Het maken van officiële statistieken wordt gecompliceerd door ontbrekende gegevens. Eén van de manieren om hier mee om te gaan is door te imputeren. Dit betekent dat schattingen worden gemaakt van ontbrekende waarden. Verschillende imputatiemethoden zijn beschikbaar zoals: nearest neighbour, ratio- en regressie-imputatie. Ook binnen *machine learning* zijn er technieken voorhanden om ontbrekende gegevens te schatten. Deze worden echter (nog?) niet wijdverbreid toegepast in de officiële statistiek. Dit artikel beschrijft een proof of concept van een nieuwe imputatiemethode voor de huishoudensstatistiek. De methode is gebaseerd op Extreme Gradient Boosting, een techniek uit machine learning.

IMPUTEREN VAN HUISHOUDSAMENSTELLINGEN MET MACHINE LEARNING

Jacco Daalmans

Jaarlijks publiceert de huishoudensstatistiek over het aantal huishoudens, uitgesplitst naar samenstelling (éénpersoonshuishoudens, paren met en zonder kinderen, etc.). De statistiek wordt afgeleid uit de huishoudensbus: een longitudinaal bestand van alle huishoudens op alle adressen. De huishoudensbus wordt gevoed vanuit verschillende bronnen. Een belangrijke is de Basisregistratie Personen (BRP). Voor een deel van de adressen levert deze bron geen eenduidige resultaten op. Voor die adressen wordt geprobeerd om de huishoudens deterministisch af te leiden. Als bijvoorbeeld twee volwassenen tegelijk naar een adres verhuizen gaat het CBS ervan uit dat het gaat om één huishouden. Een relatief klein aantal adressen kan ook niet deterministisch worden afgeleid. Die overgebleven groep, ruim 900.000 huishoudens, wordt modelmatig geïmputeerd. Omdat dit voor alle onbekende adressen gebeurt spreekt men van massa-imputatie; zie het kader verderop voor de valkuilen van die techniek. Momenteel worden ontbrekende gegevens geschat met een regressiemethode. Hierin wordt de relatie gelegd tussen huishoudsamenstellingen en hulpvariabelen, zoals leeftijdsverschil, wel of niet hetzelfde geslacht van de bewoners en stedelijkheid van de gemeente. Goede resultaten van *machine learning* methoden voor vergelijkbare classificatieproblemen geven aanleiding om te onderzoeken of Extreme Gradient Boosting ook kan worden aangewend voor de imputatie van huishoudsamenstellingen. Als eerste is er een *proof of concept* gemaakt, waarin gefocust is op een vereenvoudigd probleem: de classificatie van adressen met twee volwassen bewoners in: 1. twee (eenpersoons)huishoudens en 2. één (tweepersoons) huishouden.

Gradient Boosting

Classificatie met Gradient Boosting betekent dat een onbekende doelvariabele wordt geschat op individueel niveau. De schattingen worden gebaseerd op achtergrondkenmerken, die idealiter sterk samenhangen met de doelvariabele. Eerst wordt een model geschat met data waarvoor doel- en hulpvariabelen bekend zijn, vervolgens wordt dat model toegepast op eenheden waarvoor alleen hulpvariabelen beschikbaar zijn. Voor classificatieproblemen worden de kansen geschat op iedere categorie. Hier wordt dus voor ieder adres kansen geschat op 'twee huishoudens' versus 'één huishouden'. Op basis van deze geschatte kansen wordt er vervolgens geïmputeerd.

Gradient Boosting is een zogenaamde ensemble-techniek. Dit betekent dat de methode verschillende schattingen voor één eenheid combineert. Binnen de klasse van ensemblemethoden bestaat een onderscheid tussen bagging en boosting. Bij bagging worden verschillende schattingen onafhankelijk gemaakt. De uiteindelijke schatting wordt verkregen door het combineren van de afzonderlijke schattingen; bijvoorbeeld door te middelen. Bij boosting zijn de opeenvolgende schattingen afhankelijk. Iedere schatting probeert het resultaat van de vorige schatting (verder) te verbeteren. Zoals de naam al aangeeft, behoort Gradient Boosting tot de boosting methoden.

