Resultaten Verhuiskans Vervolgproject

Prototype verbeterslagen samengevat

Jasper Menger
Tim de Jong
José Gómez Pérez

CBS Heerlen
CBS-weg 11
6412 EX Heerlen
Postbus 4481
6401 CZ Heerlen
+31 45 570 60 00

www.cbs.nl

projectnummer 305709 / 09
BPM / CBD / DVP / SAL
September 2021

Inhoudsopgave

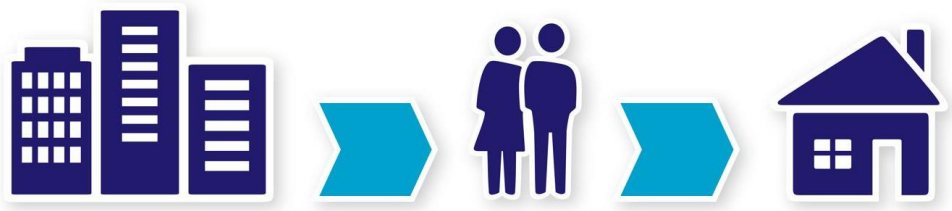
1. Samenvatting

Het ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK), de opdrachtgever van Woononderzoek Nederland (WoON), heeft het CBS gevraagd om onderzoek te doen naar alternatieve manieren om de verhuiskans zo goed mogelijk te schatten. Het CBS heeft daarom onderzoek gedaan naar de mogelijkheden om deze te bepalen op basis van registraties en met behulp van *machine learning* technieken.

Tijdens de eerste projectfase in 2017-2018 is een prototype ontwikkeld voor schatting van persoonlijke verhuiskansen. Het CBS heeft aan BZK voorgesteld om dit prototype verder te verfijnen, en op te schalen naar de volledige populatie. In het huidige project zijn hiertoe twee benaderingen gevolgd; één op individueel niveau (binaire classificatie), en één op subpopulatieniveau (regressie).

Bij de eerste benadering is uitgegaan van de oorspronkelijke gedachte van flexibele inzetbaarheid, waarbij individuele verhuiskansen door de gebruiker naar believen kunnen worden geaggregeerd naar de doelgroepen voor het onderzoek in kwestie. De kwaliteit van de schattingen is verbeterd ten opzichte van het eerste project, maar van onvoldoende nauwkeurigheid om op individueel niveau toe te passen. De resultaten zijn daarmee ongeschikt voor massa-imputatie (in Woonbase of elders); het is niet raadzaam deze micro-uitkomsten voor derden beschikbaar te stellen.

De tweede benadering kan in principe betere schattingen opleveren voor verschillende tijdsperioden en subpopulaties, en leent zich voor snelle reguliere actualisatie (bijvoorbeeld elk kwartaal of elke maand). Een eerste opzet hiervoor is gemaakt, echter binnen de beschikbare tijd was het niet mogelijk dit concept tot in detail uit te werken. Ook vereist een dergelijke aanpak dat zaken als doelgroepen en kwaliteitsmaten van tevoren worden vastgelegd; de beoogde toepassing is bepalend voor de modelkeuze. Desgewenst kan een op de toepassing toegesneden schatting van verhuiskansen in een derde ontwikkeltraject alsnog volledig worden geïmplementeerd en geëvalueerd.



2. Inleiding

2.1 Doel

Hoeveel mensen gaan op korte termijn verhuizen? Dit is een cruciale vraag voor beleidsmakers, die willen bepalen hoeveel woningen er gebouwd moeten worden. Dit wordt momenteel mede geschat via de verhuiswens; het CBS vraagt in het Woononderzoek Nederland (WoON) aan mensen of zij van plan zijn om binnen twee jaar te verhuizen.

Het ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK), de opdrachtgever van WoON, heeft het CBS gevraagd om onderzoek te doen naar alternatieve manieren om de verhuiskans zo goed mogelijk te schatten. Het CBS heeft daarom onderzoek gedaan naar de mogelijkheden om deze te bepalen op basis van registraties en met behulp van *machine learning* technieken.

Tijdens de eerste projectfase is een prototype ontwikkeld voor schatting van persoonlijke verhuiskansen. Het CBS heeft aan BZK voorgesteld om dit prototype verder te verfijnen, en op te schalen naar de volledige populatie. Voornaamste doel van dit vervolgonderzoek is nagaan of de aldus verbeterde verhuiskansen voldoende bruikbaar blijken voor beleid met betrekking tot specifieke bevolkingsgroepen en regionale uitsplitsingen.

Bij het vervolgonderzoek is in eerste instantie voortgebouwd op de gedachte van flexibele inzetbaarheid, waarbij de gebruiker zelf naar believen individuele verhuiskansen aggregeert naar de doelgroepen voor het onderzoek in kwestie. Gedurende de uitvoering van het project kwamen enkele mogelijke toepassingen voor een korte-termijn-schatting van de verhuiskans ter sprake:

1. **Lokale vooruitblik**. Korte-termijn verhuiskansen kunnen worden gepubliceerd op geaggregeerd niveau, per regio en naar achtergrondkenmerk, en zodoende ten goede komen aan geïnteresseerde beleidsmakers en onderzoekers.

2. **Scherp enquêteren**. Gerichte benadering van respondenten bij het woononderzoek, voor betere waarneming van het verband tussen verhuiswens en -gedrag. Door stratificatie van het steekproefontwerp kunnen meer personen met een hoge (of juist lage) kans op verhuizen worden aangeschreven.

3. **Betere bevolkingsprognose**. Nieuw ontdekte verbanden met voorspellende waarde kunnen worden verwerkt in de modellen voor de PBL/CBS regionale bevolkings- en huishoudens-prognose.

Het prototype voor het schatten van verhuiskansen is op diverse punten verder ontwikkeld. Het huidige document beschrijft de analyses en resultaten van dit vervolgonderzoek bondig. Op verzoek is een uitgebreide technische rapportage (in het Engels) beschikbaar.

