Algorithm Audit

Probleemstelling

Voorspellen van onverantwoord rijgedrag

Algoprudentie

ALGO:AA:2025:01:P

13 januari 2025

Managementsamenvatting

Dit document zet verschillende normatieve vragen uiteen met betrekking tot risicovoorspellingen van onverantwoord rijgedrag door een algoritme in een database van een business-to-consumer autodeelplatform. Door middel van machine learning wordt een risicomodel (balanced random forest) getraind om rijgedrag te identificeren dat in verband wordt gebracht met gebruikers die eerder van het platform zijn geblokkeerd vanwege onverantwoord rijgedrag. Het model berekent de risicoscore van een gebruiker na elke nieuwe rit. Als een risicoscore de waarschuwingsdrempel overschrijdt, wordt er een melding naar de gebruiker gestuurd met het advies om zijn rijgedrag te verbeteren en specifieke suggesties over hoe dit gedaan kan worden. Als het rijgedrag van een gebruiker niet verbetert bij volgende ritten kan het platform de gebruiker blokkeren van de diensten na een menselijke beoordeling. Dit document biedt technische, juridische en contextuele achtergrond die nodig is om de specifieke normatieve vragen te beantwoorden.

Het autodeelplatform heeft Algorithm Audit benaderd om onafhankelijk advies te geven over het verantwoorde gebruik van het risicovoorspellingsalgoritme. Om dit te vergemakkelijken heeft het platform Algorithm Audit toestemming gegeven om een due diligence-beoordeling uit te voeren. Alle informatie in dit document is gebaseerd op interviews met de data scientists van het platform en een uitwisseling van relevante documenten. Hoewel Algorithm Audit een due diligence-beoordeling heeft toegepast betreffende deze informatie, is de broncode niet onafhankelijk beoordeeld door Algorithm Audit. Het autodeelplatform is geanonimiseerd om de publicatie van deze casus te vergemakkelijken. Algorithm Audit heeft geen financiële of enige vergoeding anderszins ontvangen van het platform voor het uitvoeren van deze onafhankelijke beoordeling.

Algorithm Audit heeft een focusgroep bijeengebracht om het gebruikersperspectief in kaart te brengen omtrent dataverzameling, data-analyse en de inzet van voorspellende algoritmes, inclusief communicatie hierover bij deelplatformen. De inzichten van deze focusgroep worden gedeeld in het document Resultaten focusgroep. Op basis van deze probleemstelling en de resultaten van de focusgroep zal een normatieve adviescommissie advies uitbrengen over de hieronder beschreven vragen, samengebracht door Algorithm Audit.

Inhoudsopgave

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn drieledig:

Kennisplatform

Samenbrengen van kennis en experts voor collectief leerproces over verantwoorde inzet van algoritmes, bijvoorbeeld ons AI Policy Observatory en position papers

Normatieve adviescommissies

Adviseren over ethische kwesties in concrete algoritmische toepassingen door het samenbrengen van deliberatieve, diverse adviesommissies, met algoprudentie als resultaat

Technische hulpmiddelen

Implementeren en testen van technische methoden voor bias-detectie en -mitigatie, zoals onze bias detection tool

Projectwerk

Ondersteuning bij specifieke vragen vanuit de publieke en private sector over de verantwoorde inzet van algoritmes.

1. Inleiding

Autodeelplatformen worden steeds populairder voor autogebruik. Bedrijven nemen voertuigen af die worden verhuurd aan klanten die op het platform geregistreerd zijn. Echter, niet alle klanten gebruiken de deelauto's op een verantwoordelijke manier. Sommige gebruikers vertonen onverantwoord rijgedrag, zoals te hard rijden, te hard versnellen en scherpe bochten maken. Dergelijk gevaarlijk rijgedrag is gekoppeld aan verkeersongevallen en verwondingen, waarbij gebruikers van het autodeelplatform en andere weggebruikers betrokken zijn. Om de verkeersveiligheid te bevorderen en schadekosten te verlagen, houdt het platform het rijgedrag van zijn gebruikers bij.

De kosten voor autodeelplatformen vanwege de schade aan voertuigen variëren doorgaans tussen € 2.000 (minimale schade) en € 20.000 (total loss schade). Voor het platform bedroegen de jaarlijkse schadekosten 7% (€2M-€3M) van de totale omzet (€25M-€45M).

Het autodeelplatform traint een voorspellingsmodel om te identificeren welk rijgedrag in verband wordt gebracht met gebruikers die eerder van het platform zijn geblokkeerd vanwege onverantwoordelijk gedrag. Gegevens over rijgedrag worden verzameld in een gestandaardiseerd formaat via een Driver Behavior Monitoring Systems (DBMS) dat is geïnstalleerd in auto's van het platform. Met behulp van deze functies wordt een balanced random forest (BRF)-model getraind om een risicoscore te voorspellen die aangeeft of een gebruiker zich waarschijnlijk onverantwoordelijk zal gedragen.

De focus van deze probleemstelling is om te evalueren hoe het BRF-model moet worden gebruikt en gekalibreerd met het oog op de potentieel schadelijke impact van onnauwkeurige voorspellingen. Er wordt vooral gefocust op het nauwkeurig identificeren van onverantwoordelijk gedrag, waarbij oneerlijke en buitensporige ongerechtvaardigde verdenkingen van gebruikers worden vermeden. Dit doel wordt nagestreefd door de balans van vals-positieve en vals-negatieve voorspellingen en de selectie van functies en hyperparameters te evalueren.

Deze probleemstelling biedt de nodige technische, juridische en contextuele achtergrond om deze vragen aan te pakken. Het schetst de gegevensverzamelings-, trainings- en testprocessen van het BRF-model en biedt een korte analyse van de relevante juridische kaders die deze casus behandelt.

Bij het overwegen van het verantwoorde gebruik van algoritmen is een belangrijke eerste vraag waarom een datagestuurde aanpak geschikt is om het probleem te verhelpen. In dit geval kan het gebruik van een datagestuurde aanpak om rijgedrag te verzamelen en te monitoren om verschillende redenen gerechtvaardigd zijn. Ten eerste heeft empirisch onderzoek aangetoond dat het verzamelen van rijeigenschappen via een DBMS de verkeersveiligheid kan verbeteren.[1] Ten tweede zijn machine learning-technieken effectief gebleken bij het identificeren van onveilig rijgedrag in verzamelde DBMS-gegevens.[2]

[1] D. Lord, F. Mannering, The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives, Transportation Research Part A: Policy and Practice, Volume 44, Issue 5 (2010) https://doi.org/10.1016/j.tra.2010.02.001

[2] E. Lattanzi, V. Freschi, Machine Learning Techniques to Identify Unsafe Driving Behavior by Means of In-Vehicle Sensor Data, Expert Systems with Applications, Volume 176 (2021) https://doi.org/10.1016/j.eswa.2021.114818

Binnen de grenzen van de wet, met name op het gebied van gegevensverzameling en -verwerking, kan het daarom de moeite waard zijn om algoritmische methoden te onderzoeken voor het identificeren van onveilig rijgedrag om de verkeersveiligheid te verbeteren. Deze casus onderzoekt hoe een algoritmische aanpak op een verantwoorde manier kan worden geïmplementeerd. Algorithm Audit heeft een onafhankelijk onderzoek naar de casus uitgevoerd door de data scientists van het platform te interviewen en relevante vragen te stellen. Deze probleemstelling is opgesteld op basis van dit onderzoek. Het zet belangrijke normatieve vragen uiteen die uit de casus naar voren zijn gekomen en presenteert deze vragen samen met een bespreking van de relevante achtergrond. De probleemstelling dient als input voor een onafhankelijke normatieve adviescommissie die is bijeengeroepen door Algorithm Audit, die advies zal uitbrengen over de hieronder uiteengezette vragen.

