De Voorspellende Kracht van Regressieanalyse en het Gebruik van een *Machine Learning* Algoritme in de E-Commercemarkt

Jurriën de Gelder (454345)

Jurriendeg@hotmail.com

Master thesis: Arbeid- en Organisatiepsychologie

Erasmus Universiteit Rotterdam

Scriptiebegeleider: Marise Born

Tweede beoordelaar: Alec Serlie

Juli 30, 2021

Aantal woorden: 11,125

Abstract (in English)

The uprising e-commerce market and growing data houses give behavioural scientists new possibilities to investigate human behaviour. In line with these opportunities, in this thesis financial behaviour of consumers was studied through the biodata method and a machine learning algorithm. Investigating personal history data of customers may explain and predict bad and good financial behaviour (biodata method). Using a sample of *N*= 28,988 customers buying goods on the internet, personal history data such as one's age, gender, socio-economic status, financial habits, past behaviour, ecological purchases, and safety purchases were used to predict financial behaviour. Regression analysis results showed that one's age, financial habits, past behaviour, and ecological purchases were positive predictors of financial behaviour (R^2= .36). These findings confirm that the biodata method can be used to better understand behaviour in a big dataset. Using the same predictors and outcomes, the machine learning algorithm (Random Forest) predicted financial behaviour only a little better (R^2= .40) than the regression analysis. In the machine learning algorithm, one's past behaviour, socio-economic status, age, financial habits, and the number of placed orders, respectively in this order, where the most important predictors of financial behaviour. Socio-economic status was the second most important variable in the machine learning algorithm but was a non-significant predictor in the linear regression analysis. These results suggest that there could be a non-linear relationship between socio-economic status and financial behaviour. Therefore, a machine learning algorithm can show which variables could be focused on in future psychological research (for example socio-economic status). In conclusion, behavioural scientists can use regression analysis and machine learning algorithms in big data to predict human behaviour.

De Voorspellende Kracht van Regressieanalyse en het Gebruik van een *Machine Learning* Algoritme in de E-Commercemarkt

Vandaag de dag is er een ongekende hoeveelheid online dataopslag aanwezig. Steeds meer informatie over mensen en organisaties wordt opgeslagen in big datasets. Onderzoek naar big data wordt binnen organisaties steeds populairder (Agrawal, 2020). Met big data worden voorspellingen gedaan om beurskoersen in te schatten, de markt in kaart te brengen en het gedrag van werknemers te onderzoeken. Daarnaast zijn steeds meer organisaties en onderzoekers geïnteresseerd naar online gedrag van consumenten (Harlow & Oswald, 2016). Denk daarbij aan onderzoek naar de wensen, de behoeften en het koopgedrag van consumenten. Door een algoritme kunnen data van miljoenen consumenten tegenwoordig worden samengevat en analyseerbaar gemaakt (Harlow & Oswald, 2016). Het groeien van big datasets gaat hand in hand met de groei van de e-commerce markt (Gunasekaran, et al., 2002; Yurtkur & Bahtiyar, 2020). De e-commerce markt omvat alle handel die online plaatsvindt. Deze groei geeft de mogelijkheid om ook vanuit organisatiepsychologisch perspectief onderzoek te doen naar het gedrag van consumenten en werknemers (Guzzo, et al., 2015). Ter illustratie wordt de klanttevredenheid van consumenten in grote e-commerce organisaties bijgehouden in big datasets om zo de werkprestaties van klantenservicemedewerkers te meten. Het bijhouden van de werkprestatie maakt het mogelijk om de effecten van ingevoerde trainingen of een nieuwe bedrijfsvoering te toetsen (Mello et al., 2014). Hieruit volgt dat onderzoek naar big data implicaties heeft voor organisaties, bijvoorbeeld om de eigen bedrijfsstrategieën aan te passen voor betere organisatie uitkomsten zoals klanttevredenheid, maar ook werkbevlogenheid en werktevredenheid onder de eigen medewerkers.

Big data worden via elektronische wegen opgeslagen. Per minuut worden duizenden tot miljoenen data verzameld. De snelheid waarmee data worden verzameld, wordt ook wel *velocity* genoemd. Daarnaast worden grote hoeveelheden verschillende variabelen *variety* genoemd en grote hoeveelheden observaties worden *volume* genoemd (Guzzo et al., 2015). De uitdaging is dan ook om op basis van miljoenen datapunten verscheidene verbanden te vinden. Denk hiertoe bijvoorbeeld aan het verband tussen het bezoeken van websites en het koopgedrag van consumenten, of het verband tussen het leuk vinden van bepaalde mediaberichten en iemands persoonlijkheid (Harlow & Oswald, 2016). Echter, binnen de arbeids- en organisatiepsychologie is het niet eenvoudig om onderzoek te doen naar big data, want dergelijk onderzoek vereist

programmeervaardigheden om data uit een big dataset te extraheren en te analyseren (Cheung & Jak, 2016). Deze vaardigheden bezitten onderzoekers binnen de gedragswetenschappen veelal nog niet.

Dagelijks worden er grote hoeveelheden data opgeslagen over consumenten en hun financiële gedrag. Inmiddels zijn meer en meer onderzoekers het erover eens dat expertise uit verschillende wetenschapsdisciplines nodig is voor valide en ethisch onderzoek binnen het big data-domein (Harlow & Oswald, 2016). Daarnaast kan onderzoek van gedrag het best worden uitgevoerd door iemand met een gedragswetenschappelijke achtergrond in plaats van door een IT'er. In dit licht moet het onderzoek van deze scriptie worden gezien, waarbij door een gedragswetenschapper (in casu de schrijver van deze thesis) onderzoek wordt gedaan naar big data in de e-commerce markt. Als gezegd, maakt de groei van de e-commerce markt het mogelijk om het online financieel gedrag van consumenten te onderzoeken. Denk bij online financieel gedrag aan het bestellen van producten en de tijd tot betaling. In dit licht staat in deze scriptie de volgende vraag centraal:

> *Kan financieel gedrag van consumenten in de e-commerce markt door hun persoonlijke kenmerken worden voorspeld in een big dataset?*

Daartoe wordt hieronder eerst de zogeheten biodatamethode uiteengezet. Deze onderzoeksmethode wordt aangevuld met bespreking van verschillende theorieën vanuit de gedragswetenschappen. Omdat *machine learning* algoritmes (MLA) steeds bekender raken binnen gedragswetenschappen en veelbelovend zijn in het voorspellen van werkgerelateerde uitkomstvariabelen aan de hand van big data, wordt naast de biodatamethode een MLA ingezet ter voorspelling van financieel gedrag.

Als eerste wordt hieronder de biodatamethode besproken. De terminologie is misschien verwarrend, want biodata zijn iets anders dan big data. De biodatamethode houdt in dat onderzoek wordt gedaan naar iemands persoonlijke historische data om te voorspellen of die persoon goed of slecht gaat presteren (bv. werkprestatie; Cook, 2016, p. 194). Denk bij persoonlijke historische data aan iemands leeftijd, geslacht, eerdere ervaring, voorkeuren en ga zo maar door. Goldsmith (1922) maakte 100 jaar geleden als eerste gebruik van deze techniek in selectieprocedures. Als eerste werd door haar onderscheid gemaakt tussen goed, middelmatig en slecht presterende werknemers in een bepaalde functie. Daarna werd gekeken welke

eigenschappen deze goed, middelmatig en slechte presterende werknemers bezitten. Als laatst werd een beoordelingsformulier opgesteld aan de hand van biodata, zoals iemands echtelijke staat, leeftijd, educatie, werk, eerdere ervaring en club lidmaatschap (Goldsmith, 1922). Deze persoonlijke historische data konden onderscheid maken tussen slecht en goed presterende werknemers. Zo bleek uit haar studie een sterke predictieve validiteit tussen de gemeten biodata en werkprestatie, waardoor het gebruik van biodata een interessante onderzoeksmethode voor selectieprocedures werd (García-Izquierdo et al., 2020). Echter, deze selectiemethode kreeg tegelijkertijd veel kritiek omdat de empirische gelegde verbanden niet theoretisch onderbouwd waren. Naar aanleiding daarvan wordt er nu onderscheid gemaakt tussen rationele en empirische biodata (zie o.a. Cook, 2016, p. 196). Bij empirisch gebruik van biodata wordt gekeken naar verbanden tussen biodata en uitkomstmaten (bijvoorbeeld werkprestaties) zonder theoretische onderbouwing. Ondanks de hoge voorspellende kracht van biodata is het in de gedragswetenschappen belangrijk te begrijpen *waarom* er een relatie tussen biodata en uitkomstmaten wordt gevonden. Om deze reden zijn binnen de gedragswetenschappen rationele biodata meer geaccepteerd. Deze onderzoeksmethode vindt verbanden in datasets aan de hand van theoretisch onderbouwde voorspellingen. Vanwege de sterke predictieve validiteit van biodata en de mogelijkheid om de variabelen theoretisch te onderbouwen, wordt in deze thesis onderzoek gedaan naar de rationele biodatamethode in een big dataset. Tot op heden is de biodatamethode nog niet eerder ingezet om big data te onderzoeken in de e-commerce markt, daarom luidt de eerste onderzoeksvraag als volgt.

> *Onderzoeksvraag 1: Is de biodatamethode bruikbaar om binnen de e-commerce markt met behulp van big data voorspellingen te doen over financieel gedrag van consumenten?*

Daarnaast worden in deze scriptie *machine learning* algoritmes (MLA) gehanteerd. MLA zijn inmiddels ingezet om iemands persoonlijkheid te voorspellen, onderscheid te maken tussen verschillende bloemen, kankerherkenning te verrichten aan de hand van medische foto's en ga zo maar door (Müller & Guido, 2016, p. 1-3). De kracht van MLA ligt, aldus Müller en Guido (2016) in de predictieve validiteit. Een MLA ontwikkelt zelfstandig een voorspellend model (Chander, 2017): een MLA 'traint' zichzelf in een trainingsdataset om een zo goed mogelijk voorspellend model te maken doormiddel van voorspellende variabelen (*Input,* X*)* en een of meer uitkomstvariabelen (*output,* Y). Dit getrainde model wordt vervolgens getoetst op een nieuwe,

soortgelijke dataset. Dit model heeft één doel en dat is het voorspellen van een (of meerdere) uitkomstvariabele(n). Dit wordt ook wel voorspellend onderzoek genoemd (Yarkoni & Westfall, 2017). Deze vorm van onderzoek is erop gericht het best voorspellende model te ontwikkelen. Hoe een MLA observaties (*inputdata*) aan een uitkomstvariabele verbindt is veelal onduidelijk, dat wil zeggen oninterpreteerbaar voor mensen. Het model dat wordt gecreëerd door een MLA wordt daarom ook wel een *black box* genoemd (Chander, 2017). Echter, zoals eerder besproken is het van belang voor gedragswetenschappers om te weten *waarom* voorspellende variabelen relateren aan de uitkomstvariabele. Ondanks de *black box*, maakt de voorspellende kracht van MLA het interessant om te onderzoeken hoe MLA kan bijdragen aan gedragswetenschappelijk onderzoek in big data (zie ook Yarkoni & Westfall, 2017). Daarnaast wordt onderzocht of regressieanalyses en MLA in voorspellende kracht van elkaar verschillen.

