Technolution

Start/Stop Wind op Zee

Adviesrapport Machine Learning Model

18/04/2025
Documentversie: 1.1

Redefining solutions

Documentinformatie

Titel:	**Adviesrapport Machine Learning Model**
Klant:	Rijkswaterstaat Zee en Delta
Auteur(s):	Camilla Kuijper, Michael Dubbeldam, Winifred Roggekamp
Versie:	1.1
Datum:	18/04/2025
Project:	Start/Stop Wind op Zee

Documentversies

Versie	Datum	Auteur	Commentaar	Review	Stijl
0.9	05/02/2025	Camilla Kuijper	Initiële versie	Winifred Roggekamp, Michael Dubbeldam	
1.0	10/03/2025	Camilla Kuijper	Verwerken reviewcommentaar	Winifred Roggekamp, Michael Dubbeldam	
1.1	18/04/2025	Camilla Kuijper	Verwerken reviewcommentaar	Winifred Roggekamp, Michael Dubbeldam	

Dit rapport is gemaakt door
Technolution B.V.
Burgemeester Jamessingel 1
Postbus 2013
2800 BD GOUDA
Nederland

Contents

Samenvatting

Achtergrond en doel
Rijkswaterstaat Zee en Delta heeft een machine learning-model laten ontwikkelen door de Universiteit van Amsterdam (UvA) om vogeltrek boven windparken te voorspellen. Dit model wordt ingezet om de rotatiesnelheid van windturbines tijdelijk te beperken bij intensieve vogeltrek. Dit rapport evalueert vanuit een technisch perspectief de betrouwbaarheid van het model en onderzoekt mogelijkheden tot verbetering.

Conclusies uit de validatie
- Het huidige model is zonder grote bugs ontwikkeld en voldoet aan data science-principes, maar technisch gezien is de betrouwbaarheid van het model laag. Er moet goed worden overwogen of dit model operationeel ingezet kan worden. Ondanks dat wij op dit moment geen betere methode kunnen adviseren, dienen de gevolgen van het wel inzetten van dit model goed te worden overwogen.
- De validaties tonen aan dat het model een hoge nauwkeurigheid heeft in het voorspellen van perioden met lage vogeltrek, maar grote moeite heeft met het correct identificeren van voor de start-stop maatregel benodigde piekmomenten.
- De gebruikte dataset is beperkt en scheef verdeeld, waardoor de betrouwbaarheid van de voorspellingen over extreme vogeltrekperioden laag blijft.
- Extra analyses bevestigen dat de prestaties van het model sterk afhankelijk zijn van de beschikbare data en weersomstandigheden, wat de voorspelbaarheid vermindert. Hieruit concluderen we dat de huidige beschikbare hoeveelheid data niet voldoende lijkt om een machine learning toepassing betrouwbaar op te kunnen trainen.

Belangrijkste beperkingen
1. Gebrek aan representatieve data: De dataset bevat te weinig waarnemingen van piekmigratie, wat leidt tot overmatige focus op algemene migratiepatronen.
2. Scheve verdeling van trainingsdata: Slechts een klein percentage van de data bevat hoge vogeltrekwaarden, wat leidt tot onnauwkeurige voorspellingen.
3. Hoge ruis in de data: de beschikbare data bevat aanzienlijke ruis door verschillende factoren, wat de betrouwbaarheid verlaagt en het model belemmert in het trekken van correcte conclusies.
4. Gebruikte predictors: er worden te veel weersvariabelen gebruikt in verhouding tot de beschikbare trainingsdata waardoor de betrouwbaarheid afneemt. Daarnaast ontbreekt een vorm van feedback vanuit de live radar, wat de voorspellingen zou kunnen verbeteren.

Aanbevelingen ter verbetering
Hoewel wij niet verwachten dat de volgende punten zullen resulteren in significante verbeteringen op korte termijn, adviseren we in volgend onderzoek de focus op de volgende punten te leggen.
1. Uitbreiding van de dataset: Meer data verzamelen over langere perioden en mogelijk meerdere radarlocaties gebruiken (het huidige model is op 1 radar gebaseerd).
2. Optimalisatie van de modelstructuur: Alternatieve machine learning-technieken en methoden voor het omgaan met scheve datasets onderzoeken.
3. Gebruik van aanvullende voorspellende factoren: Integratie van real-time radargegevens kan de nauwkeurigheid verbeteren.

4. Gebruik van minder weersvariabelen: Onderzoeken of het model beter presteert bij het verwijderen van sommige inputsvariabelen.
5. Samenwerking met gespecialiseerde partijen: Bedrijven zoals Technolution, TNO en Deltares kunnen bijdragen aan de verdere ontwikkeling en validatie van het model.

Conclusie
Het huidige machine learning-model biedt waardevolle inzichten in vogeltrekpatronen, maar de gekozen technische oplossing is naar onze mening onvoldoende betrouwbaar om leidend te zijn voor het operationele aansturen van windparken. Hoewel er op dit moment geen echte “quick wins” zijn, zal de focus voor verbetering moeten liggen op uitbreiding van de dataset en verbeteringen van omgaan met scheve data in het model. Er wordt aanbevolen om verdere aandacht te besteden aan dataverzameling en modelontwikkeling, en samenwerking te zoeken met gespecialiseerde kennispartners.

1. Inleiding

Langs de Nederlandse kust bevinden zich meerdere windparken, die zich in de vliegroutes van migrerende vogels bevinden. Tweemaal per jaar passeren deze vogels de windparken tijdens hun trektocht. Eerder onderzoek heeft aangetoond dat vogels regelmatig in aanvaring komen met windturbines.

Om conflicten tussen vogels en wind turbines te verminderen, heeft een onderzoeksgroep aan de UvA een machine learning model ontwikkeld het aantal trekvogels per km/ per uur voorspelt (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). Het voorspellende model kan gebruikt worden om de rotatiesnelheid van de windturbines tijdelijk te beperken als de intensiteit van de vogeltrek heel hoog is. Op basis van weersvoorspellingen bepaalt het model de intensiteit van de vogeltrek per uur dagelijks, gedurende het trekseizoen, twee dagen vooruit. Als een drempelwaarde is bereikt, begint de start/stop procedure.

1.1. Doel

Het doel van dit rapport is om te kijken naar het functioneren van de toegepaste machine learning in dit voorspellingsmodel. We zullen eerst inzoomen op het huidige functioneren van het model, kijkend naar de analyses die al eerder zijn gedaan en hoe het huidige model reageert op de nieuw verzamelde data die sinds het publiceren van dit onderzoek is verzameld. Daarna zullen we kijken naar de technische aspecten van het model die eventuele positieve of negatieve invloeden kunnen hebben gehad op het al dan niet goed functioneren van het model. Als laatst zullen we een advies uitbrengen over het verbeteren of gebruiken van dit model.

De aspecten die we onderzoeken, zijn als volgt:

1. Kwaliteit/robuustheidcheck van de huidige toepassing van Machine Learning in het voorspellingsmodel. Het kan nu zijn dat er bugs in de huidige implementatie zitten. Tevens dient te worden onderzocht of de huidige implementatie de optimale AI-methodiek gebruikt.
2. Hoe kunnen we de kwaliteit van het model verhogen
    a. Welke andere methodes/methodieken (zowel binnen machine learning of daarbuiten) zouden gebruikt kunnen worden om een goede vogeltrekvoorspelling te doen en wat zijn de pro’s en cons van deze methodieken?
    b. Als het vogelvoorspellingsmodel opnieuw gebouwd zou worden, hoe zou dat moeten gebeuren en met welke methode?
3. Welke bedrijven zijn geschikt om een zo’n model te ontwikkelen en welke bedrijven zijn geschikt om het te beheren?

1.2. Context

Het beschouwde model is versie 1.1 zoals deze is geleverd door de UvA. Er zijn twee verschillende modellen getraind, een voor de migratiestroom die in het voorjaar plaatsvindt en een voor het najaar. Beide modellen zijn apart getraind, op aparte data. Hierdoor zijn deze dus eigenlijk los van elkaar te beschouwen en zullen ook apart bekeken worden in dit rapport. Wel zijn de ontwikkel- en trainingsmethodes van beide modellen identiek. Dus eventuele conclusies over de methodiek of de algemene beschikbaarheid van data zullen op beide van toepassing zijn. De data om het model te trainen is gebaseerd op slechts één horizontale radar dat op Luchterduinen staat opgesteld.

De weerdata die gebruikt zijn voor de training van het model, is ERA5 data. Deze data zijn gebaseerd op globale spatiale gegevens geaggregeerd over een uur (Hersbach, et al., 2020). De gebruikte data zijn verzameld over een periode van 5 jaar; van 2018 tot en met 2023. De toegevoegde voorspellende factoren die zijn gebruikt om het model op te trainen zijn gebaseerd op eerder

onderzoek naar hoe vogelmigratie beïnvloed wordt door weersinvloeden en weersveranderingen (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). De door de UvA onderzochte voorspellers per seizoen zijn uiteengezet in Tabel 1 - Gebruikte variabelen in het model. . Niet alle variabelen zijn uiteindelijk onderdeel gebleven van het model als predictors binnen het model.

Voorjaar	**Najaar**
Total precipitation at radar location	Total precipitation at radar location
Total precipitation at departure location	Total precipitation at departure location
Temperature at radar location	Temperature at radar location
Temperature at departure location	Temperature at departure location
Mean sea level pressure at radar location	Mean sea level pressure at radar location
Mean sea level pressure at departure location	Mean sea level pressure at departure location
The nightly difference in mean sea level pressure at departure location	The nightly difference in mean sea level pressure at departure location
Accumulation due to wind assistance at departure location	Accumulation due to wind assistance at departure location
The nightly difference in accumulation due to wind assistance at departure location	The nightly difference in accumulation due to wind assistance at departure location
Accumulation due to total precipitation at departure location	Accumulation due to total precipitation at departure location
The nightly difference in accumulation due to precipitation at departure location	The nightly difference in accumulation due to precipitation at departure location
Diurnal phenology	Diurnal phenology
Seasonal phenology	Seasonal phenology
Wind assistance towards ***E*** at radar location	Wind assistance towards ***W*** at radar location
Wind assistance towards ***E at United Kingdom***	Wind assistance towards ***W at northwestern Germany***
Wind assistance towards ***NE*** at radar location	Wind assistance towards ***SW*** at radar location
Wind assistance towards ***NE at northwestern France***	Wind assistance towards ***SW at northern Netherlands***
	Wind assistance towards SW at Denmark

Tabel 1 - Gebruikte variabelen in het model.

De trainingsdata (daar waar het model mee is getraind) en validatiedata (daar waarmee het getrainde model mee is geverifieerd) zijn gebaseerd op radardata afkomstig van een tracking radar (van Erp, van Loon, De Groeve, Bradarić, & Shamoun-Baranes, 2024). Deze radar is gepositioneerd op het

windpark Luchterduinen. De radar heeft twee antennes die ronddraaien en informatie verzamelen over de hoeveelheid langsgekomen vogels. Een van die antennes (de verticale antenne) levert data over de hoeveelheid vogels en hoogtes, de andere (de horizontale antenne) levert data over de hoeveelheid, de vliegrichting en de snelheid van vogels tot 500 meter hoog. In het onderzoek van de UvA is alleen gebruik gemaakt van radardata verzameld door de horizontale antenne, over een (vogeltrek)periode van 2018 tot en met 2023. De eenheid die gebruikt wordt voor migratievolume door (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024) is MTR; mean traffic rate of migration. Deze eenheid staat voor de hoeveelheid vogels per kilometer per uur die langs het windpark vliegt, gemeten door de radar. In de rest van dit rapport zullen we gebruik maken van de afkorting MTR. De radardata bevat vaak nog veel ruis. Hierdoor is in meerdere stadia van het produceren van een MTR-waarde, gebruik gemaakt van filters en andere post-processing technieken om de training data zo zuiver mogelijk te krijgen.

Het doel van het ontwikkelde model door de UvA is om het aantal vogels/km/uur te voorspellen. Rijkswaterstaat wil het model gebruiken om alle uren te identificeren waarin de migratie-intensiteit (MTR) boven de 500 ligt. Deze drempelwaarde is vastgesteld door de minister in het kavelbesluit op basis van eerder onderzoek.

De gebruikte drempelwaarden in dit onderzoek kan wat verwarring opleveren. De waarde van een vogel trekpiek die door het ministerie in de kabelbesluiten is vastgelegd is 500 vogels/km/uur. De UvA heeft echter bij het trainen en valideren van het model andere waarden toegepast. Naar alle waarschijnlijkheid heeft de UvA andere waarden gekozen omdat er weinig pieken boven de 500 zijn en omdat de UvA zich ook heeft gericht op migratiepatronen en niet alleen op migratie pieken. Aangezien dit rapport het model evalueert, gebruiken wij de drempelwaarden zoals deze ook door de UvA zijn gebruikt.

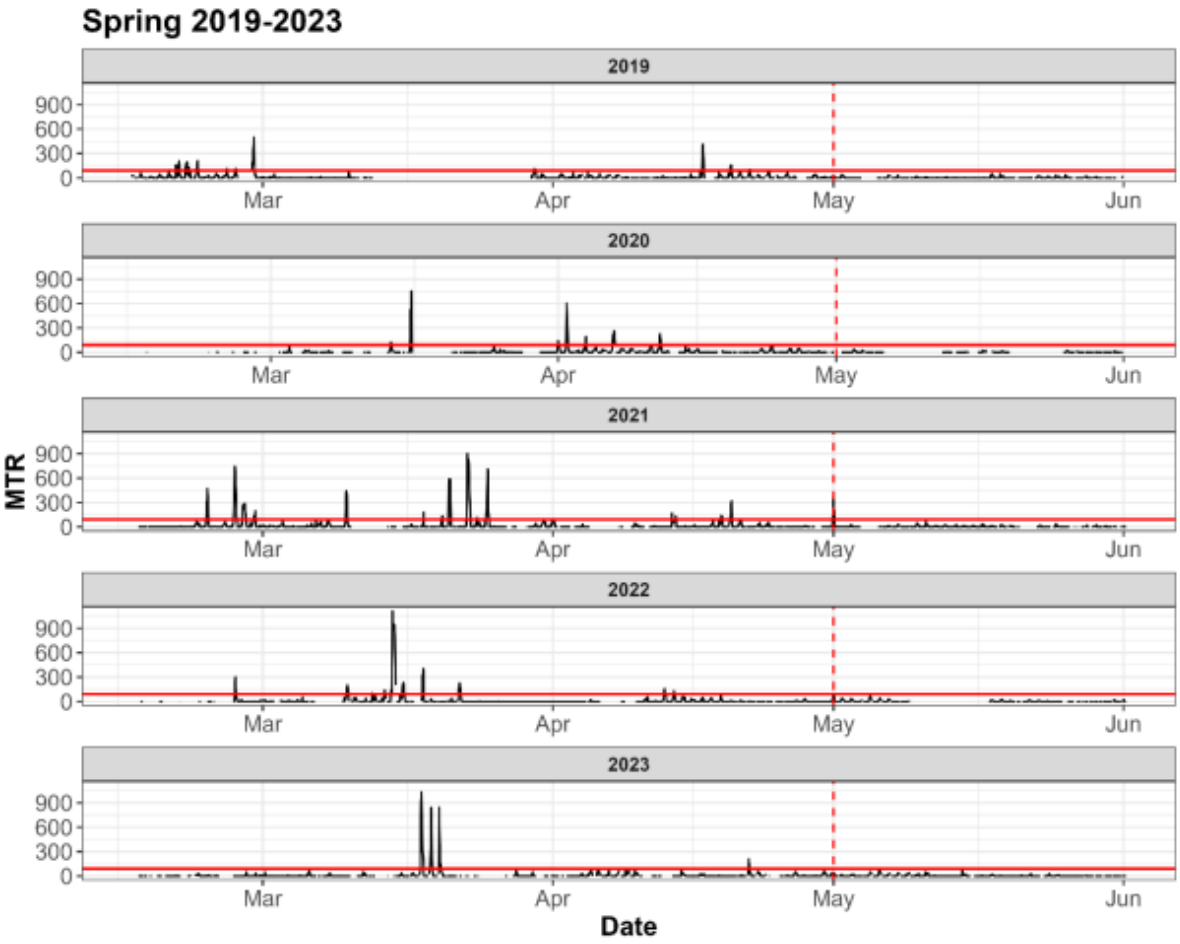
2. Validatie huidig model

In (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024) is veel aandacht besteed aan de validatie van het model. Om de prestaties te evalueren, is voor beide modellen een jaar aan data niet meegenomen tijdens het trainen, zodat de werking getest kon worden op gegevens die het model nog niet eerder had gezien.