Ter voorbereiding op de adviescommissie heeft Algorithm Audit een focusgroep gehouden, bestaande uit vier gebruikers die regelmatig gebruikmaken van de diensten die worden geleverd door deelplatformen. In dit geval zijn de gebruikers belangrijke belanghebbenden die worden beïnvloed door de implicaties van gegevensverzameling en verkeerde risicovoorspelling van het BRF-model. De focusgroep werd gehouden om vast te leggen wat gebruikers in het bijzonder vinden van verschillende aspecten van een deelplatform in de context van gegevensverwerking, algoritmische voorspellingen en communicatie. Er werden meerdere vragen gesteld over hun ervaringen met en ideeën rondom een dergelijk platform, wat ook leidde tot discussies binnen de groep. De vragen waren specifiek gericht op opvattingen over gegevensverzameling, rij- en betaalgedrag als datapunten, het gebruik van het algoritme om onverantwoordelijke gebruikers te voorspellen, de procedure rond het blokkeren van iemands account (inclusief menselijke interactie en herstel) en de afweging tussen vals-negatieven en vals-positieven. Na het uitvoeren van de focusgroep werd het interview getranscribeerd en geparafraseerd tot een rapport met vragen en antwoorden. De resultaten zijn te vinden in Resultaten focus groep. De inzichten uit de focusgroep dienen als input voor de beraadslaging van de normatieve adviescommissie. Toch is de commissie niet verplicht om de standpunten van de focusgroep in haar uiteindelijke advies op te nemen.

De structuur van deze probleemstelling is als volgt: eerst wordt het proces van gegevensverzameling besproken in 2. Gegevensverzameling, en wordt vraag 1 geïntroduceerd. Daarna in 3. Voorspellingsmodel, worden details gedeeld over het trainings- en testproces van het BRF-model, waarbij vragen 2-5 worden geïntroduceerd. Vervolgens wordt een korte juridische analyse van deze casus gegeven in 4. Juridische analyse.

De bijlagen:

> Appendix A geeft een lijst en beschrijft alle onafhankelijke variabelen die in het BRF-model worden gebruikt
> Appendix B schetst hoe gebruikers worden geïnformeerd over waarschuwingen of uitsluitingen van het autodeelplatform
> Appendix C biedt een overzicht van modelvoorspellingen voor specifieke BRF-hyperparameters
> Appendix D bevat technische details over de gevoeligheidstest van het BRF-model.

2. Gegevensverzameling

Het proces van gegevensverzameling door het autodeelplatform wordt in dit hoofdstuk beschreven.

2.1 Rijgedraggegevens

Risicovoorspellingen door het BRF-model zijn gebaseerd op datapunten met betrekking tot rijeigenschappen. Persoonlijke gegevens, zoals leeftijd, postcode of het aantal jaren in het bezit zijn van een rijbewijs, zijn niet opgenomen in de analyse. Gegevensverzameling via Driver Behavior Monitoring Systems (DBMS) is een industriestandaard voor operators om gestandaardiseerde rijeigenschappen in realtime te meten, met name bochten, remmen, versnellen en snelheidsgebeurtenissen. Tijdens het huren van een auto verzamelt het DBMS datapunten over het rijgedrag van gebruikers in de vorm van longitudinale gegevens, die gepoolde rijeigenschappen vertegenwoordigen over de gehele reisgeschiedenis van de gebruiker. Een gedetailleerde beschrijving van 32 kenmerken afgeleid van het DBMS die relevant zijn voor deze casus, is te vinden in Appendix A.

Gebruikersperspectieven op gegevensverzamelingspraktijken door gedeelde mobiliteitsplatforms zijn te vinden in het document Resultaten focusgroep.

Voorheen hield het platform rekening met de betalingsgegevens van gebruikers - zoals het aantal bankrekeningen dat aan het platform is gekoppeld en gevallen van te late betalingen - evenals strafbare feiten, met name het aantal boetes. Zorgen over mogelijke schendingen van de Algemene Verordening Gegevensbescherming (AVG) en operationele uitdagingen in verband met de vertraging tussen een verkeersovertreding en de ontvangst van de boete hebben ertoe geleid dat het platform het gebruik van deze variabelen heeft stopgezet.

2.2 Onverantwoord rijgedrag

De labels in de trainings- en testdatasets identificeren gebruikers die eerder van het platform zijn geblokkeerd. Gebruikers worden geblokkeerd wanneer een menselijke analist concludeert dat hun gedrag onverantwoord is geweest. Het volgende gedrag resulteert in het blokkeren van het gebruikersaccount:

1. Als een gebruiker niet reageert op een schademelding en er bewijs is dat hij de schade heeft veroorzaakt.
2. Als een gebruiker de auto voor de tweede of derde keer niet op de juiste locatie terugbrengt, afhankelijk van de ernst van het geval.
3. Als een gebruiker zijn facturen langer dan 6 maanden niet betaalt.
4. Als de gebruiker wordt gerapporteerd voor het roken in de auto of het vuil achterlaten van de auto, voor de tweede of derde keer, afhankelijk van de ernst van het geval.
5. Als een gebruiker tijdens een rit, die eigenlijk voor operationele doeleinden bedoeld was (bijvoorbeeld het afgeven van de auto), voor de derde keer een frauduleuze rit maakt.

Ongevallen waarbij geen sprake is van verwijtbare nalatigheid worden niet als onverantwoordelijk beschouwd.

Q1

Hoewel alleen gegevens over gedrag op het platform en niet over een individu worden verzameld en gebruikt voor het maken van risicovoorspellingen, moet een zorgvuldige beoordeling overwegen gemaakt worden over specifiek welke functies gerechtvaardigd kunnen worden gebruikt voor het doel van het voorspellen van onverantwoord rijgedrag. Het gebruik van bepaalde functies kan bijvoorbeeld mogelijk (indirecte) discriminatie, ongewenste vooringenomenheid of andere vormen van oneerlijkheid in de risicovoorspellingen met zich meebrengen.

Vraag 1a: Moeten kenmerken gerelateerd aan agressief rijgedrag (agressieve versnelling, hard remmen, zwaar bochtenwerk) of snelheidsgebeurtenissen worden ingevoerd in het BRF-model (zie Appendix A)? Zijn er redenen waarom dergelijke rijeigenschappen niet gebruikt zouden moeten worden als invoervariabelen voor het voorspellen van onverantwoord rijgedrag?

Vraag 1b: Is het gerechtvaardigd om andere functies met betrekking tot reisgebeurtenissen (met name het onlangs terugbrengen van een auto, het vuil achterlaten van de auto) en betalingsgegevens (zoals het aantal betaalmethoden dat aan een gebruikersaccount is gekoppeld, te laat betalen van facturen) te gebruiken om risicovoorspellingen te doen voor onverantwoordelijke bestuurders?

3. Voorspellingsmodel

Hier wordt ingegaan op de training, evaluatie en kalibratie van het Balanced Random Forest (BRF)-model, gevolgd door sociaal-technische vragen over normatieve beslissingen die tijdens het datamodelleringsproces moeten worden genomen.

3.1 Risicovoorspellingen

Op basis van historische gegevens voorspelt de BRF welke actieve gebruikers in de toekomst waarschijnlijk onverantwoordelijk gedrag zullen vertonen. Het getrainde BRF-model kent een risicoscore r_i toe aan elke gebruiker i, met $1 \leq i \leq n$ gebruikers in de database. Risico scores zijn onderverdeeld in 3 categorieën:

> i) **Geen actie:** de voorspelde risicoscore van de gebruiker **i** valt onder de waarschuwingsdrempel r_w, i.e., $r_i < r_w$;

ii) **Waarschuwing:** de voorspelde risicoscore van gebruiker i valt tussen de waarschuwingsdrempel r_w en de blokkeringsdrempel r_b, i.e., $r_w < r_i < r_b$;

iii) **Blokkering:** de voorspelde risicoscore van de gebruiker i overschrijdt de blokkeringsdrempel r_b, i.e., $r_i > r_b$.

Een BRF-model wordt toegepast op de longitudinale gegevens van 75.919 gebruikers om DBMS-gegevens (onafhankelijke variabelen) te koppelen aan eerder geblokkeerde accounts (afhankelijke variabele). Slechts

Box 2

Andere voorspellingsmodellen

Naast een Balanced Random Forest (BRF) is ook een logistiek regressiemodel getraind op rijeigenschappen om onverantwoord gedrag te voorspellen. Het BRF-model presteerde echter consistent beter dan het logistieke regressiemodel in alle op confusion matrix gebaseerde evaluatiemetrieken. Als gevolg hiervan besloot het platform het BRF-model te implementeren en te evalueren hoe het op de meest verantwoorde manier kan worden geïmplementeerd.