In deze scriptie wordt onderzoek uitgevoerd binnen een organisatie die het kredietrisico van verschillende webwinkels in Nederland volledig overneemt. Wanneer consumenten op deze webwinkels een bestelling plaatsen kunnen zij ervoor kiezen om pas te betalen wanneer de bestelling is geleverd. Over deze consumenten worden voorspellingen gedaan over hun online financieel gedrag. Dit onderzoek draagt op de volgende manieren bij aan theoretische ontwikkeling en praktische ontwikkelingen. Ten eerste wordt in deze scriptie onderzocht of de biodatamethode inzetbaar is voor het voorspellen van gedrag in big data. Onder de biodata die in deze scriptie worden behandeld vallen demografische gegevens, gewoontes in betaalgedrag en psychologische kenmerken van consumenten om hun betaalgedrag te voorspellen. Daarnaast wordt onderzocht hoe gedragswetenschappers MLA kunnen toepassen op een big dataset in een organisatiecontext. Daartoe worden de hierboven genoemde biodata ingevoerd in een MLA. Ook wordt de voorspellende kracht van MLA vergeleken met klassieke regressieanalyse (i.e. de traditionele wijze om biodata te gebruiken in een voorspellingsmodel). Tot slot levert dit onderzoek meer duidelijkheid over het financieel gedrag van consumenten binnen de snelgroeiende e-commerce markt.

De Biodatamethode

Demografische Variabelen en Financieel Gedrag

Zoals eerder aangegeven is de biodatamethode al 100 jaar geleden binnen de personeelspsychologie ontwikkeld om tijdens selectieprocedures te voorspellen welke kandidaten goede, middelmatige en slecht presterende medewerkers zouden worden (Goldsmith, 1922). Dit

principe wordt in deze scriptie vertaald naar het voorspellen van goed tot en met slecht betalende consumenten. Om voorspellingen te kunnen doen over betaalgedrag met behulp van big data, baseert de huidige scriptie zich op eerder gepubliceerde psychologische studies over financieel gedrag van consumenten. Uit dergelijk onderzoek blijkt dat psychologische factoren bijdragen aan het voorspellen van financieel gedrag bovenop veelgebruikte economische variabelen zoals vraag en aanbod (schaarste zorgt bijvoorbeeld voor meer impulsief aankoopgedrag), overheidsinvloed, de hoogte van de spaarrente op dat moment, etc. (Dholakia et al., 2016; Letkiewicz & Heckamn, 2019; Loibl et al., 2011). Demografische variabelen (Baek & Hong, 2004; Experian, 2013; Oksanen et al., 2015; Wang et al., 2011), financiële gewoontes (Dholakia et al., 2016; Loibl et al., 2011) en dispositie (Rustichini et al., 2016) blijken als volgt een belangrijke rol te hebben in het voorspellen van financieel gedrag van consumenten.

Allereerst zijn, in de categorie demografische variabelen, de meest onderzochte factoren naar financieel gedrag de leeftijd van de consument (Autio et al., 2009; Experian, 2013), het geslacht (Baek & Hong, 2004) en iemands sociaaleconomische status (Wang et al., 2011). Financieel gedrag wordt daarbij gedefinieerd als het uitgeven van geld, het lenen van geld, het sparen van geld, kopen van producten en het opbouwen en aflossen van schuld. Ter illustratie van het onderzoek waarbij het effect van demografische factoren op financieel gedrag werd bestudeerd, kan een studie van Oksanen en zijn collega's (2015) genoemd worden. Zij bestudeerden in Finland het kredietgebruik en het financiële gedrag van consumenten. Finland is een welvarend land waarbij werknemers over het algemeen goed betaald krijgen en gunstige arbeidsovereenkomsten hebben (Oksanen et al., 2015). Om deze reden werd door de onderzoekers verwacht dat demografische factoren weinig invloed zouden hebben op het financiële gedrag. Echter, de onderzoekers vonden weldegelijk effecten, want tussen 2005 en 2013 had 20 procent van de onderzochte steekproef schulden waarvan een groot gedeelte te verklaren was door de demografische factoren leeftijd, geslacht en sociaaleconomische variabelen (opleidingsniveau, inkomen en arbeid; Oksanen et al., 2015).

Onder de demografische factoren blijkt leeftijd een belangrijke voorspeller van financieel gedrag. De *life-cycle* route van Webley en Nyhus (2001) stelt dat er een trend is waarbij jongeren (18- 24 jaar) meer geld uitgeven, minder sparen en meer lenen. Op de middenleeftijd (35-45 jaar) wordt er veelal meer gespaard. Volgens Webley en Nyhus (2001) heeft iemands levensfase dus een grote invloed op het financieel gedrag van die persoon. Oksanen en zijn collega's (2015)

vonden bewijs voor deze *life-cycle* route, want meer dan een kwart van de schulden waren te vinden bij 19-24-jarigen. Daartegenover stond de groep 50-64-jarigen met een veel lager percentage, namelijk 12.1 procent. Het verschil tussen deze groepen was significant (Oksanen et al., 2015). Ook Experian (2013) vond in zijn onderzoek dat jongeren veelal te laat waren met het betalen van hun rekeningen en hun kredietkaart saldo vaker volledig opmaakten. Een verklaring die de ontwikkelaars van de *life-cycle* route geven, is dat jongeren van hun jeugd tot jongvolwassenheid meer controle krijgen over geld, terwijl ze er nog niet mee kunnen omgaan (Webley & Nyhus, 2001). Jongeren vinden het moeilijk om met geld om te gaan, omdat ze minder ervaring hebben in financiële zaken (Autio et al., 2009; Nelson, 2011). Daarnaast suggereerden zij dat een aantal levensgebeurtenissen invloed hadden op jonge mensen. Denk aan onverwachte familie-uitbreiding, scheidingen en werkloosheid. Het probleem bij dergelijke gebeurtenissen is dat jonge mensen vaak geen financiële reserves hebben om met deze levensgebeurtenissen om te gaan (Autio et al., 2009). In het huidige onderzoek wordt dan ook verwacht dat consumenten met een hogere leeftijd beter financieel gedrag vertonen dan consumenten met een lagere leeftijd.

Daarnaast is de invloed van geslacht op financieel gedrag onderzocht. Mannen blijken meer geldproblemen te hebben dan vrouwen (Balmer 2006; Oksanen et al., 2015; Patel, 2012). Zo tonen meerdere onderzoeken aan dat mannen vaker schulden hebben dan vrouwen (Balmer et al., 2015; Oksanen et al., 2015). Een genoemde reden hiervoor is dat mannen vaak risicovoller ondernemen dan vrouwen. Ook bevonden mannen zich vaker dan vrouwen in criminele circuits en dit correleerde met slecht financieel gedrag. Om deze reden valt ook in het huidig onderzoek te verwachten dat vrouwen beter financieel gedrag vertonen dan mannen.

Tot slot is er een onderzoek gepubliceerd naar de effecten van sociaaleconomische status (SES) op financieel gedrag. Iemands SES is opgebouwd uit diens opleiding, inkomen en werkzaamheid (Oksanen et al., 2015). Wang en zijn collega's (2011) rapporteerden dat mensen met een lage SES vaker dan mensen met een hoge SES het krediet van hun kredietkaart volledig opmaakten. Daarnaast is gevonden dat schulden het hoogst waren bij mensen die het minst verdienden (Oksanen et al., 2015). Onder laag sociaaleconomische groepen vormden bovendien levensgebeurtenissen een gevaar. Zo vond een ander onderzoek dat schadelijke levensgebeurtenissen, zoals ernstige ziekte of het verliezen van een familielid, het inkomen verlagen, maar dat deze gebeurtenissen het koopgedrag niet verminderen (Wang et al., 2011). Op

grond van deze studieresultaten wordt verwacht dat mensen met een lage SES minder verantwoord financieel gedrag vertonen dan mensen met een hoge SES. Met betrekking tot leeftijd, sekse en sociaaleconomische klasse kunnen dan ook de volgende hypothesen worden opgesteld.

> *Hypothese 1. Uit big data blijkt dat mensen met een hogere leeftijd beter financieel gedrag vertonen dan jonge mensen.*
>
> *Hypothese 2. Uit big data blijkt dat vrouwen beter financieel gedrag vertonen dan mannen.*
>
> *Hypothese 3. Uit big data is een positief verband zichtbaar tussen iemands sociaaleconomische status en diens financiële gedrag.*

Financiële Gewoontes en Financieel Gedrag

Financieel gedrag wordt niet alleen beïnvloed door demografische variabelen, maar kan ook worden verklaard door gewoontes en het plannen van gedrag (Letkiewicz & Heckman, 2019). Zo stelt Ajzen's *Theory of planned behavior* (TPB) dat gedrag wordt uitgevoerd vanuit een intentie (Ajzen, 1991; Bamberg et al., 2003; Conner, 2020). In onderzoek naar de TPB is veelal gevonden dat drie antecedenten van intentie de kans verhogen op het uitvoeren van gedrag (Bamberg et al., 2003; Conner, 2020). De antecedenten van intentie zijn iemands houding over bepaald gedrag, iemands subjectieve normen over bepaald gedrag (sociale normen) en iemands waargenomen controle over diens gedrag (Ajzen, 1991). Bijvoorbeeld bij het aangaan van een lening is het dus belangrijk om de intentie en de antecedenten van deze intentie tot het terugbetaalgedrag te begrijpen. Een veel voorkomend probleem hierbij is overigens het zogeheten "intentie-gedragsgat". Dit gat verwijst ernaar dat een intentie niet altijd leidt tot het erbij behorende gedrag (Letkiewicz & Heckman, 2019). Als voorbeeld van dit 'gat' kan het uitstelgedrag bij studenten genoemd worden. De intentie om op tijd te studeren kan sterk aanwezig zijn, maar alsnog raken studenten vaak afgeleid en vertraagt daarom hun studie. Het "intentie-gedragsgat" kan overbrugd worden door het concept: *habit* (gewoonte; Letkiewicz & Heckman, 2019). Een gewoonte ontstaat wanneer gedrag geautomatiseerd wordt. Het gedrag wordt dan geactiveerd door de context of een *cue* in plaats van door iemands gedragsintentie alleen (Letkiewicz & Heckman, 2019). Het creëren van gewoontes in financiële kwesties speelt een belangrijke rol in financieel gedrag (Letkiewicz & Heckman, 2019). Uit onderzoek blijkt

bijvoorbeeld dat spaargedrag wordt voorspeld door het creëren van financiële gewoontes (Dholakia et al., 2016; Loibl et al., 2011). Een financiële gewoonte is herhaaldelijk gedrag dat betrekking heeft op financiële kwesties. Ter illustratie van zo'n gewoonte: na het ontvangen van het maandelijks loon stort de ontvanger direct een percentage van dit loon op een spaarrekening. Letkiewicz en Heckman (2019) rapporteerden dat gewoontes een *carry-over* effect op financieel gedrag hebben. Een *carry-over* effect voor financiële gewoontes betekent dat als iemand een gewoonte heeft om op eenzelfde tijd diens schulden af te lossen (dag na het ontvangen van salaris), dat dan de kans groot is dat diegene meer geld spaart (positief financieel gedrag). Tot slot is het relevant om aan te geven dat het eenmalig verbreken van een gewoonte niet automatisch impliceert dat de gewoonte wordt stopgezet (Lally et al., 2009). In lijn met deze onderzoeksresultaten wordt in de huidige studie verwacht dat de groep consumenten met een positieve financiële gewoonte beter financieel gedrag zal vertonen dan de groep consumenten zonder een financiële gewoonte. Hypothese 4 kan als volgt hieruit worden afgeleid.

