Niet-lineaire algoritmen meest geschikt om woningwaarde van Nederlandse buurten te voorspellen

Winnaar Capital Value Scriptieprijs, 2022

Jasper Kars

CAPITAL VALUE.

SCRIPTIEPRIJS CAPITAL VALUE

Inhoudsopgave

1

Introductie

De Nederlandse woningmarkt heeft de afgelopen jaren een snelle groei doorgemaakt. Datzelfde kan worden gezegd over het gebruik van algoritmen in woningmarktonderzoek.

De laatste jaren zijn er een breed scala aan Machine Learning (ML)-technieken ontwikkeld. Veel van het onderzoek naar deze technieken richt zich (uitsluitend) op het doen van voorspellingen en niet op, zoals in de econometrie gebruikelijker, statistische inferentie.

Met inferentie wordt bedoeld het generaliseren van onderzoeksuitkomsten op basis van een steekproef naar de totale populatie. Wanneer dit principe wordt toegepast op de woningmarkt, hebben we het over de analyse van een set transacties om daarmee de totale markt te verklaren.

Dit onderzoek probeert het gat tussen de benaderingen van voorspellen en generaliseren te dichten. In het huidige, complexe (data)landschap van de woningmarkt is het belangrijk dat er inzichten ontstaan over relevante variabelen en de effecten die deze (op elkaar) hebben.

Het doel van dit onderzoek is om het snijvlak van Machine Learning en de hedonische-prijsmethode (HPM)[1] te verkennen door het vinden van belangrijke variabelen die een rol spelen in het voorspellen van woningwaarde op buurtniveau. Daarbij worden de mogelijkheden van zowel statistische als ML methoden ingezet om de potentie van data science voor woningmarktonderzoek volop te benutten. Hierin staat de vraag centraal: hoe verhouden vier machine-learning-algoritmen (PCR, SVR, RF en k-NN) zich tot elkaar bij de implementatie van de hedonische-prijsmethode op Nederlandse woningprijzen (op buurtniveau)?

[1] De hedonische-prijsmethode (HPM) is een economische waarderingsmethode waarbij men uitgaat van het feit dat er vele factoren zijn (zoals milieu en omgeving) die de waarde van een marktgoed (zoals een woning) vormen.

2

Methode

Dit hoofdstuk geeft een beeld van de data die gebruikt zijn in dit onderzoek. In drie figuren met toelichting wordt inzichtelijk gemaakt hoe de dataset is samengesteld. De figuren laten daarmee de verhoudingen op de Nederlandse woningmarkt zien. De inzichten die daaruit volgen zijn van belang voor het uitvoeren van data-analyse. Data science onderzoek vraagt immers een goed begrip van de onderzochte data en de daarbij behorende context(en).

2.1 Dataset

De dataset die is onderzocht maakt gebruik van (open) data van het Centraal Bureau voor de Statistiek (CBS). De zogeheten Kerncijfers Wijken en Buurten bieden een overzicht van de statistische gegevens van gemeenten, wijken en buurten in Nederland.

Het betreft data op buurtniveau van de jaren 2016 en 2017. Deze jaren symboliseren het (definitieve) herstel van de woningmarkt na de financiële crisis (Boelhouwer 2017; DNB 2018). De gecombineerde dataset bevat 19,696 rijen (buurten) en 108 variabelen (kenmerken van buurten). 'Gemiddelde woningwaarde' vormt de afhankelijke variabele die is gebaseerd op de WOZ-waarde van een buurt. Belangrijke onafhankelijke variabelen betreffen onder meer inkomen en buurt- en omgevingskarakteristieken.

Achtergrond: Voorbewerking van de data

Om de dataset voor te bereiden op het trainen van de ML-modellen zijn een aantal voorbewerkingstappen gezet[2]:

