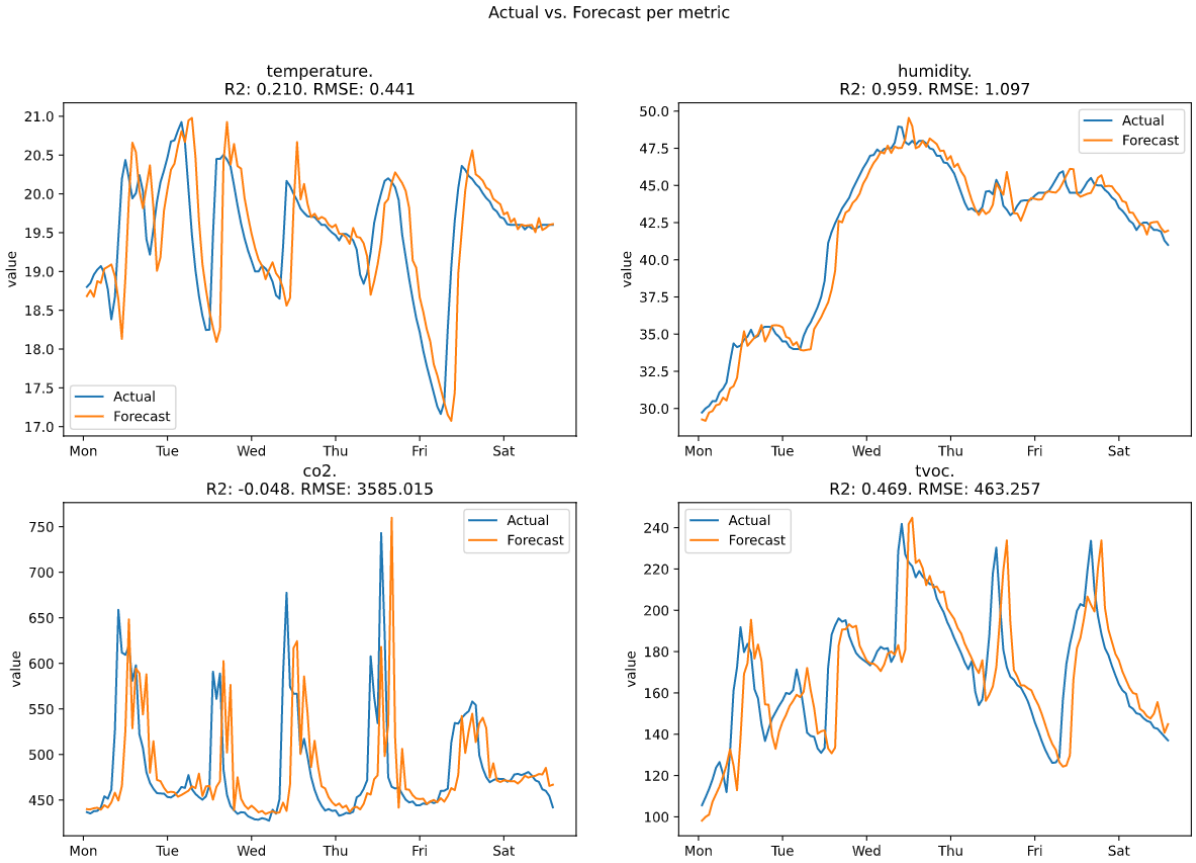
Implementatie

De in de voorgaande hoofdstukken ontwikkelde ARIMA modellen presteerde het beste en zijn daarom gebruikt om de Twindle applicatie uit te breiden. In dit hoofdstuk zal teruggekeken worden op deze implementatie.