> *Hypothese 4. In de big dataset is een positief verband zichtbaar tussen het hebben van een positieve financiële gewoonte en financieel gedrag.*

Financieel Gedrag als Index voor Consciëntieusheid

Meerdere onderzoeken hebben de relatie tussen financieel gedrag en dispositie (ofwel persoonlijkheid) onderzocht (Letkiewicz & Heckamn, 2019; Webley & Nyhus, 2001). Uit onderzoek was bijvoorbeeld een positieve relatie tussen neuroticisme en financieel gedrag zichtbaar (Letkiewicz & Heckamn, 2019). Ook bleek uit onderzoek dat dat positief financieel gedrag een gedragsindicatie was voor de persoonlijkheidseigenschap consciëntieusheid (Jackson et al., 2010). Zij legden dit uit aan de hand van de *act frequency theorie* (AFT; Buss & Craik, 1983). De AFT stelt dat wanneer specifieke gedragingen zich regelmatig voordoen dit gedrag een indicatie kan vormen voor bepaalde persoonlijkheidseigenschappen (Buss & Craik, 1983). Ter illustratie: wanneer iemand regelmatig sociale gelegenheden opzoekt, kan dit een indicatie zijn van een extraverte persoonlijkheid. Vanuit deze redenering observeerden Jackson en zijn collega's (2010) 185 verschillende gedragingen en vonden positieve correlaties tussen het tonen van deze gedragingen en iemands consciëntieusheid ($r = .30$). Een van deze gedragingen is bijvoorbeeld: het vertonen van positief financieel gedrag, zoals het sparen van geld en het ordenen van financiële zaken. Vanuit de bevinding dat positief financieel gedrag een indicator is

van consciëntieusheid (Jackson et al., 2010) en omdat persoonlijkheid over het algemeen onveranderbaar is (Cobb-Clark & Schurer, 2012), wordt in de huidige studie verwacht dat eerder vertoond positief financieel gedrag toekomstig positief financieel gedrag kan voorspellen.

Hypothese 5. In de big dataset is een positief verband zichtbaar tussen eerder vertoond financieel gedrag en huidig financieel gedrag.

Gedrag Gerelateerd aan Consciëntieusheid

In het licht van de bevinding van Jackson en zijn collega's (2010), waar positief financieel gedrag een index voor consciëntieusheid bleek te zijn, wordt gedacht dat gedragingen die *gerelateerd* zijn aan consciëntieusheid waarschijnlijk ook een positieve relatie hebben met financieel gedrag. Uit één onderzoek werd bijvoorbeeld een positieve relatie tussen consciëntieusheid en eco-vriendelijke gedrag gevonden (Kvasova, 2015). Met eco-vriendelijk gedrag worden gedragingen bedoeld die bijdragen aan het verbeteren of behouden van de natuur en het milieu. Daarom wordt verwacht dat consumenten die meer eco-vriendelijk of *fairtrade* bestellingen plaatsen beter financieel gedrag vertonen. Ook is er een relatie gevonden tussen consciëntieusheid en het vertonen van veiligheidsgedragingen (Lee & Dalal, 2014). Daarom wordt verwacht dat consumenten die producten bestellen die te maken hebben met het verbeteren van hun veiligheid, beter financieel gedrag vertonen. Voorbeelden van deze aankopen zijn anticonceptiemiddelen en veiligheidskleding.

Hypothese 6. In de big dataset is een positief verband zichtbaar tussen het bestellen van eco-vriendelijke producten en financieel gedrag.

Hypothese 7. In de big dataset is een positief verband zichtbaar tussen het bestellen van veiligheidsproducten en financieel gedrag.

Machine Learning tegenover Regressieanalyses

Op de hierboven genoemde biodata zal ook een *Machine learning algoritme* (MLA) worden toegepast. Daarbij wordt onderzocht of MLA en regressieanalyses (RA) in voorspellende kracht van elkaar verschillen. Tabel 1 geeft een overzicht van vier eerdere onderzoeken die de voorspellende kracht van RA en MLA hebben vergeleken. Onderzoek toont gemengde resultaten

over de voorspellende kracht van beide methodes (Gravesteijn et al., 2020; Piros et al., 2019; Youyou et al., 2015; Yildiz et al., 2017). Zo onderzochten Gravesteijn en zijn collega's (2020) de voorspelbaarheid van morbiditeit door middel van hersenschade. Hierbij werd morbiditeit voorspeld aan de hand van informatie over patiënten met hersenschade door zowel MLA als een RA. De AUC waren voor bijna alle MLA en de RA gelijk. Een ander onderzoek in Hongarije probeerde sterfte te voorspellen door middel van een hartinfarctdiagnose (Piros et al., 2019). In dit onderzoek werden MLA en logistische RA met elkaar vergeleken. Ook in dit onderzoek was er geen verschil in voorspellende kracht tussen MLA en RA. Youyou en zijn collega's (2015) probeerden persoonlijkheidseigenschappen te voorspellen door middel van MLA en een RA. Respondenten vulden een persoonlijkheidsvragenlijst in over zichzelf en vervolgens vulden familie, vrienden en collega's (het sociale netwerk) van elke respondent dezelfde persoonlijkheidsvragenlijst in over een respondent. Zij moesten zo accuraat mogelijk de persoonlijkheid van de respondent inschatten. Daarnaast legde een MLA verbanden tussen het leuk vinden van mediaberichten op Facebook en zijn of haar persoonlijkheidsvragenlijst. Uiteindelijk correleerden de onderzoekers de voorspelde en geobserveerde waarde met elkaar. Hieruit bleek dat het MLA hogere predictieve validiteit had dan de RA (zie Tabel 1). Dit is één van de weinige onderzoeken die kijkt naar verschil tussen MLA en RA binnen de gedragswetenschappen. Tot slot bestudeerden Yildiz, en zijn collega's (2017) het verschil tussen MLA en RA in organisaties met als uitkomstvariabele het energieverbruik van een gebouw. In hun studie probeerden ze het jaarlijkse en maandelijkse energieverbruik van een gebouw te voorspellen aan de hand van variabelen zoals geografische ligging, weersomstandigheden, hoeveelheid mensen in het gebouw en ga zo maar door. Zij concludeerden in hun onderzoek dat de voorspellende kracht van MLA even hoog was als de RA, simpelweg omdat de R^2 van beide analysemethodes niet veel van elkaar verschilden (zie Tabel 1). Aangezien drie van de vier onderzoeken geen verschil vonden tussen MLA en RA, en gezien de hoge predictieve validiteit van het gebruik van historische persoonsdata (biodatamethode) aan de hand van regressieanalyses en de hoge predictieve validiteit van MLA in het algemeen, wordt in huidig onderzoek verwacht dat de RA en MLA beide even goed het financiële gedrag van consumenten kunnen voorspellen.

> *Hypothese 8. In een big dataset is geen verschil zichtbaar tussen een machine learning algoritme en multipele regressieanalyse in het vermogen om financieel gedrag te voorspellen.*

Tabel 1.

Samenvatting van Vier Eerdere Onderzoeken: Machine Learning Versus Regressie Modellen

Onderzoekers	Vraagstuk	Voorspellende variabelen		Uitkomstvariabele	Soort model/ MLA	Vergelijking	Hoofdresultaat	
		Regressie	MLA				MLA	Regressie
Youyou, Kosinski en Stillwell, 2015	Andere beoordeling vs. ML: persoonlijkheid in gedragswetenschappen	PV ingevuld door vrienden en familie van de respondenten (*k= 5)*	Facebook *likes* van respondenten, PV ingevuld door de respondenten (*k*= 6)	PV ingevuld door de respondenten (continu)	Multipele regressieanalyse, *Computer modellen*	Correlatie geobserveerde waarde en voorspelde waarde	*r*= .56	*r*= .49
Gravesteijn, Nieboer, Ercole, Lingsma, Nelson, en Van Calster, Steyerberg, 2020	Logistieke regressie vs. ML: sterfte door hoofdletsels in medische wetenschappen	Leeftijd, hoofdletsels, glucose, natrium, hemoglobine, etc. (*k*=11)	Dezelfde variabelen	Sterfte (dichotoom)	Logistische regressieanalyse, VM, RF, GBM, ANN.	Gebied onder ROC-curve meet aantal keer goed voorspelt (AUC)	AUC= .79 - .81	AUC= .81
Yildiz, Bilbao en Sproul, 2017	Multipele regressie vs. ML: energie verbruik van grote gebouwen in technische wetenschappen	Zonnestraling, ligging gebouw, tijdstip van het jaar, vakantietijd etc. (*k*= 10)	Dezelfde variabelen	Energie verbruik op campus per uur (continu)	Multipele regressieanalyse, ANN*, RT, VM	*R*= 1 – *Sum of squares error/ sum of squares total*	R^2= .93 - .99	R^2= 0.89
Piros, Ferenci, Fleiner, Andréka, Fujita, Főző, Kovács en Jánosi, 2019	Logistieke regressie vs. ML: sterfte door hartinfarct in medische wetenschappen	Roken, hartklachten, demografische variabelen etc. (*k*= 23)	Dezelfde variabelen	Overlijden binnen een jaar (dichotoom)	Logistische regressieanalyse, ANN, RT	Gebied onder ROC-curve meet aantal keer goed voorspelt (AUC)	AUC= .81 & .70	AUC= .81

Noot. *k*= aantal variabelen, PV= persoonlijkheidsvragenlijst MLA= *machine learning* algoritmes, VM= Vector Machines, RF= Random Forest, GBM= Gradient Boosting Machines, ANN= Artificial Neural Networks, RT= Regression Trees, *in dit onderzoek wordt gebruik gemaakt van meerdere varianten ANN.