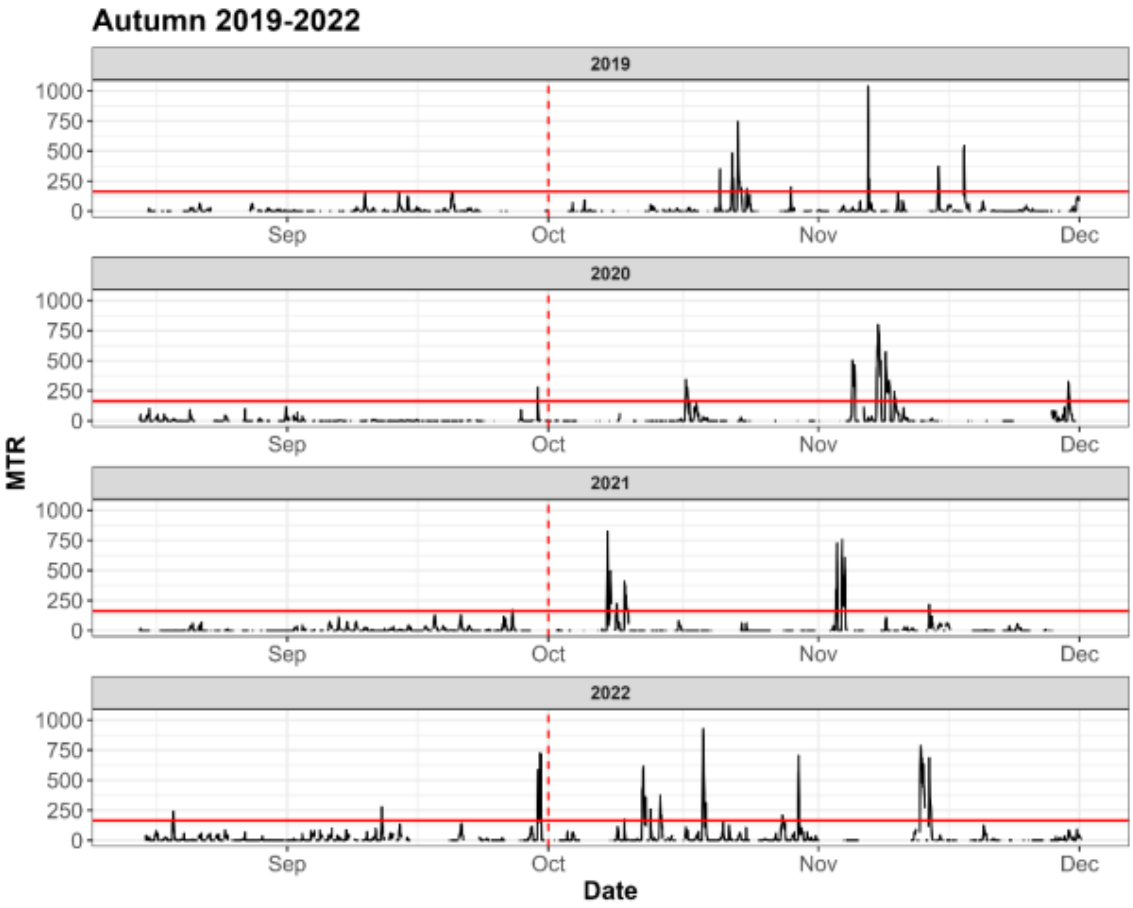
In dit hoofdstuk bespreken we eerst de bevindingen uit het oorspronkelijke UvA onderzoek en plaatsen deze in de context van het beoogde start/stop-systeem. Vervolgens vullen we deze bevindingen aan met aanvullende analyses op nieuwe jaren van weerdata die tijdens de training van het model nog niet beschikbaar was. Deze resultaten worden vergeleken met de eerdere bevindingen uit het UvA onderzoek. Tot slot proberen we de resultaten vanuit het UvA onderzoek zo nauwkeurig mogelijk te reproduceren.

2.1. Algemene dataverdeling

Een globaal overzicht van de door de UvA gebruikte data resulteert in Afbeelding 1 en Afbeelding 2. Het eerste wat opvalt is dat de data erg scheef verdeeld is; er zijn erg weinig datapunten die boven de door de UvA gekozen ecologisch relevante drempelwaarde uitkomen en nog minder datapunten die boven de algemeen gehanteerde drempelwaarde van 500 uitkomen. Met andere woorden: in de periode 2018-2023 heeft de radar, maar enkele keren een trekpiek gemeten boven de drempelwaarde tegenover heel veel metingen die onder de drempelwaarde liggen. Er is tijdens de ontwikkeling van het model rekening gehouden met deze scheve verdeling (wel piek vs geen piek) in het trainen en interpreteren van het model, dit zal besproken worden in het volgende hoofdstuk. Het is belangrijk om deze dataverhouding in acht te nemen bij het valideren van het model in de rest van dit hoofdstuk.



Afbeelding 1 - Globaal overzicht van gebruikte data in het trainen en valideren van het voorjaarsmodel uit (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). De data uit 2022 is gebruikt om het model op te testen, en is dus niet gebruikt in het trainingsproces voor dit UvA onderzoek. De rode gestippelde lijn representeert het gekozen cut off punt van het seizoen. Alle data rechts van deze lijn is niet gebruikt in het model. De doorgetrokken horizontale lijn representeert de 95% grens; 95% van de datapunten vallen onder deze lijn en 5% erboven.



Afbeelding 2 - Globaal overzicht van gebruikte data in het trainen en valideren van het najaarsmodel uit (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). De data uit 2020 is gebruikt om het model op te testen, en is dus niet gebruikt in het trainingsproces voor dit UvA onderzoek. De rode gestippelde lijn representeert het gekozen cut off (eind-)punt van het seizoen. Alle data links van deze lijn is niet gebruikt in het model. De doorgetrokken horizontale lijn representeert de 95% grens; 95% van de datapunten vallen onder deze lijn en 5% erboven.

2.2. Validatie uitgevoerd door UvA tijdens de modelontwikkeling

Zoals eerder genoemd is er voor beide modellen een jaar aan data niet gebruikt voor de training set om te gebruiken voor het valideren. De volgende validaties zijn uitgevoerd op data van 2022 voor het voorjaar en data van 2020 voor het najaar, dit noemen we de test set. In het onderzoek zijn er ook validatiestappen uitgevoerd waarbij er modellen getraind zijn met de andere jaren als test set. Deze zijn te vinden in de appendix, maar bespreken we verder niet in dit onderzoek (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024).

De RMSE (root mean square error) is 46.51 voor de lente en 60.04 voor de herfst. De RMSE is een getal dat een indicatie geeft van de performance van het model op data waar het niet op getraind is. Dit is dus een indicatie van hoe het model in de praktijk zou presteren. De RMSE is grofweg het gemiddelde verschil tussen wat het model als voorspelling geeft en wat de daadwerkelijke waarden zijn (in ons geval de gefilterde radardata). Dus in dit geval verschillen de voorspellingen van het voorjaarsmodel gemiddeld 47 vogels/km/uur (MTR) van het daadwerkelijke getal en dit is 60 vogels/km/uur voor het najaarsmodel. Deze RMSE is vrij hoog, aangezien de drempelwaardes bij de

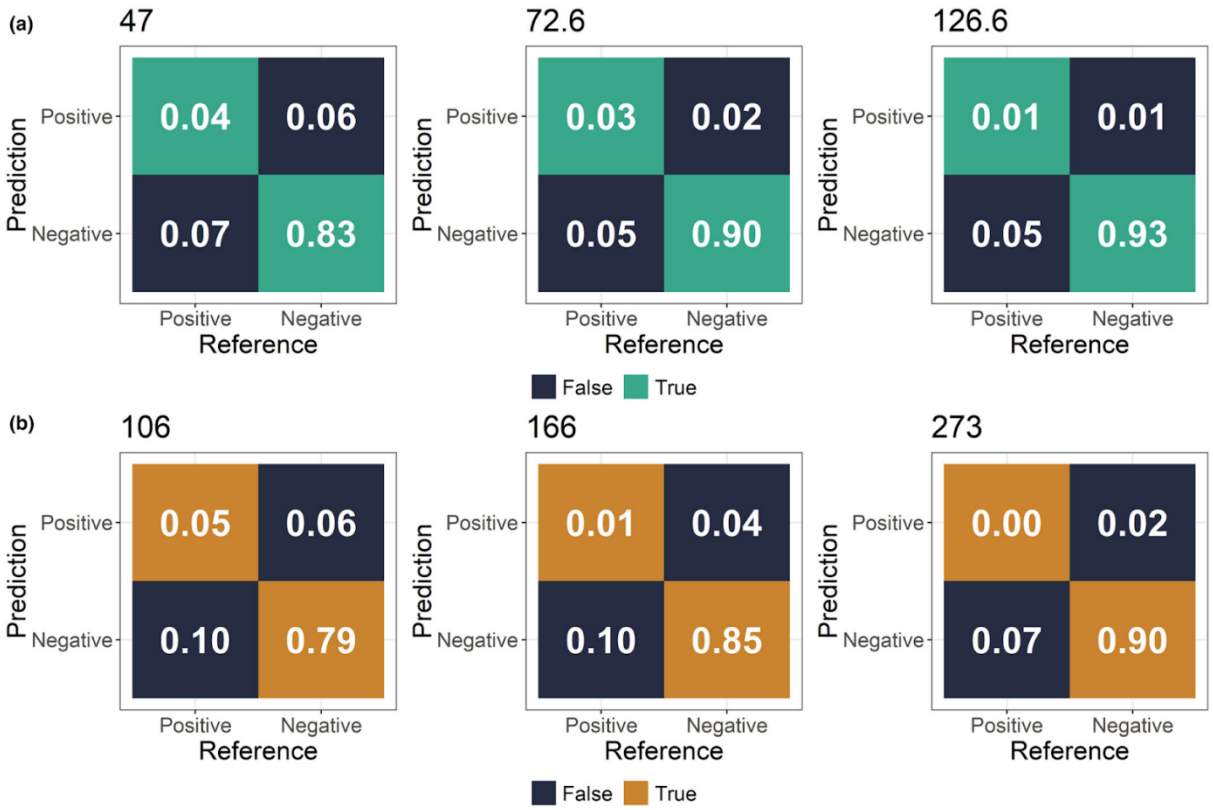
validatie van de UvA op respectievelijk 126.6 en 273 vogels/km/uur zijn gezet. Bij een verschil van ongeveer 50, zit je dan vrij snel ver boven of onder deze drempelwaarde.
De R^2 van beide modellen is ongeveer 0.71. R^2 is een getal dat weergeeft welk deel van de variantie in de data kan worden voorspeld en uitgelegd door het model. Een ideale R^2 zou een 1.0 zijn, dan wordt 100% van de vogeltrek goed voorspeld door het model. Een score van 71% is relatief goed. Dit betekent dat het model een vrij groot deel van de variantie in de data kan voorspellen.

Een relatief hoge R^2 in combinatie met een hoge RMSE kan verschillende oorzaken hebben, waarvan de scheve verdeling van de beschikbare data waarschijnlijk de grootste is. In de dataset zijn veel datapunten laag, met MTR-waarden rond de 0 tot 30. Het merendeel van de gegevens (ongeveer 95%) ligt onder de drempelwaarde. Hierdoor kan een groot deel van de data correct voorspeld worden, maar deze correcte voorspellingen bevinden zich vooral in het lage MTR-bereik. Dit kan verklaren waarom de RMSE relatief hoog blijft: terwijl het model goed presteert op de veelvoorkomende lage waarden, kan het moeite hebben met het accuraat voorspellen van zeldzame pieken in migratie. Wanneer er hoge pieken optreden bij hoge migratie-intensiteit, kunnen de voorspelfouten aanzienlijk groter worden, wat de gemiddelde fout (RMSE) opblaast. Zowel RMSE als R^2 zijn gevoelig voor scheve datasets zoals deze, waarin de meerderheid van de waarden laag is en slechts enkele extreme uitschieters voorkomen.

Een meer representatieve indicatie van de effectiviteit van het model voor het start/stop concept is te zien in de confusion matrices. Hierin zien we hoe vaak het model een voorspelling doet die leidt tot een correcte categorisatie in een van twee categorieën; een uur met normale of lage vogeltrek of een uur met hoge vogeltrek. De confusion matrices van de twee modellen uit het paper van de UvA zijn te zien in Afbeelding 3. Linksboven in elke matrix is het percentage van de testdata te zien waarop zowel het voorspellingsmodel als de radar (daadwerkelijk gemeten MTR) een indicatie geven van hoog migratievolume. Een hoog migratievolume is hierbij gedefinieerd als het overschrijden van de MTR van een specifieke drempelwaarde (te zien boven de matrix). Rechtsonder elke matrix is het percentage van de data weergegeven waarvoor zowel het model als de radar aangeven dat de MTR niet boven de drempelwaarde uit komen. De vakjes linksboven en rechtsonder representeren dus de hoeveelheid keren dat het model een ‘correcte’ voorspelling doet.

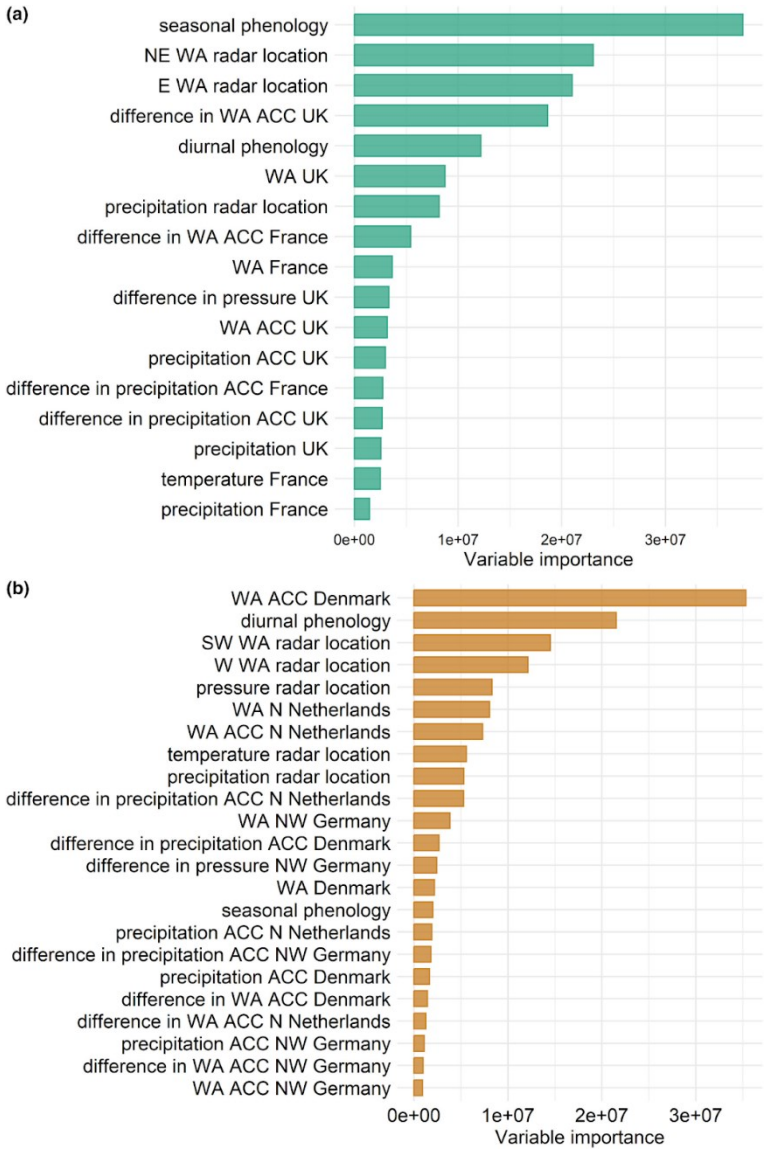
De overige twee getallen representeren het percentage van de data waarvoor het voorspellingsmodel een andere conclusie trekt dan de radar. Voor de data linksonder voorspelt het model dat de MTR niet boven de drempelwaarde uitkomt terwijl dit volgens de radar wel het geval is. Bij de waarde rechtsboven concludeert het model dat de drempelwaarde wordt overschreden terwijl de radardata dit niet bevestigt.

Kijkend naar de twee matrices corresponderend met de gekozen drempelwaarden voor het model (de rechter twee), zien we dat de modellen beide erg goed zijn in het voorspellen van momenten met lage migratie. De modellen voorspellen 98% (voorjaar) en 97% (najaar) van de tijd geen piekmigratie, waarvan het overgrote deel correct voorspeld is (*accuracy* = 0.94 voor de lente en 0.90 voor de herfst). Echter betekent dit ook dat het model bijna nooit een voorspelling geeft van piekmigratie; 2% voor beide modellen. En van de momenten dat het model een piekmigratie aangeeft, zijn er veel van deze voorspelde trekpieken incorrect; dit geldt voor ongeveer de helft van de tijden voor de lente (*precision* = 0.50) en voor bijna alle positieve voorspellingen voor het herfstmodel (*precision* = 0.00). Modellen die getraind zijn op erg scheve data laten vaker dit soort patronen zien; het model lijkt *overfit* te zijn op momenten met lage migratie en daarom erg goed in voorspellen van lage MTR-uren, maar erg slecht in voorspellen van hoge migratie. Dit geeft het volgende probleem: het stilzetten van de windturbines op basis van het model gebeurt slechts in een erg klein percentage van de gevallen op het moment dat er daadwerkelijk een vogeltrekpiek is. De overige momenten dat het model voorspelt dat er een vogeltrekpiek is, zouden de windturbines dus onterecht stilstaan.



Afbeelding 3 - verschillende confusion matrices van de twee modellen zoals getoond in het paper van de UvA (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). De matrices voor het voorjaarsmodel (a, groen) en het najaarsmodel (b, oranje) tonen het deel van de resultaten dat goed en fout voorspeld was voor verschillende drempelwaardes van het model. Voor doeleinden van ons systeem zijn de twee meest rechtse matrices het meest relevant, aangezien die dichter in de buurt komen van de drempelwaardes die gehanteerd zouden worden in productie.

In het paper van de UvA is ook gekeken naar de relevantie van de verschillende weersvariabelen in het model. De resultaten hiervan zijn zichtbaar in Afbeelding 4 - Relevantie van de toegepaste variabelen op het uiteindelijke model voor de lente (a, blauw) en de herfst (b, oranje) . De betekenis van de afkortingen op de y-as zijn terug te vinden in het paper van de UvA.. Er is duidelijk zichtbaar dat niet alle variabelen even veel meewegen in het beslisproces van het model.



Afbeelding 4 - Relevantie van de toegepaste variabelen op het uiteindelijke model voor de lente (a, blauw) en de herfst (b, oranje) (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). De betekenis van de afkortingen op de y-as zijn terug te vinden in het paper van de UvA.

Voor de doeleinden van de start/stop maatregel is het minder relevant hoeveel van de modelvoorspellingen correct zijn. Het is voornamelijk belangrijk dat het model de trekpieken correct kan voorspellen. Uit de validatie van het model lijkt het model getraind te zijn op het tegenovergestelde; het is heel goed in het correcte voorspellen van momenten dat er geen trekpiek is. Daarmee zakt de performance voor de sporadische momenten in het jaar dat er heel veel vogels tegelijk langskomen. In het volgende hoofdstuk zullen we uitzoeken waarom het model dit soort resultaten lijkt te tonen.