2,7% van de gebruikers in de trainingsdataset heeft een geblokkeerde accountstatus. De dataset is verdeeld in een verhouding van 80:20 voor training en testen. Voor elke gebruiker i in de testdataset voorspelt de BRF een risicoscore r_i tussen 0 en 1. Op basis van dit model krijgen bestaande gebruikers in de dataset een risicoscore. Gebruikers met een risicoscore tussen de waarschuwing en een blokkeringsdrempel krijgen een waarschuwing, terwijl gebruikers met een score boven de blokkeringsdrempel worden doorverwezen naar een menselijke analist. Na menselijke beoordeling ontvangt de gebruiker een waarschuwingsbericht of wordt hij uitgesloten van de autodeelservice. De exacte procedures voor het uitgeven van waarschuwingen en het blokkeren van gebruikers worden beschreven in Appendix B.

3.2. Evalueren van het voorspellingsmodel

Voorspellingen van het BRF-model op de trainings- en testdataset worden geëvalueerd volgens confusion matrix evaluatiemetrieken. Indien de positieve klasse bestaat uit onverantwoordelijke bestuurders (geblokkeerde gebruikers in het verleden), bestaat de confusion matrix uit de volgende elementen:

- **True positive (TP):** Onverantwoordelijke bestuurder wordt geclassificeerd als onverantwoordelijk;
- **False positive (FP):** Verantwoordelijke bestuurder wordt geclassificeerd als onverantwoordelijk;
- **True negative (TN):** Verantwoordelijke bestuurder wordt geclassificeerd als verantwoordelijk;
- **False negative (FN):** Onverantwoordelijke bestuurder wordt als geclassificeerd verantwoordelijk.

De volgende evaluatiemetrieken worden gebruikt op basis van de voorspellingen van het BRF-model:

- **Precision:** Fractie van TP's ten opzichte van alle voorspelde onverantwoordelijke bestuurders, i.e. TP/(TP+FP);
- **Recall:** Fractie van TP's ten opzichte van alle waargenomen onverantwoordelijke bestuurders, i.e. TP/(TP+FN).

Om de socio-technische implicaties van de BRF-voorspellingen te beoordelen moet de impact van FN- en FP-voorspellingen worden afgewogen, aangezien beide typen voorspellingen bepaalde risico's met zich meebrengen.

Voor FN's ontstaan materiële risico's voor het bedrijf doordat gebruikers die eerder schade aan het voertuig veroorzaken dan verantwoordelijke bestuurders niet worden gedecteerd. Dit drijft de schadekosten op. Immateriële risico's van FN's omvatten het onopgemerkt laten van onverantwoordelijk gedrag, wat kan worden samengevat met het gezegde: *"don't be gentle, it's a rental"*. Dit kan er bijvoorbeeld toe leiden

dat de voertuigen van het platform worden geassocieerd met onverantwoordelijk rijgedrag op de weg. Het minimaliseren van het aantal FN's draagt bij aan de verkeersveiligheid voor iedereen.

Bij FP's ontstaan materiële kosten voor het bedrijf doordat menselijke analisten handmatig verdachte gebruikers beoordelen voordat waarschuwings- en blokkeringsberichten worden verzonden. Een risico is dat gebruikers ten onrechte worden beschuldigd van onverantwoordelijk rijgedrag. Dit kan ertoe leiden dat ze overstappen naar een concurrerend platform, wat materiële kosten voor het bedrijf met zich meebrengt. Dit risico wordt momenteel beperkt door de handmatige beoordeling door menselijke analisten die talloze vals-positieve resultaten eruit filteren. Het kan zo zijn dat bepaalde groepen meer onderhevig zijn aan FP's dan gemiddeld, wat een risico op vooringenomenheid en ongelijke impact met zich meebrengt. Een ander risico van vals-positieve resultaten is dat gebruikers het gevoel kunnen hebben dat ze constant in de gaten worden gehouden terwijl ze de diensten van het platform gebruiken.

Q2

Rekening houdend met de acties die volgen op een toegekende risicoscore zoals beschreven in Appendix B, is het even schadelijk om verantwoordelijke bestuurders als onverantwoordelijk (vals positief) te classificeren als het is om onverantwoordelijke bestuurders als verantwoordelijk (vals negatief) te classificeren? Welke uitkomst weegt zwaarder en waarom?

Op basis van de confusion matrix worden de volgende evaluatiemetrieken voor het BRF-model overwogen:

> False Positive Rate (FPR): fractie van FP's ten opzichte van alle waargenomen verantwoordelijke bestuurders, d.w.z. FP/(FP+TN);
> False Negative Rate (FNR): fractie van FN's ten opzichte van alle waargenomen onverantwoordelijke bestuurders, d.w.z. FN/(TP+FN).

Op dezelfde manier kunnen, met meer nadruk op TP's, de volgende gerelateerde evaluatiemetrieken worden overwogen:

> Recall, ook wel True Positive Rate (TPR): fractie van TP's ten opzichte van alle waargenomen onverantwoordelijke bestuurders, d.w.z. TP/(TP+FN);
> Precision: fractie van TP's met betrekking tot alle voorspelde onverantwoordelijke bestuurders, d.w.z. TP/(TP+FP).

In de ideale uitkomst zouden zowel FP- als FN-voorspellingen geminimaliseerd worden. In de praktijk zou er echter een evenwicht gevonden moeten worden. Deze afweging wordt beïnvloed door de keuze van hyperparameters die geselecteerd zijn voor het BRF-model.

3.3 Hyperparameter-afstemming van het BRF-model

Het BRF-model, zoals geïmplementeerd in het Python scikit.learn-pakket, wordt afgesteld op de trainingsdataset met 32 kenmerken voor rijgedrag. Dit model maakt gebruik van de volgende hyperparameters:

- Aantal bomen (n_est);
- Maximale diepte per boom (max_depth);
- Minimaal aantal observaties per split (min_samples_split);
- Minimaal aantal vereiste samples in een blad (min_samples_leaf);
- Aantal functies waarvan wordt overwogen om een splitsing te maken (max_features);
- Of de minderheidsklasse in evenwicht is (sampling_strategy)
- Bomen worden getraind op bootstrap-samples, d.w.z. willekeurig geselecteerde samples uit de trainingsdataset (bootstrap);
- Datapunten voor de bootstrap-sample worden willekeurig geselecteerd met of zonder vervanging (replacement);
- Waarschuwingsdrempel (r_w).

Een uitgebreide uitleg van deze parameters is te vinden in de scikit.learn-documentatie.[3] Een gevoeligheidsanalyse van het BRF-model is te vinden in Appendix D.

Q3 Welke gevoeligheidstesten* moeten worden uitgevoerd om n_est, min_samples_split , min_samples_leaf en max_features te selecteren?

** gevoeligheidstesten verwijzen naar het evalueren hoe gevoelig de prestaties van het BRF-model zijn voor veranderingen in zijn hyperparameters. Het doel is om de stabiliteit en robuustheid van het model te beoordelen onder verschillende omstandigheden. Dit kan helpen identificeren hoe veranderingen in het ontwerp van het model zijn voorspellingen kunnen beïnvloeden. Een voorbeeld van gevoeligheidstesten voor een BRF-model is te vinden in Appendix D.*

Stel dat de volgende parameters zijn geselecteerd: n_est=100, min_samples_split=2, min_samples_leaf=1, max_features=√32, sampling_strategy= 'not majority', bootstrap=False, replacement=True. De maximale diepte per boom (max_depth) is ingesteld op geen, wat betekent dat splitsingen worden uitgebreid totdat alle bladen zuiver zijn of totdat alle bladen minder dan 2 (min_samples_split) samples bevatten. Voor deze set hyperparameters worden de evaluatiemetrieken van het BRF-model besproken.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Box 3