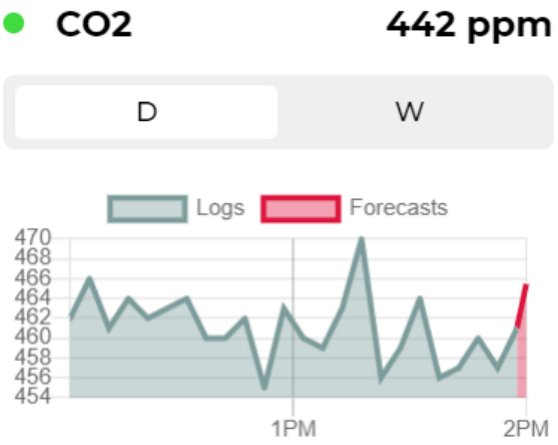
De onderstaande afbeelding geeft weer welke stappen de machine learning (ML) pipeline doorloopt. Op deze manier wordt stapsgewijs het model geüpdatet en nieuwe voorspellingen gemaakt.



Het resultaat van deze pipeline is in de onderstaande afbeelding te zien. Hieruit blijkt dat het model altijd een stap achter loopt. Tijdens een oplevering voor de stakeholders is naar voren gekomen dat dit mogelijk eigenschap is van het ARIMA model. Daarnaast zijn de R2 & RMSE scores niet hoog genoeg om aan de eisen van de stakeholders te voldoen



Een ander minpunt is dat dit model maximaal een uur in de toekomst kan kijken, hierna wordt het te onnauwkeurig. Hierdoor is de impact die het heeft op de Twindle applicatie minimaal. In de onderstaande afbeelding valt dit te zien.



Het zou beter zijn dat er een aantal uur aan voorspellingen aan de grafiek toegevoegd kunnen worden. Zo valt het beter op in de grafiek en kunnen gebouwbeheerders beter inspelen op de voorspellingen.

Ook is er een probleem met de grote van de bestanden. In de onderstaande tabel vallen de bestandsgrootte van de ARIMA modellen te vinden.

TVOC	550 MB
Luchtvochtigheid	2 MB
CO2	263 MB
Temperatuur	81 MB

Dit zijn alleen de modellen voor een enkele ruimte. Iedere locatie waar Twindle actief is heeft meerdere ruimten waar modellen voor nodig zijn. Dit zou betekenen dat er voor, bijvoorbeeld, een hotel honderden GB’s aan opslag benodigd zouden zijn.

Conclusie

Tijdens het onderzoeken van de samenstelling van de data was geconstateerd dat er sterke autocorrelaties, seizoensgebondenheid en trends aanwezig zijn. Hierdoor zijn standaard least-squares regressie methoden niet toepasbaar.

Na onderzoek te hebben verricht naar mogelijk toepasbare technieken is er geëxperimenteerd met de volgende technieken.

1. Lineaire regressie; voorspellingen van veranderingen in de meetwaarden.
2. Exponential smoothing; Forecasting gebaseerd op voorgaande observaties.
3. ARIMA; Forecasting d.m.v. een combinatie van autoregressie en moving average.

Vanuit de product owner is het doel gesteld dat deze technieken tenminste een R2 score behalen van 0,9. Momenteel behaalt geen enkel model dit doel, de ARIMA modellen komen hier het dichtst bij in de buurt.

Deze modellen zijn toegepast om de Twindle applicatie uit te breiden. De voorspellingen van deze modellen lopen altijd 1 uur achter. Dit is een eigenschap van de modellen waardoor ze niet toepasbaar zijn voor dit probleem.

Daarnaast kan er maximaal een uur vooruit accuraat voorspelt worden. De impact die dit heeft op de applicatie is te klein.

Om dit doel te behalen en de implementatie te verbeteren kunnen de aanpassingen gemaakt worden:

1. Onderzoeken wat invloed heeft op veranderingen in de meetwaarden en deze data toevoegen aan de modellen.
2. Verder in de toekomst kijken, 3 uur of meer.
3. Betere controleren voor welk tijdstip de voorspelling gemaakt worden.
4. Bestandsgrootte modellen verminderen om schaalbaarheid te vergroten.