2.2 Voorgeschiedenis

In 2019 heeft het CBS een eerste onderzoek naar een schatting van verhuiskansen van Nederlanders opgeleverd; het resultaat kan worden geraadpleegd via een interactief histogram, en het prototype wordt beschreven in Burger et al. (2018).

In het woononderzoek krijgen respondenten de vraag "Bent u van plan om binnen twee jaar te verhuizen?". Als alternatief is een model ontwikkeld om aan de hand van registergegevens de kans te schatten dat een persoon binnen twee jaar zal verhuizen. De doelgroep bestond uit alle personen die op 31 december van een bepaald jaar in Nederland woonden. Hieruit werd een steekproef van 100 duizend personen getrokken, aangevuld met de respondenten van het onderzoek. De doelvariabele was een (binaire) indicator voor verhuizingen, die na twee jaar bekend zijn uit het bevolkingsregister. Domeindeskundigen leverden potentiële voorspellers uit registergegevens over personen, huishoudens, woningen en regio's. Naast de huidige status is informatie uit het verleden opgenomen, zoals de tijd sinds de laatste wijziging van achtergrondkenmerken en het aantal wijzigingen.

Een aantal modellen werd vergeleken, waaronder ten eerste logistieke regressie (gegeneraliseerd lineair model met logit-link en binomiale foutverdeling) met derdegraads polynomen om rekening te houden met niet-lineaire relaties, met of zonder regularisatie (*lasso or ridge*) om *overfitting* te voorkomen, met of zonder tweerichting-interacties; en ten tweede *random forest*. Modellen werden getraind door de relatie tussen kenmerken over 1996-2012 en verhuizingen in 2013-2014 in kaart te brengen, en getest door verhuizingen in 2015-2016 te schatten op basis van kenmerken over 1998-2014, en deze schattingen te vergelijken met waargenomen verhuizingen.

In een aanvullende analyse werden modelprestaties vergeleken tussen verschillende subpopulaties (Burger 2020). Het model presteerde het beste in subpopulaties met een tussentijds waargenomen verhuispercentage en een hoge standaardafwijking in geschatte verhuiskans (mogelijk lijken individuen binnen groepen met gemiddeld een hoge verhuiskans te veel op elkaar in kenmerken voor het model om hun individuele verhuiskans goed te schatten). Modelprestaties waren slechter voor subpopulaties met een lage verhuiskans; zoals leeftijdsgroep 55-74, getrouwde paren zonder kinderen, en huiseigenaren met kinderen. Merk op dat het voor schattingen op geaggregeerd niveau niet nodig is om schattingen op individueel niveau te maken.

Sommige indicatoren voor modelprestaties zijn gevoelig voor klasse-onbalans (er zijn veel meer blijvers dan verhuizers). Dit maakt het moeilijk om subpopulaties, datasets of toepassingen te vergelijken die verschillen in het waargenomen verhuispercentage; Burger en Meertens (2020) gaan hier nader op in.

2.3 Toepassingsgebied

2.3.1 Woonbase

Het CBS werkt in samenwerking met het ministerie van Binnenlandse Zaken en Koninkrijksrelaties aan de ontwikkeling van een woononderzoek op basis van integrale gegevensbronnen; de Woonbase. De ontwikkelingen op de woningmarkt vragen om recentere, integraal beschikbare cijfers. De Woonbase zal deze integrale gegevens, aangevuld met informatie uit enquêtes, omvatten. Doel is om de verhuiskansen, zodra van voldoende kwaliteit, toe te voegen aan de populatiebestanden van Woonbase.

2.3.2 Projectplan

De hoofdopdracht zoals opgesteld in het projectplan is “het oorspronkelijke verhuiskansmodel te verbeteren, doelgroeponderzoek te doen en aan de hand daarvan extra verbeteropties uit te werken.” Het voorgaande ambitieuze doel om een modelmatig alternatief voor de verhuiswens zelf te ontwikkelen verschuift daarmee wat naar de achtergrond, en de nadruk komt te liggen op het zo goed mogelijk schatten van de verhuismobiliteit van personen.

Gaandeweg de uitvoering zijn de doelen uit dit projectplan van 2019 op enkele punten bijgesteld. Voornamelijk vanwege beperkingen op het vlak van IT bleek het niet mogelijk om alle van te voren voorgenomen doelen tijdig te behalen. De vijf stappen uit het projectplan zijn gaandeweg als volgt bijgesteld:

1. **Extra kenmerken toevoegen**. Toegevoegd zijn; afstand tot geboorteplaats, afstand ouder-kind, woningkenmerken, en medicijn- en zorggebruik. Baaninformatie blijft achterwege.
2. **Validatie voor doelgroepen**. Gedaan. Indelingen voor het afbakenen van doelgroepen zijn meegenomen bij de resultaten.
3. **Model verbeteren voor specifieke doelgroepen**. Analyse is verricht. Zie de nota *Kwaliteit geschatte verhuiskans per doelgroep* van December 2020.
4. **Varianten van verhuiskansen**. De bestemming van de verhuizing wordt niet nader uitgesplitst (type woning, naar instituut, binnen/buiten gemeente). De algemene verhuiskans heeft prioriteit.
5. **Recentere schatting**. De verhuiskansen worden geschat voor 2017, zijnde het meest recente onderzoeksjaar ten tijde van de start van dit onderzoek.

De opdracht is dus in essentie voltooid, maar wel met enkele concessies. De zaken die zijn blijven liggen kunnen in een eventueel vervolgtraject alsnog weer worden opgepakt, gegeven dat de huidige resultaten veelbelovend worden bevonden en de benodigde (reken)faciliteiten dan voorhanden zijn.

2.3.3 Doelpopulatie

Wat de doelpopulatie betreft is ervoor gekozen om aan te sluiten op de definitie van verhuizingen zoals gebruikt bij Demografie; verhuizingen binnen en tussen gemeenten tellen mee, bewegingen over de landsgrenzen (immigratie en emigratie) niet. Dit vergemakkelijkt vergelijkingen met cijfers over bevolkingsontwikkeling. Hier bovenop wordt een leeftijd selectie van 15+ toegepast (gepeild op 1 januari van het onderzoeksjaar). Kinderen onder deze leeftijd worden verondersteld samen te wonen met hun ouders of verzorgers, en zullen zodoende dezelfde verhuiskans als de referentiepersoon in het huishouden hebben.