Een formelere manier om de boosting methode te beschrijven is als volgt. Stel dat we geïnteresseerd zijn in een doelvariabele $\mathbf{y}$. In dit geval staat y_i voor de kans op twee huishoudens op adres *i*. De kans op één huishouden is gelijk aan één minus deze kans. Om $\mathbf{y}$ te schatten hebben we hulpinformatie tot onze beschikking, die wordt aangeduid met $\mathbf{X}$. In eerste instantie wordt de best mogelijke schatting voor $\mathbf{y}$ bepaald, zeg $f_1(\mathbf{X})$. Dit is de schatting die optimaal is volgens een wiskundig criterium. Deze eerste schatting noteren we met $\hat{\mathbf{y}}_1 = f_1(\mathbf{X})$. De schatting $f_1(\boldsymbol{X})$ kan op verschillende manieren worden gemaakt. Gradient Boosting gebruikt beslisbomen, maar het zou bijvoorbeeld ook met regressie kunnen. Waar veel andere methoden hier ophouden, voert Gradient Boosting vervolgstappen uit om de eerste schatting $\hat{\mathbf{y}}_1$ te verbeteren. Iedere stap is erop gericht om de schattingsfout na afloop van de voorgaande stap, het zogenaamde residu, zo goed mogelijk te voorspellen. De gedachte is dat als het lukt om de fouten te schatten dat men daarvoor ook kan corrigeren. In de tweede stap wordt dus geprobeerd om de residuen, $\mathbf{r}_1 = \mathbf{y} - \hat{\mathbf{y}}_1$ te schatten. De uitkomst van stap 2 zijn schattingen $f_2(\mathbf{X})$ voor $\mathbf{r}_1$. De schatting voor de doelvariabele $\mathbf{y}$ na stap 2, is de som van de schatting $f_1(\mathbf{X})$ na stap 1 en de geschatte correctie $f_2(\mathbf{X})$ We krijgen dus dat $\hat{\mathbf{y}}_2 = f_1(\mathbf{X}) + f_2(\mathbf{X})$. Vervolgens wordt in stap 3 een schatting $f_3(\mathbf{X})$ gemaakt van het residu na stap 2. Op deze manier volgt dat de schatting na stap **J** te schrijven is al een som $\sum_{j=1}^{J} f_j(\mathbf{X})$. In de praktijk wordt vaak een gewogen som toegepast, maar daar gaan we omwille van de eenvoud niet verder op in.

Zoals eerder opgemerkt, worden de schattingen $f_j(\mathbf{X})$ gemaakt met beslisbomen. Bomen bestaan uit knopen en uit vertakkingen (zie figuur 1). In binaire beslisbomen wordt de dataset in iedere knoop verdeeld in twee delen, die beide zo homogeen mogelijk zijn met betrekking tot de te schatten doelvariabele. Er wordt een stopcriterium gehanteerd dat bepaalt wanneer er wordt gestopt met het verder vertakken van de boom (oftewel: opsplitsen van de data). Wanneer men te ver doorgaat met vertakken loopt men het risico op overfitting. Dit betekent dat de boom goede schattingen geeft voor de data die zijn gebruikt om de boom af te leiden (de zogenaamde trainingset), maar minder goed werkt op een onafhankelijke dataset, waarop men het model wil toepassen (de testset). Om overfitting te voorkomen wordt het aantal takken van de boom beperkt. Het algoritme probeert dus om zo goed mogelijke schattingen te maken met zo eenvoudig mogelijke beslisbomen. Bij veel toepassingen van Gradient Boosting wordt een groot aantal, soms wel honderden, relatief eenvoudige beslisbomen gecombineerd. Een

Zoals eerder opgemerkt is het beschreven advies een toepassing van massa-imputatie: het grootschalig schatten van ontbrekende gegevens op microniveau. De bedoeling is om die schattingen te gebruiken voor verschillende doeleinden. Deze werkwijze is niet onomstreden.