Methode

Respondenten

Het onderzoek werd uitgevoerd binnen een organisatie die het kredietrisico overnam van verschillende webwinkels in Nederland. De consumenten hoefden bij deze webwinkels niet direct hun bestelling te betalen. De consumenten konden bij deze webwinkels ervoor kiezen om pas te betalen nadat ze hun bestellingen hadden ontvangen. Deze consumenten werden dan opgedragen om binnen twee weken het bedrag van de bestelling te betalen. Het financieel gedrag werd onderzocht van consumenten die ervoor kozen om pas te betalen nadat ze hun bestellingen hadden ontvangen. Bestellingen van consumenten werden opgeslagen en bijgehouden in een big dataset. Voor deze scriptie werd een steekproef van 28,988 Nederlandse unieke consumenten geanalyseerd over heel 2020. De voorspellende variabelen financiële gewoonte en eerder vertoond gedrag vereisen meerdere opname momenten om wat zinnigs te kunnen zeggen over hun relaties met de uitkomstvariabele. Om deze reden werden alleen consumenten geanalyseerd die twee of meer bestelling hadden geplaatst in 2020. Verder was het mogelijk om consumenten te analyseren die tot op heden hun bestellingen uit 2020 nog niet betaald hebben. Echter, de kans is groot dat dit om fraudegevoelige situaties gaat. Fraudeonderzoek was niet de focus van deze thesis, daarom werden alleen consumenten geanalyseerd die hun bestelling uit 2020 hadden betaald.

Materialen

De data werden opgevraagd via een programma genaamd Tableau. Daarnaast werd voor deze scriptie gebruik gemaakt van de programmeertaal R. Binnen R werd gebruik gemaakt van de *package* Tidyverse (Wickham et al., 2019). Dit is een bekende analyse *package* die het mogelijk maakte om grote hoeveelheden data te transformeren, te analyseren en te visualiseren. Daarnaast werd gebruik gemaakt van de *package* Caret (Kuhn, 2020). Deze *package* maakte het mogelijk om verschillende statistische modellen toe te passen op de dataset. Van deze modellen werd het Random Forest (RF) model en multipele regressieanalyse (RA) toegepast. Als laatst werd een computer gebruikt die 8 GB werkgeheugen had.

Het RF-model is geschikt als MLA voor het voorspellen van financieel gedrag in big data, omdat het RF-model makkelijker te begrijpen is dan andere MLA (Müller & Guido, 2016). Het RF-model is makkelijker te begrijpen omdat het model is opgebouwd uit verschillende besluit-bomen (*decision trees*) die over het algemeen voor mensen makkelijk te begrijpen zijn. Ook zorgt

het RF-model ervoor dat de algoritme niet *overfit* op de getrainde dataset waardoor het RF-model ongeveer dezelfde voorspellende kracht heeft op de testdataset (Müller & Guido, 2016). Daarnaast is het mogelijk om continue en categorische variabelen te voorspellen. Ook is het mogelijk om categorische, dichotome en continue variabelen als voorspellende variabelen in te voeren in het model (Müller & Guido, 2016). Deze redenen gaven de voorkeur om in dit onderzoek het RF-model toe te passen.

Procedure

De data werden in een aantal stappen opgevraagd vanuit een dataopslagplaats die in bezit is van de organisatie waar dit onderzoek plaatsvond. Een dataopslagplaats dient als verzamelplaats voor alle elektronische informatie die digitaal wordt uitgewisseld tussen de consument en een organisatie. Bijvoorbeeld informatie over alle geplaatste bestellingen, de betaalgeschiedenis en de persoonlijke informatie van consumenten die is uitgewisseld met een organisatie. Om de data op te vragen uit de digitale opslagplaats werd de applicatie Tableau gebruikt. In Tableau werden een aantal ingebouwde opties aangevinkt om specifieke data op te vragen. Ter illustratie werd in een keuzemenu aangevinkt dat data moesten worden opgevraagd over consumenten uit Nederland die twee of meerdere bestelling hadden geplaatst in 2020, met inhoudelijke informatie over hun bestellingen, de betaaltijd, de leeftijd en het geslacht van deze consumenten. Nadat de data werden opgevraagd, werden de data opgeslagen in een *comma-separated-values*-bestand (CSV; dat is een compact databestand).

Vervolgens werd het opgeslagen CSV-bestand ingeladen in R. R is een programmeertaal waarin een oneindig aantal mogelijkheden zijn met betrekking tot het analyseren van data, het transformeren van data, het opbouwen van datatabellen en ga zo maar door (R Core Team, 2021). In R werden de data getransformeerd naar analyseerbare data. Als eerste werd de formatie van data per bestelling omgezet naar een formatie waarin data per consument werden weergegeven. Daarna werden data zoals leeftijd in de vorm van iemands geboortedatum (bijvoorbeeld 12-03-1985) naar iemands leeftijd in jaren omgezet. Ook werden variabelen met tekst omgezet naar dichotome getallen. Bijvoorbeeld, de letters van geslacht (bijv. voor vrouwen: F, V, W etc.) werden omgezet naar een 1 of een 0. Als laatst werden bijzondere tekens (denk aan het euroteken) verwijderd uit de dataset.

Voordat de analyses werden uitgevoerd, moesten er een aantal stappen worden genomen om de anonimiteit van de consumenten te waarborgen. De variabelen geslacht,

sociaaleconomische status, leeftijd, bestelgeschiedenis en betaaltijd werden geëxtraheerd uit de dataopslagplaats. Persoonlijke gegevens zoals naam, email-adres, woonplaats en telefoonnummer werden niet geëxtraheerd uit de dataopslagplaats. Sociaaleconomische status werd bepaald aan de hand van de postcode. Volgens Guzzo en zijn collega's (2015) is het mogelijk om door middel van geslacht, leeftijd en postcode iemands persoonlijke gegevens volledig te achterhalen. Om deze reden werd actie ondernomen om de postcode van de consumenten te maskeren. Dit wordt onder het kopje sociaaleconomische status verder uitgelegd. Verder werden data opgevraagd van individuele consumenten van januari 2020 tot en met december 2020 met betrekking tot alle hieronder benoemde variabelen. Er was geen *informed Consent* afgenomen bij de consumenten. Dit is in lijn met de richtlijnen van APA wanneer er maatregelen worden genomen om de consument volledig anoniem te maken (Guzzo et al., 2015). De consumenten werden in deze scriptie volledig anoniem gemaakt door enkel de volgende persoonlijke gegevens te gebruiken in de analyses: het geslacht van de consumenten en de leeftijd van de consumenten. Vervolgens werd de biodatamethode toegepast. Eerst werd onderscheid gemaakt tussen slecht en goed betalende consumenten (zie kopje uitkomst variabele: financieel gedrag). Daarna werd onderzocht of de historische persoonlijke data van consumenten waren gerelateerd aan hun financieel gedrag. Als laatste werden over dezelfde voorspellende variabelen en de uitkomstvariabele het RF-model toegepast.

Variabelen

Uitkomst Variabele: Financieel Gedrag

Financieel gedrag werd gemeten aan de hand van de gemiddelde betaaltijd van een particuliere consument. De gemiddelde betaaltijd betekent de gemiddelde tijd (in dagen) die consumenten nodig hadden om hun bestelling te betalen. De consumenten werd opgedragen binnen twee weken het openstaande bedrag te betalen. Wanneer een consument nalatig was in het betalen van een bestelling dan kreeg deze consument een boete. Deze boete stijgt binnen een termijn van 90 dagen van 40 euro naar maximaal 270 euro. Consumenten kregen tot 90 dagen de tijd om het bedrag alsnog te betalen. Consumenten werden hiervan op de hoogte gesteld per mail. Later betaalgedrag is om twee redenen een indicatie van slecht financieel gedrag. Ten eerste maakt een consument vooraf bij het plaatsen van een bestelling een overweging of het bedrag op tijd betaald kan worden. Gemiddelde latere betaling is een indicatie dat de consument laks of onzorgvuldig handelt, of een onverstandige overweging heeft gemaakt en de consument riskeert

hierbij een boete. Daarnaast kan na het ontvangen van een bestelling het product alsnog worden teruggestuurd. Een consument kan een geplaatste bestelling dus rectificeren. Bijvoorbeeld wanneer een consument beseft dat er deze maand onvoldoende geld overblijft om de bestelling te betalen. Na het ontvangen van de bestelling kan dus alsnog een overweging worden gemaakt: "zal ik op tijd het verschuldigde bedrag kunnen betalen of niet?". Ongeacht de reden voor niet nakoming was het aan de consument om een inschatting te maken van de eigen betaalmogelijkheden. Om deze redenen is de gemiddelde betaaltijd in dagen een indicatie van financieel gedrag. Hierbij is een lagere gemiddelde betaaltijd (dus gemiddeld eerder betalen) een indicatie van goed financieel gedrag en een hogere gemiddelde betaaltijd (dus gemiddeld later betalen) een indicatie van slecht financieel gedrag.

De gemiddelde betaaltijd van een consument werd berekend aan de hand van diens betaalde bestellingen in 2020. Het aantal bestellingen dat de consumenten plaatste in 2020 varieerde per consument. Bijvoorbeeld een consument kon vier bestellingen hebben geplaatst en een andere consument kon 41 bestellingen hebben geplaatst. Over het eerste aantal (eerste 50%) geplaatste bestellingen werd de gemiddelde betaaltijd berekend. Deze gemiddelde betaaltijd valt onder de voorspellende variabele: 'eerder vertoond financieel gedrag'. Over het tweede gedeelte van de geplaatste bestellingen (overige 50%) werd ook de gemiddelde betaaltijd berekend. Deze gemiddelde betaaltijd valt onder de uitkomstvariabele: 'financieel gedrag'. Zoals in het voorbeeld hierboven met vier bestellingen werd over de eerste twee bestelling de gemiddelde betaaltijd berekend en over de overige twee geplaatste bestelling werd ook de gemiddelde betaaltijd berekend. Voor de tweede consument gold hetzelfde principe: over de eerste twintig bestellingen werd de gemiddelde betaaltijd berekend en over de tweede 21 geplaatste bestellingen werd ook de gemiddelde betaaltijd berekend. Het aantal bestellingen werd dus per consument in tweeën gesplitst om zo een voorspellende variabele te berekenen en om een uitkomstvariabele te berekenen. Verder werd ervan uitgegaan dat het aantal bestellingen geen invloed had op het financieel gedrag. Om deze reden werd het aantal bestellingen van een consument meegenomen als controlevariabele in de analyses.