2.4 Status quo

In deze paragraaf bespreken we de modelprestaties van het eerste onderzoek.

2.4.1 Verhuiswens

Om de wens om te verhuizen in te kunnen schatten, vraagt het CBS mensen of ze van plan zijn binnen twee jaar te verhuizen. Bij dit onderzoek wordt door middel van een vragenlijst informatie verzameld over de huidige woning en de tevredenheid met de woning én de woonomgeving. Daarnaast wordt gevraagd naar de verhuiswens en woonwensen. Het WoON-onderzoek wordt eens in de drie jaar uitgevoerd.

De zelfgerapporteerde verhuiswens kan worden gebruikt als proxy voor de verhuiskans op korte termijn. Echter aangezien een wens tot verhuizen niet per sé hoeft te culmineren in een daadwerkelijke verhuizing, en omdat het niet makkelijk is om de eigen persoonlijke situatie te voorspellen, blijkt de nauwkeurigheid van deze schatting in de praktijk beperkt. Ook zijn er op basis van een enquête-steekproef minder gedetailleerdere uitsplitsingen naar doelgroepen en regio's te maken dan via registraties. Het uitvragen van de verhuiswens zal interessant zijn als doel op zich (wie zouden willen verhuizen en waarom?), maar is dus beperkt toepasbaar voor het schatten van de mobiliteit (wie gaan verhuizen?).

2.4.2 Prestaties eerste prototype

Het model schat voor elke persoon in de testverzameling een waarschijnlijkheid dat de persoon binnen twee jaar zal verhuizen. Hierbij worden de waarschijnlijkheden in een binaire variabele omgezet; dat wil zeggen dat "persoon gaat verhuizen" wordt geschat indien de waarschijnlijkheid een grenswaarde overschrijdt, en anders "persoon blijft".

	Positive Prediction	Negative Prediction
Positive class	9192 (TP)	3989 (FN)
Negative class	18695 (FP)	45360 (TN)

***Tabel 1** – Voorbeeld van confusie-matrix voor het model uit de eerste projectfase. Voor beste RDG2-classificator, geoptimaliseerd voor G-mean bij drempel 0.11 en met score 0.70 (exclusief respondenten woononderzoek).*

Conclusie van dit eerste onderzoek: de verhuiskans van een individu kan met behulp van registerinformatie even betrouwbaar worden geschat als met behulp van de verhuiswensvraag van het woononderzoek, en dat een deel van de kenmerken de schattingen verbeteren ten opzichte van gokken. De kwaliteit van de individuele schattingen is echter onvoldoende om op individueel niveau een verhuizing te voorzien.

3. Methode

In het huidige project zijn twee benaderingen gevolgd voor het schatten van verhuiskansen:

1. **Binaire classificatie**. Schatting op individueel niveau. Er is opgeschaald van steekproef naar volledige doelpopulatie, en extra kenmerken zijn toegevoegd.
    a. **Eerste fase.** Eerst in de geest van het oorspronkelijke project.
    b. **Tweede fase**. Vervolgens met aangepaste methodiek (wijzigingen in voorbewerking en *machine learning* aanpak).

2. **Regressie**. Schatting op niveau van subpopulaties (doelgroepen, regio's, en achtergrondkenmerken). Niet voorzien bij aanvang van het project, maar gaandeweg de uitvoering parallel geïntroduceerd naar aanleiding van voortschrijdend inzicht. Deze aanpak kan leiden tot betere nauwkeurigheid bij lagere benodigde rekencapaciteit.

Voor beide benaderingen worden de methode en belangrijkste uitkomsten uitgelicht.

Indeling	Subpopulaties	Aantal
Regio	Nederland	1
	G4 (wel/niet onderdeel van)	2
	Landsdeel	4
	Provincie	12
	COROP	40
	Arbeidsmarktregio	42
	Gemeente	388
Persoon	Man/Vrouw	2
	Leeftijd (volwassenen, leeftijd 15-17/18-34/35-54/55-74/75+)	6
	Starter (wel/geen starter)	2
Huishoudens	Samenstelling (éénpersoons / ongetrouwd paar zonder kinderen / ongetrouwd paar met kinderen / getrouwd paar zonder kinderen / getrouwd paar met kinderen / éénoudergezin / overig)	7
	Eigendom (eigen woning / huur)	2

***Tabel 2** – Overzicht van gebruikte subpopulaties.*

3.1 Gebruikte data

Voor beide benaderingen is dezelfde brondata gebruikt. Deze zijn met name afkomstig uit het stelsel van sociaal-statistische bestanden (SSB). Op grond hiervan zijn op individueel niveau de achtergrondkenmerken en het *label* (verhuist wel/niet) bepaald. Levensgebeurtenissen zijn waar mogelijk vastgesteld op basis van zogenaamde bus-bestanden, waarbij per gebeurtenis een start- en einddatum wordt gegeven. Daarnaast wordt er gebruik gemaakt van diverse jaargegevens uit zogeheten tab-bestanden. Een uiteenzetting van de benutte gegevens:

- **Persoonskenmerken**. Waaronder leeftijd, geslacht, burgerlijke staat, geboorteplaats en -land, eventuele sterfdatum, immigraties en emigraties.
- **Huishoudkenmerken**. Waaronder plaats binnen en samenstelling van het huishouden, geregistreerde partners, en ouder-kind relatie (ook gebruikt voor bepalen van afstand). Ook percentielscores voor welvaart, inkomen, en vermogen, onderscheid tussen huur of eigen woning, en voornaamste inkomensbron van het huishouden.
- **Gezondheidskenmerken**. Uitkeringen met betrekking tot ziekte en arbeidsongeschiktheid, en ook jeugdhulp (mogelijk doch niet noodzakelijkerwijs gelinkt aan ziekte).
- **Adreskenmerken**. Locatie van het woonadres, waaronder afstand ouder-kind en afstand tot geboorteplaats.
- **Buurtkenmerken**. Zaken als stedelijkheid van de buurt, typische verblijftijd in de buurt, en percentages 65-plussers, kinderen, éénpersoonshuishoudens, buitenlanders, lage inkomens, en huiseigenaren.
- **Verhuizer**. Wanneer het verblijfadres van een persoon verandert in basisregistratie personen (BRP), en deze wijziging van adres gemeten in geo-coördinaten groter is dan nul meter, dan is sprake van een verhuizing. In *machine learning* terminologie wordt of iemand wel of niet verhuist het *label* genoemd. Verschillende *labels* zijn gebruikt voor de gebruikte perioden (variërend van 1 maand tot 2 jaar).