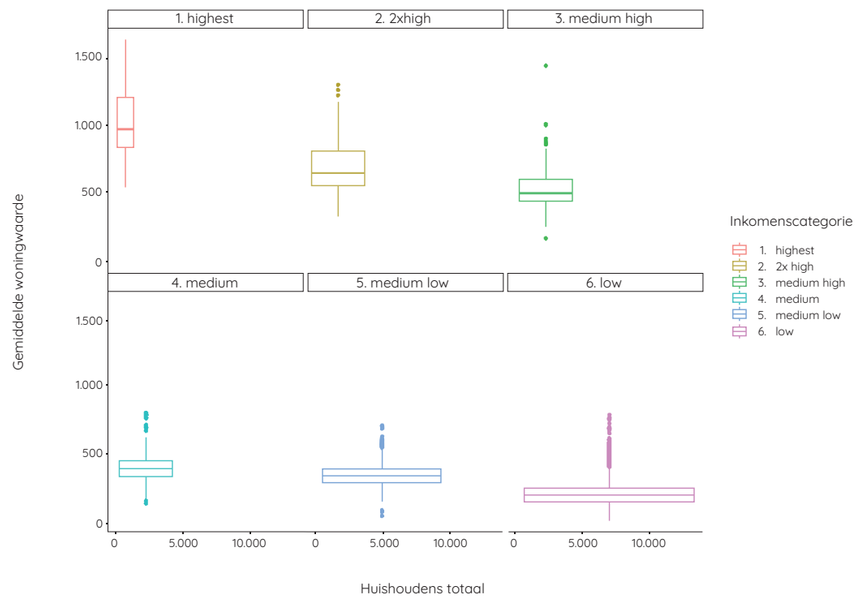
- Rijen die geen woningwaarde bevatten zijn verwijderd
- Missende waarden in variabelen die een afstand bevatten zijn verwijderd
- De dataset is opgedeeld in twee gedeelten: een trainingset (70% van de rijen) en een testset (30% van de data).
- Variabelen die meer dan 20% missende waarden bevatten zijn verwijderd (Enders 2003). Op variabelen die onder dit percentage bleven is imputatie[3] toegepast om de missende waarden in te vullen (Van Buuren 2012).

2.2 De dataset in beeld: drie weergaven van de Nederlandse woningmarkt

Figuur 1, 2 en 3 bieden inzicht in belangrijke verhoudingen op de Nederlandse woningmarkt en helpen om de data(set) inzichtelijk te maken. Juist dit begrip is van groot belang bij het trainen en vergelijken van ML-modellen, zoals gebeurt in hoofdstuk 3.

Allereerst is de verhouding tussen inkomensgroepen en WOZ-woningwaarde verkend. Dit is gedaan door voor zes afzonderlijke inkomensgroepen te kijken welke woningwaarden in deze groep voorkomen (Figuur 1). Daarbij worden de woningwaarden afgezet in de Y- as (staande as) en het aantal buurten in de X-as (liggende as). Zodoende valt direct een patroon op in de zes boxplots: de hoge inkomens zijn een relatief kleine groep, waarbinnen de spreiding van woningwaarden boven de 500.000 euro ligt. Circa 75% van deze buurten heeft een woning die meer waard is dan 750.000 euro. In de laagste inkomensgroep is het aantal buurten veel groter, en ligt de woningwaarde voor 75% van deze buurten onder de 300.000 euro. Daarnaast valt op dat de groep buurten met middeninkomens de kleinste onderlinge verschillen heeft wat betreft woningwaarde.

Figuur 1. Verdeling van huishoudens over en binnen inkomensklassen

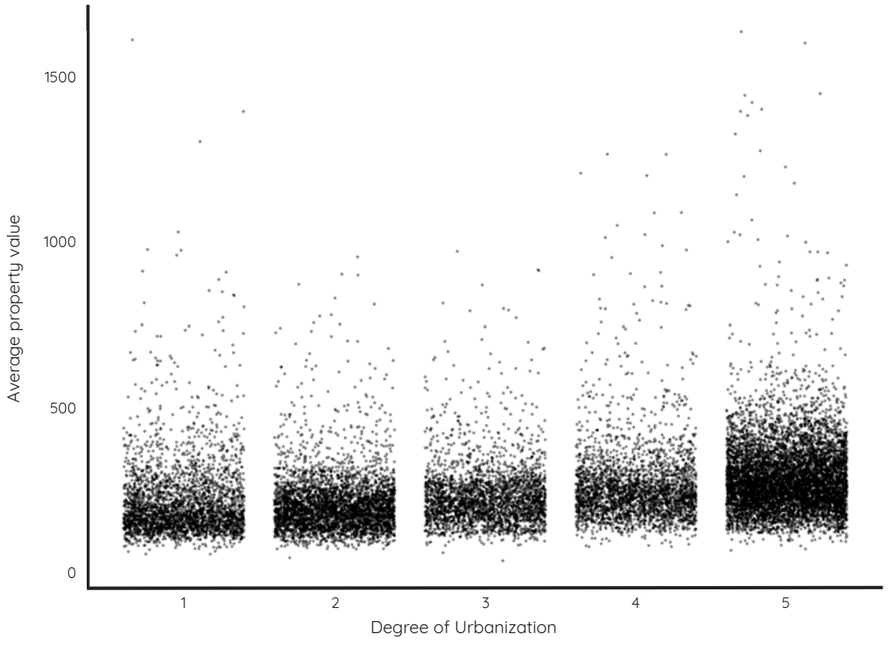


[2] Voor de geïnteresseerde lezer is de code te vinden op github: https://github.com/jasper-ck/-ML-and-HPM-in-The-Dutch-Neighborhood-Housing-Market.