2.2.1. Conclusie van hoofdstuk Validatie uitgevoerd door de UvA

De data die is gebruikt om het model te trainen bevat heel weinig trekpiekmomenten ten opzichte van momenten zonder trekpiek. Deze scheve data zorgt ervoor dat een machine learning model moeilijk te trainen is. Daarnaast is het model voornamelijk voor lagere pieken getraind zodat de trekpieken

minder goed voorspeld worden. In de validatie van de UvA is dit terug te zien. Hierbij is het voor de start-stop maatregel van belang om te kijken naar hoeveel trekpieken het model goed voorspelt en is het minder belangrijk om te kijken naar hoeveel keer het model de andere migratiemomenten goed voorspelt.

2.3. Reproductie en cross validation

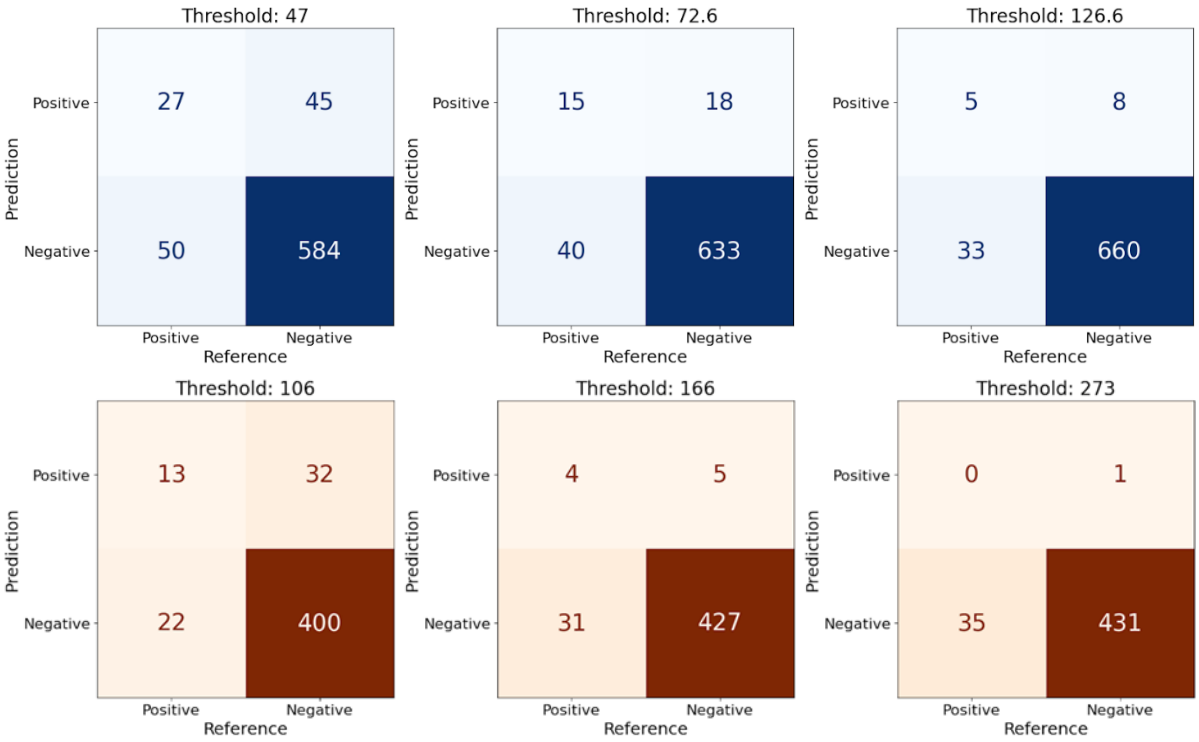
Op basis van de gepubliceerde stukken en data en op basis van verkregen code en data vanuit Rijkswaterstaat en de onderzoeksgroep aan de UvA, hebben we enkele reproducties en extra validatiestappen uitgevoerd. De resultaten hiervan zijn uiteengezet in deze paragraaf en appendix A.

2.3.1. Reproductie van het UvA model

Het UvA model maakt gebruik van het zogenaamde random forest algoritme. Vanuit de UvA hebben we een representatief stuk code ontvangen waarin uitgelegd werd hoe het random forest algoritme toegepast is. Dit heeft ons in staat gesteld om de onderzoekresultaten van de UvA tot op zekere hoogte te reproduceren en te controleren of wij tot vergelijkbare resultaten kunnen komen. We hebben niet alle code volledig ontvangen en konden daardoor niet exact achterhalen hoe alle stappen (cleaning, training en validatie) precies uitgevoerd zijn. Ook zijn er meerdere versies van het model ontwikkeld en het is ons niet altijd helemaal duidelijk geworden voor elke versie welke stappen toegepast en/of veranderd waren en welke data gebruikt is. Door deze twee factoren zullen de resultaten in deze paragraaf licht afwijken van de resultaten gepubliceerd in het paper van de UvA en ook van de resultaten die zichtbaar zijn vanuit het datadashboard van het productiesysteem. Echter, de verkregen resultaten zijn met grote mate vergelijkbaar met de resultaten gepubliceerd in het paper en daarmee lijkt ons het onderzoek voldoende reproduceerbaar.

Voor beide modellen hebben wij de data gepubliceerd bij (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024) gebruikt om zelf een model te trainen en te valideren met dezelfde methodiek. We hebben hierbij dezelfde hyperparameters en dezelfde training en test datasets gebruikt. Voor het lentemodel zijn wij uitgekomen op een RMSE van 45.75 en een R^2 van 0.71. Voor het herfstmodel is dit een RMSE van 61.65 en een R^2 van 0.70. Deze zijn erg vergelijkbaar met de resultaten besproken in de vorige paragraaf.

We hebben ook de confusion matrices van ons reproductiemodel gegenereerd. Deze zijn zichtbaar in Afbeelding 5. Er is hier gekozen om de absolute aantallen weer te geven, waar in het paper de relatieve hoeveelheid uren worden gebruikt. Wij hebben dit gedaan omdat dit de lage hoeveelheid uren soms beter weergeeft. Voor het lentemodel zien we een *accuracy* van 0.94 en een *precision* van 0.38. Voor het herfstmodel zien we een *accuracy* van 0.92 en een *precision* van 0.00. Deze matrices en metrieken zijn erg vergelijkbaar met de resultaten uit het paper en besproken in het vorige hoofdstuk. Hiermee kunnen we stellen dat de code om het model te trainen geen significante bugs bevat.



Afbeelding 5 - Verschillende confusion matrices van de twee modellen resulterend uit de reproductiestappen. De matrices voor het voorjaarsmodel (a, blauw) en het najaarsmodel (b, oranje) tonen het deel van de resultaten dat goed en fout voorspeld was voor verschillende drempelwaardes van het model. Er is gekozen voor absolute waardes in plaats van de relatieve waardes in deze afbeeldingen.

2.3.2. Cross validation

De voorgaande besproken validatiestappen zijn allemaal uitgevoerd op een test set welke is weggelaten bij het trainen van het model. Deze test set is 2022 voor het voorjaarsmodel en 2020 voor het najaarsmodel. Om te onderzoeken wat de invloed is van de specifiek gekozen test- en training sets en om te kijken in hoeverre het model afhankelijk is van de specifiek gemeten datapunten, hebben wij een vorm van cross validatie uitgevoerd. Hierbij hebben we meerdere keren een model getraind, elke keer met een ander jaar gebruikt voor de validatie (en dus niet voor het trainen). Daarna hebben we dezelfde validatiestappen uitgevoerd op elk getrainde model. De resultaten hiervan zijn te zien in appendix A.

We zien dat er een groot verschil zit in de resultaten tussen de verschillende modellen. De RMSE varieert tussen 42.0 en 48.6, de R^2 tussen 0.71 en 0.78, de *accuracy* varieert tussen 0.85 en 0.96 en de *precision* varieert tussen 0.09 en 0.50 voor het voorjaarsmodel. De RMSE varieert tussen 58.7 en 61.6 en de R^2 tussen 0.70 en 0.77, de *accuracy* varieert tussen 0.81 en 0.97 en de *precision* varieert tussen 0.00 en 0.09 voor het najaarsmodel. Deze grote spreiding in de metrieken is een indicatie dat de werking en de validatie van de resultaten van het model erg afhankelijk zijn van fluctuaties in de specifiek verzamelde datapunten. Dit soort resultaten op cross validatie is typerend voor modellen die getraind zijn op relatief weinig datapunten en daarmee *overfit* zijn op de specifieke data waar het wel op getraind is. Dit resulteert in een lage nauwkeurigheid voor toekomstige

voorspellingen op jaren die het model nog niet gezien heeft, zoals in het productiesysteem voor start/stop.

2.3.3. Conclusie van hoofdstuk reproductie en crossvalidatie

De resultaten en methodes van de UvA zijn voldoende te reproduceren. Daarmee is het aannemelijk dat er geen grote fouten zitten in de code en de modelgeneratie.
Bij de cross validatie is gebleken dat er een grote variatie zit in welke data er gebruikt wordt voor training en voor validatie. Dit betekent dat de nauwkeurigheid voor het voorspellen van toekomstige jaren laag is.

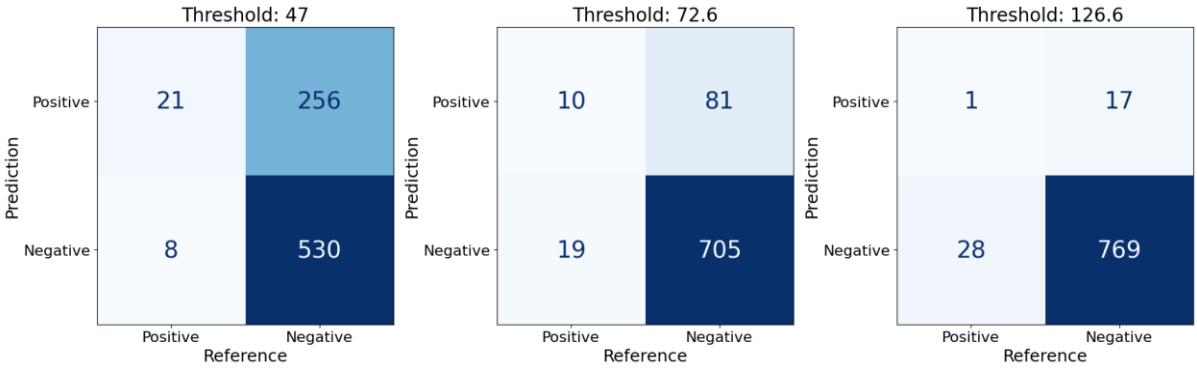
2.4. Validatie over data 2024

Sinds het trainen en uitbrengen van het model, is er ongeveer een jaar aan extra data verzameld (data uit 2024). Dit geeft ons de kans om extra validatiestappen uit te voeren over de werking van het model. We zullen hier kijken naar wat algemene validaties van het model op basis van radar- en weersdata verzameld van 15 februari tot 30 mei voor het voorjaarsmodel en van 15 augustus tot 30 november voor het najaarsmodel. Als eerste zullen we vergelijkbare stappen nemen als ook in het UvA paper genomen zijn (zie ook het vorige hoofdstuk).
Het is belangrijk om te vermelden dat deze validaties zijn uitgevoerd op iets andere data dan waarop het model is getraind. Het voorjaarsmodel is getraind op gegevens tot en met 30 april, en het najaarsmodel is gebaseerd op data vanaf 1 oktober. Desondanks is het doel om het model ook toe te passen op de eerder genoemde dagen, waardoor het essentieel is om deze data mee te nemen in de validatie. De eerder besproken validatie is met ERA5 weerdata gedaan. Dit is data die achteraf is gegenereerd en daarmee de hoogste betrouwbaarheid heeft. Echter, het daadwerkelijke model zal moeten voorspellen met (ECMWF) weerdata die 2 dagen van tevoren is verzameld. Daarna zullen we dus de resultaten van het model valideren op weerdata van 2 dagen van tevoren. We hebben deze validaties uitgevoerd op de meest recente versie (v1.1) van het model dat wij tot onze beschikking hadden. De validaties zijn op basis van data opgevraagd vanuit het wind op zee datadashboard begin december 2024.

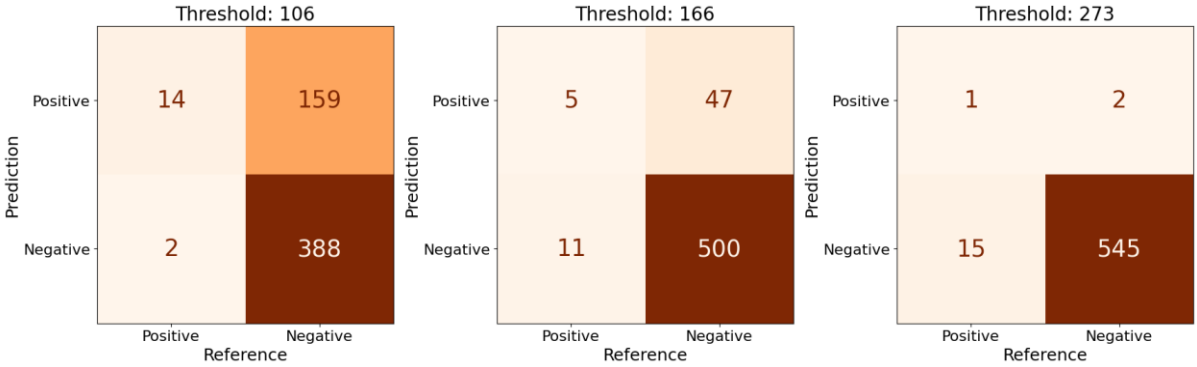
2.4.1. Validatie model op basis van weerdata achteraf (ERA5)

Voor het model wat data produceert met ERA5 weerdata voor de zelfde dag zijn er confusion matrices gemaakt met dezelfde drempelwaardes als in het paper van de UvA over de nieuwe data (jaar 2024). Deze zijn zichtbaar in Afbeelding 6 en Afbeelding 7. Voor deze matrices zijn alle nachtelijke uren tussen 15/2/2024 en 31/5/2024 voor het voorjaar en tussen 15/8/2024 en 30/11/2024 voor het najaar gebruikt. Hierbij is een uur bestempeld als een uur met hoge MTR in de referentiedata als de horizontale radar een MTR van 126.6 rapporteerde in de lente en 273 in de herfst.

Voor beide modellen zien we dezelfde verschijnselen als bij de validaties in het paper en in het vorige hoofdstuk. Voor de lente zien we in alle matrices een erg hoge accuracy (0.94 voor de matrix met de hoogste drempelwaarde) en een erg lage precision (0.06 voor de matrix met de hoogste drempelwaarde). Voor het najaarsmodel zien we ook steeds erg hoge accuracy voor alle modeldrempelwaardes (0.97 voor de matrix met de hoogste drempelwaarde), en lage precision (0.33 voor de matrix met de hoogste drempelwaarde).



Afbeelding 6 - confusion matrices met absolute aantallen voor het voorjaarsmodel over data 0 dagen van tevoren. Let op dat deze matrix gemaakt is op nieuwe data sinds het paper van de UvA is uitgekomen, dus het komt niet volledig overeen met voorgaande matrices.



Afbeelding 7 - confusion matrices met absolute aantallen voor het najaarsmodel over data 0 dagen van tevoren. Let op dat deze matrix gemaakt is op nieuwe data sinds het paper van de UvA is uitgekomen, dus het komt niet volledig overeen met voorgaande matrices.

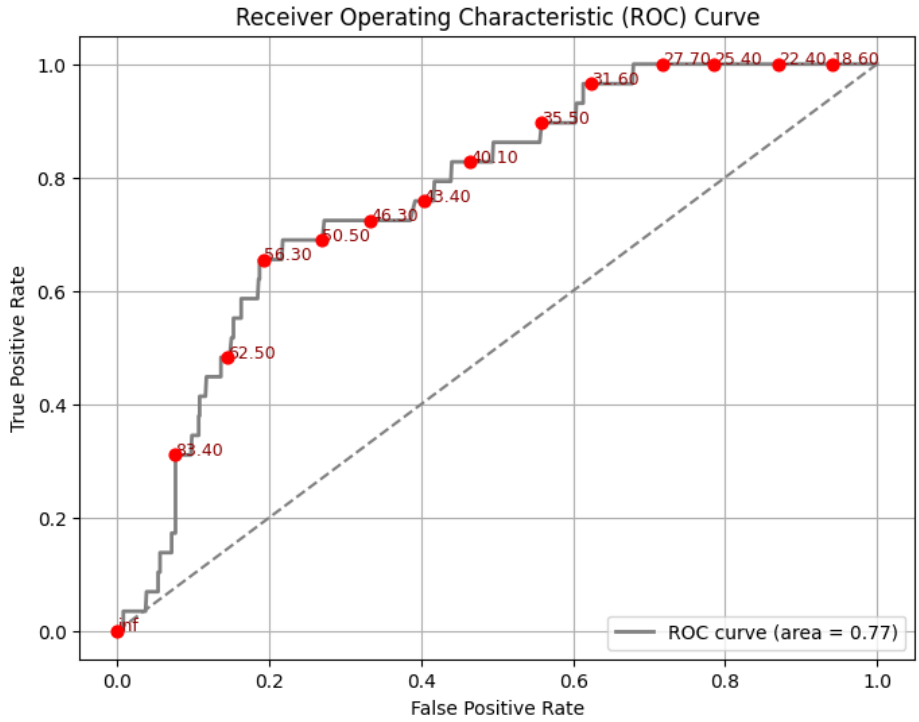
Op vergelijkbare manieren als gedaan in de validaties in het paper, zijn er ROC curves gemaakt over de nieuw verzamelde data om de het model op te kunnen evalueren. ROC curves zijn visuele weergaves van het presteren van een model op verschillende drempelwaardes in termen van 'true positive rate' (TPR) en 'false positive rate' (FPR). De TPR is de hoeveelheid correct positief bestempelde observaties van alle correct bestempelde observaties door het model. De FPR is de hoeveelheid incorrect positief bestempelde observaties ten opzichte van alle negatieve observaties. Een perfect werkend classificatiemodel heeft een TPR van 1.0 en een FPR van 0.0. Hierdoor is het over het algemeen een goede vuistregel voor het lezen van ROC curves dat hoe dichter een model in de buurt komt van de linkerbovenhoek, hoe beter het model presteert. Een classificatiemodel dat compleet willekeurig waardes toekent zal ongeveer een ROC curve hebben die de diagonaal volgt, over het algemeen als stippellijn aangegeven in ROC curves.