Ook worden er enkele macro(economische) indicatoren vanuit StatLine en DNB (met name de hypotheekrente) gebruikt.

3.2 Voorbewerking

Om een goede schatting te kunnen doen op een bepaald moment in de tijd (bijvoorbeeld vandaag) en voor een bepaalde periode in de toekomst (1 maand, 6 maanden, 1 jaar, …) is het zaak dat informatie over het individu zo nauwkeurig en actueel mogelijk is. Hiertoe is er het voorbewerkingstraject.

3.2.1 Generieke voorbewerking

Om een overzicht te genereren van de meest recente informatie op een bepaald tijdstip, moeten de gebeurtenissen worden gecombineerd met indicatoren op jaarbasis. Daartoe worden de jaargegevens ingevoegd als gebeurtenissen, waarbij de begindatum van de gebeurtenis wordt gesteld op 1 januari van het jaar, en de einddatum op 31 december; hiermee is alle informatie in één bestand aanwezig. De huidige toestand voor een bepaalde datum kan vervolgens op de gebruikelijke wijze worden afgeleid.

Tijdens de generieke voorbewerking zijn in de registers verschillende inconsistenties aangetroffen die moesten worden opgelost voor verdere verwerking kon plaatsvinden. Onder

deze inconsistenties waren niet-bestaande data, personen met meer dan één ouderpaar, personen die een sociale uitkering hadden voordat zij werden geboren, en data van gebeurtenissen die een begindatum hadden later dan de einddatum.

3.2.2 Eerste fase: voorbewerking conform eerste project

In het begin van het project werd de voorbewerking van het vorige project voortgezet, met als enig opmerkelijk verschil de grotere hoeveelheid van de gebruikte gegevens. Bij de eerste fase is één groot busbestand gemaakt, waarin op elke gewenste datum ingeprikt kan worden. Dit bestand is uniek per persoon en mutatie. Alle software is opnieuw geïmplementeerd wegens noodgedwongen overstap naar een nieuw IT platform (*Greenplum*) en programmeertalen (*SQL* en *Python*).

3.2.3 Tweede fase: verbeterde voorbewerking

Bij de tweede fase werd de voorbewerking vereenvoudigd, door te kijken naar de toestand van een individu op een bepaalde datum. De voorgeschiedenis is ingekort tot vijf jaar en kenmerken zijn gefilterd op hun correlatie met het verhuisgedrag. De voorbewerking kan dan worden beperkt tot het opzoeken van de periodes in de grote verzameling gebeurtenissen die overlappen met de huidige situatie, en het samenvoegen van data. Hierbij worden de volgende afleidingen toegepast:

- **Leeftijd van de persoon**. Leeftijd in jaar op basis van de geboortedatum.
- **Afstanden**. Afstanden worden bepaald op basis van coördinaten “zoals de vogel vliegt”; geografische afstand tot vader, tot moeder, tot geboorteplaats, en verhuisde afstand.
- **Aantal verstreken dagen sinds laatste verandering**. Verandering van burgerlijke staat van het individu, van burgerlijke staat van de referentiepersoon, van partner en van adres.
- **Aantal veranderingen in laatste vijf jaar**. Het aantal wisselingen van huishouden of adres.
- **Binaire samenvatting**. Of het individu Nederlandse nationaliteit heeft, in Nederland is geboren, en of diens ouders in Nederland zijn geboren. Ook of de persoon maatschappelijke ondersteuning (WMO) of jeugdhulp ontvangt.
- **Verhuizer** Voor elke relevante classificatie-periode (1 maand, 3 maanden, 6 maanden, 1 jaar, 2 jaar) wordt een wel/niet verhuisd kolom toegevoegd.

3.2.4 Derde fase: regressie

Voor de regressie wordt voortgebouwd op de tweede fase voorbewerkte gegevens. Deze gegevens worden nu opgedeeld in een aparte dataset per maand en per subpopulatie, en met één rij per subpopulatie en maand. Macro-economische cijfers over hypotheek(rentes) worden hieraan toegevoegd.

Feature variable	Description
status_month	month summarized by the row (1-12)
imputatiecodehh_perc	% of people whose household is imputed
inverkoop_perc	% of people whose house is register as "for sale"
marital_H_perc	% of married people
marital_O_perc	% of single people
number_of_addreeses_5y_mean	Mean of the # of addresses during the last 5 years
number_of_days_address_mean	Mean of the # of days living in the same place
person_age_mean	Mean age
hh_wealth_mean	Mean of the percentile groups of private household wealth
ownership_E_perc	% of people who live in their own house (based on *typeeigendom*)
ownership_H_perc	% of tenants (based on *typeeigendom*)
ownership2_1_perc	% of people who live in their own house (based on *inhehalgr*)
ownership2_2_perc	% of tenants without rent allowance (based on *inhehalgr*)
hh_income_type_1_perc	% of private households with income observation (no student household)
hh_income_type_2_perc	% of private student households with income observation
properties_1_perc	% of people living in properties that comprise one accommodation
mortgages	# of mortgages
mortgages_ratio	# of mortgages / total # of people
mortgages_inc	# of new mortgages since last month
mortgages_inc_ratio	# of new mortgages since last month / total # of people
mortgages_growth_yearly	% of mortgages yearly growth
mortgages_growth_monthly	% of mortgages monthly growth
mortgages_existing_interests	Interest rate of the existing mortgages
mortgages_new_interests	Interest rate of the new mortgages