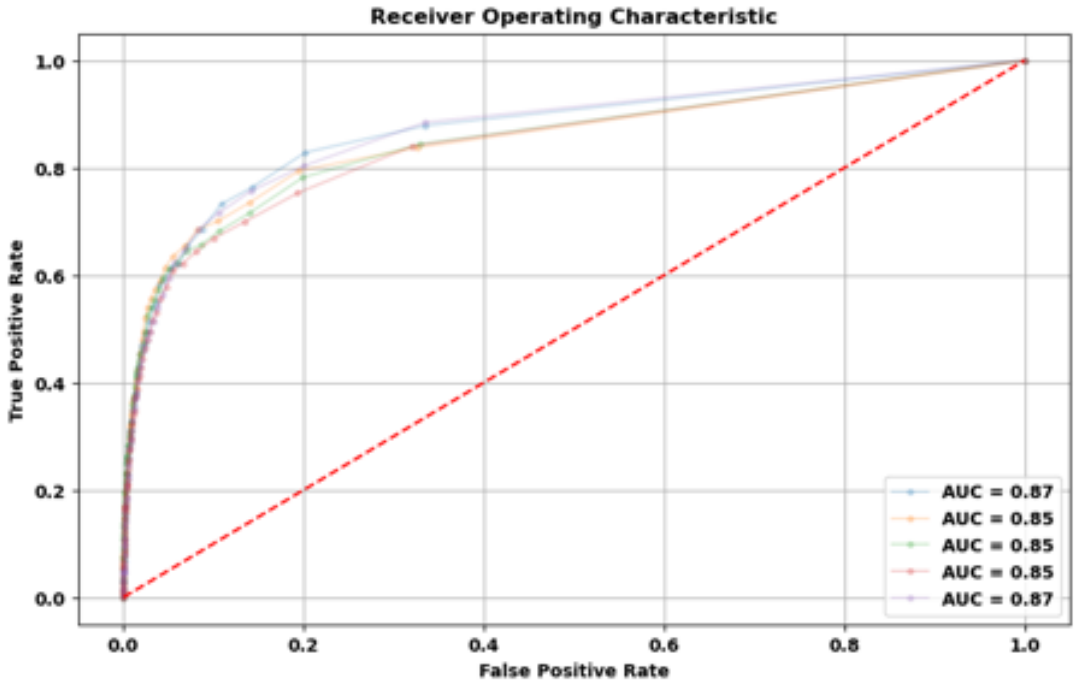
K-fold cross validation

K-voudige kruisvalidatie verwijst naar het verdelen van de dataset in k gelijke, niet-overlappende subsets of "vouwen". Elke vouw wordt één keer gebruikt als testset, terwijl de resterende k-1 vouwen worden gebruikt voor training. Deze methode wordt veel gebruikt om overfitting te voorkomen, ervoor te zorgen dat elk deel van de dataset wordt gebruikt voor zowel training als testen, en een betrouwbare schatting van de modelprestaties te bieden. Nadat het model is getraind met behulp van k-voudige kruisvalidatie, wordt het uiteindelijke model toegepast op de testdataset, die niet was opgenomen in het k-voudige kruisvalidatieproces.

Meer informatie over kruisvalidatie is te vinden in Hoofdstuk 5 van An Introduction to Statistical Learning, G. James, D. Witten, T. Hastie and R. Tibshirani

3.4 ROC curve

De balans tussen FN's en FP's hangt af van de selectie van drempelwaarde r_w. Met behulp van 5-voudige kruisvalidatie worden voor de hyperparamters TPR en FPR berekend voor $0.01<r_w<0.9$ in intervals van 0.01 (zie Box 3). Als alternatief kunnen precisie en recall worden gebruikt als evaluatiemetrieken. Deze metrieken leggen echter minder nadruk op de impact van FP's (zie 3.2 Evalueren van het voorspellingsmodel). Omdat het platform streeft naar een balans tussen het vastleggen van zoveel mogelijk risicovolle gebruikers en het vermijden van buitensporige valse vermoedens van verantwoordelijke bestuurders, wordt de ROC-curve de voorkeur gegeven voor het evalueren van dit model. In een ROC-curve wordt de TPR uitgezet tegen de FPR. Zie Figuur 1.



Figuur 1 - ROC curve

De confusion matrix van het BRF model voor drempelwaarden r_w=0.25 and r_w=0.5 kan worden gevonden in Appendix C.

Op basis van de resultaten in Figuur 1 worden drie scenario's overwogen om een evenwicht te bereiken tussen de TPR en de FPR:

> **Scenario blauw:** r_w=0.20 met een TPR van 0,4240 en een FPR van 0,0157. De TPR is 27 keer hoger dan de FPR.
> **Scenario oranje:** :r_w=0.40 met een TPR van 0.2100 en een FPR van 0,0028. De TPR is 75 keer hoger dan de FPR.
> **Scenario groen:** r_w =0.60 met een TPR van 0,0740 en een FPR van 0,0003. De TPR is 219 keer hoger dan de FPR.

Een afbeelding met de precisie en recall-percentages voor dezelfde 5-voudige kruisvalidaties kan worden gevonden in Appendix C.

Q4 Welk scenario is het meest wenselijk, gezien de sociaal-technische impact van de FP- en FN-classificaties?

3.5 Belang van de kenmerken

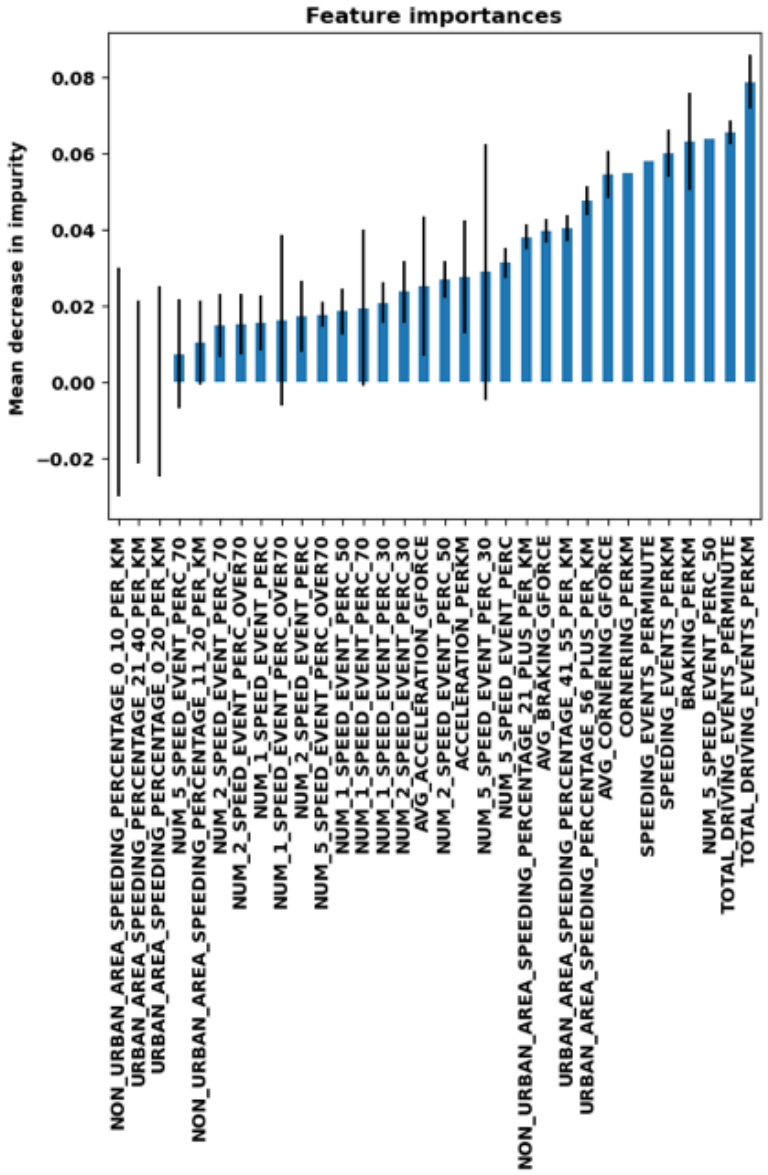
Stel r_w =0.25. Voor dit model is TPR 0.3640, FPR 0.0096, TNR 0.9904 en FNR 0.6360. De waarschuwingsdrempel r_b is ingesteld op 0,4, wat overeenkomt met een TPR van 0.2250 en FPR 0.0034 . Voor dit model wordt het BRF-model afgestemd op de testgegevens. De 31 datapunten met de meest voorspellende waarde in termen van mean decrease in impurity (gedefinieerd door de Gini-index) worden weergegeven in Figuur 2. De top-3 variabelen worden als kenmerkend beschouwd voor onverantwoord rijgedrag, welke zijn:

> **TOTAL_DRIVING_EVENTS_PERKM:** Totaal aantal bochten- en remgebeurtenissen gedeeld door het aantal gereden kilometers;
> **TOTAL_DRIVING_EVENTS_PERMINUTE:** Totaal aantal bochten- en remgebeurtenissen gedeeld door de totale rijtijd in minuten per gebruiker;
> **BRAKING_PERKM:** Het aantal geregistreerde remgebeurtenissen over alle ritten per gebruiker gedeeld door het aantal gereden kilometers.