Op basis van de ROC curves kan het oppervlakte onder de curve berekend worden (AUC). Een andere goede vuistregel voor het beoordelen van modellen in termen van hun ROC is dat elke AUC boven de 0.8 gezien kunnen worden als goed, en elke score boven 0.9 als erg goed.

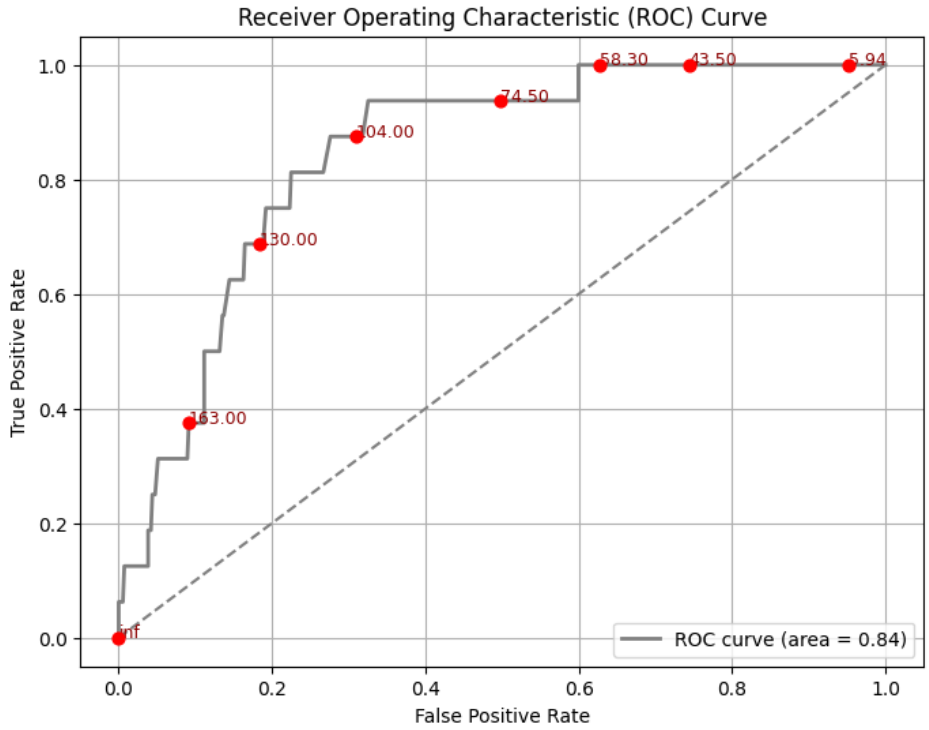
Voor de radardata in de lente worden, net als voor de confusion matrices, alle uren met een MTR hoger dan 126.6 bestempeld als hoge MTR uren. Dit is 273 voor het model voor de herfst. Dit betekent dat de ROC curves goed vergeleken kunnen worden met de confusion matrices in Afbeelding 6 en Afbeelding 7.

De ROC curve voor de lente (Afbeelding 1Afbeelding 8) komt niet in de buurt van de linkerbovenhoek. Dit betekent dat er geen drempelwaarde gekozen kan worden waarvoor zowel genoeg waardes als

correct worden bestempeld als waar ook geen sprake is van veel vals positieve bestempelingen. De ROC curve voor de herfst (Afbeelding 9) lijkt een stukje dichterbij de linkerbovenhoek te komen voor drempelwaardes tussen de 104 en de 130. Dit suggereert dat er voor het najaarsmodel drempelwaardes te kiezen zijn waarvoor het model de testdata relatief goed kan voorspellen. Hetzelfde zien we in de AUC waardes. Een waarde van 0.77 voor het voorjaarsmodel is redelijk maar niet exceptioneel goed. Een AUC van 0.84 voor het najaarsmodel suggereert dat het model de waardes goed kan voorspellen.



Afbeelding 8 - De ROC curve voor het voorjaarsmodel voor weerdata van de dag zelf. Een aantal drempelwaardes zijn als rode stippen aangegeven als referenties. De AUC is rechtsonder zichtbaar.



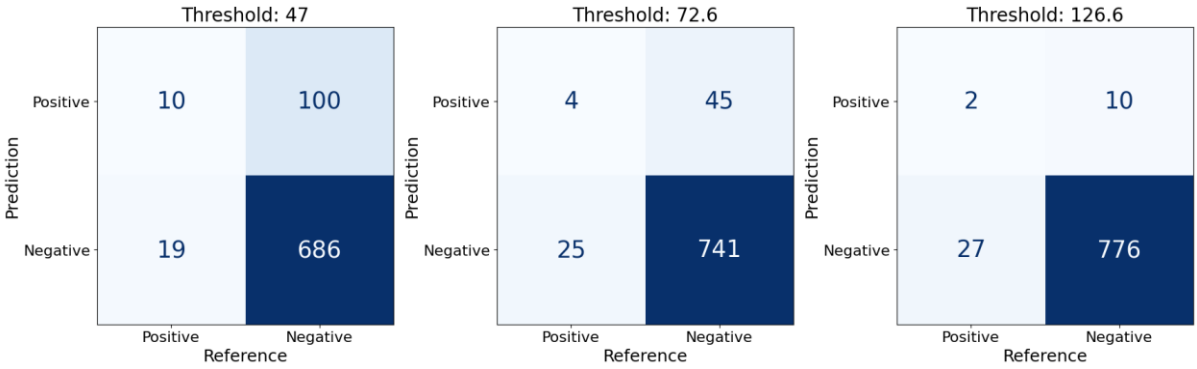
Afbeelding 9 - De ROC curve voor het najaarsmodel voor weerdata van de dag zelf. Een aantal drempelwaardes zijn als rode stippen aangegeven als referenties. De AUC is rechtsonder zichtbaar.

Wat wel belangrijk is om in gedachten te houden is dat er in de data gebruikt voor de confusion matrices en de ROC curves erg weinig punten zitten die boven de drempelwaarde uitkomen (29 uren voor de lente en 16 voor de herfst). Dit betekent dat de waarde van de uitgevoerde validaties zeer beperkt is. Omdat er best variatie in de hoge MTR uren kan zitten, is dit slechts een indicatie van het presteren van het model.

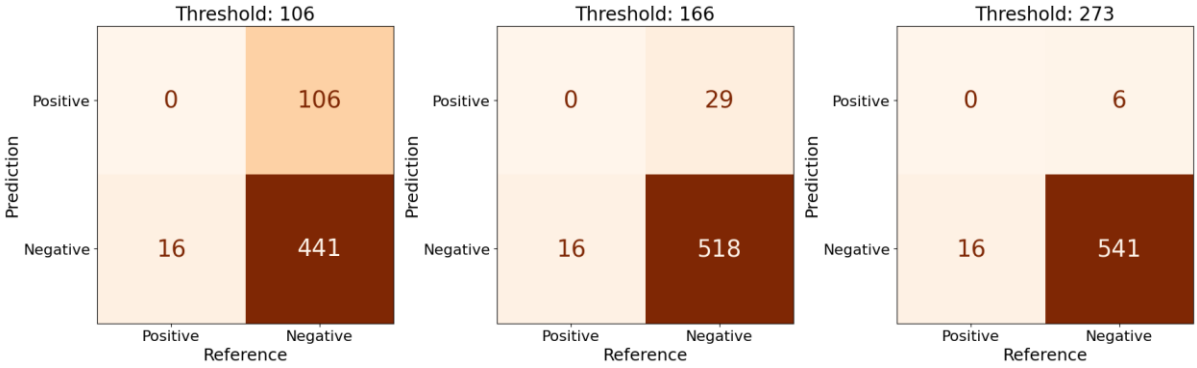
2.4.2. Validatie model op basis van weerdata van 2 dagen van tevoren (ECMWF)

Op dezelfde manier als het vorige hoofdstuk, zijn er confusion matrices en ROC curves gemaakt, waarbij het model voorspellingen heeft gedaan over de data van 2024 en voorspellingen deed met weersvoorspellingen van 2 dagen van tevoren. Deze resultaten zijn waarschijnlijk representatiever met hoe het model in de werkelijkheid zou presteren, omdat de keuze om de windturbines uit te zetten in werkelijkheid 2 dagen van tevoren gemaakt moet worden.

De confusion matrices zijn te zien in Afbeelding 10 en Afbeelding 11 voor het voorjaars- en najaarsmodel respectievelijk. De accuracy is hoog voor alle matrices (0.95 en 0.96 voor de matrices met de hoogste drempelwaardes). De precision is laag (0.17 en 0.00 voor de matrices met de hoogste drempelwaardes). Interessant om op te merken is dat er nu voor alle drempelwaardes boven de 106 MTR geen uur wordt gevonden waarbij het model correct een hoge MTR voorspelt in het najaarsmodel.

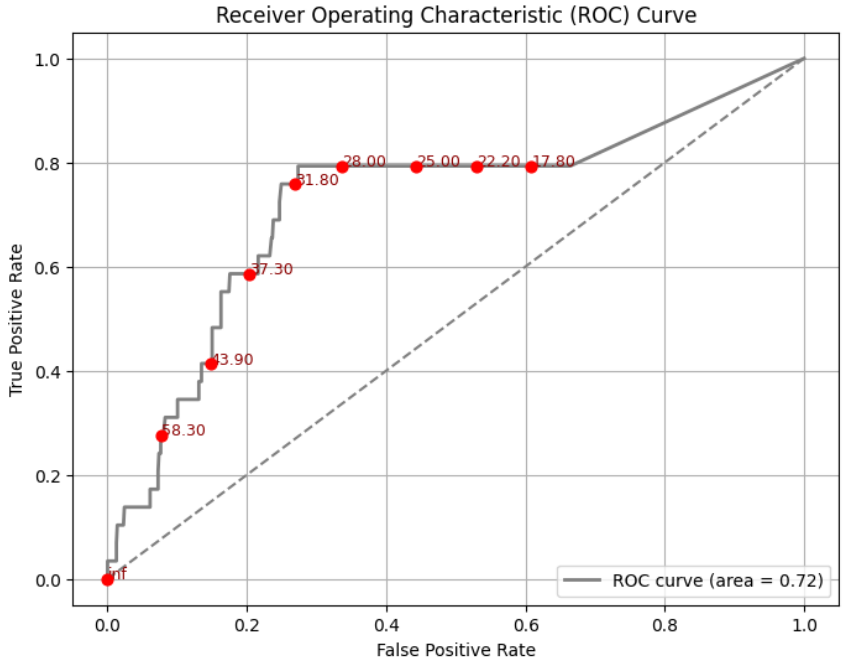


Afbeelding 10 - confusion matrices met absolute aantallen voor het voorjaarsmodel over data 2 dagen van tevoren.

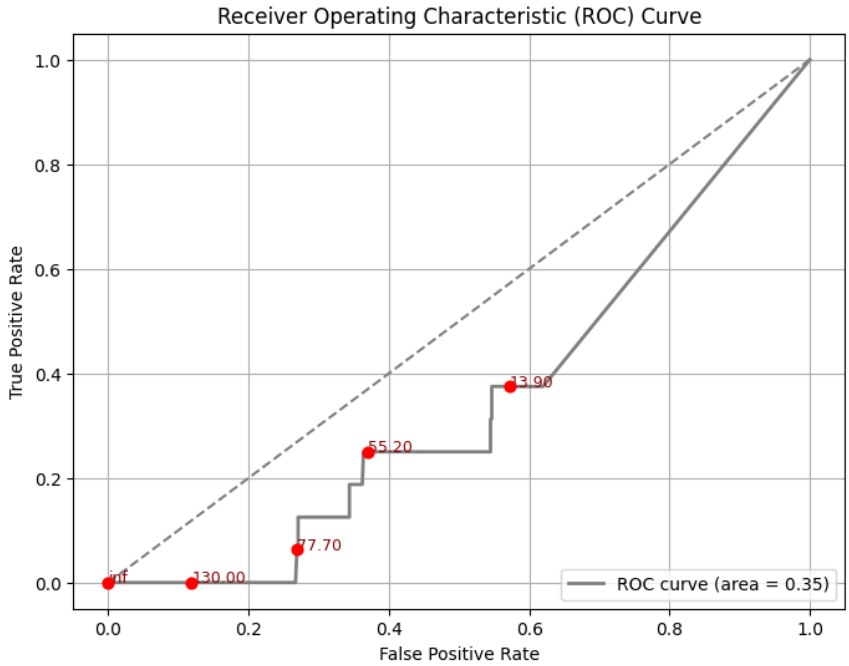


Afbeelding 11 - confusion matrices met absolute aantallen voor het najaarsmodel over data 2 dagen van tevoren.

De ROC curves van de resultaten met weerdata 2 dagen van tevoren zijn te zien in Afbeelding 12 en Afbeelding 13. Beide komen niet in de buurt van de bovenhoek en hebben ook een lage AUC. Voor het voorjaarsmodel moeten er erg lage drempelwaardes aangehouden worden om in de buurt te komen van een bruikbare TPR, namelijk drempelwaardes van rond de 40 MTR of lager. De ROC van de herfsttijd lijkt zelfs minder te presteren dan een willekeurig classificatiemodel. Deze resultaten zijn echter waarschijnlijk erg vertekenend, omdat er bij de in de praktijk toegepaste drempelwaardes nooit een correct geclassificeerde hoge MTR voorkomt.



Afbeelding 12 - De ROC curve voor het voorjaarsmodel voor weerdata 2 dagen van tevoren. Een aantal drempelwaardes zijn als rode stippen aangegeven als referenties. De AUC is rechtsonder zichtbaar.



Afbeelding 13 - De ROC curve voor het najaarsmodel voor weerdata 2 dagen van tevoren. Een aantal drempelwaardes zijn als rode stippen aangegeven als referenties. De AUC is rechtsonder zichtbaar.

Ook deze resultaten zijn gedaan over een beperkte periode met een beperkt aantal hoge MTR uren. Ondanks dat, kan er wel geconcludeerd worden dat dit model niet goed kan voorspellen wat de hoge pieken gaan zijn 2 dagen van tevoren. Voor beide modellen moeten er erg lage drempelwaardes

genomen worden om momenten te krijgen waarop de windturbines uitgezet zouden moeten worden en met zulke lage drempelwaardes gaat het aantal uren waarbij het model foutief uitgezet zou worden enorm omhoog, aangegeven door de lage precision waardes. Dit resulteert in een lage nauwkeurigheid bij het correct classificeren voor het inperken van de windturbinesnelheid.

2.4.3. Conclusie hoofdstuk Validatie over data 2024

De validatie van de nieuwe data die in 2024 is verzameld resulteert in vergelijkbare en soms zelfs minder goede resultaten dan de eerdere validaties uit dit rapport.
De vergelijking van de modelresultaten bij weersvoorspellingsdata van de zelfde dag met gebruik van weersvoorspellingsdata van 2 dagen vooraf levert geen significantie verschillen op.

2.5. Conclusie validatie

De validaties uit het paper van de UvA, de reproductie analyse en de validaties over de nieuwe (2024) data zijn over het algemeen vergelijkbaar in conclusie, namelijk beide modellen lijken een prima weergave te zijn van algemene migratiestromen. Migratietrends lijken in zekere zin voorspelbaar te zijn op basis van de gebruikte inputvariabelen. Dit komt overeen met het doel wat de UvA zichzelf heeft gesteld bij het maken van het model.

Voor het gebruik binnen de start-stop maatregel is het nodig om momenten van intense migratie te voorspellen. Voor het voorspellen van intense piek migratie heeft het model een zeer lage nauwkeurigheid. Ons advies is dat het model een te lage nauwkeurigheid heeft om in een productieomgeving te draaien. In het komende hoofdstuk zullen we factoren van het systeem en het domein bespreken die potentieel bijdragen aan deze resultaten.

3. Werking huidig model

In dit hoofdstuk zullen we kijken naar de werking van het model, de keuzes die gemaakt zijn bij het opzetten en trainen en andere factoren die invloed hebben op de resultaten van het vorige hoofdstuk. We zullen vooral kijken naar hoe de keuzes in het proces bij kunnen hebben gedragen aan de resultaten gevonden in het vorige hoofdstuk en hoe dit invloed kan hebben op het systeem in productie.

3.1. Hoeveelheid en distributie van de data

De getallen in dit hoofdstuk zijn gebaseerd op de dataset uit (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). Hier zit de data die recent verzameld is in 2024 dus nog niet bij.

Het voorjaarsmodel is getraind op data van 2019 tot 2023 en het najaarsmodel op data van 2019 tot 2022. De weerdata en (geaggregeerde) radardata zijn beide over tijden van een uur. Alleen de data 's nachts zijn gebruikt en alleen data van de betreffende maanden in het seizoen zijn gebruikt (15 februari tot 30 april voor de lente en 1 oktober tot 30 november voor de herfst). Dit resulteert in ongeveer 5000 (5 jaar * ~13 uur per dag * ~75 dagen per jaar) en ongeveer 3200 (4 jaar * ~13 uur per dag * 61 dagen) potentiële datapunten voor respectievelijk de lente en herfst.
Van die datapunten zijn er ook nog wat datapunten afgevallen tijdens het cleaning en filtering proces (van Erp, van Loon, De Groeve, Bradarić, & Shamoun-Baranes, 2024). Ook waren er momenten waarop de radardata niet beschikbaar was door technische mankementen. Hierdoor bleven er in werkelijkheid ~3450 datapunten over voor het voorjaar en ~2650 voor het najaar (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024).
Voor beide modellen is er 1 jaar (2022 voor de lente en 2020 voor de herfst) apart gehouden om het model op te valideren voor het onderzoek. Dit betekent dat voor training voor het voorjaarsmodel slechts 4 jaar en voor het herfstmodel slechts 3 jaar aan data is gebruikt in (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). Voor het productiemodel is wel de gehele dataset gebruikt. In dit hoofdstuk beschouwen we de test set dus als onderdeel van de training set, terwijl dat in het validatie hoofdstuk niet is gedaan. Op deze manier kunnen we zowel het onderzoek van de UvA als het productiemodel optimaal tot hun recht laten komen en eerlijk onderzoeken.