***Tabel 3** – Kenmerken gebruikt bij de regressie datasets.*

3.3 Training voor binaire classificatie

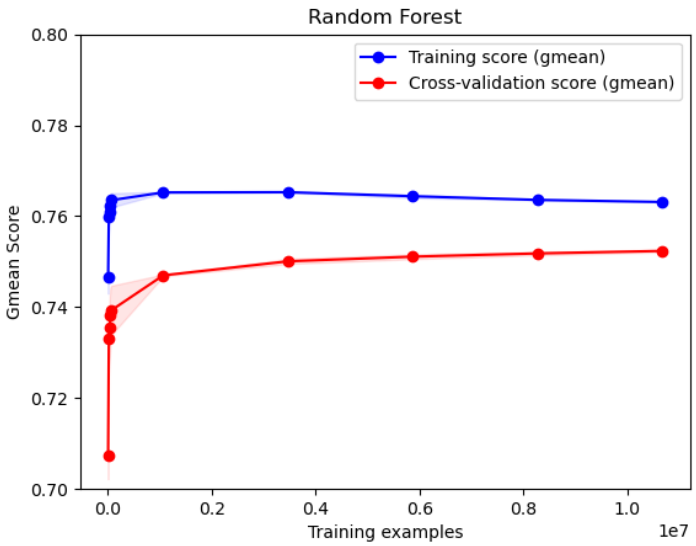
In de tweede fase van het project werden de schatting op individueel niveau uitgevoerd. Voor elk individu moest het model aangeven of dat individu in een bepaalde tijdsperiode zou verhuizen of niet. Daarom werd het model getraind om een binaire classificatie uit te voeren met *supervised learning* methoden; diverse modellen uit het arsenaal van *scikit-learn* zijn hierbij beproefd, waaronder logistische regressie en *random forest*. In deze methodologie (zoals beschreven door Jason Brownlee) worden de gegevens eerst geanalyseerd en worden methoden voor selectie van relevante kenmerken toegepast. Daarna worden verschillende modellen getraind met standaardparameters in een methode die *spot checking* wordt genoemd en die een eerste idee geeft van welke modellen geschikt zijn om het probleem op te lossen. Vervolgens worden de modellen verfijnd voor het probleem met behulp van een *random search*, en ten slotte kunnen strategieën voor *oversampling / undersampling* worden gebruikt om de prestaties van het model verder te optimaliseren.

De binaire classificatie annex schatting op individueel niveau is een onevenwichtig leerprobleem (*unbalanced machine learning problem*). De individuen die besluiten te verhuizen zijn in de

minderheid, en de mensen die blijven zijn in de grote meerderheid. Omdat de meest gangbare algoritmen op dit terrein juist uitgaan van goed gebalanceerde data, is onderzocht of het model verbetert als met dit onevenwicht expliciet rekening wordt gehouden bij respectievelijk het kiezen, trainen, en evalueren van de modellen..

Pearson's correlaties zijn gebruikt om inzicht te krijgen in de relaties van afzonderlijke variabelen met de verhuizingen (de zogenaamde *labels*). Bovendien is *Recursive Feature Elimination* gebruikt om te begrijpen welke variabelen en combinaties van variabelen het meest bijdragen tot het juist schatten van verhuizingen. Naast het aantal kenmerken moet ook de hoeveelheid gegevens worden bepaald die nodig is om een succesvol *machine learning* model te trainen. Vaak verzadigt de zogenaamde leercurve op een bepaald punt, wat betekent dat het toevoegen van meer gegevens niet veel invloed meer zal hebben op de leerprestaties, terwijl het wel veel rekentijd kost. Door het verzadigingspunt te bepalen (voorbeeld in onderstaande figuur), kan de noodzakelijke grootte van de training-set worden bepaald.

Voor de modellen die bij de *spot checking* als veelbelovend uit de bus kwamen zijn vervolgens de hyper-parameters geoptimaliseerd middels *random search*. Hierbij werd voor elke combinatie van hyperparameters de data willekeurig opgesplitst in een deel voor training (80%), en een deel voor evaluatie (20%). Net als bij het eerste project speelt de confusie-matrix een cruciale rol bij evaluatie van de resultaten, vaak samengevat in een voor ongebalanceerde data geschikte kwaliteitsmaat; op basis van recente literatuur is gekozen voor *G-mean, F0.5,* of *F1* (in plaats van de eerder gebruikte MCC).



Figuur 1 *– Representatief voorbeeld van verzadigende leercurve, voor periode van 2 jaar.*

3.4 Regressie

Elke combinatie van subpopulatie en voorspelperiode vormt een regressieprobleem. Voor elke combinatie wordt het percentage verhuizers geschat. Voor het evaluatiejaar 2017 wordt een kwart achtergehouden voor validatie, drie kwart blijft voorbehouden voor training. Vier verschillende *ensemble machine learning* algoritmes zijn beproefd: *AdaBoost*, *Gradient Boosting*, *Random Forest* en *Extra Trees*.

De volgende kenmerken zijn geselecteerd na trainen op een periode van 12 maanden op de volledige doelpopulatie, gebaseerd op minimalisatie van de *mean squared error* (MSE). De selectie is gemaakt op de gehele populatie, de schattingen zijn bepaald per subpopulatie en maand.

- Percentage geïmputeerde huishoudens (*imputatiecodehh_perc*).
- Percentage mensen met huis in verkoop (*inverkoop_perc*).
- Percentage getrouwde mensen (*marital_h_perc*).
- Percentage alleenstaanden (*marital_o_perc*).
- Gemiddeld aantal adressen de afgelopen vijf jaar (*number_of_addresses_5y_mean*).
- Gemiddeld aantal dagen geregistreerd partnerschap (*number_of_days_partner_mean*).
- Gemiddeld aantal dagen woonachtig op adres (*number_of_days_address_mean*).
- Gemiddelde leeftijd (*person_age_mean*).
- Percentage personen woonachtig in eigen huis (*ownership_2.1_perc*).
- Jaarlijkse groei hypotheekrente (*mortgages_growth_yearly*).