Een uitleg van alle kenmerken kunnen worden gevonden in Appendix A.

Q5 Als we ervan uitgaan dat het wenselijk is om gebruikers feedback te geven wanneer ze een waarschuwing voor onverantwoord rijgedrag ontvangen, wat houdt in dit geval een zinvolle uitleg in?

1) Een algemene melding over de detectie van agressief rijgedrag en te hard rijden van de gebruiker, zonder verdere specificatie;
2) Specificatie van de waarde 'total driving events' (aantal bochten- en remgebeurtenissen, per minuut of km), versnellings- en snelheidsgebeurtenissen van de gebruiker;
3) Gedetailleerde specificatie van de bochten- en remgebeurtenissen van de gebruiker, en van verschillende categorieën snelheidsgebeurtenissen.



Figuur 2 - Kenmerkbelang van BRF-model dat onverantwoordelijk gedrag voorspelt. Gemiddelde afname in onzuiverheid is gebaseerd op de Gini-index. Een hogere gemiddelde afname in onzuiverheid geeft een groter belang van het kenmerk aan. De drie meest voorspellende kenmerken zijn 1. TOTAL_DRIVING_EVENTS_PERKM, 2. TOTAL_DRIVING_EVENTS_PERMINUTE and 3. BRAKING_PERKM.

Box 4

Disclaimer juridische analyse

Deze probleemstelling geeft inzichten weer uit verschillende domeinen om normatieve vragen te beantwoorden die verband houden met de casus die wordt beoordeeld, waarvoor momenteel geen duidelijke juridische richtlijnen bestaan. De statistische en juridische achtergronden worden zo nauwkeurig mogelijk beschreven, maar kan wellicht wat tekortschieten. Er kunnen geen rechten worden ontleend aan dit document.

4. Juridische analyse

Verschillende juridische aspecten zijn van toepassing op het besluitvormingsproces op basis van machine learning om onverantwoordelijke bestuurders van het platform te weren. Een samenvatting van de relevante passages van het privacybeleid van een specifiek autodeelplatform en de relevante wetgeving die van toepassing is op deze casus worden hier uiteengezet.

4.1 Privacybeleid van het autodeelplatform

De klanten gaan akkoord met het privacybeleid door gebruik te maken van de diensten van het platform. In het privacybeleid informeert het autodeelplatform de gebruikers dat DBMS-gegevens worden verzameld en gebruikt voor een risicovoorspellingsmodel. Het platform beschrijft dat de gegevens die van een klant via zijn account worden verwerkt kunnen worden beschouwd als persoonsgegevens onder de definitie van de AVG. De privacyverklaring geeft aan dat een klant op verzoek kan nagaan op welke manier zijn gegevens zijn gebruikt. De data die aan een klant worden verstrekt zijn uitsluitend de data die zijn toegepast om een bepaalde beslissing over een gebruiker te nemen. Het platform werkt ook samen met particuliere organisaties en overheidsinstanties om wanbetalingen, fraude en strafbare feiten aan te pakken. Gebruikersgegevens kunnen worden bewaard nadat een account is verwijderd, maar niet langer dan strikt noodzakelijk. Als het bewaren van gegevens de termijn van 'strikt noodzakelijk' overschrijdt, wordt ervoor gezorgd dat deze niet kunnen worden herleid tot een voormalige gebruiker door de toepassing van aggregatie of anonimisering.

4.2 Data Privacy Impact Assessment

Op verzoek van het platform heeft een extern advocatenkantoor in het najaar van 2024 een Data Privacy Impact Assessment (DPIA) uitgevoerd. Op basis van de bevindingen van deze DPIA is de lijst met functies herzien die in het BRF-model worden gebruikt, wat heeft geleid tot een lijst van de 32 functies die in Appendix A worden vermeld. Een impactbeoordeling gericht op informatiebeveiliging staat gepland voor 2025, wat mogelijk kan leiden tot wijzigingen in het privacybeleid van het platform en de BRF-train-testpijplijn.

4.3 Algemene Verordening Gegevensbescherming (AVG)

In deze paragraaf worden enkele van de meest relevante aspecten van de AVG in verband met deze casus behandeld en hoe deze verplichtingen door het platform worden nagekomen. Het biedt echter slechts een korte juridische analyse en is niet uitputtend (zie ook Box 4).

Art. 4(1) van de AVG stelt dat persoonsgegevens 'alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon'. De gegevens die door het platform worden gebruikt vallen onder de definitie van persoonsgegevens. De reden hiervoor is dat tijdens het algoritmische proces rijgedrag en betalingsgegevens

herleidbaar zijn tot een specifieke gebruiker via zijn account. Om een profiel op het platform aan te maken moet een gebruiker zijn eigen persoonsgegevens verstrekken (zoals naam, adres en geboortedatum). Bovendien stelt het platform persoonsgegevens beschikbaar die gekoppeld zijn aan de persoon in kwestie wanneer een gebruiker wil inzien op welke manier zijn persoonsgegevens zijn verwerkt in het kader van algoritmische besluitvorming.

Art. 6 AVG betreft de rechtmatigheid van de verwerking. Volgens lid 1 onder a is de verwerking van gegevens rechtmatig wanneer de betrokkene daarvoor toestemming heeft gegeven. Door de algemene voorwaarden van het platform te accepteren gaat een gebruiker akkoord met het privacybeleid, inclusief welke persoonsgegevens worden verwerkt, voor welke doeleinden en met wie deze worden gedeeld. Om gebruikers bewust te maken van het feit dat DBMS-gegevens worden verzameld, wordt aan het begin van elke rit een pop-upscherm getoond met de melding dat het rijgedrag wordt gemonitord en een link naar het privacybeleid met meer details. Dit moedigt de bestuurder aan het privacybeleid te lezen.

Volgens het Hof van Justitie van de Europese Unie (HvJ-EU) omvat een besluit dat uitsluitend is gebaseerd op geautomatiseerde verwerking ook profilering. In de zaak Schufa wordt profilering zelf beschouwd als een 'geautomatiseerde besluitvorming' in de zin van de AVG, zelfs als de uiteindelijke beslissing over het verstrekken van krediet aan alleen geschikte kredietverstrekkers wordt genomen na menselijke tussenkomst. De definitie van 'profilering' valt onder het bepalen van de geautomatiseerde vaststelling van een waarschijnlijkheidswaarde die is gebaseerd op persoonsgegevens van de betrokkene en die betrekking heeft op het vermogen van een persoon om in de toekomst een lening af te lossen.[4]

Volgens artikel 22(3) AVG moet de verwerkingsverantwoordelijke passende maatregelen treffen, zoals het recht van de betrokkene op menselijke tussenkomst om zijn standpunten kenbaar te maken en eventueel beroep aan te tekenen.

Artikel 15 van de AVG stelt dat een verwerkingsverantwoordelijke de betrokkene moet informeren over zijn rechten in het kader van de verwerking van persoonsgegevens. Dit omvat het bekendmaken van het bestaan van geautomatiseerde besluitvorming. In overeenstemming met artikel 15(1)(h) moet het platform informatie verstrekken over de werking van het algoritme, met inbegrip van de logica die ten grondslag ligt aan de geautomatiseerde verwerking van de persoonsgegevens en de gevolgen van een dergelijke verwerking. De gebruiker kan vervolgens een weloverwogen beslissing nemen om zijn recht op toegang of zijn recht om bezwaar te maken tegen de verwerking van zijn persoonsgegevens uit te oefenen.