Bronnen

Brownlee, J. (2019, augustus 28). How To Backtest Machine Learning Models for Time Series Forecasting. Machine Learning Mastery. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting

Brownlee, J. (2019b, augustus 21). How to Convert a Time Series to a Supervised Learning Problem in Python. Machine Learning Mastery. https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/

Hyndman, R. J., & Athanasopoulos, G. (2018). 8.1 Stationarity and differencing. In Forecasting: Principles and Practice (2de editie, p. 1). OTexts.

Bijlagen

Bijlage 1: Linear Regression (versie 1)

Lineaire regressie wordt toegepast om relaties tussen twee of meer kwantitatieve variabelen in te schatten. In het geval van het Twindle 3.0 project zal deze techniek toegepast worden om de toekomstige luchtvochtigheid, temperatuur, CO2 en TVOC waarden in te schatten.

Data voorbereiding

Om lineaire regressie toe te kunnen passen zal de sequential data omgezet moeten worden naar paren van inputs en outputs (Brownlee, 2019b). Hiervoor zal de Pandas `shift()` methode toegepast worden. Onderstaand valt een voorbeeld te vinden van deze methode

	humidity	temperature
0	49.5	12.9
1	49.5	12.8
2	49.5	12.8
3	49.5	12.8
4	49.5	12.8

Afbeelding 1: Originele data

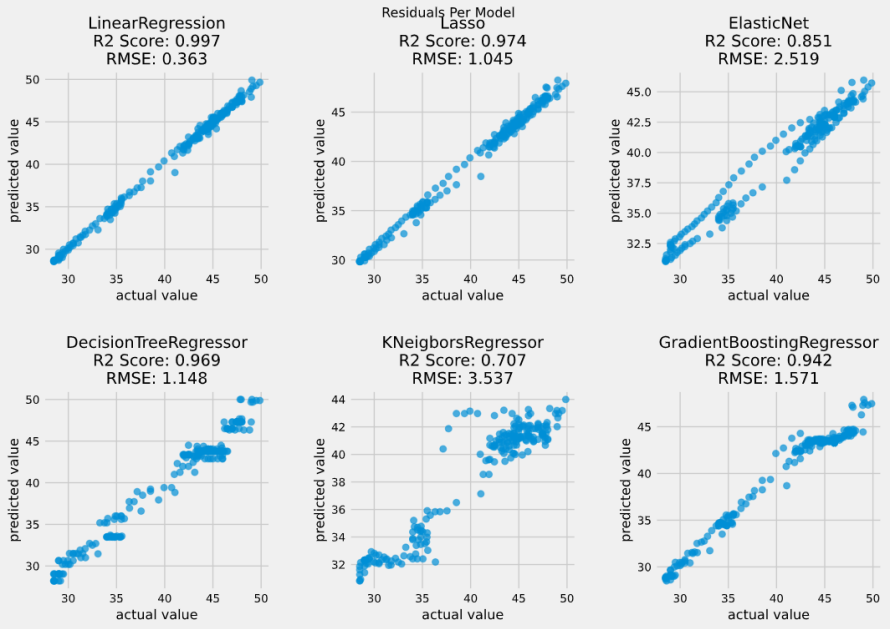
	humidity(t-1)	temperature(t-1)	humidity(t)	temperature(t)	humidity(t+1)	temperature(t+1)
5758	43.0	20.2	43.0	20.3	43.0	20.3
4629	33.0	19.1	33.0	19.1	33.0	19.1
4469	33.5	19.8	33.5	19.8	33.5	19.8
2921	39.5	21.4	39.5	21.5	39.5	21.6
3271	41.5	19.6	41.0	19.5	41.0	19.5