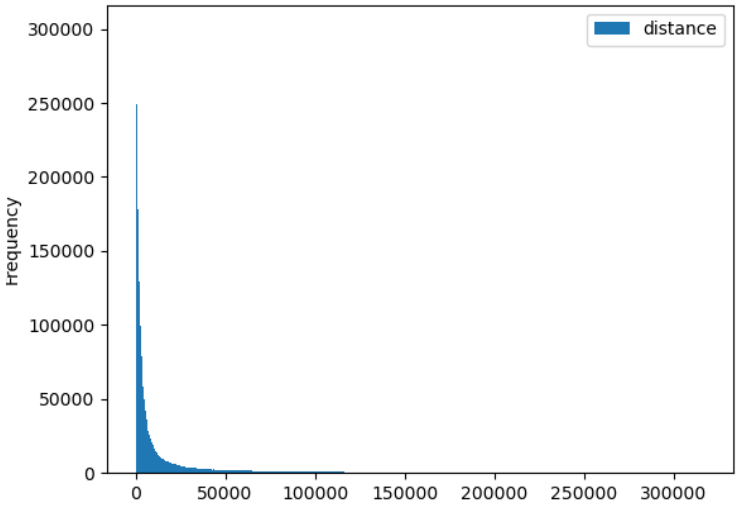
De prestaties van de regressiemodellen worden samengevat middels de *root mean squared error (RMSE)* per subpopulatie en maand. De schattingen van het regressiemodel wordt telkens vergeleken met een basisschatting (*baseline*), bepaald als het gemiddelde verhuispercentage van de personen in de trainingsset.

4. Resultaten

4.1 Analyse voorbewerkte data

Na data-exploratie is gebleken dat de verdeling van sommige kenmerken continu is, en voor anderen juist zeer discreet (slechts enkele specifieke waarden komen voor). Voor sommige algoritmen waren aanvullende verwerkingsstappen nodig om ze goed met de discrete variabelen om te laten gaan.

De fractie van verhuizers is uiteraard afhankelijk van de periode van observatie. Binnen een maand verhuist typisch 1,4% - 1,7%, en binnen twee jaar verhuist zo'n 15% - 17% van de mensen. Naast het aantal mensen dat verhuist, wordt dit gedrag ook gekarakteriseerd door de afstand die mensen afleggen. Onderstaande figuur toont de verhuisafstanden voor 2016. Hieruit valt direct op dat een meerderheid van de personen niet ver verhuist. Meer specifiek, 50% van de individuen verhuist een afstand van 3,2 km, terwijl 75% binnen een afstand van 12 km blijft.



***Figuur 2** – Verdeling van verhuisde afstand in meters voor 2016.*

4.2 Modelkwaliteit binaire classificatie

Samengevat bleken de volgende drie modellen het nauwkeurigst in hun schattingen van het wel of niet verhuizen van alle individuen:

1. **Random Forest** presteerde over het geheel genomen het beste (eerste plaats voor zowel *G-mean* als *F1*, derde plaats voor *F0.5*).
2. **XGBClassifier** deed het slechts een fractie minder goed (vierde plek voor *G-mean*, eerste voor *F0.5*, en derde voor *F1*).
3. **Support Vector Classifier** presteerde ook best aardig (goede scores voor *G-mean* en *F1*).

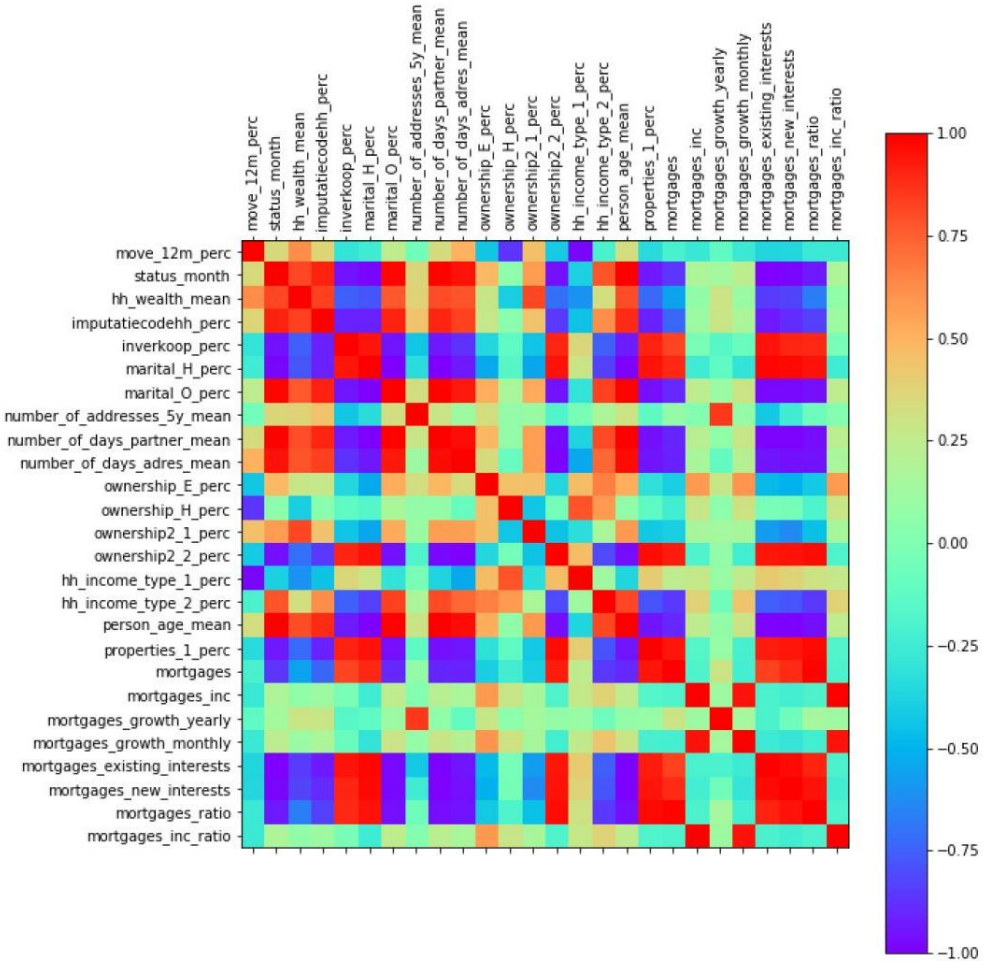
In het algemeen wordt de confusie-matrix en de optimale drempelwaarde, zoals verwacht, grotendeels bepaald door de gekozen kwaliteitsmaat. Elke metriek optimaliseert een ander probleem en resulteert daarom in een andere verhouding van vals-positieven en vals-negatieven. Zo heeft *G-mean* het laagste aantal vals-negatieven, F0.5 het laagste aantal vals-positieven, en balanceert *F1* de afweging tussen beide. Afhankelijk van de vraag of vals-negatieven of vals-positieven belangrijker zijn, moet uit deze drie een geschikte metriek worden gekozen. Uit alle gemaakte ijkgrafieken blijkt echter een tendens om het aantal positieven te onderschatten. Welke maat ook wordt gekozen, dit blijft een probleem bij het schatten van de positieve klasse, oftewel de verhuizers, om wie het nu juist net allemaal te doen is.