In dit geval zijn algoritmische risicovoorspellingen een vorm van profilering, aangezien er kansen worden gegenereerd om te bepalen of een gebruiker geschikt is om de diensten van het platform te gebruiken op basis van persoonsgegevens. Naar aanleiding van de Schüfa-uitspraak zou dit kunnen worden beschouwd als geautomatiseerde besluitvorming, zelfs als er menselijke betrokkenheid aanwezig is bij het besluitvormingsproces. Omdat de menselijke analist discretionaire bevoegdheid toekomt, die hij 40-50% van de tijd gebruikt om af te wijken van de voorspelling van het algoritme (zie ook Box 5 in Appendix B), het is waarschijnlijk dat er geen sprake is van volledig geautomatiseerde besluitvorming, zoals gedefinieerd in artikel 22 van de AVG. Het privacybeleid benadrukt dat de beslissing om een account te blokkeren op basis van

[4] ECLI:EU:C:2023:957 §47

voorspellingen van een algoritme uiteindelijk wordt genomen door een menselijke analist. Het privacybeleid biedt een beschrijving van het algoritme op hoog niveau, maar er zijn geen specifieke richtlijnen over het niveau van detail dat vereist is om de werking ervan uit te leggen om te voldoen aan de verplichtingen van artikel 15 van de AVG. Daarom is de verstrekte uitleg waarschijnlijk voldoende.

Het platform werkt aan een volledige update van het privacybeleid zodat veel gedetailleerder kan worden beschreven hoe het algoritme werkt en welke gegevens worden gebruikt om het rijgedrag van een gebruiker te evalueren.

Gebruikers kunnen bezwaar maken tegen beslissingen met betrekking tot hun blokkering van het platform. Nadat een gebruiker op de hoogte is gesteld dat zijn account wordt geblokkeerd, kan hij bezwaar maken door een e-mail te sturen naar de klantenondersteuning om toegang te vragen tot zijn gegevens. De opgevraagde gegevens worden vervolgens verstrekt. Als gebruikers onjuiste gegevens vaststellen en bewijs kunnen leveren ter ondersteuning hiervan, die dus hebben bijgedragen aan de beslissing om hen te blokkeren, kunnen ze deze informatie gebruiken om de beslissing van het platform aan te vechten.

Op basis van het due diligence-proces dat voorafgaand aan het opstellen van dit document is uitgevoerd, waaronder een analyse door Algorithm Audit van het privacybeleid van het platform, lijkt het erop dat er enkele waarborgen zijn getroffen om ervoor te zorgen dat gebruikers toegang hebben tot menselijke tussenkomst, hun mening kunnen uiten en beslissingen die over hen worden genomen kunnen aanvechten.

4.4 AI Verordening

De AI Verordening stelt eisen aan het gebruik van kunstmatige intelligentie (AI) binnen de Europese Unie (EU). Voor toepassingen van AI-systemen met een hoog risico zijn verplichte controlemaatregelen vereist om de veiligheid, gezondheid en fundamentele rechten van EU-burgers te beschermen.

Het algoritmesysteem dat wordt beoordeeld kwalificeert als een AI-systeem omdat het machine learning gebruikt om voorspellingen te genereren op basis van invoergegevens, wat voldoet aan de definitie in artikel 3 van de AI Verordening en wat verder is uitgewerkt in overweging 12. Volgens bijlage III van de Verordening valt het systeem echter niet in de categorie AI-toepassingen met een hoog risico, aangezien het gebruiksscenario niet is opgenomen in de genoemde categorieën met een hoog risico.

De dichtstbijzijnde relevante toepassing met een hoog risico die in Bijlage III is vermeld valt onder categorie 5: *"Toegang tot en gebruik van essentiële particuliere en publieke diensten en uitkeringen"*. Het gebruik van een auto van een autodeelplatform kwalificeert echter niet als een essentiële privédienst, aangezien meerdere opties voor autodelen en alternatieven voor openbaar vervoer (zoals bussen en treinen) over het algemeen beschikbaar zijn voor gebruikers.

Als gevolg hiervan zijn de verplichte controlemaatregelen voor toepassingen met een hoog risico niet van toepassing op het algoritmische systeem dat wordt beoordeeld. Bovendien wordt verwacht dat de controlemaatregelen, zoals verzocht door de Europese Commissie om te worden ontwikkeld door standaardisatieorganisatie CEN-CENELEC, in de loop van 2026 worden afgerond en daarom momenteel niet beschikbaar zijn om gebruikt te worden als standaard.

Appendix A – Verzamelde rijeigenschappen

Tabel 1 geeft een overzicht van alle rijeigenschappen die door het BRF model meegewogen om een voorspelling te genereren.

Tabel 1. Overzicht van het Data Behavior Monitoring System (DBMS) en betalingsgegevens die aan het BRF-model worden gevoed.

#	Variable name	Description
Cornering events		
1	AVG_CORNERING_GFORCE	The average G-force of cornering events based on all trips per user
2	CORNERING_PERKM	The number of cornering events recorded over all trips per user divided by the number of driven kilometers
Breaking events		
3	AVG_BRAKING_GFORCE	The average G-force of braking events based on all trips per user
4	BRAKING_PERKM	The number of braking events recorded over all trips per user divided by the number of driven kilometers
Acceleration events		
5	AVG_ACCELERATION_GFORCE	The average G-force of acceleration events based on all trips per user
6	ACCELERATION_PERKM	The number of acceleration events recorded over all trips per user divided by the number of driven kilometers
Driving events		
7	TOTAL_DRIVING_EVENTS_PERKM	Total number of driving events (cornering and braking events) per user divided by the number of driven kilometers
8	TOTAL_DRIVING_EVENTS_PERMINUTE	Total number of driving events (cornering and braking events) per user divided by the total driving time in minutes per user
Speed events		
9	SPEEDING_EVENTS_PERKM	Total number of speeding events (>15 km/h above the limit) per user divided by the number of driven kilometers
10	SPEEDING_EVENTS_PERMINUTE	Total number of speeding events (>15 km/h above the limit) per user divided by the total driving time in minutes per user

11	NUM_1_SPEED_EVENT_PERC	Percentage of trips of a user with at least 1 speeding event
12	NUM_2_SPEED_EVENT_PERC	Percentage of trips of a user with at least 2 speeding events
13	NUM_5_SPEED_EVENT_PERC	Percentage of trips of a user with at least 5 speeding events
14	NUM_1_SPEED_EVENT_PERC_30	Percentage of trips of a user with at least 1 speeding event in a limit up to 30 km/h
15	NUM_2_SPEED_EVENT_PERC_30	Percentage of trips of a user with at least 2 speeding events in a limit up to 30 km/h
16	NUM_5_SPEED_EVENT_PERC_30	Percentage of trips of a user with at least 5 speeding events in a limit up to 30 km/h
17	NUM_1_SPEED_EVENT_PERC_50	Percentage of trips of a user with at least 1 speeding event in a limit between 30 and 50 km/h
18	NUM_2_SPEED_EVENT_PERC_50	Percentage of trips of a user with at least 2 speeding events in a limit between 30 and 50 km/h
19	NUM_5_SPEED_EVENT_PERC_50	Percentage of trips of a user with at least 5 speeding events in a limit between 30 and 50 km/h
20	NUM_1_SPEED_EVENT_PERC_70	Percentage of trips of a user with at least 1 speeding event in a limit between 50 and 70 km/h
21	NUM_2_SPEED_EVENT_PERC_70	Percentage of trips of a user with at least 2 speeding events in a limit between 50 and 70 km/h
22	NUM_5_SPEED_EVENT_PERC_70	Percentage of trips of a user with at least 5 speeding events in a limit between 50 and 70 km/h
23	NUM_1_SPEED_EVENT_PERC_OVER70	Percentage of trips of a user with at least 1 speeding event in a limit at least 70 km/h
24	NUM_2_SPEED_EVENT_PERC_OVER70	Percentage of trips of a user with at least 2 speeding events in a limit at least 70 km/h
25	NUM_5_SPEED_EVENT_PERC_OVER70	Percentage of trips of a user with at least 5 speeding events in a limit at least 70 km/h
26	URBAN_AREA_SPEEDING_ PERCENTAGE_0_20_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was between 0 and 20%
27	URBAN_AREA_SPEEDING_ PERCENTAGE_21_40_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was between 21 and 40%

28	URBAN_AREA_SPEEDING_ PERCENTAGE_41_55_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was between 41 and 55%
29	URBAN_AREA_SPEEDING_ PERCENTAGE_56_PLUS_PER_KM	Number of trips per total km driven in an urban area, where the speed limit was up to and including 50 km/h and the speeding percentage above the limit was at least 56%
30	NON_URBAN_AREA_SPEEDING_ PERCENTAGE_0_10_PER_KM	Number of trips per total km driven in an non-urban area, where the speed limit was greater than 50 km/h and the speeding percentage above the limit was between 0 and 10%
31	NON_URBAN_AREA_SPEEDING_ PERCENTAGE_11_20_PER_KM	Number of trips per total km driven in an non-urban area, where the speed limit was greater than 50 km/h and the speeding percentage above the limit was between 11 and 20%
32	NON_URBAN_AREA_SPEEDING_ PERCENTAGE_21_PLUS_PER_KM	Number of trips per total km driven in an non-urban area, where the speed limit was greater than 50 km/h and the speeding percentage above the limit was at least 21%

Appendix B – Communicatie met onverantwoordelijke bestuurders

Hier worden de procedures beschreven wanneer klanten de waarschuwings- of blokkeringsdrempel overschrijden. Elke keer dat een gebruiker een nieuwe reis voltooit wordt er een risicoscore berekend met behulp van het BRF-model.