Idealiter houdt men bij het schatten van ontbrekende gegevens rekening met het doel waarvoor men de imputaties wil gebruiken. Stel dat iemand geïnteresseerd is in het verband tussen het aantal adressen met twee eenpersoonshuishoudens en de stedelijkheid van de gemeente. Als men dan bij het imputeren geen rekening houdt met deze relatie dan kan het zo zijn dat de geïmputeerde data een onjuist beeld geven over de samenhang tussen beide variabelen. Een berucht voorbeeld is het zogenaamde hondenbrokkenprobleem. Stel dat we een data set hebben met daarin de uitgaven aan hondenvoer. Alle ontbrekende waarden worden geïmputeerd. Het gegeven of iemand wel of geen hond heeft wordt echter niet meegenomen bij het maken van de schattingen. Deze informatie is niet beschikbaar of wordt niet relevant beschouwd. Een onderzoeker koppelt vervolgens de geïmputeerde data aan een tweede data set, waarin het hebben van een hond (wel) is opgenomen. Omdat bij het imputeren van uitgaven aan hondenbrokken geen rekening is gehouden met het feit of iemand een hond heeft, kunnen er vele hondenbezitters worden gevonden die geen geld aan hondenbrokken uitgeven, omgekeerd kunnen er veel niet-hondenbezitters zijn die toch geregeld hondenbrokken kopen. Dit kan dus leiden tot een verkeerde conclusie over de relatie tussen twee variabelen.

Een beter alternatief voor een generieke vorm van massa-imputatie is dat alle gebruikers zelf imputaties afleiden, die specifiek bedoeld zijn voor het doel waarvoor zij de data nodig hebben. In de praktijk kan dit echter lastig zijn omdat de gebruikers niet altijd alle data hebben. Bovendien betekent het veel werk voor de gebruiker en kan het inconsistente uitkomsten opleveren als verschillende gebruikers verschillende imputatiemethoden toepassen. Vanwege de bovenstaande argumenten en vanwege het relatief lage aantal imputaties voor huishoudens, is ervoor gekozen om, ondanks de bezwaren toch massa imputatie toe te passen. Bovenstaande betekent echter niet dat massa imputatie in het algemeen is aan te bevelen voor iedere statistiek.

Figuur 1. Een fictief voorbeeld van een beslisboom

eindknoop van een boom, een zogenaamd blad, hoort bij een specifieke, homogene groep. Bijvoorbeeld alle duo's op één adres van hetzelfde geslacht met minder dan 15 jaar leeftijdsverschil. Voor ieder blad wordt er ofwel een waarde van de doelvariabele geschat (hier: de kans op twee huishoudens op één adres), of een correctie ten opzichte van een eerdere schatting.

Als resultaat van Gradient Boosting krijgen we voor ieder adres een geschatte kans voor één versus twee huishoudens. Deze kansen kunnen we gebruiken voor het afleiden van imputaties. Dit is gedaan door het trekken van random getallen tussen nul en één. Stel dat we voor een bepaald adres afleiden dat er 60% kans is op één huishouden en 40% op twee huishoudens. We trekken dan vervolgens uit een uniforme verdeling op het interval [0,1]. Als de uitkomst kleiner of gelijk is aan 0,6 dan imputeren we 'één huishouden', is de uitkomst groter dan imputeren we 'twee huishoudens'. Deze stochastische methode om imputaties af te leiden wijkt af van de gangbare methode bij machine learning die kansen afrondt. Als er een kans van 0,6 is geschat op één huishouden, dan wordt die afgerond naar 1 en wordt er dus 'één huishouden' geïmputeerd. Hoewel deze benadering op individueel niveau de kleinste schattingsfout oplevert, heeft deze als effect dat de verdeling van de doelvariabele sterk af kan wijken van die van de geobserveerde waarden. Stel bijvoorbeeld dat voor alle adressen 60% kans wordt geschat op 'één huishouden'. Afronden zou dan betekenen dat voor alle adressen 'één huishouden' wordt geïmputeerd. De categorie 'één huishouden' wordt dan dus geobserveerd in 60% van de adressen, maar komt voor in 100% van de imputaties. Voor veel toepassingen bij statistische bureaus is dit zeer ongewenst, aangezien een juiste verdeling op geaggregeerd niveau belangrijker is dan een nauwkeurige schatting op individueel niveau.

Resultaten

Gradient Boosting blijkt betere schattingen te geven dan de huidige regressiemethode. Door toepassing van de methode op woningen met een bekende huishoudsamenstelling (zogenaamde cross-validatie) is een inschatting te maken van de precisie van de imputaties. Gradient Boosting classificeert 77% van de adressen correct, terwijl dit percentage voor de huidige regressiemethode op 74% ligt. Daarnaast is er ook gekeken naar de zogenaamde AUC (Area Under the Receiver Operating Characteristics (ROC) Curve). Eén van de interpretaties hiervan is hoe waarschijnlijk het is dat een grotere kans voor klasse 1 wordt voorspeld voor iemand uit klasse 1 vergeleken met iemand uit klasse 0. De score ligt tussen 0,5 en 1. Een score van 0,5 betekent dat een model willekeurig gokt en 1 staat voor een perfecte discriminatie tussen de twee groepen. De AUC/ROC voor het regressiemodel en Gradient Boosting bedragen respectievelijk 0,87 en 0,90.