	Positive Prediction	Negative Prediction
Positive class	895 (TP)	35 (FN)
Negative class	22909 (FP)	16053 (TN)

Figuur 3 *– De confusie matrix voor de beste Random Forest classificator geoptimaliseerd voor G-mean en periode van 2 maanden; een voorbeeld van de tendens voor misclassificatie van de positieven (verhuizers).*

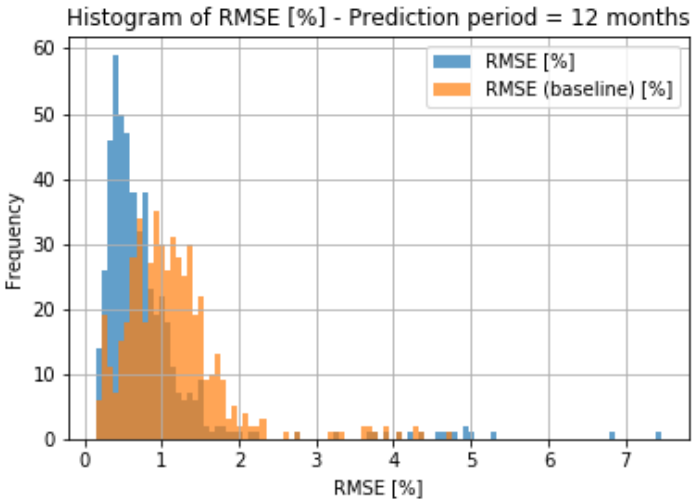
4.3 Regressie

Zoals eerder vermeld loont het bij regressie om eerst een gedetailleerd inzicht in de gegevens te krijgen. Hierbij zijn de Pearson correlaties tussen doel- en kenmerkvariabelen een goede eerste indicator van de potentiële voorspellingskwaliteit. Hoe hoger de absolute waarde van die correlaties, hoe beter.

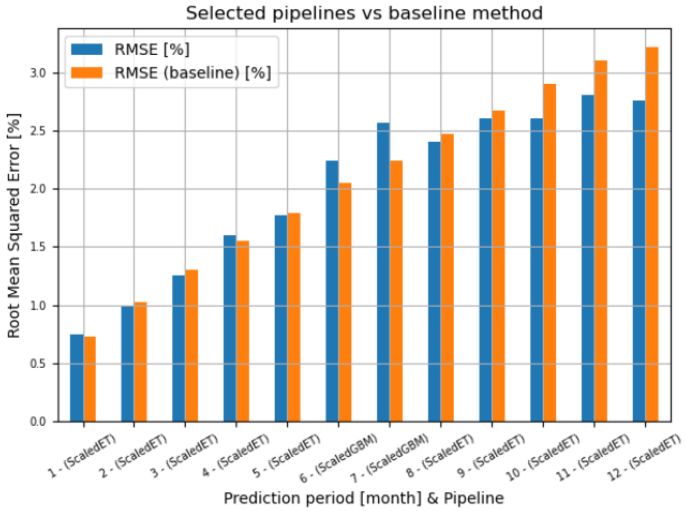


Figuur 4 *– Pearson correlaties van de trainingsset voor de volledige doelpopulatie.*

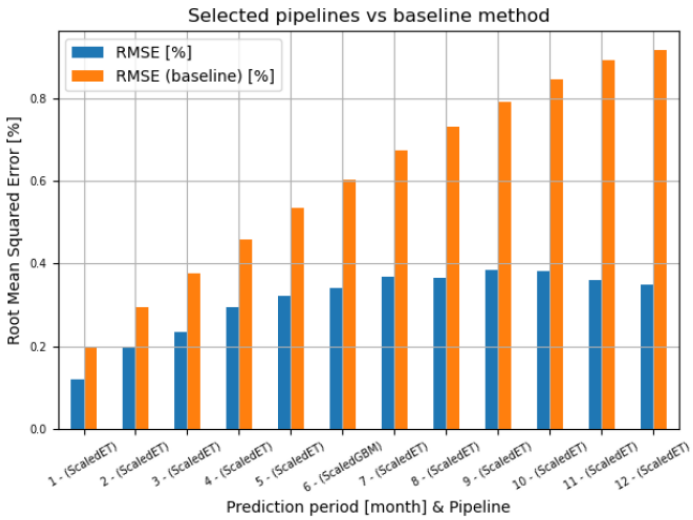
In totaal zijn er 6096 regressiemodellen gemaakt en beproefd, een voor elke subpopulatie en maand. Gedetailleerde uitkomsten hiervan kunnen worden geraadpleegd in het technische rapport (sectie 3.2 en bijlage H). De kwaliteit van de modelschattingen is over het algemeen goed (zie onderstaande figuur). In zijn algemeenheid gaat hier op; hoe kleiner de betreffende subpopulatie, des te lastiger het is om hun verhuisgedrag goed te schatten. En des te langer de voorspelperiode, hoe beter de regressiemodellen in relatieve zin presteren ten opzichte van de simpele basisprognose.