Afbeelding 2: Resultaat Shift methode

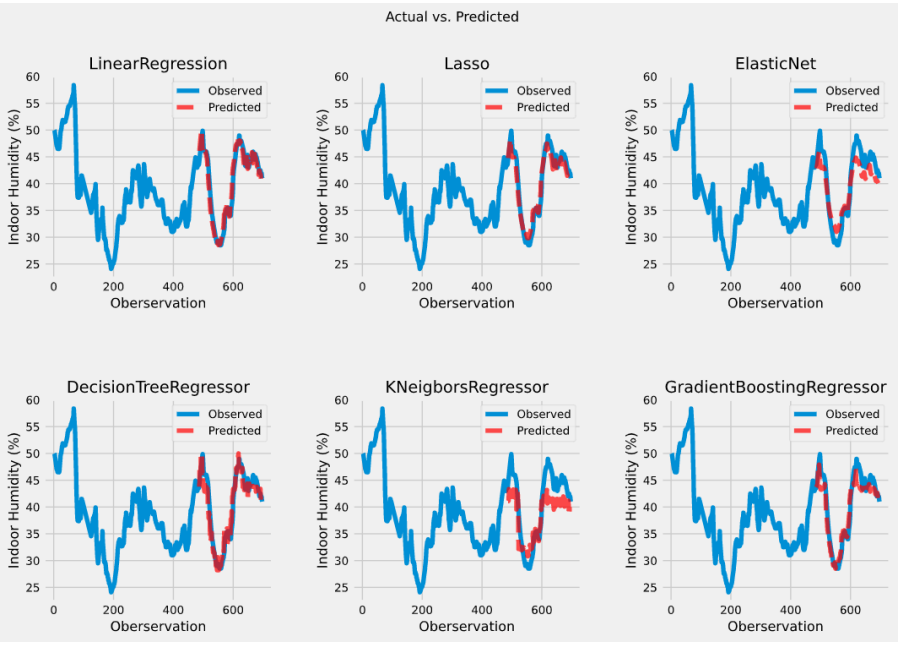
Modellen

luchtvochtigheid

Uit het data requirements onderzoek is gebleken dat de luchtvochtigheid binnenshuis afhankelijk is van de buitentemperatuur en luchtvochtigheid en binnentemperatuur. Deze gegevens zijn gebruikt om de onderstaande modellen te trainen.



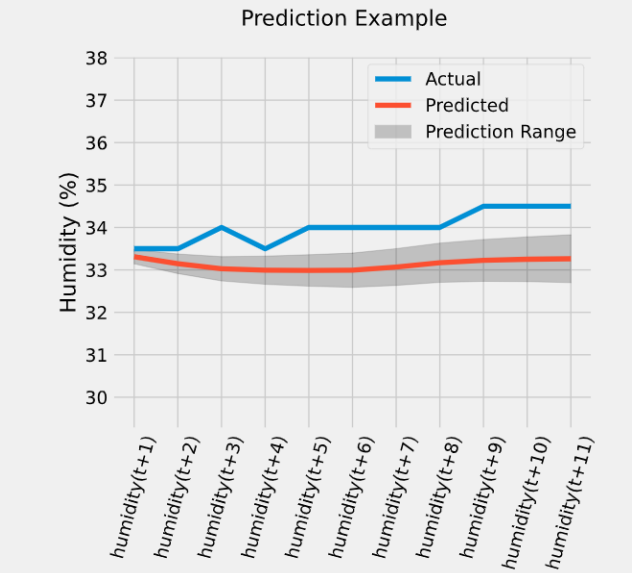
Afbeelding 3: Luchtvochtigheid model resultaten



Afbeelding 4: Luchtvochtigheid model resultaten tijdlijn

Zoals te zien valt in de bovenstaande afbeelding presteren de model goed. De LinearRegression, Lasso en DecisionTreeRegressor modellen behalen de hooste scores. Deze modellen hebben als nadeel dat ze alleen 1 uur in de toekomst kunnen voorspellen. Het zou beter zijn als er meerdere voorspellingen over een kortere tijdsinterval gemaakt

kunnen worden. Na de informatie toevoeging van de variabelen te hebben bekeken, zie onderstaande afbeelding, kan vastgesteld worden dat de buiten- temperatuur en luchtvochtigheid geen informatie toevoegen en weggelaten kunnen worden. Hierdoor kunnen modellen gemaakt worden met een tijdsinterval van vijf minuten.



Evaluatie

Bijlage 2: Linear Regression voorbeeld