Machine learning modellen zijn erg afhankelijk van welke data er gebruikt wordt voor het trainingsproces. De hoeveelheid data die nodig is om een model correct te laten voorspellen is gebaseerd op de complexiteit van de verdeling (positives vs negatives) en op de hoeveelheid variabelen die gebruikt worden als input (dimensionaliteit) (Domingos, 2012). Een veelgebruikte vuistregel voor het trainen van machine learning-modellen is dat er minimaal tien keer zoveel datapunten nodig zijn als het aantal vrije variabelen in het model. Deze richtlijn is echter een ondergrens; idealiter is er aanzienlijk meer data beschikbaar. Bovendien geldt deze vuistregel alleen onder specifieke omstandigheden: wanneer alle relevante factoren bekend zijn en als inputvariabelen worden meegenomen, wanneer de onderliggende verdeling relatief eenvoudig is en wanneer de inputvariabelen onderling onafhankelijk zijn. In de praktijk zijn deze voorwaarden zelden volledig vervuld, waardoor vaak een grotere hoeveelheid trainingsdata nodig is om een robuust en generaliseerbaar model te ontwikkelen.

Voor de lente zijn er 17 variabelen gebruikt en voor de herfst 23 (Afbeelding 4). Volgens deze vuistregel zou dit model dus genoeg data moeten hebben om een enigszins correcte voorspelling te kunnen doen over de hoogte van de MTR. Dit model gaat echter binnen de start-stop maatregel voornamelijk gebruikt worden om tijden van hoge migratie te voorspellen. In het vorige hoofdstuk hebben we al gezien dat er sprake is van een erg scheve dataset, er is veel minder data beschikbaar van tijden met hoge migratie en veel meer voor tijden met een lage migratie.

Door toepassing van het stratified sampling algoritme (meer uitgelegd in de volgende paragraaf) is dit model meer gefocust op uren waar de MTR hoger is dan 91.2 en 163.5 voor respectievelijk de lente en de herfst. In de originele dataset zien we slechts 225 momenten van deze intensiteit in de lente en 208 in de herfst. Als we dezelfde vuistregel als eerder aanhouden is deze hoeveelheid data niet genoeg of net op het randje van de ondergrens. Echter, als we kijken naar de categorisatiegrens die gehanteerd worden in het paper van de UvA of zelfs naar de grens gekozen in het kavelbesluit, is er veel te weinig data beschikbaar om logische voorspellingen over te doen, zeker omdat er ook geen sprake is van de voorwaarden besproken in de vorige paragraaf voor de ondergrens van 10 keer de hoeveelheid inputvariabelen. De hoeveelheid data van uren met verschillende grenzen van hoge migratie is uiteengezet in tabel Tabel 2.

	Traindrempelwaardes voor stratified sampling	Hoogste drempelwaardes gebruikt in validatie	Kavelbesluit drempelwaardes
Voorjaar	225 (> 91.2 MTR)	173 (> 126.6 MTR)	33 (> 500 MTR)
Najaar	208 (> 163.5 MTR)	134 (> 273 MTR)	50 (> 500 MTR)

Tabel 2 - De hoeveelheid beschikbare datapunten van hoge migratie in de training dataset per drempelwaarde.

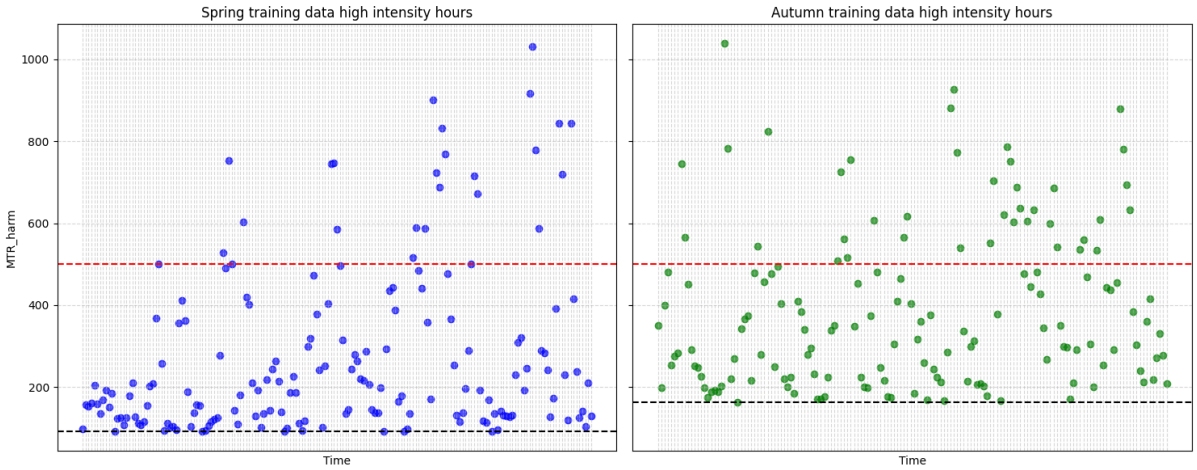
3.2. Stratified sampling

Om de effecten van de scheefheid van de dataset tegen te gaan, is er een vorm van stratified sampling toegepast om de training data mee te selecteren. Er is gekozen om voor het stratified sampling algoritme 5% van de data als data met hoge intensiteit (trekpiek) te beschouwen, dit komt overeen met drempelwaardes van 91.2 en 163.5 MTR voor respectievelijk de lente en de herfst. Er is een gewicht van 20 gegeven aan alle uren boven deze drempelwaarde en 1 aan alles er onder voor het trainen van het model. Dit heeft tot gevolg dat de uren met een trekpiek 20 keer zo veel gebruikt zullen worden in het trainen van het model dan die van lage intensiteit. Hierdoor ontstaat er een verdeling van ongeveer 50% data met trekpiek en 50% data zonder trekpiek.

Deze methode is een erg logische en toegankelijke keuze. De gebruikte R package biedt al een makkelijke oplossing om dit te implementeren (Wright & Ziegler, 2017). Er zijn ook andere opties om met scheve datasets om te gaan. Deze zijn minder geschikt omdat de dimensionaliteit van de dataset en de predictors erg hoog is. Eventueel zou er wel gekeken kunnen worden naar andere methodieken. Hier hebben we het in het volgende hoofdstuk over.

Afbeelding 14 is een grafische weergave van de verdeling van datapunten die in het stratified sampling algoritme aangeduid worden met een hoger gewicht. Er is te zien dat de MTR van het grootste gedeelte van de datapunten aan de lage kant liggen, dichter bij de ondergrens dan bij een hogere MTR zoals 500. In het stratified sampling algoritme worden deze punten veel vaker gebruikt

voor het trainen van het model dan de andere punten. Echter, voor het soort voorspellingen dat voor de start-stop maatregel relevant is, zijn vooral de bovenste waardes interessant. Het is niet de bedoeling dat het systeem de windturbines stil zet voor de waardes rond de ondergrens van deze classificatie, alleen voor die aan de bovenkant. Door de ondergrens van de classificatie te zetten op de hoogste 5% van de totale dataset, legt het model veel focus op de data tussen de zwarte en rode lijnen. Het model wordt hierdoor wel getraind om pieken te herkennen, maar vooral om middelgrote en kleine pieken te herkennen.



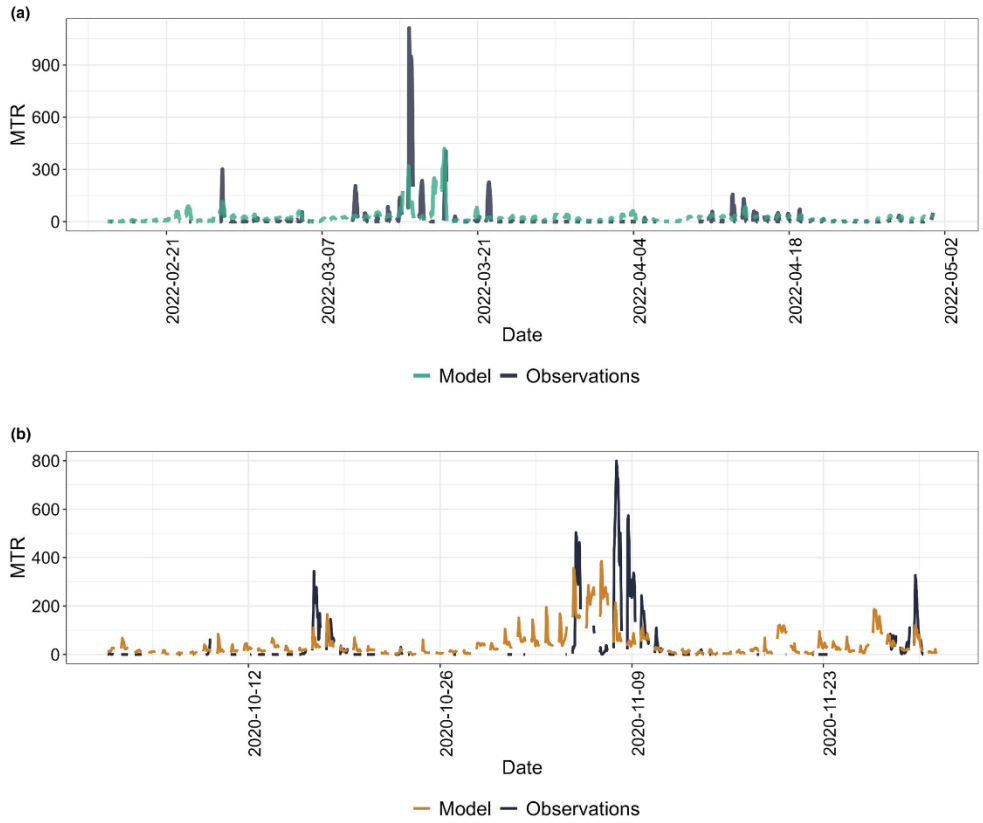
Afbeelding 14 - grafiek van alle punten gecategoriseerd als 'hoge intensiteit' voor het stratified sampling algoritme. De zwarte lijn geeft de ondergrens voor de classificatie van hoge intensiteit aan, dat is 91.2 voor de lente en 163.5 voor de herfst. De rode lijn geeft een MTR van 500 aan.

3.3. Keuze drempelwaarde

Er is meerdere keren in het paper gekeken naar de juiste drempelwaarde om te gebruiken om de resultaten van dit model mee te interpreteren. De drempelwaardes die gebruikt zijn voor het stratified sampling algoritme zijn het laagst, namelijk 91.2 voor het voorjaar en 163.5 voor het najaar. Deze drempelwaardes zijn gekozen zodat 5% van de dataset in de categorie met hoge intensiteit (trekpiek) zou vallen en 95% in de andere.

Voor het analyseren van het model zijn er meerdere verschillende drempelwaardes gebruikt. In de confusion matrices zijn er al een paar zichtbaar. In het paper van de UvA is geen duidelijk advies gegeven over welke drempelwaarde aangehouden moet worden in de windparken.

Opmerkelijk is dat geen van de genoemde drempelwaardes in de buurt komen van de aangeraden MTR van 500. Ook al zien we in de achtergehouden validatieset 7 datapunten voor de lente en 11 voor de herfst die deze waarde overschrijden, is het model nooit uitgekomen op een resultaat hoger dan ~400. Dit is te zien in Afbeelding 15. Om dit probleem tegen te gaan, is er dus voor gekozen om de drempelwaarde zo te kiezen dat er wel overschrijdingen te zien zijn in de voorspellingen. Dit heeft echter tot het gevolg dat het model veel vaker een MTR voorspelt op basis waarvan de windturbines uitgezet moeten worden, terwijl het eigenlijk heel weinig voorkomt. Dit laatste is goed zichtbaar in de confusion matrices in het vorige hoofdstuk, namelijk het getal rechtsboven in elke matrix.



Afbeelding 15 - De resultaten van het model en de daadwerkelijk gemeten datapunten door de radar op de achtergehouden test set. De zwarte lijnen zijn de gemeten punten en de gekleurde lijnen zijn de voorspellingen van het model. De blauwe lijn (a) correspondeert met het voorjaarsmodel en de oranje lijn (b) met het najaarsmodel.

3.4. Processing radardata

De processing en post-processing van de radardata valt lichtelijk buiten de scope van dit rapport, aangezien het geen onderdeel is van het machine learning model. Echter, omdat het voor het trainen gebruikt wordt om te bepalen wanneer er een vogeltrek heeft plaatsgevonden, is het wel een belangrijke factor die invloed kan hebben op het training proces en de uiteindelijke resultaten van het model. In (van Erp, van Loon, De Groeve, Bradarić, & Shamoun-Baranes, 2024) staat uitgebreid uitgelegd hoe de radardata ingewonnen en verwerkt is. In dit rapport gaan we niet in op de technische details van dit onderzoek. Er zijn daarna nog post processing stappen ondernomen (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024), om de data voor te bereiden voor gebruik voor het machine learning model. Het belangrijkste resultaat daarvan is dat er onbetrouwbare uren van de dataset verwijderd zijn. Dit is een logische stap voor het trainen van een machine learning model. Om het model zo goed mogelijk conclusies te laten trekken uit een dataset, is het nodig dat de data zo betrouwbaar mogelijk is. Enige vorm van data cleaning en processing is daardoor te verwachten. Dat er veel extra cleaning stappen nodig zijn is het gevolg van de accuraatheid van de radar. Het is moeilijk te zeggen hoe accuraat die radar echt is.

Er wordt kort uitgelegd hoe de post-processing is verlopen (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). Veel van de cleaning stappen hebben te maken met het verminderen van de invloed van *clutter* op de radardata. *Clutter* zijn alle objecten in het vizier van de radar die iets

anders zijn dan vogels. Een van de stappen die misschien invloed kan hebben op de werking van het model, is de volgende; tijdens uren waar de hoeveelheid *clutter* erg hoog is, wordt de radardata als te onbetrouwbaar bestempeld. Deze uren worden dan uit de dataset gefilterd. Er is een correlatie tussen de hoeveelheid *clutter* en de weersomstandigheden; Als er veel wind is, zal er meer *clutter* zijn (van Erp et al; WE reports). Dit betekent dat het zou kunnen dat de hoogte van het migratievolume in de training set anders is dan deze in de werkelijkheid is en dat er een vorm van *selection bias* is geïntroduceerd (Wang & Singh, 2021). Alhoewel dit een ongelukkig resultaat is van de filtering, is het wel een noodzakelijke stap om de data zo betrouwbaar mogelijk te krijgen.

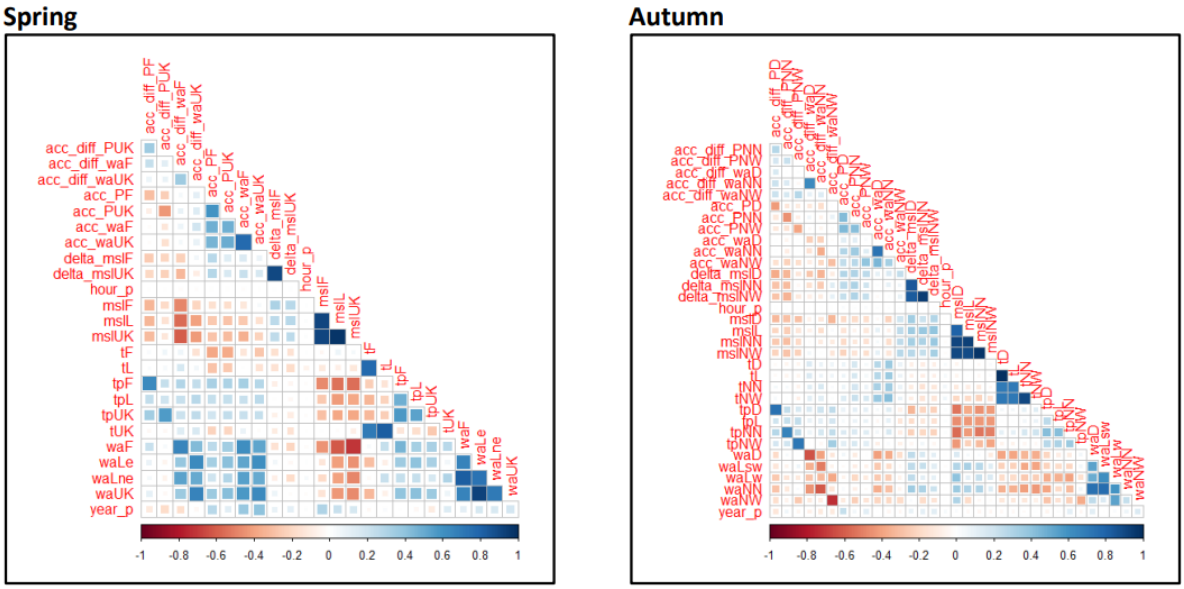
Tabel 3 is een uiteenzetting van de hoeveelheid beschikbare datapunten in de volledige dataset die na het filtering proces nog overgebleven zijn, weergegeven per uur van de dag. Er lijkt een kleine discrepantie te zijn in hoeveel data er per uur verzameld of overgebleven is. Zo zien we bijvoorbeeld data aanwezig rond 07:00 uur 's ochtends, terwijl deze eigenlijk niet gebruikt hoort te worden voor het trainen van het model. Dit soort verschijnselen kunnen wijzen op fouten in het verzamelen van de data of het cleaning proces. Wij hebben wat code ontvangen over hoe het cleaning en training proces is verlopen. Echter kunnen we hier niet in traceren waar deze eventuele fouten vandaan komen. Wij vermoeden daardoor dat wij een andere versie van dit proces in handen hebben dan gebruikt is voor het onderzoek zelf. Hierdoor hebben wij niet volledig uit kunnen zoeken waar deze discrepanties aan kunnen liggen, en of er nog andere aannames zijn gedaan die niet uitgelegd zijn. Naar aanleiding van een analyse van de code vermoeden we dat dit soort discrepanties echter weinig invloed hebben over het uiteindelijke productiemodel.