[3] Imputeren is het inzetten van geobserveerde data om een (betrouwbare) schatting van de missende waarden te maken.

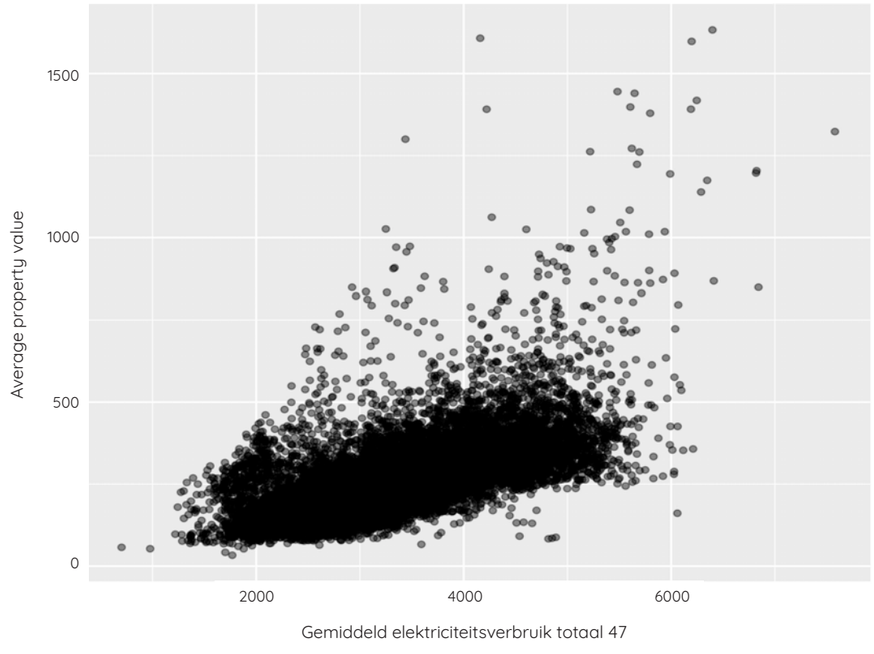
Als tweede factor is gekeken naar de verhouding tussen woningwaarde en urbanisatiegraad van de wijk of buurt waar de woning zich bevindt. In figuur 2 is te zien hoe de spreiding van de waarde van woningen per urbanisatiegraad verandert: de hoogst verstedelijkte wijken en buurten (categorieën 1 en 2) laten relatief weinig spreiding in woningwaarden zien, terwijl de spreiding in de minst verstedelijkte wijken en buurten (categorie 5) veel groter is. Daarnaast lijken woningwaarden in hoogstedelijke buurten relatief laag te liggen, tussen 150.000 euro en 300.000 euro. Dit kan mogelijk worden verklaard door de grote hoeveelheid sociale huurwoningen in deze buurten. De hogere woningwaarden in de minst verstedelijkte gebieden kan worden verklaard door het soort locaties die in deze groep zitten, zoals villawijken, buurten aan de randen van steden of plekken waar de woningwaarde hoog is doordat de woningen er groot zijn, zoals in rurale gemeenten.

Figuur 2. Spreiding van buurten naar mate van verstedelijking en gemiddelde woningwaarde



Als derde factor is gekeken naar de verhouding van elektriciteitsverbruik en de waarde van woningen. In figuur 3 is de gemiddelde woningwaarde (y-as) tegenover het gemiddeld elektriciteitsverbruik (in kWh) (x-as) in een buurt in beeld gebracht. Hier valt op dat de uitbijters – buurten met veel elektriciteitsverbruik – ook veelal de buurten zijn met de hoogste woningwaarde. In buurten met een lagere gemiddelde woningwaarde is het elektriciteitsverbruik lager.

Figuur 3. Elektriciteitsverbruik en woningwaarde van buurten



3

Bevindingen

Dit hoofdstuk gaat in op de stappen die in de analyse in de modellen zijn uitgevoerd en de resultaten van de getrainde modellen.

3.1 Data voorsorteren: de Principale Componentenanalyse
Principale componenten analyse (PCA) is een techniek die wordt ingezet op grote datasets om de dataset samen te vatten en de samenhang tussen gegevens inzichtelijker te maken. PCA reduceert het aantal factoren in het onderzoek door variabelen te groeperen in samenhangende clusters. Hieronder worden de verschillende principale componenten (PCs oftewel samenhangende clusters) uiteengezet en wordt bepaald hoe verschillende variabelen hebben bijgedragen aan het vormen van deze clusters.