Figuur 5 *– Histogram van schattingsfout van regressiemodellen (blauw) versus basisschatting (oranje) voor alle beproefde subpopulaties bij een periode van 12 maanden. De fout van de regressiemodellen is doorgaans een stuk lager dan die van de basisschatting.*



Figuur 6 *– Subpopulatie met relatief grote schattingsfout: gemeente Ameland, N = 3573.*



***Figuur 7** – Subpopulatie met relatief kleine voorspelfout: getrouwd zonder kinderen, N = 3343129.*

5. Bevindingen

5.1 Bevindingen betreffende binaire classificatie

Een verzameling classificatie-methoden voor machinaal leren werd geëvalueerd op hun voorspellend vermogen of een individu al dan niet zou verhuizen. Na een eerste modeltest werden verschillende modellen gekozen en verder geoptimaliseerd in een willekeurige zoekactie. Het bleek dat voor de binaire classificatie, drie modellen globaal de beste prestaties hadden; *Random Forest Classifier, XGBClassifier, Support Vector Classifier*.

Hoewel ze het best presteerden in vergelijking met de andere *machine-learning classifiers*, waren de algemene prestaties op individuele schattingen nog steeds niet hoog. *G-mean* waarden piekten op 0,771, *F0.5* waarden op 0,511, en *F1* op 0,527 (op een schaal van 0 tot 1). Verschillende metrieken optimaliseren verschillende afwegingen, en hoewel *F0.5* minder vals-positieven heeft dan de andere, en *F1* zowel vals-positieven als vals-negatieven optimaliseert, zijn voor elk van deze maten het aantal vals-positieven en vals-negatieven hoog. In de technische documentatie worden bijbehorende confusiematrices en calibratieplots gegeven.

De resultaten voor de verschillende voorspellingsperioden ondersteunen deze bevindingen. Voor zowel *F0.5* als *F1* neemt de kwaliteit toe naarmate de periode toeneemt. Dit zou verklaard kunnen worden door een groter aantal positieven in de beschouwde periode; het model heeft meer positieve voorbeelden, en presteert daarom beter. Uit de leercurves volgt echter dat een toename van de steekproefgrootte, en dus van het aantal positieven, de prestaties van het model niet veel verder doet stijgen. Het toevoegen van meer kenmerken uit de beschikbare set, meer observaties, of het veranderen van de voorspellingsperiode lijkt dit niet te veranderen. Alle ijkgrafieken vertonen een algemene tendens om de positieve klasse te onderschatten en het aantal negatieven te overschatten. Dit wijst er op dat de voorspellingskwaliteit vooral voor de positieve klasse te wensen overlaat, en niet kan worden verbeterd met de huidige gegevens.

De resultaten van *under- & oversampling* lijken dit verder te bevestigen. Het *Random Forest* classificatiemodel werd geëvalueerd met verschillende bemonstermethoden, omdat die de beste

algemene prestaties combineerden met trainingsgemak en snelle trainingstijden. Alleen random oversampling presteerde iets beter dan het model zonder bemonstering.

In vergelijking met het vorige project zijn de *G-mean* waarden verbeterd van 0,70 - 0,71 tot 0,77. Bovendien heeft de confusie-matrix voor optimale *G-mean* een lager aantal vals-positieven en vals-negatieven. Evenzo was de *F1*-waarde voor het beste model 0,48 in het vorige project en 0,522 in dit project. Hoewel beide projecten niet rechtstreeks kunnen worden vergeleken wegens de verschillende doelen, datasetgroottes en positief/negatief-verhoudingen, is aannemelijk dat de kwaliteit van individuele schattingen laag was voor beide projecten en alle beschouwde datasets. Naar aanleiding hiervan is overgegaan op het modelleren van het verhuisgedrag op een geaggregeerd niveau.

5.2 Bevindingen betreffende regressie

Zoals verwacht blijken schattingen gebaseerd op geaggregeerde data kwalitatief een stuk beter aan te sluiten op de beoogde belangrijkste toepassing. Pearson correlatiecoëfficiënten bleken hierbij een nuttige indicator van het voorspel-potentiaal van de dataset.

Echter gegeven de beperkt beschikbare tijd kon het potentiaal van de regressie-methoden niet ten volle worden verkend. Naast enkele voorzichtige bevindingen, leverde deze exercitie met name aanknopingspunten op voor verdere verbeteringen.

5.3 Aanbevelingen

Adviezen voor toekomstige projecten en voortzetting van de verhuiskans-prognoses:

- **Algemeen**. Voor het verrichten van individuele schattingen in toekomstige projecten kan het zich lonen om tijdens de beginfase van het project het voorspellend potentieel van de beschikbare data expliciet te toetsen. Als deze laag blijkt kan tijdig worden overwogen om over te stappen naar een alternatieve aanpak, zoals schattingen op groepsniveau.

- **Verhuiskansen binaire classificatie**. De voorspelprestaties op individueel niveau worden onvoldoende nauwkeurig geacht voor massa-imputatie (in Woonbase of elders); het is niet raadzaam de micro-uitkomsten in directe vorm voor derden beschikbaar te stellen.

- **Verhuiskansen regressie**. Deze aanpak levert in principe betere schattingen voor verschillende tijdsperioden en subpopulaties, en leent zich voor snelle reguliere actualisatie (bijvoorbeeld elk kwartaal of elke maand). Een eerste opzet hiervoor is gemaakt, echter binnen de beschikbare tijd was het niet mogelijk dit concept tot in detail uit te werken. Ook vereist een dergelijke aanpak dat zaken als doelgroepen en kwaliteitsmaten van tevoren worden vastgelegd; de beoogde toepassing is bepalend voor de modelkeuze. Desgewenst kan een op de toepassing toegesneden schatting van verhuiskansen in een derde ontwikkeltraject alsnog volledig worden geïmplementeerd en geëvalueerd.