Tijd	Hoeveelheid data in "spring.csv", waar MTR aanwezig is	Hoeveelheid data in "autumn.csv", waar MTR aanwezig is
00:00	290	152
01:00	286	184
02:00	288	187
03:00	289	191
04:00	291	188
05:00	172	184
06:00	88	141
07:00	0	35
15:00	0	34
16:00	0	132
17:00	94	176
18:00	229	175
19:00	279	172
20:00	285	170
21:00	283	168
22:00	286	170
23:00	291	178
Gemiddelde over alle uren	246.5	155.1

Tabel 3 - Hoeveelheid beschikbare data gebruikt in (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024). Alleen de uren waarvoor een MTR waarde beschikbaar is na post-processing, zijn weergegeven.

3.5. Predictors

In Afbeelding 16 zijn correlatiematrices voor alle potentiële weersvariabelen te vinden die als invloedsmogelijkheden naar boven kwam in voorgaand onderzoek (Bradarić, Bouten, Fijn, Krijgsveld, & Shamoun-Baranes, 2020) (Manola, et al., 2020). Er is te zien dat sommige variabelen erg sterk correleren met anderen. Omdat het weinig extra waarde heeft voor het trainen van het model zijn deze verwijderd uit de trainingsdata en zijn niet gebruikt als variabelen in het model. Hierdoor zijn er 7 variabelen verwijderd uit de set voor het voorjaar en 12 uit die van de herfst. Dit resulteerde in uiteindelijk 17 variabelen voor het voorjaarsmodel en 23 voor het najaarsmodel.



Afbeelding 16 - Correlatiematrices voor alle potentiele weersvariabelen zoals gerapporteerd door de UvA (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024).

Door de allerhoogst gecorreleerde variabelen te verwijderen wordt het probleem van de complexiteit van het systeem (het aantal piekmomenten t.o.v. het aantal potentiële variabelen) al een stukje verholpen. Het model zal sneller en met minder data tot een oplossing komen omdat het makkelijker de bijdrage van elk van de variabelen individueel kan achterhalen. Echter is het wel duidelijk aan de correlatiematrices dat de overgebleven variabelen ook niet compleet onafhankelijk zijn. Dit is een indicatie dat het voorspellen van de migratiestromen een complex probleem is en dat er waarschijnlijk veel data nodig gaat zijn om het model naar de juiste schattingen van de bijdragen van de variabelen te laten komen. Helaas kan hier waarschijnlijk niet veel aan gedaan worden en is het ook logisch dat dit soort resultaten zichtbaar zijn. Het gedrag en de invloeden van migratiepatronen zijn moeilijk te doorgronden.

Naast de mogelijkheid dat bepaalde variabelen in het model overbodig zijn, kan het ook zo zijn dat essentiële variabelen ontbreken die een significante invloed hebben op het gedrag van migratievogels. Bij Technolution zijn we geen experts op het gebied van ecologie of vogeltrek, en we hebben geen diepgaand onderzoek gedaan naar de onderliggende drijfveren van deze

migratiestromen. Vanuit een buitenstaandersperspectief achten we het echter goed mogelijk dat er cruciale factoren ontbreken die de nauwkeurigheid van het model zouden kunnen verbeteren.

Een variabele die logisch lijkt om toe te voegen aan het systeem, en in zekere mate ook al onderdeel is van het productiesysteem, is een vorm van feedback van de live radardata tot het punt van de voorspelling. Vogels die al overgevlogen zijn drie dagen geleden, zullen vandaag niet opnieuw overvliegen. Dit is zelfs het geval als de weersomstandigheden vandaag gunstiger zijn dan de omstandigheden drie dagen geleden. Ook kan het misschien zo zijn dat vogels meer druk ervaren om te gaan overvliegen later in het seizoen, maar alleen als ze tot dan toe nog niet zijn gaan vliegen. Dat kan tot het gevolg hebben dat vogels later in het seizoen minder gunstige omstandigheden zullen accepteren, gegeven dat je weet dat ze nog niet eerder zijn gaan migreren. Als laatste zou het ook heel goed kunnen zijn dat er meer vogels overvliegen op een dag dat er al veel vogels langs zijn gekomen, of juist minder. Dit omdat ze misschien door elkaar beïnvloed worden, dat alle omstandigheden (ook die niet onderdeel van de training data) gemiddeld die dag gunstiger zijn of dat ze juist al langs zijn gekomen en dat daardoor de piek lager is dan voor andere momenten met dezelfde weersomstandigheden.

Op dit moment is er wel enige vorm van feedback dat terug het model wordt ingestuurd. Dit gebeurt op basis van aannames over het gedrag van vogels over de afgelopen veertien dagen in de vorm van accumulatie. Deze voorspellingen worden als extra training variabele in het productiesysteem voor het model gebruikt. Dit zorgt voor een relatief makkelijke vorm van herintroductie van recent historische data. Echter zal dit wel tot het gevolg hebben dat deze variabelen verantwoordelijk zijn voor het introduceren van extra ruiskansen, namelijk, een foute voorspelling geeft direct een fout in de accumulatie en het model gaat vervolgens trainen en voorspellen op basis van deze foute factor. Het lijkt een logische en bereikbare oplossing te zijn om in plaats van deze factor, hier directe feedback vanuit de radar te gebruiken. Deze radar is constant metingen aan het uitvoeren, mede om daarna als nieuwe trainingsdata voor het model te kunnen dienen. De data van deze radar zou in *real-time* schoongemaakt en gefilterd moeten worden, om vervolgens terug als variabele voor het model gebruikt te kunnen worden. Het toevoegen van deze vorm van feedback van de werkzame radar zou hiermee een waardevolle informatiebron kunnen zijn. Er zou eventueel een *sliding window* effect toegepast kunnen worden, waarbij de cumulatieve MTR van de afgelopen paar uur, van de afgelopen week of een ander tijdsbereik als variabele meegenomen wordt. Er moet dan wel rekening gehouden worden met dat de voorspelling twee dagen van tevoren gedaan moet worden, dus alleen radardata van langer dan twee dagen geleden zou gebruikt kunnen worden. De voordelen en nadelen van zo'n radar-terugkoppeling zal verder moeten worden onderzocht.

3.6. Ruis in radardata en weerdata

Een belangrijke factor om te onderzoeken bij elk domein dat van machine learning gebruik maakt, is in hoeverre het systeem zo ingericht is dat het model kan voorspellen wat daadwerkelijk beïnvloed wordt door de input variabelen. In eerder onderzoek is gebleken dat de migratietrends van de vogels beïnvloed worden door de weersvariabelen. In ideale omstandigheden zou er dan een model getraind kunnen worden dat in een vacuüm, naast de vogels en de weersvariabelen, een perfecte voorspelling doet. Echter is dit in werkelijkheid niet het geval.

Er zijn veel factoren die de voorspelling, en daarmee de correctheid van de voorspelling, beïnvloeden. De uiteenzetting van de totstandkoming van de voorspelling is te uiten in de volgende formule:

$$V = W + f(w2) + f(test\_radar) + f(model) + \varepsilon$$

Waarbij *f(model)* staat voor:

$$f(model) = f(train\_radar) + f(w0) + lim$$

- *V*: De uiteindelijke voorspelling in MTR, gemaakt door het machine learning model.
- *W*: De werkelijke verdeling van de migratietrends in MTR. Dit is de verdeling die het model probeert te benaderen en in theorie voorspelbaar is op basis van weersvariabelen. Dit is in de ideale situatie gelijk aan *V*.
- *f(w2)*: De imperfectie in de voorspelling van de weersvariabelen 2 dagen van tevoren, dus de foutmarge in de inputvariabelen. Omdat deze voorspelling 2 dagen van tevoren gedaan moet worden, is deze vaak significant.
- *f(test_radar)*: De imperfectie in de radardata die gebruikt wordt om de resultaten van de voorspellingen van het model te toetsen en daarmee bijvoorbeeld de grenswaarde voor het interpreteren bepaald wordt. Door de limitaties van de radar zelf, het bereik van de radar, technische mankementen, minder correcte voorspelling bij heftig weer, etc., wordt er een significante foutmarge geïntroduceerd.
- *ε*: Een factor van inherente willekeur. In elke machine learning toepassing is er altijd een oprecht onvoorspelbare factor. Soms zullen vogels nou eenmaal niet dezelfde keuze maken onder dezelfde omstandigheden.
- *f(model)*: De imperfectie van het model. Deze is dan weer beïnvloed door het training proces, en daarmee door de volgende factoren:
    - *f(train_radar)*: De imperfectie in de historische radardata die gebruikt is om het model op te trainen.
    - *f(w0)*: De imperfectie in de historische weerdata van 0 dagen van tevoren, gebruikt om het model op te trainen. De foutmarge van deze data is waarschijnlijk erg klein, echter kunnen er altijd fouten voorkomen.
    - *lim*: De limitaties van het model waar we het vooral over hebben in dit rapport; limitaties door de hoeveelheid/scheve data, het gekozen model, de gekozen variabelen, etc. Dit is de fout die verminderd wordt door het model te verbeteren of meer te trainen.

Zoals er gezien kan worden, worden de resultaten van dit proces door erg veel factoren beïnvloed, naast alleen de werking van het model. Dit geld niet alleen voor het ontwikkelen en trainen van het proces (dus de onderste formule), maar ook voor het voorspellen zelf en de waarnemingen in de werkelijkheid (de bovenste formule). In het proberen te verbeteren van de werking van dit model zijn er erg veel mogelijkheden die de *lim* factor kunnen verminderen, maar dit zal weinig positieve resultaten opleveren als de andere factoren grote fouten veroorzaken. Deze factoren noemen we in dit geval ruis, dat zijn wisselingen in de werkelijke uitkomst die buiten de controle en buiten het bereik van het model liggen.

Het kan mogelijk zijn om tot zekere hoogte de invloed van deze ruis te bagatelliseren. Dit kan bijvoorbeeld voor externe factoren die in een vorm afhankelijk zijn van de inputvariabelen. Een voorbeeld hiervan is dat de radardata minder betrouwbaar is als de weersomstandigheden heftiger zijn. Echter ben je dan niet meer alleen het model aan het trainen om voorspellingen te doen over de werkelijke MTR, maar ook impliciete externe factoren te voorspellen over de werking van bijvoorbeeld de radar. Dit betekent dat er nog meer data nodig is om dat soort distributies te voorspellen. En

daarnaast is het niet mogelijk om dit te doen voor sommige andere factoren, zonder extra inputvariabelen toe te voegen. Een voorbeeld hiervan is de accuraatheid van de weersvoorspellingen 2 dagen van tevoren. Deze zijn namelijk nauwelijks voorspelbaar vanuit de inputvariabelen en daarom dus niet impliciet te leren voor het model.

Dat er zoveel factoren zijn die extra ruis introduceren in dit voorspelproces, is een belangrijk knelpunt voor het doel om een precies genoeg systeem te ontwikkelen dat met vrij hoge zekerheid kan voorspellen dat er sprake is van daadwerkelijk genoeg vogels om de windturbines in te perken. Sommige factoren zouden eventueel verbeterd kunnen worden om de hoeveelheid ruis te verminderen (bijvoorbeeld meer accurate radars plaatsen), maar dit kunnen technisch complexere en dure oplossingen zijn. En daarnaast blijft er zelfs met andere oplossingen ook een vorm van ruis, en zijn er factoren die niet oplosbaar zijn (bijvoorbeeld de weersvoorspellingen 2 dagen van tevoren).

3.7. Conclusies uit hoofdstuk: Huidige werking model

Om het model goed te trainen is de beschikbare data ongeschikt. Dit is voornamelijk omdat er te weinig trekpieken zijn ten opzichte van de perioden zonder piektrek. Door het toepassen van datavoorbereiding (stratified sampling) kan dit enigszins worden verbeterd, maar in de praktijk blijkt dat te weinig om een hoge nauwkeurigheid te behalen. Daarnaast is het verschil in waarvoor het model is ontwikkeld (het zien van trends in vogeltrek) en waar het voor wordt toegepast (het op de juiste tijd voorspellen van een grote vogeltrek) anders. Daarvoor zou het model anders moeten worden getraind, echter is dit niet mogelijk door de lage hoeveelheid representatieve data.

De vogelradarsystemen worden gebruikt om te bepalen wanneer er een vogeltrek plaatsvindt. Dit heeft echter ook beperkingen. Zeker omdat de radar maar een beperkt zicht heeft op de Noordzee (een vogeltrek kan makkelijk langs het radarzicht vliegen), de radarsystemen veel uitstaan, veel last hebben van clutter, enzovoorts, is niet alle data even betrouwbaar.

Het aantal predictors (inputdata waarop het model zijn voorspelling baseert) is relatief hoog ten opzichte van de vuistregels over wat mogelijk is met dit type model en data. Dit is echter niet eenvoudig te verminderen. Verbeteringen hierop dienen verder te worden onderzocht.

Alles bij elkaar is de hoeveelheid ruis in de uitkomst van het model hoog. Daarmee is de nauwkeurigheid van het model voor het voorspellen van trekpieken laag.

4. Eventuele verbeteringen

Zoals gezien in de vorige hoofdstukken, lijkt de grootste oorzaak van het suboptimaal presteren van het model te zijn dat er te weinig data beschikbaar is en dat de beschikbare data erg scheef is, waardoor het probleem met weinig data nog verergert voor het voorspellen van piekmigratie. In dit hoofdstuk zullen we kijken of er manieren zijn om de voorspelling te verbeteren.

Hierbij tekenen we aan dat er geen garantie is dat het doorvoeren van deze suggesties de knelpunten wegnemen. Wij verwachten dat de kans groot is dat geen van deze verbeterpunten genoeg zal zijn om het model resultaten te laten bereiken van het significant beter precisieniveau. Het modelleren en voorspellen van gebeurtenissen in een complex systeem zoals wat hier van toepassing is, is erg moeilijk en het zou goed kunnen dat met de huidige hoeveelheid data en ruis er geen manier is om een model te creëren dat significant beter is dan het huidige model.

4.1. Meer data

De grootste beperking bij het trainen van dit model is het gebrek aan voldoende data. De meest voor de hand liggende oplossing is daarom het verzamelen van meer gegevens. Hoewel dit vanzelf zal gebeuren doordat de radar continu meet en data verzamelt, zou het significant lang duren (waarschijnlijk 10+ jaar) voordat er voldoende data beschikbaar is om representatieve resultaten te bereiken. Het is lastig om hier een exacte tijdsinschatting voor te geven, maar als we uitgaan van de ondergrens van tien keer het aantal weersvariabelen, zou het nog minstens 10 tot 20 jaar duren voordat deze ondergrens wordt bereikt voor data met een hoge MTR (>500).

Als er alternatieve methoden worden gevonden om meer data te verzamelen—zoals het uitlezen van andere radars of het plaatsen van extra radars op nabijgelegen locaties—kan dit proces mogelijk worden versneld. Hoeveel sneller de ondergrens van benodigde data dan bereikt kan worden, hangt echter sterk af van de genomen maatregelen en is lastig exact te bepalen. Van belang hierbij is dat de extra verzamelde data minder onafhankelijk zal zijn. Bijvoorbeeld, radars die dicht bij elkaar staan, zullen waarschijnlijk vergelijkbare metingen opleveren, wat de effectieve toename van unieke informatie beperkt. Bovendien kan het noodzakelijk zijn om extra variabelen aan het model toe te voegen, zoals de afstand tot de vertreklocatie per datapunt, wat de complexiteit vergroot. Daarom is het essentieel om deze alternatieve methoden zorgvuldig te onderzoeken voordat maatregelen worden genomen, zodat de toegevoegde waarde van extra dataopslag en -verwerking goed wordt afgewogen tegen de potentiële verbeteringen in modelprestaties. Daarbij zal de benodigde investering ook een belangrijke rol spelen.

Een andere vorm van toevoegen van extra data aan het systeem zou het imputeren (toevoegen van ontbrekende data op basis van een algoritme) van missende variabelen kunnen zijn. Op dit moment mist er ongeveer 27% aan radardata in de dataset voor de lente en 43% voor die in de herfst. Deze missende uren aan data kan door verschillende dingen komen, zoals uren voor onderhoud van de radar of uren waarbij het weer zo heftig was dat de radar geen betrouwbare waardes kon opleveren. Er zou kunnen worden onderzocht of deze uren aan data opgevuld kunnen worden door middel van imputatie. Imputatie van deze waardes is echter wel lastig met data in hoge dimensionaliteit en met zo weinig datapunten om het mee te vergelijken in hoge MTR. Daardoor zal het veel minder

betrouwbaar zijn. Omdat er al veel ruis in de dataset zit en het een complexe verdeling is, is het belangrijk om voorzichtig te zijn met het implementeren van dit soort algoritmes.