PC1: Structurele buurtkenmerken
PC1 toont een matig negatieve correlatie met structurele buurtkenmerken zoals ‘woningvoorraad’ en ‘totaal aantal huishoudens’. Dit wijst erop dat de gemiddelde vastgoedprijs lijkt te dalen wanneer deze variabelen hoge waarden hebben. Daartegenover staat dat PC1 een zwakke positieve correlatie heeft met ‘afstand tot supermarkt’ en ‘afstand tot huisarts’. Deze principale component wordt daarom geïnterpreteerd als een variabele die verschillende structurele buurtkenmerken combineert.

PC2: Woningverhuur versus eigenwoningbezit
PC2 wijst op een contrast tussen huur- en koopwoningen. Er is een positieve correlatie waar te nemen met het ‘totaal aantal huurwoningen’ en een negatieve correlatie met het ‘totaal aantal koopwoningen’.

PC3: Inkomensniveau
PC3 toont een duidelijke positieve correlatie met variabelen gerelateerd aan inkomen.

PC4: Mate van dichtbevolktheid
PC4 bevat een duidelijke positieve correlatie met de variabele ‘percentage bewoond’ tegenover een negatieve correlatie van dezelfde sterkte met ‘percentage onbewoond’. Daarnaast tonen ‘totaal landoppervlakte’ en ‘totaaloppervlakte’ negatieve correlaties.

PC5: Ouderen versus niet-westers
Deze PC draait om positief gecorreleerde variabelen op het gebied van vergrijzing (bijv. 'totaal sterftecijfer'). Hoewel het ook een negatieve correlatie heeft met variabelen die het aantal mensen met een niet-westerse achtergrond weergeven. Daarom wordt PC5 geïnterpreteerd als een variabele die contrasteert met variabelen gerelateerd aan leeftijd en culturele achtergrond.
PC6: Bouwjaar
PC6 vertoont een duidelijk onderscheid tussen ’bouwjaar voor 2000’ en 'bouwjaar na 2000'. De eerste variabele (bouwjaar voor 2000) betreft een negatieve correlatie, terwijl de tweede (bouwjaar na 2000) een positieve correlatie bevat.

PC7: Arbeidsparticipatie
PC 7 bevat een positieve correlatie tussen woningwaarde en het percentage actieve mensen op de arbeidsmarkt (o.a. de variabele ‘percentage van het inkomen afkomstig uit werk of onderneming’). Daartegenover staat een negatieve correlatie tussen woningwaarde en de variabele ‘relatief sterftecijfer’. PC7 wordt daardoor gezien als een variabele die gaat over arbeidsparticipatie.

PC8: Buurtveiligheid
Vanwege een matig negatieve correlatie met alle variabelen die verband houden met misdaad, wordt PC8 geïnterpreteerd als draaiend om veiligheid. Dit is daarnaast gebaseerd op het feit dat er een positieve correlatie met 'totaaloppervlakte' en 'totale grond' is. Buurten met een groter oppervlakte zijn immers vaak minder dichtbevolkt en hebben lagere criminaliteitscijfers.

PC9: Diverse stadskenmerken
PC9 heeft een matig positieve correlatie met 'percentage bezet' en een vergelijkbare negatieve correlatie met 'percentage onbewoond'. Daarnaast is hier sprake van een positieve correlatie met verschillende misdaad gerelateerde variabelen, net als 'eigendom' door een wooncorporatie'. Aangezien dit variabelen zijn die bij stedelijk leven passen wordt PC9 geïnterpreteerd als een variabele die informatie over verschillende stadskenmerken bevat.

3.2 Performance van de modellen

Deze paragraaf gaat nader in op de prestaties van de verschillende (getrainde) modellen. Een eenvoudig regressiemodel fungeert als baseline om de meer geavanceerde modellen mee te vergelijken.

PCR (Principal Component Regression)

PCR kan worden gezien als een (standaard) lineair regressiemodel met daarin door middel van PCA gegenereerde variabelen. Het PCR-model presteert het beste wanneer alle 9 PCs (nieuwe variabelen) worden ingezet. Toch dragen PC8 en PC5 weinig bij aan het best presterende model. Daartegenover staat dat PC3, PC2 en PC4 een grote bijdrage leveren aan het model. Dit suggereert dat het PCR-model vooral leunt op volgende attributen (ofwel PCs): inkomensniveau, woningverhuur versus eigenwoningbezit en mate van dichtbevolktheid. Een model waarin de minst relevante attributen werden weggelaten bracht vergelijkbare resultaten voort, echter onvoldoende om het model met alle 9 PCs te overtreffen.