B1. Waarschuwing gebruikers

De volgende procedure wordt gevolgd wanneer de risicoscore van een gebruiker de waarschuwingsdrempel overschrijdt.

Het platform stuurt een eerste waarschuwingsmail naar de gebruiker om te melden dat zijn rijgedrag signalen van onverantwoordelijkheid vertoont. Een menselijke analist zal de DBMS-gegevens van de gebruiker beoordelen, afwijkingen identificeren en deze observaties aan de gebruiker communiceren om zijn rijgedrag te helpen verbeteren.

Als een nieuwe rit resulteert in een risicoscore die nog steeds de waarschuwingsdrempel overschrijdt, wordt het rijprofiel van de gebruiker doorgestuurd naar een menselijke analist die het profiel van de gebruiker handmatig controleert (zie Box 5). De analist beslist dan of er een tweede, strengere waarschuwing wordt gestuurd. Als een volgende rit resulteert in een lagere risicoscore wordt er geen extra waarschuwing gestuurd.

B2. Gebruikers blokkeren

De volgende procedure wordt gevolgd wanneer de risicoscore van een gebruiker de blokkeringsdrempel overschrijdt.

Het profiel van de gebruiker wordt naar een menselijke analist gestuurd voor een handmatige beoordeling (zie Box 5). Als een gebruiker de blokkeringsdrempel overschrijdt zonder eerder een waarschuwing te hebben ontvangen, kan de analist ervoor kiezen om eerst een waarschuwingsbericht te versturen. In gevallen van ernstige schending van de algemene voorwaarden van het platform, zoals te hard rijden, kan de analist echter besluiten om de gebruiker onmiddellijk te blokkeren.

De gebruiker wordt transparant behandeld omtrent de variabelen die bijdragen aan hun blokkering. Een menselijke analist beoordeelt de DBMS-gegevens van de gebruiker om de afwijkende kenmerken te identificeren die hebben geleid tot een hoge risicoscore, die vervolgens met de gebruiker worden gedeeld. De e-mailmelding die gebruikers informeert over hun blokkering biedt gebruikers ook de mogelijkheid om in beroep te gaan tegen de beslissing van het platform. Als gebruikers zich zorgen maken over de nauwkeurigheid van de over hen verzamelde DBMS-gegevens kunnen ze contact opnemen met het platform voor verduidelijking.

Box 5

Beoordeling van een gebruikersprofiel door een menselijke analist

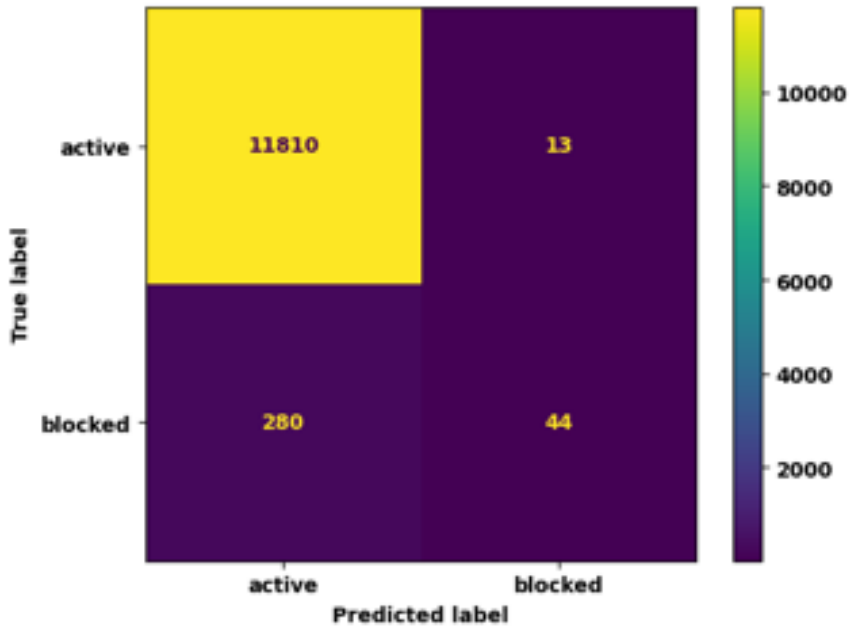
Handmatige controles vereisen menselijke tussenkomst, e.g. menselijke analisten controleren of de gegevens die aan een gebruiker zijn gekoppeld betrouwbaar en volledig lijken. Deze beoordeling omvat geen analyse van persoonlijke gegevens, alleen een database-ID wordt weergegeven in combinatie met rijeigenschappen. Er worden geen namen, geboortedata, geregistreerde adressen en andere persoonlijke gegevens weergegeven wanneer rijeigenschappen worden geanalyseerd. Menselijke analisten volgen een training voordat ze beoordelingen uitvoeren. Gemiddeld besteedt een menselijke analist 30-60 minuten per dag aan het inspecteren van ongeveer 20 waarschuwingsgevallen en 5 blokkeringsgevallen. 50-60% van de gebruikers wordt geblokkeerd na menselijke inspectie waarvoor een risicoscore wordt voorspeld die de blokkeringsdrempel overschrijdt.

Appendix C – Confusion matrices en PR curve

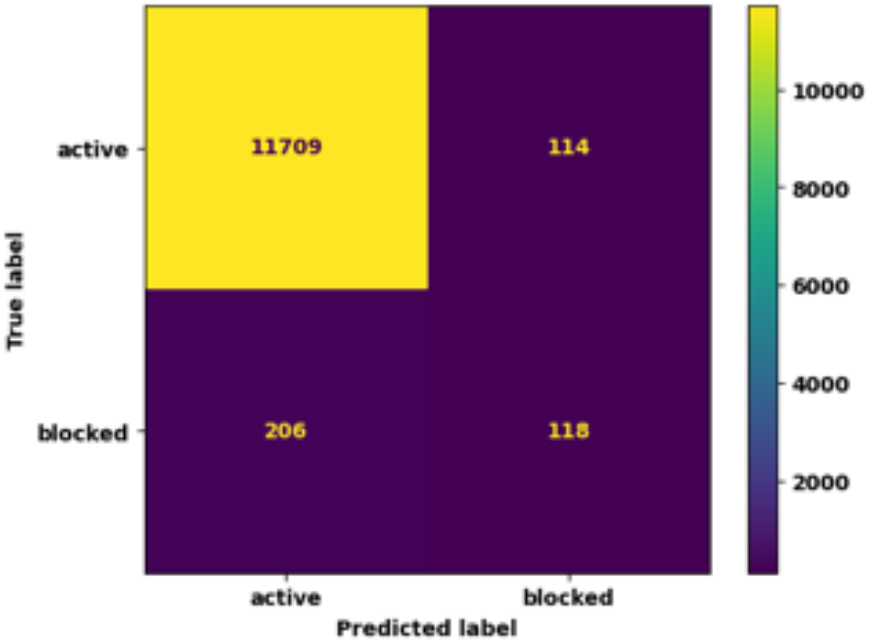
De confusion matrices van het BRF-model voor drempelwaardenr_w=0.25 and r_w=0.5 op de testdataset met 12.147 datapunten, met hyperparameters zoals gespecificeerd in 3.3 Hyperparameter-afstemming van het BRF-model , worden weergegeven in Figuur 3 en Figuur 4.

De waarden in de verwarringsmatrices komen de volgende betekenis toe:

> **True Positive (TP):** True label blocked, predicted label blocked;
> **False Positive (FP):** True label active, predicted label blocked;
> **True Negative (TN):** True label active, predicted label active;
> **False Negative (FN):** True label blocked, predicted label active.