4.2. Vermindering van fouten veroorzaakt door predictors

In Afbeelding 4 is te zien dat niet alle resterende variabelen even belangrijk zijn in de werking van het model. Nu deze kennis beschikbaar is, is het een mogelijkheid om modellen te trainen waarbij er steeds minder variabelen zijn meegenomen, waarbij steeds de meest belangrijke variabelen wel in het model worden gehouden. Over het algemeen zal een model met minder variabelen met minder resultaten een correcte voorspelling kunnen doen (Domingos, 2012), ervan uitgaande dat de weggelaten variabele niet essentieel was in het modelleren van de werkelijkheid. Afbeelding 4 doet vermoeden dat dit bij een aantal variabelen minder relevant zal zijn in het ecosysteem waar we naar aan het kijken zijn. Door een test set apart te houden kunnen dan de verschillende modellen gevalideerd worden op dezelfde manier als dat nu voor dit model is gedaan. Daaruit kan een keuze gemaakt worden over het aantal variabelen dat behouden moet worden om de migratie zo correct mogelijk te kunnen voorspellen. De eerder genoemde absolute ondergrens van de hoeveelheid vereiste data wordt significant verlaagd op het moment dat er minder inputvariabelen worden gebruikt, waardoor er al vrij snel een beter werkend model bereikt zou kunnen worden, gegeven dat deze variabelen daadwerkelijk weinig invloed uitoefenen op de beslissing van de trekvogels.

Een andere overweging in de keuze voor de inputvariabelen van het model is het meenemen van een vorm van feedback vanuit de vogelradars, zoals besproken in het vorige hoofdstuk. Door het toevoegen van een feedbackloop vanuit de gemeten MTR over de afgelopen paar uur, of de afgelopen paar dagen, of een combinatie van meerdere tijdsvensters, kan er gekeken worden of dit invloed heeft op de resultaten van het model. Het is onze verwachting dat dit een waardevolle en relatief makkelijke toevoeging is aan het systeem, maar het zou ook een negatief effect kunnen hebben. Daarbij moet wel in de gaten worden gehouden dat er nu al veel ruis in het systeem wordt geïntroduceerd doordat de radar vaak inaccurate data levert, en daardoor kan een extra afhankelijkheid van de radar het systeem ook minder accuraat maken. Desalniettemin lijkt het een erg belangrijke factor in de voorspelling van de migratie en is het waardevol om dit te onderzoeken.

4.3. Ruisvermindering

In het vorige hoofdstuk zagen we dat er vrij veel externe factoren meespelen in het proces, die elk in meer of mindere mate ruis introduceren in de resultaten van het model. Het verminderen van deze ruis zou tot gevolg kunnen hebben dat het model minder data nodig heeft om juiste inferenties te maken over de migratiedistributie en over de interpretatie van het ecosysteem in zijn geheel. Het verminderen van deze ruis is echter een moeilijk en wellicht kostbaar proces. Er zou bijvoorbeeld geïnvesteerd kunnen worden in het neerzetten van meer accurate radars, die een groter bereik en een hogere technische betrouwbaarheid hebben. Ook kan er onderzoek gedaan worden naar het gebruiken van andere technieken voor het detecteren van de vogels, in plaats van het semi-handmatig filteren en schoonmaken van de data, zoals meer geavanceerde computer vision technieken. Het implementeren van dit soort maatregelen zou echter een vrij lang en wellicht kostbaar proces kunnen zijn en slechts een minimale verbetering opleveren, aangezien er dan nog steeds veel ruis in het systeem zal blijven bestaan wat niet of veel moeilijker te bestrijden is, zoals geïntroduceerd door de weersvoorspellingen twee dagen van tevoren.

4.4. Andere soorten modellen

In (Bradarić, Kranstauber, Bouten, & Shamoun-Baranes, 2024) is niet heel diep ingegaan op waarom er gekozen is voor een random forest model. Na gesprek met de onderzoekers in het kader van dit rapport is duidelijk geworden dat ze dit vooral gekozen hebben door praktische overwegingen en omdat het een gangbaar model is voor dit soort toepassingen. Verder vonden de onderzoekers het ook erg belangrijk dat het resulterende model interpreteerbaar is; dat het duidelijk is in hoeverre elke weervariabele bij heeft gedragen aan de keuze van een bepaalde voorspelling. Dit kan significante bijdrages leveren aan verder begrijpen en verder onderzoek van deze migratiestromen. Bij een random forest model is dit erg duidelijk, zoals te zien aan de correlatiematrices.

Over het algemeen lijkt ons de keuze voor een random forest model een goede keuze; het is een relatief simpel algoritme dat geschikt is voor vraagstukken met veel dimensies. Toch is het de moeite waard om te onderzoeken of er alternatieven zijn. De volgende modellen lijken ons de moeite waard om te onderzoeken.

- **Gradient boosting.** GBMs zijn over het algemeen beter in het omgaan met scheve datasets dan random forest modellen, omdat ze meer focussen op moeilijk te voorspellen datapunten (Louk & Tama, 2022). Echter wordt de kans op overfitting op de specifieke momenten met een hoge MTR dan wel groter en is het lastiger om de resultaten te interpreteren in termen van ecologische betekenis. Bij het huidige model is nu heel duidelijk hoe belangrijk elke weersvariabele was voor het trainen van het model, wat meer inzicht in de werking geeft.
- **Neurale netwerken.** Neurale netwerken zijn heel goed in het interpreteren en trainen op situaties met erg complexe distributies en veel variabelen; iets wat hier zeker het geval is. Echter vermoeden we dat dit met de huidige hoeveelheid data niet realistisch is, omdat er voor een neuraal netwerk nog veel meer data nodig is dan voor een random forest model. Ook zal hetzelfde probleem als bij gradient boosting ontstaan; neurale netwerken zijn ‘black boxes’. De betekenis en werking van het model zelf is niet meer duidelijk na training.
- **Een model dat gespecialiseerd is op tijdsgebaseerde data, zoals SARIMA.** Bijna alle variabelen in dit model zijn gecorreleerd met de waardes die in de uren ervoor of erna komen; niet alleen de hoeveelheid vogels is erop gebaseerd of er in de voorgaande uren of dagen al veel vogels zijn geweest, maar de weersvariabelen zijn natuurlijk ook sterk gecorreleerd op andere uren van dezelfde dag. Modellen als SARIMA zijn specifiek ontwikkeld om seizoensgebonden trends in de data te herkennen. Dit soort algoritmes zijn echter waarschijnlijk niet complex genoeg om om te gaan met de complexe en multidimensionale situatie die zich hier voordoet, maar eventueel kan er gekeken worden naar het samenvoegen van een tijdsgebaseerd model en andere machine learning modellen, zoals random forest.

4.5. Andere technieken voor stratified sampling en drempelwaarde

Het is essentieel om het model zo te trainen dat het zich richt op de MTR-uren die cruciaal zijn voor voorspellingen in de praktijk. Uit de validatieresultaten blijkt echter dat dit momenteel niet het geval is. Het model lijkt zich vooral te concentreren op uren met een lage MTR of op kleine pieken in de trends, wat waarschijnlijk het gevolg is van de scheve verdeling binnen de dataset. Hierdoor worden de meest kritieke migratiemomenten mogelijk niet nauwkeurig genoeg voorspeld, wat de effectiviteit van het model in de productieomgeving kan beperken. Het is moeilijk om het probleem dat de pieken niet goed worden voorspeld tegen te gaan; als we de drempelwaarde van het stratified sampling

algoritme verhogen naar de gewenste 500 MTR, hebben we niet genoeg datapunten over om artificieel toe te voegen. Als we dat doen, zal het systeem enorm overfitten op de enkele datapunten die in de afgelopen paar jaren gezien zijn en zal dan niet genoeg kunnen generaliseren om in de toekomst ook correct momenten van hoge MTR te voorspellen. Als we helemaal geen stratified sampling toepassen, houden we het probleem van dat de dataset scheef is en zal het model waarschijnlijk convergeren naar resultaten van lage MTR en zal het (bijna) nooit een hoge MTR aangeven. Dit is een erg lastig probleem, en wellicht niet op te lossen voor de huidige hoeveelheid en distributie van data. Zoals eerder besproken, kan dit probleem mogelijk worden verminderd door aanzienlijk meer data te verzamelen. Afhankelijk van de hoeveelheid en de nauwkeurigheid van de verzamelde data, kan vervolgens worden overwogen om de grenswaarde voor categorisatie binnen het stratified sampling-algoritme aan te passen.
Hoewel dit het probleem niet volledig zal oplossen, kan het de moeite waard zijn om ook andere technieken te overwegen om de invloed van de scheefheid in de data te verminderen. De volgende methoden zouden mogelijk onderzocht kunnen worden. Het is lastig om vooraf in te schatten hoeveel resultaat deze onderzoeken zullen opleveren. Ons advies is om voor elke optie een kort onderzoek uit te voeren door een vergelijkbaar model te trainen en deze vervolgens te vergelijken met zowel elkaar als het huidige model. Op basis daarvan kan worden bepaald of deze aanpakken waardevolle verbeteringen bieden.

- **SMOTE.** SMOTE is een techniek die toegepast kan worden om extra datapunten te creëren op basis van nabijheid tot al bestaande datapunten. Net als bij stratified sampling zal dan de distributie van hoge en lage MTR uren een klein beetje gelijk getrokken worden. Eventueel zouden deze twee technieken simultaan kunnen resulteren in het significant vergroten van de hoge MTR dataset. Er moet echter wel rekening gehouden worden bij het toepassen van dit soort technieken dat de training dataset minder betrouwbaar gemaakt wordt omdat er datapunten toegevoegd worden op basis van bestaande punten, wat waarschijnlijk niet heel accuraat gedaan kan worden omdat er zo weinig uren zijn met de gewenste MTR. Ook is een techniek zoals SMOTE niet heel goed geschikt voor situaties met heel veel te trainen variabelen.
- **ADASYN.** Net als SMOTE is ADASYN een techniek om synthetisch meer datapunten te creëren van de gewenste minderheidsgroep. ADASYN doet dit op basis van de datadistributie van de groep van datapunten met hoge MTR. Deze techniek zal dezelfde voor- en nadelen hebben als SMOTE.
- **Een aangepaste loss function.** In de ranger function in R, welke gebruikt is om dit model te trainen, is de loss function de MSE (Wright & Ziegler, 2017). Een loss function wordt gebruikt om te bepalen hoe goed de voorspelde MTR is ten opzichte van de MTR in de training data. Voor een gebalanceerde dataset is de MSE een prima maatstaf om het model op te beoordelen en dus de juiste richting op te trainen. Het zou echter kunnen helpen om een andere soort loss function te gebruiken; een die verkeerde voorspellingen op hoge MTR uren meer afstraft en die grote fouten voor de momenten met minder vogels minder belangrijk acht of zelfs negeert. Dit heeft dan wel tot resultaat dat het model niet meer goed kan zijn in het voorspellen van algemene migratietrends, en alleen gefocust is op hoge MTR voorspellingen. Voor het doeleinde van de windturbines is dit echter niet heel belangrijk. Wel zal hierdoor ook de kans op overfitting van de weinige momenten van hoge MTR significant zijn. Deze custom loss function zal dan ook zorgvuldig gekozen moeten worden.

4.6. Conclusie van hoofdstuk: eventuele verbeteringen

Er zijn niet één of twee voorstellen die de nauwkeurigheid van het model significant verbeteren. Er zijn een aantal mogelijkheden geboden die verder moeten worden onderzocht.

- Verzamel meer, en vooral ook kwalitatievere data
- Onderzoek of het aantal predictors verminderen (zodat het model minder complex wordt en daarmee beter presteert met minder data) een beter resultaat oplevert
- Onderzoek of het mogelijk is om een vorm van feedback vanuit de radar toe te voegen
- Onderzoek of andere modellen beter functioneren
- Onderzoek nieuwe of andere technieken om beter met scheve data om te gaan

Het verbeteren van de nauwkeurigheid van de voorspellingen zal een combinatie van bovengenoemde punten zijn. Onderzoek hiernaar is belangrijk. Ons advies is om dit te onderzoeken met de focus op data science, in aanvulling op de huidige ecologische kennis.

5. Conclusie

5.1. Advies

Dit rapport heeft als doel een advies uit te brengen over de werking van het vogeltrekvoorspellingsmodel versie 1.1, de plaatsing ervan binnen het bredere systeem waarvoor het is ontworpen, en de mogelijke vervolgstappen. Uit het onderzoek blijkt dat de accuraatheid van het voorspellingsmodel laag is. Het huidige machine learning-model biedt waardevolle inzichten in vogeltrekpatronen, maar is voor het voorspellen van een trekpiek, vanuit een technisch perspectief in de huidige vorm onvoldoende betrouwbaar om leidend te zijn voor het operationele aansturen van windparken.

Op basis van de voor dit rapport uitgevoerde analyses hebben we de volgende oorzaken geïdentificeerd van de incorrecte voorspellingen van het model:

- Een tekort aan data en een scheve verdeling binnen de dataset
- De hoeveelheid ruis (fouten in de data, naast de echte informatie) en de omvang van de foutmarges in de factoren die het systeem beïnvloeden
- De keuze en samenstelling van de gebruikte inputvariabelen

Vervolgonderzoek naar manieren om deze knelpunten te verminderen, of naar alternatieve variabelen die de impact van deze beperkingen kunnen minimaliseren, zou de moeite waard kunnen zijn. Echter, wij achten het onwaarschijnlijk dat met de huidige stand van de techniek en de beschikbare data, dergelijke maatregelen voldoende zullen zijn om de effectiviteit van het systeem substantieel te verbeteren.

Tijdens dit onderzoek zijn we tegen enkele moeilijkheden aangelopen, met name vanwege onduidelijkheden over welke data waar is gebruikt en op welke momenten verbeteringen aan het model zijn doorgevoerd. Daarnaast was het lastig te verifiëren of we alle relevante gegevens en code in hun volledigheid hebben ontvangen. Voor een mogelijk vervolgonderzoek adviseren wij daarom samen te werken met een partij die een effectieve versiebeheerstrategie hanteert. Dit zou de traceerbaarheid en reproduceerbaarheid van het onderzoek aanzienlijk verbeteren en mogelijke misverstanden of dataverlies in toekomstige iteraties voorkomen.

5.2. Onderzoeksvragen

In sectie 1.1 zijn een aantal onderzoeksvragen gedefinieerd. Nu kunnen we terugkomen op deze vragen en een uiteindelijke conclusie definiëren.

De volgende onderzoeksvragen zijn onderzocht en geëvalueerd:

1. Kwaliteit/robuustheidcheck van de huidige toepassing van Machine Learning in het huidige systeem. Het kan nu zijn dat er bugs in de huidige door de UvA geproduceerde implementatie zitten. Tevens dient te worden onderzocht of de huidige implementatie de optimale AI-methodiek gebruikt.

Wij zijn van mening dat het machine learning-model, gegeven de omstandigheden, zo goed mogelijk is toegepast. Gedurende het hele proces—van het filteren van de data en de modelkeuze tot de

artificiële datatoevoeging en de interpretatie van de resultaten via onderzoek van verschillende grenswaarden van categorisatie—is er zorgvuldig rekening gehouden met de beperkingen en uitdagingen van de dataset.
In dit onderzoek is er slechts onderzocht wat de technische aspecten zijn van het machine learning model, in termen van data science en intelligente systemen. Overige overwegingen die bij kunnen dragen aan de werking van het model, zoals ecologische factoren, vallen buiten de expertise van Technolution en zijn daarom buiten scope van dit onderzoek gehouden.

Toch blijft het voorspellen van vogeltrek een complex vakgebied, waarin zowel de verdeling van de trekpieken over de gehele trekperiode, als omgevingsfactoren (weersvoorspellingen, radarprecisie, etc.) een aanzienlijke rol spelen. Veel van deze factoren zijn niet eenvoudig te beïnvloeden op een manier die nodig is om de vereiste nauwkeurigheid te bereiken voor een effectief start/stop-systeem voor de windturbines. Op basis van de uitgevoerde validaties in dit rapport concluderen wij daarom dat het model in de huidige vorm een lage betrouwbaarheid heeft.

Het is belangrijk op te merken dat de conclusies uit dit rapport niet alleen gebaseerd zijn op algemene data science best practices (ervaringen en aanbevelingen van data scientists met dit type data), maar vooral op de binnen dit rapport uitgevoerde validaties die zijn uitgevoerd op de beschikbare data. De beschikbare data is echter beperkt door radardata op slechts 1 plek en doordat het aantal geconstateerde vogeltrek piekuren relatief klein is ten opzichte van het aantal niet-trek perioden. Door deze zogenaamde scheve verdeling binnen de dataset bestaat echter de mogelijkheid dat de gebruikte testdata niet volledig representatief is voor de werkelijke migratiepatronen. Het is mogelijk dat zowel de test set uit het oorspronkelijke onderzoek als onze onafhankelijke test set toevallig een periode beslaan die niet volledig representatief is. Daardoor kan het model in de praktijk, over meerdere jaren heen, beter of slechter presteren dan onze huidige analyses suggereren. Echter, gezien de beperkte hoeveelheid data waar we nu over beschikken, kunnen we hier op dit moment geen definitieve uitspraken over doen.

2. Hoe kunnen we de kwaliteit van het model verhogen
    a. Welke andere methodes/methodieken (zowel binnen machine learning of daarbuiten) zouden gebruikt kunnen worden om een goede vogeltrekvoorspelling te doen en wat zijn de pro’s en cons van deze methodieken?

Er zijn in het rapport meerdere technieken besproken die een positieve invloed zouden kunnen hebben op de nauwkeurigheid van de voorspellingen. Het is moeilijk vast te stellen hoe groot de impact van deze maatregelen zou zijn en of de bijdrage voldoende rechtvaardigt om extra onderzoekstijd te investeren.

De belangrijkste toevoeging die noodzakelijk is om verbeteringen aan te brengen aan het model, is het vergroten van de hoeveelheid data. Zonder maatregelen te nemen om sneller meer data te verzamelen, verwachten wij dat het nog minstens 10 tot 20 jaar duurt voordat er genoeg data verzameld is om het model een hoge betrouwbaarheidsgraad te laten bereiken.

    b. Als het vogelvoorspellingsmodel opnieuw gebouwd zou worden, hoe zou dat moeten gebeuren en met welke methode?