In vergelijking met de andere drie niet-lineaire modellen wordt duidelijk dat het baseline-PCR-model minder presteert dan Random Forest (SF), Support Vector Regression (SVR) en k-Nearest Neighbours (k-NN). Dit kan gedeeltelijk worden verklaard doordat het voorspellen van woningwaarde (veelal) een niet-lineair verband betreft. Het gaat hier immers vaak om modellen met een grote complexiteit die veel heterogene variabelen bevatten. PCR is echter een lineair regressiemodel op basis van PCA, dit maakt het minder geschikt voor niet-lineaire voorspellingen, blijkt ook uit het onderzoek van bijvoorbeeld Ceh et al. 2018.

K-nn (k nearest neighbor)

K-nn is een algoritme dat de mogelijkheid biedt om datasets in groepen te verdelen op grond van de dichtstbij gelegen datapunten. Vandaar dat K-nn ook staat voor Nearest Neighbor: dichtstbijzijnde buur. Bij een regressiemodel – zoals ook in dit onderzoek – is eigenschapswaarde het gemiddelde van de waarden van k naaste buren. K-nn is goed in staat om non-lineaire taken uit te voeren. De waarde van k hangt voor een groot gedeelte af de omvang van de training set (16.636 rijen). Het model is meermaals getraind met verschillende waarden voor k. De hoogste performance werd bereikt met k = 4. K staat hier voor een parameter die verwijst naar het aantal naaste buren dat moet worden opgenomen in de afweging. Nadat k = 50 neemt de performance af en komt het in de buurt van het PCR-model.

Support Vector Regression (SVR)

Het algoritme SVR maakt een scheiding tussen twee klassen door een hypervlak[4] te genereren waarbij de datapunten zo ver mogelijk van deze lijn of dit vlak verwijderd zijn. De datapunten die zich het dichtstbij bevinden zijn uiteindelijk bepalend voor de positie van het hypervlak. SVR bleek een robuust niet-lineair algoritme te zijn als het gaat om het voorspellen van woningwaarde in Nederlandse buurten. Tijdens het trainen van verschillende SVR-modellen werd duidelijk dat SVR-modellen waarin non-lineaire regressie als basis fungeerde, beter presteerde dan zijn lineaire tegenhangers. Dit onderzoek bevestigt ook de gevoeligheid van SVR wat betreft uitbijters: buurten in gemeenten met een zeer hoge woningwaarde, vooral in het westen van Nederland, worden vaak verkeerd voorspeld. Omgekeerd doet het algoritme het heel goed als het gaat om woningwaarden die rond de mediaan vallen (225.000 euro).

Random Forest (RF)

Een supervised learning algoritme dat vaak wordt toegepast op het gebied van ML woningmarktonderzoek is RF. Dit model creëert meerdere beslissingsbomen om voorspellingen te combineren. Bij een regressiemodel wordt de voorspelling van het ensemble vastgesteld door het gemiddelde te nemen van de resultaten van de individuele bomen (Ceh et al. 2018). Een boom wordt in deze benadering gevormd door het steeds verder vertakken van bij elkaar horen sets waarnemingen, die onderling steeds minder van elkaar verschillen. Ook het RF-model in dit onderzoek bleek betrouwbaar in het uitvoeren van niet-lineaire voorspellingen. Een mogelijke verklaring voor RF's hoge performance kan het vermogen van dit algoritme zijn om hoge variantie te verminderen door het gemiddelde te nemen van een groot aantal bomen op de trainingsset. Naast het feit dat RF een goede voorspellende kwaliteit heeft bij niet-lineaire problemen, biedt ook het open karakter van dit algoritme voordelen. Zo valt goed te zien dat ook in dit model - net als bij PCR - PC3 en PC2 de belangrijkste variabelen zijn. Opmerkelijk genoeg zijn PC1 en PC6 (structurele buurtkenmerken en) echter veel belangrijker dan in het best presterende PCR-model.

Concluderend overzicht

Over het geheel genomen blijkt de Random Forest methode de beste voorspellingen te leveren. Dit blijkt uit de analyse van de verklaarde variantie (R^2) en de onderliggende gemiddelde kwadratische fout (RMSE).[5] In tabel 1 worden deze parameters voor elk getraind model weergegeven, waarbij gezocht wordt naar een hoge verklaarde variantie (R^2) en een lage kwadratische fout (RMSE).

Tabel 1: Resultaten van de getrainde ML-modellen na het toepassen van PCA. In andere woorden, hoe goed is het model in staat om gemiddelde woningwaarde te voorspellen? Een hoge R^2 en een lage RMSE vormen de beste combinatie.