Leeftijd

Tableau extraheerde de geboortedata van de consumenten uit de dataopslagplaats. Vervolgens werden deze geboortedata in R omgezet naar leeftijd in jaren.

Geslacht

Daarnaast extraheerde Tableau geslacht uit de dataopslagplaats. Geslacht werd in R omgezet naar een dichotome variabele waarbij man (1) en vrouw (0) in de dataset werd weergegeven.

Sociaaleconomische Status

Sociaaleconomische status (SES) werd bepaald aan de hand van de variabelen inkomen, educatie, en werkzaamheid (Oksanen et al., 2015). De SES van een consument werd ingeschat door de postcode van de consument. Bij het Centraal Bureau voor de Statistiek (CBS; 2021) zijn namelijk datasets beschikbaar over de sociaaleconomische kenmerken per postcode in Nederland. In twee stappen werden de postcodes van consumenten omgezet naar SES-scores. Ten eerste werden in de CBS-dataset de drie sociaaleconomische variabelen getransformeerd naar sociaaleconomische scores. De drie variabelen werden eerst gestandaardiseerd en vervolgens bij elkaar opgeteld. Elke postcode in Nederland kreeg hierdoor haar eigen unieke SES-score. Vervolgens werden de SES-scores per postcode gekoppeld aan de postcodes van de consumenten. Wanneer de SES-scores waren gekoppeld aan de consumenten, via hun postcode, werd de variabele postcode door een algoritme automatisch verwijderd uit de dataset. Zo bleef alleen de geschatte SES-scores over en waren er geen data over iemands woonadres aanwezig in de dataset. Een lage SES-score betekende dat een consument in een wijk woonde met meer laagopgeleiden, een gemiddeld laag inkomen en waar veel werkloosheid voorkwam. Een hoge score betekent dat er op een bepaalde postcode veelal meer hoog opgeleiden woonden, meer mensen werkzaam waren en dat er meer mensen met een hoog inkomen woonden.

Financiële gewoonte

De variabele financiële gewoonte had betrekking op de regelmaat waarmee betalingen plaatsvonden rond hetzelfde tijdstip. Te denken valt aan een consument die altijd binnen de eerste twee weken na het ontvangen van een bestelling betaalde. Al deze consumenten kregen een score van ‘1’ en betekent dat de consument een betaalgewoonte had. De consumenten die niet betaalden op eenzelfde tijdstip of rond hetzelfde tijdstip kregen een score van ‘0’. Een score van 0 betekent dat een consument geen betaalgewoonte had.

Eerder vertoond financieel gedrag

Het indexgedrag voor consciëntieusheid is het eerder vertoonde financieel gedrag van een consument. Het financieel gedrag werd, net als bij de uitkomstvariabele, bepaald door de gemiddelde betaaltijd van een consument in dagen. Zoals eerder is aangegeven werd dit per

consument berekend over de eerste 50% van diens geplaatste bestellingen in 2020.

Consciëntieusheid-gerelateerde Gedragingen

Om te bepalen of consumenten eco-vriendelijke of veiligheid producten hadden besteld, werden ongeveer duizend webwinkels beoordeeld die zijn aangesloten bij de organisatie waar dit onderzoek werd uitgevoerd. Het was niet mogelijk om de producten per bestelling te analyseren. Wel was het mogelijk de webwinkels te analyseren waar de bestellingen geplaatst werden. Daarom werden de webwinkels gecategoriseerd in eco-vriendelijke webwinkels, veiligheid gerelateerde webwinkels en alle overige webwinkels. Wanneer een consument één of meer bestellingen had geplaatst op een eco-vriendelijke webwinkel kreeg deze consument een score "1" op eco-vriendelijke overwegingen. Wanneer een consument geen bestelling had geplaatst bij een eco-vriendelijke webwinkel dan kreeg deze consument een score van "0". De beoordeling van eco-vriendelijke webwinkels zijn gedaan aan de hand van de volgende criteria: de webwinkel mocht alleen eco-vriendelijke producten verkopen en deze producten moesten positief bijdragen aan het milieu. Voorbeelden van eco-vriendelijke webwinkels zijn vegetarische webwinkels, webwinkels die duurzame producten verkopen en biologische webwinkels. Verder, wanneer een consument een bestelling had geplaatst op en webwinkel met alleen veiligheidsproducten dan kreeg deze consument een score "1" op de variabele veiligheidsproducten. Wanneer een consument geen bestelling had geplaatst bij webwinkels die veiligheidsproducten verkochten dan kreeg deze consument een score van "0". De beoordeling van webwinkels met veiligheidsproducten zijn gedaan aan de hand van de volgende criteria: een bestelling moest zijn geplaatst op een webwinkel die alleen veiligheidsproducten verkoopt en deze producten moesten bijdragen aan iemands veiligheid. Voorbeelden van veiligheidsproducten zijn veiligheidskleding en condooms. Eco-vriendelijke overwegingen en het bestellen van veiligheidsproducten zijn dus dichotome variabelen.

Statistische Analyses

Om te toetsen of de regressieanalyse (RA) en het MLA financieel gedrag kon voorspellen, werd voor beide analysetechnieken dezelfde dataset gebruikt. Dit betekent dat het MLA en de biodatamethode (doormiddel van een RA) aan de hand van dezelfde inputvariabelen dezelfde uitkomstvariabele (gemiddelde betaaltijd) voorspellen (zoals hierboven beschreven). Op advies van Müller en Guido (2016) werden de consumenten in de dataset willekeurig verdeeld in twee groepen. Twee derde van de consumenten werden geplaatst in de trainingsgroep en één derde in

de testgroep (belangrijk voor de cross-validatie van het MLA). Met een multipele regressieanalyse werd het verband tussen demografische variabelen, gewoonte, eerder vertoond gedrag, en consciëntieus-gerelateerd gedrag enerzijds, en betaalgedrag anderzijds onderzocht. In het eerste blok werd de controlevariabele geplaatst. In het tweede blok werden de variabelen leeftijd, geslacht en SES toegevoegd. In het derde blok werd de variabele financiële gewoonte geplaatst. In het vierde blok werden de variabelen eerder vertoond betaalgedrag (als indicatie voor consciëntieusheid), ecologische aankopen en veiligheid aankopen geplaatst.

Voor deze scriptie werd het ML-algoritme *Random Forest* (RF) gebruikt (Müller & Guido, 2016). Het RF-model analyseert dezelfde dataset met daarbij dezelfde verdeling in een trainingsgroep en een testgroep als bij de multipele RA. Het MLA kiest het best passende model in plaats van een regressielijn (Müller & Guido, 2016). Het RF-model leerde in de trainingsgroep de voorspellende variabelen te verbinden aan de gemiddeld betaaltijd van een consument. In de testgroep werd vervolgens gecontroleerd of het geleerde model de gemiddelde tijd tot betaling kon voorspellen. Om een vergelijking te maken tussen het RF-model en de regressieanalyse werd het percentage verklaarde variantie in de uitkomstvariabele (R^2) door beide modellen berekend (Rosenbusch et al., 2021; Yildiz et al., 2017). Een hoge R^2 betekent dat de voorspelde waardes dicht bij de daadwerkelijke waardes zitten (bijv. Rosenbusch et al., 2021). Voor het MLA betekent dit dat de R^2 een X % variantie verklaart aan de hand van de variabelen die worden toegepast in het voorspellende model. Ook kan het RF-model laten zien welke variabelen belangrijk zijn voor het voorspellen van de uitkomst variabele (een voorbeeld hiervan is te zien in Figuur 1). Alle hieronder beschreven resultaten zijn gemeten in de testgroep (N= 7,860).

Resultaten

Allereerst is er gecontroleerd op uitschieters. In totaal zijn 720 observaties verwijderd uit de dataset, omdat deze observaties onmogelijke waardes bleken te bevatten. Een reden hiervoor was het verkeerd invullen van het bestelformulier door consumenten. Dit kwam voornamelijk voor bij de variabele leeftijd. In totaal werden de data van 28,988 consumenten geanalyseerd waarvan twee-derde (N= 16,540) gebruikt werd als trainingsgroep en één-derde (N= 7,860) als testgroep. In Tabel 2 zijn de gemiddelden, de standaarddeviaties en de onderlinge correlaties tussen de voorspellende variabelen en de uitkomstvariabele weergegeven. In Tabel 2 zijn relatieve zwakke

Tabel 2.

Gemiddelden, Standaarddeviaties en Intercorrelaties tussen alle Variabelen (N= 7,860).

	M	*SD*	1	2	3	4	5	6	7	8
1. Aantal bestellingen	8.24	8.18	-							
2. Geslacht	0.72	0.45	0.06**	-						
3. Leeftijd	43.14	12.99	0.04**	0.12**	-					
4. SES-score	5.98	2.62	0.01	-0.01	0.06**	-				
5. Gewoonte	0.61	0.49	-0.09**	-0.01	0.14**	0.04**	-			
6. Eerder vertoond gedrag	23.60	29.37	0.08**	-0.02	-0.18**	-0.07**	-0.34**	-		
7. Ecologisch	0.20	0.40	0.07**	-0.08**	0.01	-0.02	-0.01	-0.01	-	
8. Veiligheid	0.17	0.37	0.03**	0.21**	0.12**	0.02	0.02	-0.03**	-0.11**	-
9. Gemiddelde betaaltijd	20.92	23.79	0.08**	-0.01	-0.17**	-0.06**	-0.30**	0.59**	-0.01	-0.02*

Noot. **. Correlatie is significant, $p < 0.01$ (2-zijdig). *. Correlatie is significant, $0.05 < p < 0.01$ (2-zijdig). Geslacht: man= 1, vrouw= 0. SES= sociaaleconomische status score. Gewoonte: geen gewoonte= 0, wel een gewoonte= 1. Ecologisch: bestelling geplaatst op ecologische webwinkels= 1, geen bestelling geplaatst op ecologische webwinkels= 0. Veiligheid: bestelling geplaatst op en webwinkel die veiligheidsproducten verkoopt= 1, geen bestelling geplaatst op webwinkels die veiligheidsproducten verkoopt= 0. ‘Gemiddelde betaaltijd’ en ‘eerder vertoond gedrag’= gemiddelde tijd tot betaling in dagen.

Tabel 3.