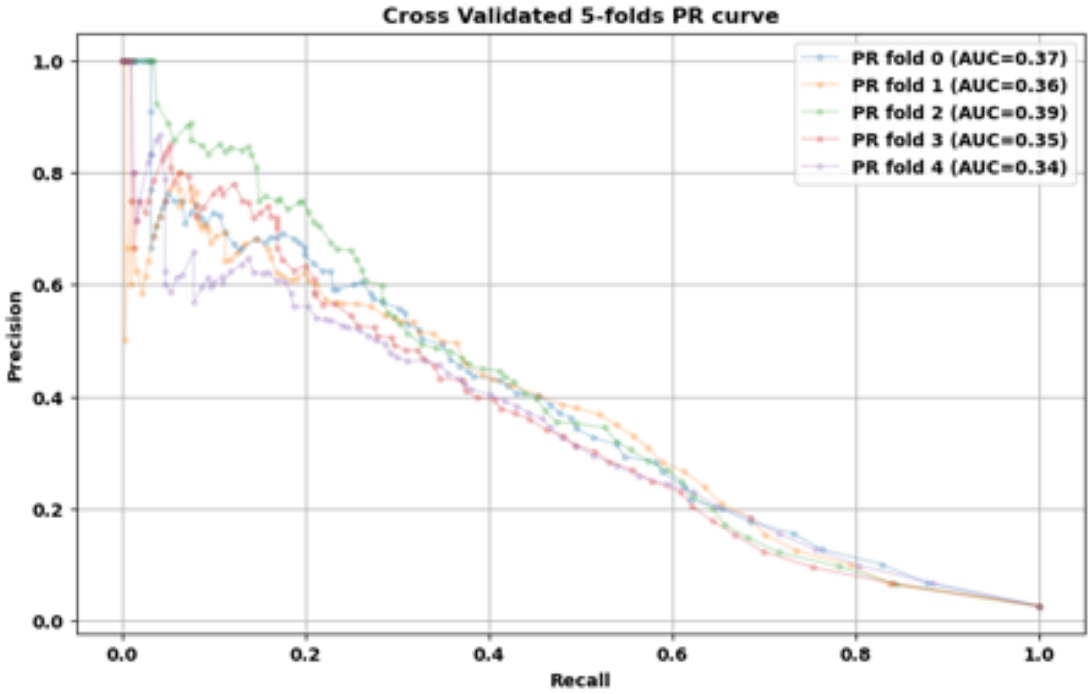


Figuur 3 - Confusion matrix voor drempelwaarde 0.5



Figuur 4 - Correlation matrix voor drempelwaarde 0.25

The Precision-Recall (PR) curve voor 5-voudige kruisvalidatie van het BRF model is weergegeven in Figuur 5.



Figuur 5 - PR-curven voor 5-voudige kruisvalidatie van het BRF-model

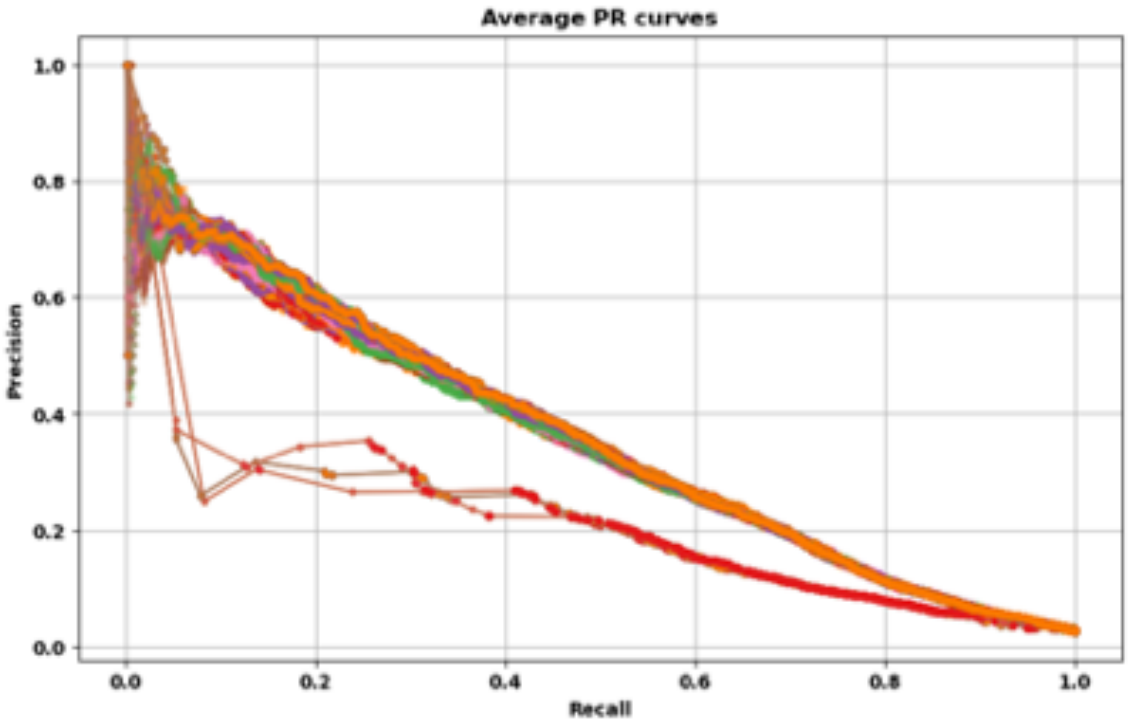
Appendix D – Gevoeligheidstesten

De gevoeligheid van het BRF-model wordt getest door het model te trainen met behulp van 3*3*4*3*1=108 verschillende hyperparameterspecificaties:

> Aantal bomen (n_est): 100, 200, 300
> Maximale diepte per boom (max_depth): None, 1, 5
> Minimaal aantal observaties per split (min_samples_split): 2, 3, 5, 10
> Minimaal aantal vereiste samples in een blad (min_samples_leaf): 1, 3, 5
> Aantal functies waarvan wordt overwogen om een splitsing te maken (max_features): 32.

Voor elke set hyperparameters genereert het BRF-model voorspellingen met behulp van 5-voudige kruisvalidatie.[5] De gemiddelden van de 5-voudige kruisvalidatievoorspellingen voor alle 108 combinaties worden weergegeven in Figuur 6.

Uit deze analyse kan worden geconcludeerd dat het BRF-model niet erg gevoelig is voor veranderingen in hyperparameters, aangezien de PR-curve grotendeels onveranderd blijft.



Figuur 6 - De gemiddelden van de 5-voudige kruisvalidatievoorspellingen voor alle 108 combinaties van een BRF model.

[5] Voor een open-sourcevoorbeeld van een vergelijkbare dataset, zie sectie 5 van dit notitieboek: https://github.com/NGO-Algorithm-Audit/ML-pipeline/blob/main/BRF_pipeline.ipynb

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn drieledig:

Kennisplatform	Samenbrengen van kennis en experts voor collectief leerproces over verantwoorde inzet van algoritmes, bijvoorbeeld ons AI Policy Observatory en position papers
Normatieve adviescommissies	Adviseren over ethische kwesties in concrete algoritmische toepassingen door het samenbrengen van deliberatieve, diverse adviesommissies, met algoprudentie als resultaat
Technische hulpmiddelen	Implementeren en testen van technische methoden voor bias-detectie en -mitigatie, zoals onze bias detection tool
Projectwerk	Ondersteuning bij specifieke vragen vanuit de publieke en private sector over de verantwoorde inzet van algoritmes.

Structurele partners van Algorithm Audit

SIDN Fonds

Het SIDN Fonds staat voor een sterk internet voor iedereen. Het Fonds investeert in projecten met lef en maatschappelijke meerwaarde, met als doel het borgen van publieke waarden online en in de digitale democratie.

European AI&Society Fund

Het European AI&Society Fund ondersteunt organisaties uit heel Europa die AI beleid vormgeven waarin mens en maatschappij centraal staan. Het fonds is een samenwerkingsverband van 14 Europese en Amerikaase filantropische organisaties.

Opbouwen van *publieke kennis*

over verantwoorde AI *zonder winstoogmerk*

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Algorithm Audit

Stichting Algorithm Audit is geregistreerd bij de
Kamer van Koophandel onder nummer 83979212