Model	R^2	RMSE
PCR	0.79	51.98
SVR	0.84	45.44
k-NN	0.84	44.55
RF	0.87	41.45

In data science is het gebruikelijk om de getrainde modellen te vergelijken met een (zo min mogelijk bewerkt) baseline model. Bij het vergelijken van het baseline model met de vier ML-algoritmen valt op dat de beste voorspeller in het baseline model (sociaaleconomische variabelen) beter presteert dan PRC en in de buurt komt van complexere modellen SVR en k-NN (tabel 2). Dit bewijst het belang van inkomensgerelateerde variabelen bij het voorspellen van woningwaarde (op buurtniveau). Het succes van lineaire regressie met sociaaleconomische variabelen kan deels worden verklaard door het lineaire karakter van deze taak. In tegenstelling tot de modellen die PCA-variabelen bevatten, oogt de relatie tussen woningwaarde en de socio-economische variabelen (meer) lineair. Dit suggereert dat het inkomen stijgt met hogere woningwaarden of andersom.

Tabel 2: Resultaten van de baseline regressie modellen met originele variabelen uit de 'Wijken en Buurten' dataset van CBS, opnieuw met de R^2 als indicator voor de sterkte van het verband tussen de genoemde factoren en de woningwaarde.

Type variabelen	R^2
1.1 variabelen met uiteenlopende populatie karakteristieken	0.209
4 variabelen met informatie over energieverbruik	0.474
7 variabelen met informatie over socio-economische status	0.831
10 variabelen met informatie over omgevingskarakteristieken	0.291

[4] Een hypervlak is een scheidingslijn tussen twee gegevensklassen in een hogere dimensie dan de werkelijke dimensie. In SVR wordt het gedefinieerd als de lijn die helpt bij het voorspellen van de afhankelijke variabele (gemiddelde woningwaarde).

[5] RMSE is de standaardafwijking van de voorspellingsfouten (het verschil tussen de daadwerkelijke observaties en voorspelling). RMSE geeft aan hoe groot een (voorspellings)fout gemiddeld is. Dit wordt gedaan door de voorspellingsfouten die gemaakt zijn in de testset te kwadrateren en daarna te middelen (hoe lager de RMSE hoe minder voorspellingsfouten).

Stabiel-belangrijke variabelen selecteren met MINREM

Om een brug te slaan tussen de data science methodiek waarbij het veelal draait om modellen die zo goed mogelijk iets kunnen voorspelen en de statistische methode waarbij wordt gekeken naar variabelen die statistisch significant zijn,[6] is voor dit onderzoek ook gekeken naar een recent ontwikkelde methode die beide werelden bij elkaar probeert te brengen: de MINREM-methode. Deze methode stelt statistisch significante variabelen centraal en draait om de kwaliteit van variabelen en niet om maar zoveel mogelijk (soms niet statisch significante) variabelen in een model te stoppen.

De MINREM-methode, die voor het eerst werd geïntroduceerd door Pérez-Rave et al. in 2019, werd in dit onderzoek ingezet om belangrijke stabiel-belangrijke variabelen te vinden (die statisch significant zijn), waarna met deze variabelen vervolgens de vier algoritmen zijn getraind. Uiteindelijk werden er vijf variabelen aangemerkt. Het betreft de volgende vijf variabelen: 'mensen van 45 tot 65 jaar', 'gemiddeld elektriciteitsverbruik', 'percentage mensen met laag inkomen', 'percentage huishoudens met laag inkomen' en 'percentage huishoudens met hoog inkomen'. Er is dus een aanzienlijk verschil tussen de meer dan 30 variabelen die statistisch significant waren in de baseline modellen en de vijf stabiel-belangrijke variabelen onder MINREM. Tabel 3 toont aan dat de vier modellen met (tabel 3) en zonder (tabel 1) de stabiel-belangrijke variabelen slechts een klein verschil in performance laten zien.

Daarnaast kan na de inzet van de variabele selectieprocedure MINREM worden gesteld dat buurten met minder inwoners van 45-65 jaar, hoog elektriciteitsverbruik en veel huishoudens binnen het hoogste kwintiel van de inkomensverdeling (top 20%), een grote(re) kans hebben op een hoge woningwaarde. Het toont ook het cruciale effect van inkomen aan bij het bepalen van woningwaarde, zoals in veel eerdere studies al is aangetoond (Boelhouwer 2000).

Tabel 3: Resultaten van de getrainde ML-modellen na het toepassen van een speciale variabele selectieprocedure MINREM (met alleen stabiel-belangrijke en statisch significante variabelen in de modellen opgenomen). Opvallend is het beperkte verschil met tabel 1, wanneer gekeken wordt naar de hoogte van R^2 en de waarde van RMSE.

Model	R^2	RMSE
PCR	0.71	60.92
SVR	0.78	53.69
k-NN	0.77	54.36
RF	0.79	51.05

[6] Als een resultaat statistisch significant is, betekent dit dat het onwaarschijnlijk is dat het onderzochte fenomeen slechts door toeval of willekeurige factoren kan worden verklaard.