Multipele Regressieanalyse met Gemiddelde Betaaltijd in Dagen als Uitkomstvariabele

	Model 1		Model 2		Model 3		Model 4	
	B	*β*	B	*β*	B	*β*	B	*β*
Controle variabele								
Aantal bestellingen	0.27*	0.09*	0.29*	0.10*	0.19**	0.07**	0.12**	0.04**
Demogr. variabelen								
Geslacht			0.77	0.01	0.23	0.00	0.31	0.01
Leeftijd			-0.29**	-0.16**	-0.22**	-0.12**	-0.09**	-0.05**
SES			-0.35**	-0.04**	-0.27*	-0.03*	-0.06	-0.01
Gewoonte								
Financiële gewoonte					-13.77**	-0.28**	-5.1**	-0.10**
Dispositie								
Eerder vertoond gedrag							0.43**	0.54**
Ecologische aankopen							-1.45*	-0.02*
Veiligheid aankopen							-0.21	0.00
R^2		0.01**		0.03**		0.11**		0.36**
R^2 *changed*				0.02**		0.08**		0.25**

Noot. **. Significantie < 0.001 (2-zijdig), *. Significantie 0.05 – 0.001 (2-zijdig), SES= sociaaleconomische status.

correlaties te zien. Opvallend is de positieve correlatie tussen leeftijd en geslacht. Dit betekent dat de mannen een hogere leeftijd hadden dan de vrouwen. Ook was er een correlatie tussen veiligheid en geslacht. Dit zou kunnen betekenen dat mannen meer veiligheidsproducten bestelden dan vrouwen. Echter waren deze gevonden relaties zwak.

Om hypothesen 1 tot en met 7 te toetsen is een multipele regressieanalyse uitgevoerd. Uit de resultaten bleek een significant negatieve relatie tussen leeftijd en de gemiddelde betaaltijd van de consument (*β*= -0.05, C.I. 95% [-0.07, -0.03]). Oudere consumenten betaalden eerder hun geplaatste bestellingen dan jongere consumenten. Hypothese 1 werd hierdoor bevestigd. Daarentegen bleek uit de resultaten dat geslacht geen significante voorspeller was voor financieel gedrag (*β*= 0.01, C.I. 95% [-0.01, 0.02]). Ook sociaaleconomische status bleek geen significante voorspeller voor financieel gedrag te zijn (*β*= -0.01, C.I. 95% [-0.02, 0.01]). Daarom werden de

hypothesen 2 en 3 verworpen. Verder was er een significant negatieve relatie tussen het hebben van een financiële gewoonte en de gemiddelde betaaltijd (β= -0.10, C.I. 95% [-0.12, -0.09]). Consumenten met een positieve betaalgewoonte betaalden hun bestelling eerder dan consumenten zonder een betaalgewoonte. Hypothese 4 werd hierdoor bevestigd. Daarnaast was er een significant positieve relatie tussen eerder vertoond financieel gedrag en de gemiddelde betaaltijd (β= 0.54, C.I. 95% [0.52, 0.56]). Consumenten die hun eerste aantal bestellingen regelmatig vroegtijdig betaalden, betaalden hun latere geplaatste bestelling ook eerder. Hierdoor werd hypothese 5 bevestigd. Ook was er een significant negatieve relatie tussen ecologische aankopen en de gemiddelde betaaltijd (β= -0.02, C.I. 95% [-0.04, -0.01]). Consumenten die ecologische keuzes maakten, betaalden eerder hun geplaatste bestellingen dan consumenten die geen ecologische keuzes maakten. Hypothese 6 werd daarom ook bevestigd. Als laatst was er geen significante relatie gevonden tussen veiligheidsaankopen en de gemiddelde betaaltijd van de consumenten (β= 0.00, C.I. 95% [-0.02, 0.01]). Hypothese 7 moest daarom worden verworpen. Tot slot bleek de controle variabele ‘aantal geplaatste bestellingen’ een significante voorspeller voor de gemiddelde betaaltijd (β= 0.04, C.I. 95% [0.02, 0.06]). Consumenten die meer bestellingen hadden geplaatst deden langer over het betalen van hun bestellingen dan consumenten die minder bestellingen hadden geplaatst.

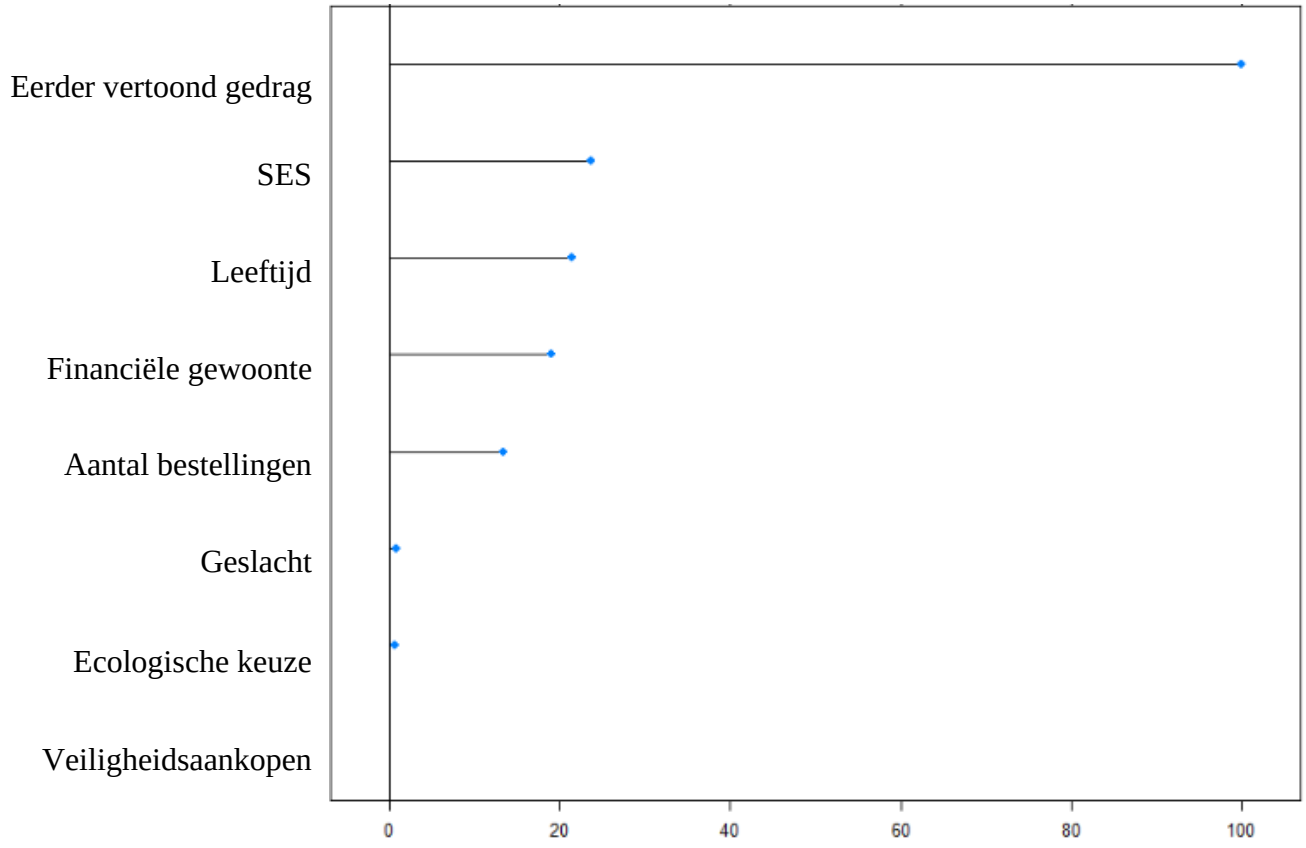
Verder werd onderzocht of de biodatamethode een geschikte onderzoeksmethode is voor het voorspellen van goed en slecht financieel gedrag (onderzoeksvraag 1). Uit de resultaten bleek dat model 4 (Tabel 3) een significante voorspeller was voor financieel gedrag in de testgroep ($F(8, 7851) = 557.7$, $p < .001$). Vier van de zeven variabelen, namelijk leeftijd, financiële gewoonte, eerder vertoond financieel gedrag en het bestellen van ecologische producten, konden onderscheid maken tussen laat en vroeg betalende consumenten. De variabelen verklaarden samen 36.24% van de variantie in dit gedrag. Deze resultaten geven aan dat historische persoonsdata goed en slecht betalende consumenten deels kunnen onderscheiden.

Om te toetsen of MLA en RA het financiële gedrag van consumenten even goed kunnen voorspellen (hypothese 7), is de voorspellende kracht van het Random Forest (RF) model vergeleken met de voorspellende kracht van het RA-model 4 aan de hand van het *least sum of squares* principe. Het RF-model is gebaseerd op keuzebomen en maakte een schatting van de gemiddelde betaaltijd aan de hand van dezelfde dataset en variabelen die zijn gebruikt in het RA-model 4 (Tabel 3). Voor de meest optimale *least sum of sqauares* werden in het RF-model 500

bomen gecreëerd. Om te toetsen of het MLA evengoed voorspelt als de multipele regressieanalyse is op advies van Kuhn (2020) een *t*-toets uitgevoerd. Uit de resultaten blijkt dat het RF-model een hogere voorspellende kracht heeft dan de multipele regressie analyse ($t(24)= -8.98$, $p < .001$). Dit resultaat houdt in dat de *least sum of squares* van het MLA significant lager is dan de *least sum of squares* van de RA: het MLA verklaarde 39.99% en de multipele regressieanalyse 36.24% van de variantie in het betaalgedrag. Hypothese 8 moest dan ook worden verworpen. De voorspellende kracht van het MLA was iets hoger dan de voorspellende kracht van de RA.

Figuur 1.

Belang van de Variabelen in het Random Forest Model ($R^2= .40$)



Noot. Belang van de variabelen is bepaald aan de hand van het aantal keer dat een variabele in de bomen ($k= 500$) is gebruikt om de gemiddelde betaaltijd te voorspellen, uitgedrukt in percentages.

Discussie

In deze scriptie is een big dataset (*N*= 28,988) gebruikt om het financieel gedrag van consumenten te voorspellen aan de hand van hun biodata, namelijk demografische variabelen (leeftijd, geslacht, en sociaaleconomische status), en psychologische variabelen (financiële gewoontes, eerder vertoond gedrag, ecologische keuzes en het bestellen van veiligheidsproducten). Als voorspellingsmodellen is gebruik gemaakt van een regressieanalyse (RA) en een *machine learning* algoritme (MLA). Per alinea worden hieronder respectievelijk de variabelen uit model 4 (Tabel 3) de resultaten uit de RA besproken en vervolgens worden de resultaten uit het RF-model besproken (Figuur 1).