In het onderzoek hebben wij geen methodieken of processen gevonden die gezien de huidige data set en invloeden een significante verbetering van de resultaten kunnen gaan opleveren. Hoewel er kleine aanpassingen zijn voorgesteld, achten wij de impact daarvan onvoldoende om een nieuw ontwikkelproces te rechtvaardigen.

3. Welke bedrijven zijn geschikt om zo’n model te ontwikkelen en te beheren?

Omdat we op dit moment geen voldoende waarde zien in het op basis van dezelfde data ontwikkelen van een geheel nieuw model, achten wij het niet nuttig om daar verder advies over uit te brengen. Echter, het kan zeker de moeite waard zijn om dieper in te gaan op de in dit rapport voorgestelde technische verbeteringen en meerdere modellen te trainen om deze met elkaar te vergelijken.

Wij raden aan het onderzoek naar een (datatechnisch) verbeterd voorspellingsmodel uit te laten voeren door partijen die zowel expertise hebben in de wetenschap en techniek achter data science en kunstmatige intelligentie, als zich kunnen verdiepen in domeinspecifieke kennis over de complexiteit van het ecosysteem. De volgende bedrijven zouden geschikte kandidaten kunnen zijn om deze stappen verder te onderzoeken en uit te voeren.

- Technolution
- TNO
- Deltares

6. Bibliography

Bradarić, M., Bouten, W., Fijn, R. C., Krijgsveld, K. L., & Shamoun-Baranes, J. (2020). Winds at departure shape seasonal patterns of nocturnal bird migration over the North Sea. *Journal of Avian Biology 51(10)*.

Bradarić, M., Kranstauber, B., Bouten, W., & Shamoun-Baranes, J. (2024). Forecasting nocturnal bird migration for dynamic aeroconservation: The value of short-term datasets. *Journal of Applied Ecology*.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55(10)*, 78-87.

Hersbach, H.,et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146(730)*, 1999–2049.

Louk, M. H., & Tama, B. A. (2022). Revisiting gradient boosting-based approaches for learning imbalanced data: A case of anomaly detection on power grids. *Big Data and Cognitive Computing, 6(2)*, 41.

Manola, I., Bradarić, M., Groenland, R., Fijn, R., Bouten, W., & Shamoun-Baranes, J. (2020). Associations of synoptic weather conditions with nocturnal bird migration over the North Sea. *Frontiers in Ecology and Evolution*.

van Erp, J. A., van Loon, E. E., De Groeve, J., Bradarić, M., & Shamoun-Baranes, J. (2024). A framework for post-processing bird tracks from automated tracking radar systems. *Methods in Ecology and Evolution*, 130-143.

Wang, Y., & Singh, L. (2021). Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics 12.2*, 101-119.

Wright, M. N., & Ziegler, A. (2017). A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77(1)*, 1-17.

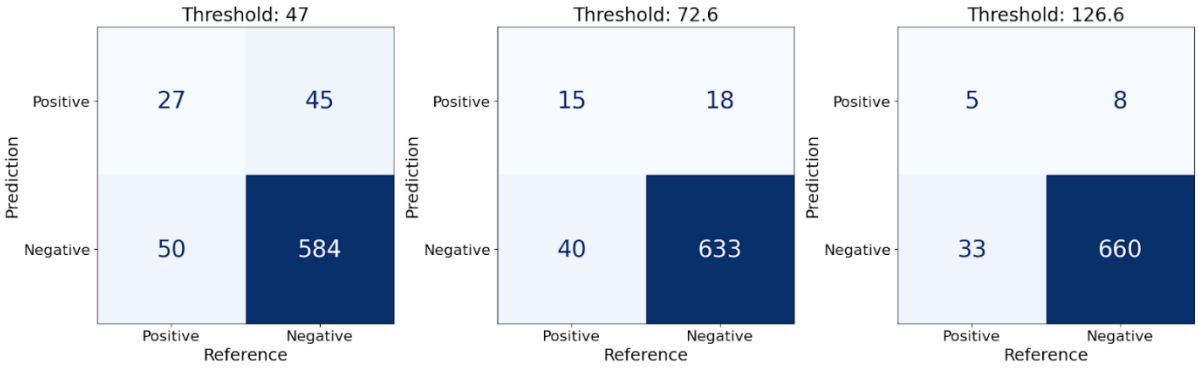
A. Appendix: Cross validation

De accuracy en precision zijn berekend over de resultaten van de hoogste drempelwaarde, dat is die die correspondeert met de meest rechter matrix.

1. Voorjaarsmodel

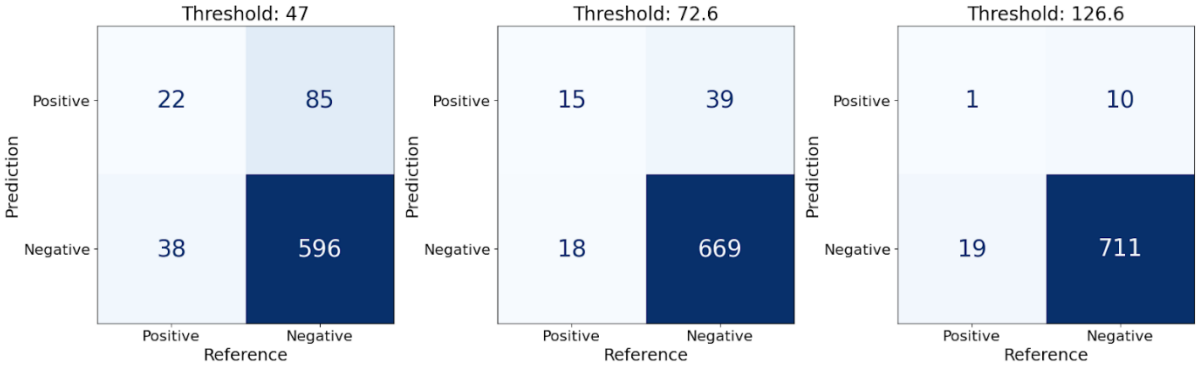
- Testset = 2022 (gebruikt in onderzoek UVA)

```
OOB prediction error (MSE):       2093.171
R squared (OOB):                  0.7137541
Accuracy:                          0.9419
Precision:                        0.3846
```



- Testset = 2023:

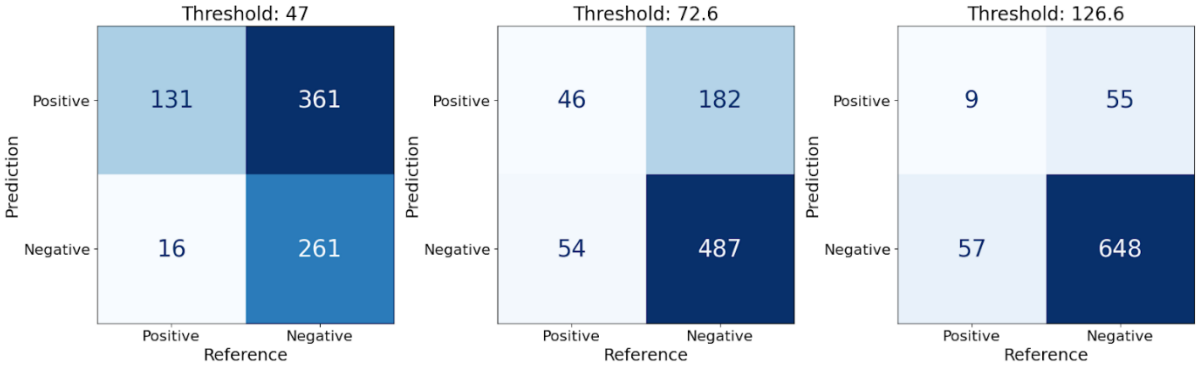
```
OOB prediction error (MSE):       1765.981
R squared (OOB):                  0.7808963
Accuracy:                          0.9638
Precision:                        0.0909
```



- Testset = 2021:

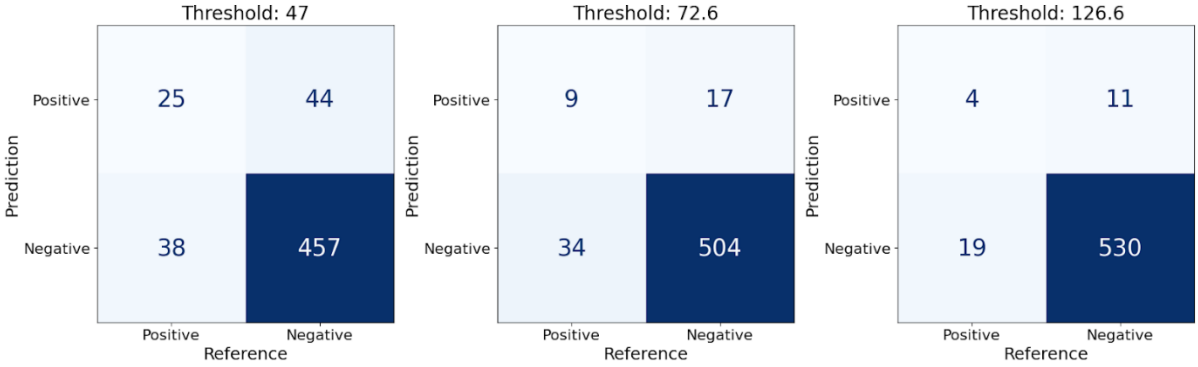
```
OOB prediction error (MSE):       1769.89
R squared (OOB):                  0.7268021
```

```
Accuracy:                       0.8544
Precision:                      0.1406
```



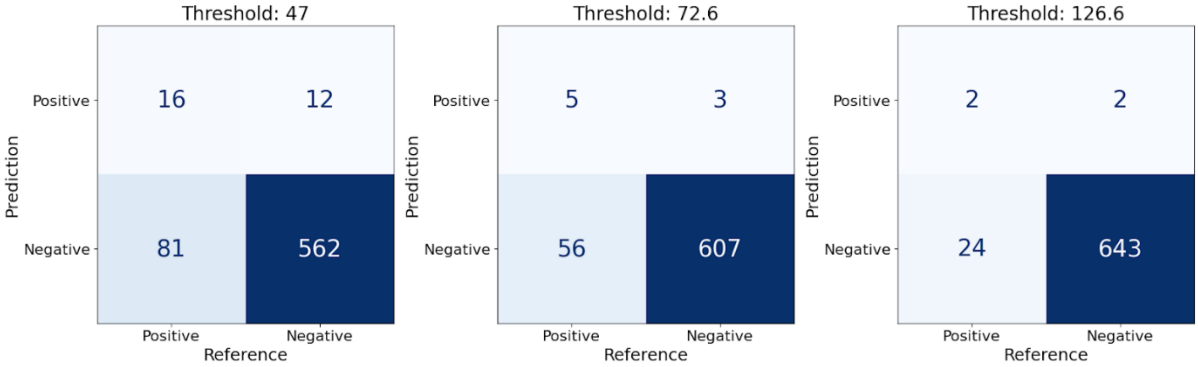
- Testset = 2020:

```
OOB prediction error (MSE):     2173.942
R squared (OOB):                0.7470692
Accuracy:                       0.9468
Precision:                      0.2667
```



- Testset = 2019

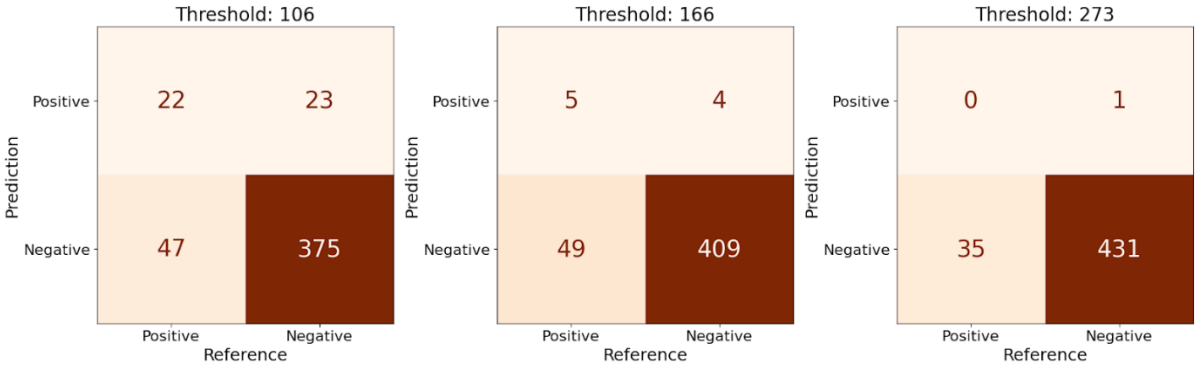
```
OOB prediction error (MSE):     2365.55
R squared (OOB):                0.745822
Accuracy:                       0.9613
Precision:                      0.5000
```



2. Najaarsmodel

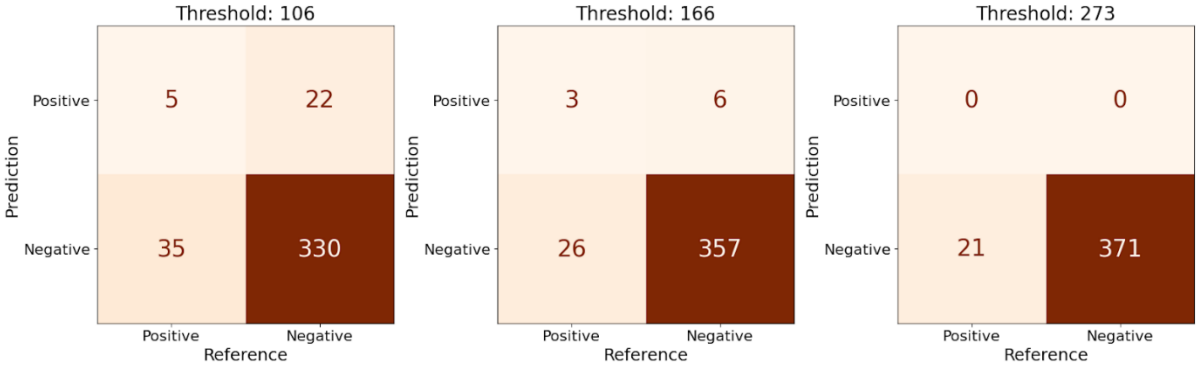
- Testset = 2020 (gebruikt in onderzoek UVA)

```
OOB prediction error (MSE):       3800.676
R squared (OOB):                  0.7010118
Accuracy:                          0.9229
Precision:                        0.0000
```



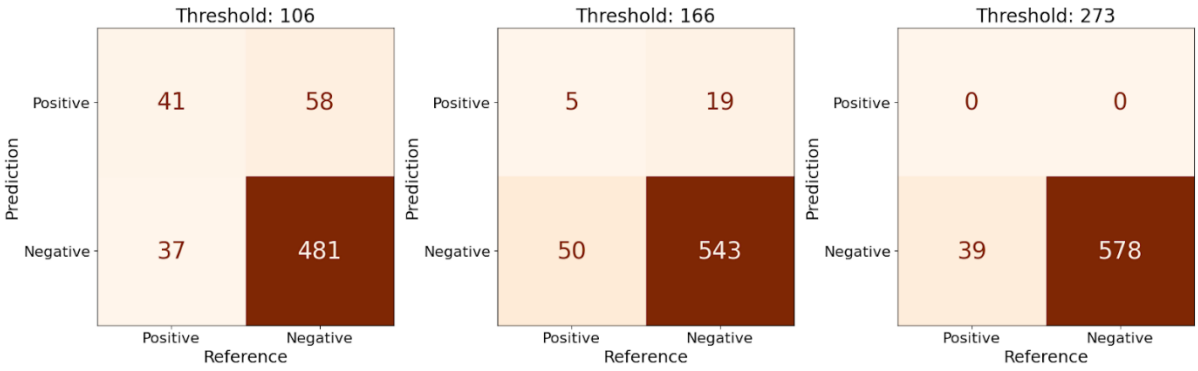
- Testset = 2023:

```
OOB prediction error (MSE):       3684.453
R squared (OOB):                  0.7238678
Accuracy:                          0.9464
Precision:                        -
```



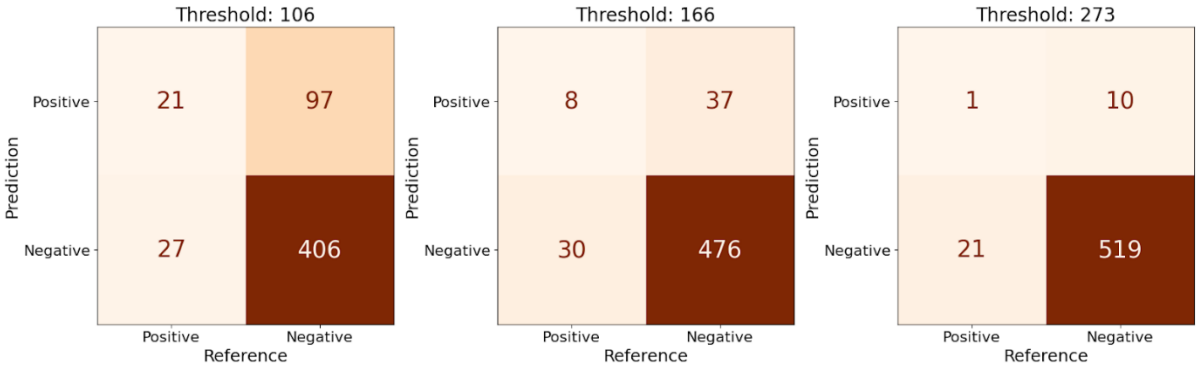
- Testset = 2022

```
OOB prediction error (MSE):       3604.513
R squared (OOB):                  0.6980381
Accuracy:                          0.9368
Precision:                        -
```



- Testset = 2021

```
OOB prediction error (MSE):       3526.279
R squared (OOB):                  0.7505728
Accuracy:                          0.8100
Precision:                        0.0909
```



- Testset = 2019

```
OOB prediction error (MSE):       3451.472
R squared (OOB):                  0.7672282
Accuracy:                          0.9721
Precision:                        -
```