4

Beleidssuggesties

Dit hoofdstuk bevat beleidssuggesties die op grond van de uitkomsten van de verschillende ML-modellen zijn opgesteld.

Bouw meer (betaalbare) woningen voor middeninkomensgroepen
Uit de dataset en de modellen blijkt een behoefte aan meer (betaalbare) woningen voor middeninkomens. Ter illustratie, de verhouding tussen woningvoorraad en aantal inwoners is bij deze groep lager dan in sommige hogere inkomensgroepen. Veel middeninkomensgroepen verdienen te veel voor sociale huur maar komen (vaak) niet in aanmerking voor een hypotheek (PBL, 2017). Het is daarom aan te raden om te kijken naar inkomensgroepen als een belangrijk uitgangspunt bij het plannen van bouwprojecten, waarbij middeninkomensgroepen speciale aandacht verdienen. Dit wordt ondersteund door het belang van inkomensvariabelen in MINREM en alle vier de ML-modellen.

Stimuleer een regionale aanpak van woningmarktbeleid
Er zijn grote regionale verschillen tussen vastgoedwaarden en woningvoorraad in Nederland. Dit vraagt van onder meer beleidsmakers dat zij de woningmarkt ook vanuit regionaal perspectief bekijken. 'Inkomensniveau' – een door middel van PCA gegeneerd kenmerk dat inkomen bevat - geeft een duidelijk verschil in vermogensverdeling aan tussen de Randstad en de rest van Nederland. Verder is er een verschil in woningwaarde tussen de Randstad (vaak hoger) en overig Nederland zichtbaar. Hierop vormen een aantal armere stadsbuurten in bijvoorbeeld Rotterdam of Den Haag een uitzondering. Vanwege deze regionale verschillen is een aanbeveling om meer maatwerk met betrekking tot het bouwen van woningen (in verschillende regio's). Zo zou, net als Boelhouwer (2020) betoogt, een (beleids- of bouw)verdeling tussen verschillende gebiedstypen een geschikte oplossing kunnen vormen. Denk daarbij bijvoorbeeld aan een gebied met duurdere buurten in grotere steden, een met goedkopere buurten in grotere steden, een met duurdere buurten buiten de kernregio's en een met goedkopere buurten buiten de kernregio's.

Besteed speciale aandacht aan voor jongere leeftijdsgroepen
Aangezien leeftijd een cruciale factor is in het domein van de woningmarkt, kan beleid dat specifiek op leeftijdsgroepen is gericht nuttig zijn. Dit blijkt ook uit de modellen in dit onderzoek. Zo bevestigen de variabele selectieprocedure en de ML-modellen dat leeftijd een belangrijk kenmerk is dat bijdraagt aan het voorspellen van woningwaarde. Daarbij ervaren veelal jongere mensen op dit moment problemen met het vinden van woonruimte in steden vanwege hoge huur- en koopprijzen. Deze leeftijdscategorie en de (jongere) gezinnen vormen de cohorten met de (relatief) hoogste woonlasten (PBL 2017). Dit kan worden verbeterd door bijvoorbeeld woningcorporaties of grote verhuurders een vast percentage aan jongeren te laten verhuren, dit wordt ook voorgesteld door bijvoorbeeld Hochstenbach (2017). Uiteindelijk zou dit starters toegang kunnen bieden tot de markt en creëert het meer mogelijkheden voor specifieke leeftijdsgroepen.

[7] Gemeenten als Amsterdam hebben de afgelopen jaren geëxperimenteerd met soortgelijk beleid.

5

Conclusie

Machine-Learning (ML)-technieken zijn zeer geschikt om de woningwaarde van Nederlandse buurten te voorspellen. Vooral de ML-modellen die niet-lineaire situaties aankunnen presteren goed. Dit geldt zeker voor Support Vector Regression, k-Nearest Neighbour en in het bijzonder Random Forest. Dit algoritme is een zeer betrouwbare regressiemethode om woningwaarde te voorspellen. Uit de ML-modellen blijkt bovendien welke variabelen van belang zijn als het gaat om het voorspellen van woningwaarde (en bij het opstellen van woningmarktbeleid). Uit alle modellen – dus zowel de niet-lineaire als de lineaire modellen en de modellen getraind met de stabiel-belangrijke variabelen[8] – blijkt de relatie tussen woningwaarde en inkomens een van de sterkste te zijn, gevolgd door diverse omgevings- en sociaaleconomische variabelen.