Vanuit de RA is een positieve relatie gevonden tussen leeftijd en gemiddelde betaalsnelheid van de consument. Dit geeft bewijs voor de *life-cycle* route van Webley en Nyhus (2001). Zij stelden dat iemands levensfase diens financieel gedrag voorspelt: Jongeren hadden over het algemeen minder geld te besteden en komen daardoor eerder in financiële problemen (i.e. schulden). Daarnaast bleek dat naarmate iemand ouder werd diegene meer ervaring had met financiële zaken en daardoor verstandiger leken om te gaan met hun financiën. Aansluitend op de RA maakte ook het MLA gebruik van de variabele leeftijd om financieel gedrag te voorspellen. Echter, zoals in de inleiding werd uitgelegd, was vanuit het MLA de relatie tussen leeftijd en de gemiddelde betaaltijd onduidelijk.

Daarentegen vond de RA geen bewijs voor de relatie tussen geslacht en financieel gedrag. Het idee dat mannen minder goed financieel gedrag vertonen is in dit onderzoek niet bevestigd. Ook maakt het MLA geen gebruik van de variabele geslacht om voorspellingen te doen over de gemiddelde betaaltijd van de consument (Figuur 1). Een nadeel voor het gebruik van de variabelen leeftijd en geslacht in big data is dat deze variabelen amper geverifieerd kunnen worden. Het kan zijn dat consumenten de verkeerde leeftijd of geslacht hebben ingevuld op het bestellingsformulier. Bijvoorbeeld omdat er geen consequenties of beloningen vast zaten aan het correct invullen van geslacht of leeftijd. De verificatieproblemen in dit onderzoek kunnen van invloed zijn geweest op de relatie tussen demografische variabelen (leeftijd en geslacht) en financieel gedrag.

Verder bleek uit de RA dat er geen positieve significante relatie was tussen sociaaleconomische status (SES) en financieel gedrag. In tegenstellingen tot de bevindingen van Oksanen en zijn collega's (2015) suggereren deze resultaten dat consumenten met een lage SES

niet per se later betaalden dan consumenten met een hogere SES. Ondanks het ontbreken van een lineaire relatie tussen SES en financieel gedrag bleek uit het MLA dat er wel een relatie aanwezig was tussen SES en financieel gedrag. In Figuur 1 is te zien dat het RF-model SES als tweede beste voorspeller gebruikt. Het MLA suggereert dat er mogelijk een kwadratische relatie, interactie of indirect effect aanwezig is. Deze mogelijke relaties kunnen met RA verder worden onderzocht. Bijvoorbeeld, wellicht vertoonden consumenten met een lage SES slechter financieel gedrag, consumenten met een gemiddelde SES beter financieel gedrag en consumenten met een hoge SES weer slechter financieel gedrag (kwadratische relatie). Een mogelijke verklaring voor een eventuele kwadratische relatie is dat consumenten met een hoge SES zich minder bekommerden om kleine boetes. Hierdoor kan het zijn dat consumenten met een hoge SES later betalen in vergelijking met consumenten met een gemiddelde SES. Een U-vormige relatie kan verder worden onderzocht. Daarnaast is gebruik gemaakt van de postcode-4 dataset (de vier cijfers van iemands postcode) om diens SES te bepalen. Wellicht wordt er een positief significante relatie tussen SES en financieel gedrag gevonden wanneer er een onderzoek wordt uitgevoerd met de postcode-6 dataset (vier cijfers en twee letters). Een reden hiervoor is dat met vier cijfers en twee letters van een postcode meer nauwkeurige schattingen kunnen worden gedaan over iemands SES. Door gebrek aan financiële bronnen is in deze thesis gebruik gemaakt van een postcode-4 dataset in plaats van een postcode-6 dataset. Een postcode-6 dataset met demografische gegevens kost 2,000 euro en de postcode-4 dataset met demografische gegevens is gratis (Centraal Bureau voor de Statistiek, 2021).

Naast leeftijd vond de RA ook een positieve relatie tussen het hebben van een financiële gewoonte en financieel gedrag. Consumenten die een positieve financiële gewoonte hadden betaalden gemiddeld sneller dan consumenten die geen financiële gewoonte hadden. Ook vormt deze bevinding een onderbouwing voor het *carry-over* effect, want de financiële handeling ‘gewoonte’ was gerelateerd aan de financiële handeling ‘gemiddelde betaaltijd’ (Letkiewicz & Heckman, 2019). Ook het RF-model maakte gebruik van de variabele financiële gewoonte om de gemiddelde betaaltijd te voorspellen. Dit resultaat geeft een extra bevestiging dat financiële gewoontes een belangrijke rol spelen in het voorspellen van financieel gedrag.

Daarnaast is dit onderzoek een aanvulling op de eerdere bevindingen van Jackson en zijn collega’s (2010). Zij vonden in hun onderzoek dat het vertonen van bepaalde gedragingen indexen vormen voor persoonlijkheidseigenschappen. Financieel gedrag bleek in hun onderzoek

een index te zijn voor consciëntieusheid (Jackson et al., 2010). In deze thesis is een van die indexen, namelijk eerder vertoond positief financieel gedrag, gerelateerd aan later vertoond positief financieel gedrag. De resultaten van dit onderzoek lieten zien dat wanneer consumenten dit gedrag vertoonden zij op later moment dit gedrag weer vertoonden. Ter onderbouwing, wanneer consumenten eerder vroegtijdig betaalden, binnen een jaar, betaalden zij later ook vroegtijdig. Daarnaast was eerder vertoond financieel gedrag de belangrijkste voorspeller voor de RA (Tabel 3) en het MLA (Figuur 1). Deze bevinding tonen aan dat indexen van persoonlijkheid bruikbaar zijn om gedrag te voorspellen in big data.

Daarnaast werd verwacht dat gedragingene zoals het bestellen van ecologische producten en het bestellen van veiligheidsproducten, positief gerelateerd zijn aan financieel gedrag. Aangezien Jackson en zijn collega's (2010) constateerden dat positief financieel gedrag een index was voor consciëntieusheid, werd verwacht dat ecologische overwegingen en het kopen van veiligheidsproducten ook positief gerelateerd zouden zijn aan financieel gedrag. Zo bleek uit de RA dat consumenten die ecologische producten hadden besteld gemiddeld sneller betaalden dan consumenten die geen ecologische producten hadden besteld. Het MLA vond het bestellen van ecologische producten geen belangrijke voorspeller voor financieel gedrag (Figuur 1). Er is zowel door de RA als door het MLA geen bewijs gevonden voor de relatie tussen het betalen van veiligheidsproducten en de gemiddelde betaaltijd van consumenten. Op basis van deze resultaten kan er daarom geconcludeerd worden dat het mogelijk is om gedrag in big data te onderzoeken aan de hand van gedragingen die gerelateerd zijn aan indexengedragingen van persoonlijkheid.

Al met al lijkt de biodatamethode een bruikbare onderzoeksmethode om te differentiëren tussen goed en slecht betalende consumenten. Dit betekent dat de ideeën van Goldsmith (1922) na 100 jaar nog steeds bruikbaar zijn om gedrag te voorspellen. De RA wist namelijk maar liefst ten minste één derde van de variantie in financieel gedrag te verklaren. Vier van de zeven voorspellers, namelijk leeftijd, het hebben van een financiële gewoonte, eerder vertoond financieel gedrag en ecologische overwegingen, hadden een positieve relatie met de uitkomstvariabele. Dus het onderzoeken van goed en slecht financieel gedrag doormiddel van historische persoonsdata heeft belangrijke implicaties voor het voorspellen van gedrag van consumenten. Vergeleken met de RA kon het RF-model net iets meer variantie in financieel gedrag verklaren. Mogelijk is het zo dat er indirecte relaties, interacties of U-vormige relaties aanwezig waren. Hier houdt het RF-model rekening mee, maar bij een multipele regressieanalyse

moet dit handmatig worden ingevoerd. Echter, een limitatie van het RF-model was dat ondanks het belang van de variabelen getoond wordt in Figuur 1, het verder onduidelijk blijft waarom deze variabelen belangrijk waren. Zo legt de RA, aan de hand van theorie en eerdere onderzoeken, wel uit waarom sommige consumenten beter financieel gedrag vertoonden dan andere consumenten. Dit heeft belangrijke implicaties voor de ontwikkeling van theorie binnen de gedragswetenschappen.

Beperkingen van de Huidige Studie en Suggesties voor Toekomstig Onderzoek

Een beperking van de methode van deze scriptie is dat er van de 49,082, beschikbare unieke Nederlandse consumenten met twee of meer bestellingen in 2020, slechts 28,988 unieke consumenten werden geanalyseerd. De reden hiervoor was een te kort aan werkgeheugen (8 GB) in het systeem van de gebruikte computer. Het realiseren van meer werkgeheugen kost geld, maar de beschikbare middelen hiervoor ontbraken. Daarnaast kan men zich afvragen of er bij 28,988 observaties wel gesproken kan worden van big data. Als alle aankopen in beschouwing zouden worden genomen, dan blijkt het dat er in het jaar 2020 meer dan 1 miljoen bestellingen zijn geplaatst bij de organisatie waar dit onderzoek plaatsvond. Door alleen Nederlandse individuele consumenten in de dataset op te nemen die twee of meer bestellingen plaatsten in het jaar 2020, is het aantal observaties in de dataset gereduceerd tot 49,082 unieke consumenten. Belangrijk om hieruit mee te nemen is dat wanneer data geanalyseerd worden, om hypothese te toetsen, de dataset gereduceerd kan worden tot een klein gedeelte van de beschikbare data. Deze reductie van data zorgt er overigens wel voor dat gedragswetenschappers big data kunnen samenvatten, analyseren en vertalen naar de maatschappij of organisaties. Ook is het belangrijk dat onderzoekers het juiste gereedschap gebruiken om grote hoeveelheden data te analyseren. Aangeraden wordt om gebruik te maken van een laptop of computer met minimaal 32 RAM GB werkgeheugen bij data ter grootte van 200,000 observaties. Dit is voornamelijk een vereiste voor het RF-model, want klassieke regressieanalyses kunnen 200,000 observaties analyseren met 8 GB werkgeheugen.

Een andere beperking van dit onderzoek is dat de variabele ‘aantal bestellingen’ invloed leek te hebben op het financiële gedrag van de consumenten. Vanwege de opbouw van de methodesectie was het van belang dat de controlevariabele geen relatie had met het financiële gedrag van de consumenten. Het aantal geplaatste bestellingen verschilde per consument. Zo had de ene consument in 2020 20 bestellingen geplaatst terwijl de andere consument vier bestellingen

had geplaatst. Een reden dat het aantal geplaatste bestellingen invloed had op het financiële gedrag kan zijn dat consumenten, naarmate consumenten vaker gebruik maakten van de keuze om later te betalen, door gewenning een eventuele boete als minder intimiderend ervaarden. Door deze gewenning worden consumenten lakser in het vroegtijdig betalen van hun bestelling. Hier kan verder onderzoek naar worden gedaan. Ook kan in de toekomst dezelfde hoeveelheid observaties per unieke consument, participant of kandidaat in een big dataset onderzocht worden.