Gradient Boosting versus regressie

Zoals hierboven beschreven geeft Gradient Boosting nauwkeurigere schattingen dan de huidige regressiemethode. Gradient Boosting heeft echter ook andere voordelen. Zo is deze methode makkelijker toepasbaar. Bij regressie moet van te voren worden bepaald wat de relatie is tussen de doelvariabele en de verklarende variabelen. Dit kan bijvoorbeeld een lineair verband zijn, maar ook een exponentieel verband. Bij Gradient Boosting is het niet nodig om dit van te voren te vast te leggen; dit wordt door de methode bepaald. Ook kan Gradient Boosting eenvoudiger dan regressie overweg met ontbrekende waarden in verklarende variabelen. Een nadeel van Gradient Boosting is dat de uitkomsten lastiger zijn te duiden. Het is niet erg eenvoudig om achteraf te achterhalen hoe specifieke schattingen tot stand zijn gekomen. Bij regressie is dit makkelijker. Vanwege de bovenstaande voordelen is geadviseerd om Gradient Boosting te implementeren. Momenteel wordt de methode verder uitgewerkt en worden er voorbereidingen getroffen om de methode in het productieproces op te nemen.

Jacco Daalmans heeft econometrie gestudeerd aan Tilburg University. Hij werkt als methodoloog voor het Centraal Bureau voor de Statistiek en is in 2019 gepromoveerd op toepassingen van macro-integratie in de officiële statistiek.
E-mail: j.daalmans@cbs.nl

Gerard Sierksma

column

Over P≠NP en een Eeuwige Student

In het decembernummer van het tijdschrift *NewScientist* staat een mooi verhaal met de titel 'P=NP?'. Een vraag als titel dus. Die vraag betreft een van de zeven beroemde millenniumproblemen met een miljoen dollar voor elke eerste oplosser. Het antwoord laat al ruim 50 jaar op zich wachten en Tamara Florijn, de auteur, twijfelt of het er ooit van komt. Ze eindigt nogal pessimistisch met zinsneden als: '(...) dat het nog wel honderd jaar kan duren voordat (...)' en 'Misschien zullen we wel nooit weten of (...)'. Je moet kennelijk wel een beetje gek zijn om er tijd en energie in te steken. Ik heb zo'n 'gek' gekend, een eeuwige student.

Iedereen heeft wel een beeld van een eeuwige student, maar wat P≠NP of P=NP betekent is minder bekend, zeker ook doordat de uitleg ervan een behoorlijke dosis wiskundige voorkennis vereist. Dat de P en NP niets van doen hebben met parkeren of zo lijkt me duidelijk, maar met wat dan wel? Om te beginnen, de P staat voor 'polynomiaal'. Je zou denken dat NP dan voor 'niet-polynomiaal' staat, maar dat is dan weer niet het geval. Schiet niet op dus, zelfs als ik toevoeg dat NP staat voor 'niet-deterministisch polynomiaal'. Dan maar kort-door-de-bocht uitgelegd, wat Tamara ook doet.

De letter P staat voor de klasse van wiskundige problemen die (met een algoritme) zijn op te lossen in polynomiale tijd, wat kort-door-de-bocht wil zeggen 'snel op te lossen'. En wat is 'snel' dan wel? Ik kom daar zo op terug. Eerst even naar NP. Dat is de klasse van problemen waarvan 'snel kan worden gecheckt' of een gevonden 'oplossing' echt wel oplossing is. Oké, nu 'snel'. Interessant in deze context is dat er een helder onderscheid is tussen 'snel' en 'niet snel'. Een probleem heet 'snel', sommigen zeggen zelfs 'makkelijk', oplosbaar als voor dat probleem een algoritme bestaat waarmee, voor alle mogelijke input data, een oplossing wordt geproduceerd binnen een beperkte tijdspanne, ook wel *real-time* genoemd. Een voorbeeldje. Als om vier uur 's ochtends de vrachtwagens moeten beginnen met het rijden van de routes en alle bezoekadressen zijn bekend om, zeg, drie uur in de morgen, dan heeft het algoritme maximaal een