Verschillende ML-technieken (bijvoorbeeld Random Forest en Principal Component Regression) blijken daarnaast waardevol als het gaat om het achterhalen van de belangrijkste variabelen voor het voorspellen van woningwaarde. Hierbij bleken structurele en socio-economische (buurt)kenmerken van groot belang. De recent ontwikkelde variabelen selectie techniek MINREM – waarin het draait om het voorspellen van stabiel-belangrijke variabelen die vanuit statisch perspectief ook significant zijn – blijkt daarnaast zeer waardevol voor het minimaliseren van het aantal belangrijke variabelen (uiteindelijk blijven inkomen, energieverbruik en leeftijd over) in de modellen en voor het combineren van statistiek met ML. Dit maakt het trainen van (woningmarkt) modellen minder ingewikkeld en minder verwerkingsintensief.

Toekomstig data science-onderzoek op de Nederlandse woningmarkt zou kunnen putten uit een soortgelijke gecombineerde (methodologische) aanpak. Daarnaast liggen er ook op het gebied van deep learning grote mogelijkheden voor woningmarktonderzoek, juist ook omdat deze modellen geschikt zijn voor complexere (niet)lineaire verbanden (zoals het voorspellen van woningwaarde).[9] Daarbij zou een onderzoek over een langere tijdsperiode (met nog meer data) nieuwe waardevolle inzichten kunnen verschaffen over de aanpak van problemen op de woningmarkt: een zeer prangend beleidsvraagstuk in Nederland. Een laatste aanbeveling voor vervolgonderzoek richt zich op de aard van de te voorspellen waarde: dit onderzoek richt zich op het voorspellen van WOZ-waarde, maar om nog beter inzicht te krijgen in de dynamiek van de woningmarkt zou een data science analyse van (daadwerkelijke) marktwaarde een waardevolle toevoeging zijn.

[8] Dit betreft op basis van de MINREM-methode dus allemaal statistisch significante variabelen.

[9] Deep learning maakt het voor computers mogelijk om (via neurale netwerken) nieuwe dingen te leren van grote hoeveelheden data.

6

Bronvermelding

Boelhouwer, P. 2000. Development of house prices in the Netherlands: an international perspective. Journal of Housing and the Built Environment, 15(1):11–28.

Boelhouwer, P. 2017. The role of government and financial institutions during a housing market crisis: a case study of the Netherlands. International Journal of Housing Policy.

Ceh, M., M. Kilibarda, A. Lisec, and B. Bajat. 2018. Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5), 168.

DNB. 2018. The housing market in major Dutch cities. De Nederlandsche Bank, https://www.dnb.nl/media/ykmhc2el/201705_nr_1_-2017-_the_housing_market_in_major_dutch_cities.pdf.

Hochstenbach, C. 2017. State-led gentrification and the changing geography of market-oriented housing policies. Housing, Theory and Society.

Pérez-Rave, J.I, J.C. Correa-Morales, and F. González-Echavarría. 2019. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. Journal of Property Research, 36:1:59–96.

PBL. 2017. Middeninkomens op de woningmarkt: Ruimte op een krap speelveld. Planbureau voor de Leefomgeving, https://www.pbl.nl/sites/default/files/downloads/pbl-2017-middeninkomens-op-de-woningmarkt-2602.pdf.

Capital Value Scriptieprijs
Capital Value is specialist en marktleider op de woningbeleggingsmarkt. Wij vinden het belangrijk om op te hoogte te blijven van wat er speelt en investeren veel in onderzoek. Omdat wij graag kansen bieden aan jong talent hebben wij de Capital Value Scriptieprijs in het leven geroepen.

De Capital Value Scriptieprijs wordt jaarlijks uitgereikt aan een scriptie vanuit een hbo of universitaire studie over de woning(beleggings)markt in Nederland. De winnaar ontvangt twee tickets naar New York.

Meer informatie
Voor meer informatie over de Capital Value Scriptieprijs en de voorwaarden om deel te nemen verwijzen wij graag naar capitalvalue.nl/scriptieprijs of kunt u contact opnemen met Thijs Konijnendijk, Head of Research & Data Intelligence.

Deze publicatie is een samenvatting van een onderzoek van Jasper Kars naar algoritmen die het meest geschikt zijn om de woningwaarde van Nederlandse buurten te voorspellen. Jasper deed dit onderzoek in het kader van zijn masteropleiding Data Science and Society aan de Tilburg University. Jasper won met zijn scriptie de Capital Value Scriptieprijs 2022. De volledige scriptie van is op aanvraag beschikbaar.

T.J. (Thijs) Konijnendijk MSc
Head of Research & Data Intelligence
030 72 71 700
t.konijnendijk@capitalvalue.nl

Uitreiking Capital Value Scriptieprijs met Jasper Kars.

CAPITAL
VALUE.

Maliebaan 85 • 3581 CG Utrecht • +31 (0)30 72 71 700 • info@capitalvalue.nl • capitalvalue.nl