Een laatste beperking aan dit onderzoek is het ontbreken van een effect grootte voor het verschil tussen RA en MLA (bijv. Cohens d). Voor dit onderzoek is geen analysemethode gevonden om de sterkte van het verschil tussen MLA en RA te toetsen (het verschil in effectgrootte tussen de twee modellen). Hierdoor blijft het onduidelijk of het gevonden significante verschil een groot of klein verschil was. Al met al kan wordt er geconcludeerd dat het MLA ietwat beter voorspeld dan de RA, maar op het oog lijken beide modellen evengoed inzetbaar om gedrag te voorspellen in een big dataset. Voor toekomstig onderzoek is het daarom belangrijk om analysemethodes te ontwikkelen die niet alleen het significantieverschil tussen modellen meet, maar ook de sterkte van dit verschil meet. Hierdoor kunnen onderzoekers statistisch onderbouwde inferenties maken over de grootte van het verschil van de voorspellende kracht tussen analysemethodes.

Implicaties

Een eerste implicatie van het onderzoek is het inzetten van de biodatamethode in big data. Wanneer er gedifferentieerd moet worden tussen goed en slecht presterende mensen kan onderzoek worden gedaan naar hun historische persoonsdata (Goldsmith, 1922). Een opmerking hierbij is dat er voldoende variabelen in de dataset aanwezig moeten zijn. Zo kunnen er manieren bedacht worden om meer variabelen te verzamelen in big datasets. Dit kan bijvoorbeeld door het toekennen van beloningen aan consumenten wanneer zij meer persoonlijke informatie vrijgeven via elektronische wegen. Een andere implicatie is dat er verder onderzoek kan worden gedaan naar SES en financieel gedrag. Het RF-model toonde namelijk aan dat SES van belang was voor het verklaren van financieel gedrag. Echter, het is verder onduidelijk waarom SES financieel gedrag verklaarde. Zonder het RF-model zou er geconcludeerd zijn dat er geen relatie was tussen SES en financieel gedrag, terwijl de resultaten van het RF-model aantoonden dat er wel degelijk iets aan de hand was. Daarbovenop kon MLA de gevonden resultaten vanuit de RA bevestigen (Figuur 1). Ook suggereerde MLA dat er indirecte- of interactie-effecten aanwezig waren in de

big data. Een volgende stap is om met RA meer complexe verbanden te onderzoeken.

Nu er bewijs is gevonden dat de biodatamethode bruikbaar is om gedrag te voorspellen in big data kunnen organisaties, managers of ondernemers in hun bestaande data de biodatamethode toepassen om het gedrag van hun consumenten of werknemers te onderzoeken. Daardoor kan er onderscheid worden gemaakt tussen goed en slecht presterende personen. Echter, voor onderzoek naar big data is enig consult van een data-analist nodig, want het transformeren en analyseren van big data vereist kennis en vaardigheden met betrekking tot programmeertalen zoals R of Python. Met enige training, bijvoorbeeld met behulp van ITers of nieuwe toevoegingen van keuzevakken op universiteiten, wordt het voor gedragswetenschappers en organisaties mogelijk om gebruik te maken van deze programmeertalen.

Conclusie

Uit dit onderzoek kan geconcludeerd worden dat leeftijd, het hebben van een financiële gewoonte, eerder vertoond financieel gedrag en ecologische overwegingen belangrijke voorspellers zijn voor financieel gedrag. Dit betekent dat biodata behulpzaam zijn bij het differentiëren tussen goed en slecht presterende consumenten binnen het big data-domein. De RA wist maar liefst 36.24% van de variantie in financieel gedrag te verklaren. Echter, tegen de verwachting in wist het MLA met dezelfde dataset net iets meer variantie in financieel gedrag te verklaren (39.99%). Ondanks de *black box* van MLA is uit de resultaten van dit onderzoek gebleken dat een MLA suggesties kan geven voor toekomstig psychologisch onderzoek. Bijvoorbeeld, gerelateerd aan het huidige onderzoek dat sociaaleconomische status een interessante voorspeller is voor toekomstig psychologisch onderzoek. Daarom kan er geconcludeerd worden dat een combinatie van wetenschappelijk onderbouwde hypothesen, klassieke regressieanalyses en *machine learning* algoritmes krachtige middelen zijn om in big data onderzoek te doen naar gedrag.

Literatuurlijst

Agrawal, R. (2020). Technologies for handling big data. *Handbook of Research on Big Data Clustering and Machine Learning*, 34–49. https://doi.org/10.4018/978-1-7998-01061.ch003

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision*

Processes, 50, 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Autio, M., Wilska, T.-A., Kaartinen, R., & Lähteenmaa, J. (2009). The use of small instant loans among young adults - A gateway to a consumer insolvency? *International Journal of Consumer Studies*, 33(4), 407–415. https://doi.org/10.1111/j.1470-6431.2009.00789

Balmer, N., Pleasence, P., Buck, A., & Walker, H. C. (2006). Worried sick: The experience of debt problems and their relationship with health, illness and disability. *Social Policy and Society*, 5(1), 39–51. https://doi.org/10.1017/S147474640500271X

Bamberg, S., Ajzen, I., & Schmidt, P. (2003). Choice of travel mode in the theory of planned behavior: The roles of past behavior, habit, and reasoned action. *Basic and Applied Social Psychology*, 25(3), 175–187. https://doi.org/10.1207/s15324834basp2503_01

Cobb-Clark, D. A., & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1), 11–15. https://doi.org/10.1016/j.econlet.2011.11.015

Centraal Bureau voor de Statistiek. (2021, 19 maart). Kerncijfers per postcode. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

Chander, A. (2017). The Racist Algorithm? *Michigan Law Review*, 115(6), 1023–1045. http://repository.law.umich.edu/mlr/vol115/iss6/13

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, 7, 1–13. https://doi.org/10.3389/fpsyg.2016.00738

Conner, M. (2020). Theory of planned behavior. *Handbook of Sport Psychology*, 1–18. https://doi.org/10.1002/9781119568124.ch1

Cook, M. (2016). *Personnel selection: adding value through people - A changing picture* (6de editie). Wiley-Blackwell.

Gravesteijn, B. Y., Nieboer, D., Ercole, A., Lingsma, H. F., Nelson, D., Van Calster, B., & Steyerberg, E. W. (2020). Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of Clinical Epidemiology*, 122(0895–4356), 95–107. https://doi.org/10.1016/j.jclinepi.2020.03.005

García-Izquierdo, A. L., Ramos-Villagrasa, P. J., & Lubiano, M. A. (2020). Developing biodata for public manager selection purposes: A comparison between fuzzy logic and

traditional methods. *Revista de Psicología del Trabajo y de las Organizaciones*, 36(3), 231–242. https://doi.org/10.5093/jwop2020a22

Goldsmith, D. B. (1922). The use of the personal history blank as a salesmanship test. *Journal of Applied Psychology*, 6, 149-155.

Gunasekaran, A., Marri, H. B., McGaughey, R. E., & Nebhwani, M. D. (2002). E-commerce and its impact on operations management. *International Journal of Production Economics*, 75(1–2), 185–197. https://doi.org/10.1016/s0925-5273(01)00191-8

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology*, 8(4), 491–508. https://doi.org/10.1017/iop.2015.40

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4), 447–457. https://doi.org/10.1037/met0000120

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). *Journal of Research in Personality*, 44(4), 501–511. https://doi.org/10.1016/j.jrp.2010.06.005

Kvasova, O. (2015). The big five personality traits as antecedents of eco-friendly tourist behavior. *Personality and Individual Differences*, 83, 111–116. https://doi.org/10.1016/j.paid.2015.04.011

Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2009). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, 40(6), 998–1009. https://doi.org/10.1002/ejsp.674

Lee, S., & Dalal, R. S. (2014). Climate as situational strength: Safety climate strength as a cross-level moderator of the relationship between conscientiousness and safety behaviour. *European Journal of Work and Organizational Psychology*, 25(1), 120–132. https://doi.org/10.1080/1359432x.2014.987231

Letkiewicz, J. C., & Heckman, S. J. (2019). Repeated payment delinquency among young adults in the united states. *International Journal of Consumer Studies*, 43(5), 417–428. https://doi.org/10.1111/ijcs.12522

Max Kuhn (2020). *Caret: Classification and regression training.* R Package Version 6.0-86. https://CRAN.R-project.org/package=caret

Mello, R., Leite, L. R., & Martins, R. A. (2014). Is big data the next big thing in performance measurement systems? *IIE Annual Conference.Proceedings*, 1837-1846. https://www-proquest-com.eur.idm.oclc.org/scholarly-journals/is-big-data-next-thing-performance-measurement/docview/1622307688/se-2?accountid=13598

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python.* Van Duuren Media.

Piros, P., Ferenci, T., Fleiner, R., Andréka, P., Fujita, H., Főző, L., Kovács, L., & Jánosi, A. (2019). Comparing machine learning and regression models for mortality prediction based on the hungarian myocardial infarction registry. *Knowledge-Based Systems*, 179, 1–7. https://doi.org/10.1016/j.knosys.2019.04.027

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, https://www.R-project.org/

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass*, 15(2), 1–25. https://doi.org/10.1111/spc3.12579

Rustichini, A., DeYoung, C. G., Anderson, J. E., & Burks, S. V. (2016). Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation. *Journal of Behavioral and Experimental Economics*, 64, 122–137. https://doi.org/10.1016/j.socec.2016.04.019

Wang, L., Lu, W., & Malhotra, N. K. (2011). Demographics, attitude, personality, and creditcard features correlate with creditcard debt: A view from China. *Journal of Economic Psychology*, 32(1), 179–193. https://doi.org/10.1016/j.joep.2010.11.006

Webley, P., & Nyhus, E. K. (2001). Life-cycle and dispositional routes into problem debt. *British Journal of Psychology*, 92(3), 423–446. https://doi.org/10.1348/000712601162275

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–

1122. https://doi.org/10.1177/1745691617693393

Yildiz, B., Bilbao, J., & Sproul, A. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, 73, 1104–1122. https://doi.org/10.1016/j.rser.2017.02.023

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Yurtkur, A. K., & Bahtiyar, B. (2020). An empirical study on the relationship between economic growth and e-commerce. *Tools and Techniques for Implementing International E-Trading Tactics for Competitive Advantage*, 71–86. https://doi.org/10.4018/978-1-7998-0035-4.ch004