Nederlandse samenvatting

Dit proefschrift onderzoekt de mogelijkheden van het gebruik van voorspellingsmodellen op justitiële strafzaakdata. Deze strafzaakdata wordt geput uit de onderzoeks- en beleidsdatabase justitiële documentatie, de OBJD. Sinds 1999 is, met de OBJD als bronbestand, de recidivemonitor ontwikkeld. In de recidivemonitor wordt de data verder veredeld zodat het geschikt is voor gestandaardiseerd (recidive-)onderzoek. De recidivemonitordata wordt al sinds 2005 gebruikt om de landelijke recidive te rapporteren. Gerapporteerde ruwe recidive is een nuttige statistiek, maar kan ook leiden tot misinterpretaties. Bijvoorbeeld, als de recidive van een bepaalde interventie wordt vergeleken met die van een niet-willekeurig verkregen vergelijkbare groep die geen of een andere interventie heeft gekregen, is het verleidelijk het verschil te interpreteren als effectiviteit. Echter, als de daders die de interventie hebben gekregen een meer omvangrijke strafrechtelijke carrière hebben gehad dan diegenen die dat niet hebben gehad, kunnen ze puur op dat instroomverschil een hogere recidive hebben, ongeacht de effectiviteit.

Ook bij het rapporteren van recidive over de tijd, kunnen variaties in instroom in verschillende jaren de werkelijke recidivetrend vertroebelen. Om te corrigeren op instroomverschillen is men al snel aangewezen op vormen van (statistisch) modelleren, waarbij het unieke effect van achtergrondvariabelen uitgepartionaliseerd wordt.

Een andere mogelijkheid die de strafzaakdata van de recidivemonitor bieden, is het inschatten van een individuele recidivekans. Omdat niet ieder kenmerk evenveel invloed op de uitkomst zal hebben, zal er een vorm van weging per kenmerk plaats moeten vinden. Om deze gewichten optimaal te schatten is men eveneens aangewezen op het gebruik van een — al dan niet — statistisch model. Voor dit doel wordt in het veld van risicotaxatie bijna exclusief logistische regressie gebruikt. Dit is eveneens gebruikt bij het construeren van de Statrec-schaal, die onderzocht wordt in hoofdstuk 2. Voor het opstellen van een logistisch regressiemodel of een ander statistisch model moet de onderzoeker er zelf voor zorgen dat een model juist gespecificeerd is. Indien er sprake is van nonlineariteit en interactie-effecten moeten deze handmatig gespecificeerd worden, waarbij interacties veelal beperkt zijn tot multiplicatieve interactietermen. Sinds de jaren zestig zijn er echter modellen ontwikkeld die automatisch nonlineaire verbanden schatten en/of automatisch interactie-effecten schatten. Deze modellen, veelal ontwikkeld in de computerwetenschappen, specifiek de subgebieden data mining en artificiële intelligentie, zijn veelal geschikt om om te gaan met rommelige data en/of data waarbij er meer variabelen zijn dan datapunten. Deze modellen/algoritmen moeten wel meestal getuned ('ingeregeld') worden voor iedere nieuwe dataset, door middel van het vastzetten van één of meerdere tuningparameters. Deze technieken houden de belofte in zich om verbeterde voorspellingen op te leveren, tegen de kosten van verminderde interpreteerbaarheid en

transparantie van het resulterende model. In hoofdstuk 3 en 4 wordt op de Nederlandse strafrechtgegevens getest of deze technieken een verbetering op kunnen leveren in de voorspelling, enerzijds in het geval waar een binaire uitkomst is (recidive ja/nee, hoofdstuk 3 en 4), dan wel een recidivekans over de tijd (survival, hoofdstuk 4).

Het maken, herontwerpen en testen van een risicotaxatie-instrument
In dit onderzoek wordt in hoofdstuk 2 een recidivevoorspellingmodel gepresenteerd dat de basis vormt van schaal 1 van een risicotaxatie-instrument, de Quickscan. Dit model is de inschatting van het statische 4-jaarsrecidiverisico, dat omgevormd wordt in een score van 0 tot 100. De StatRec-schaal werd al jaren toegepast in de reclasseringspraktijk in de Quickscan. Dit instrument wordt gebruikt om per individu tot de beslissing te komen of een uitgebreidere risicotaxatie met het de RISc nodig is. Omdat de schaal al jaren ongewijzigd werd gebruikt, ontstond de vraag of dat het instrument geupdate moest worden en hoe vaak dit zou moeten gebeuren. In dit hoofdstuk wordt gekeken of en in hoeverre de schaal zijn voorspellende waarde behoudt over de tijd (d.i. temporele validatie). Ook wordt gekeken of de schaal geen last heeft van lokale effecten door te valideren naar arrondissement (d.i. geografische validatie). Omdat de schaal soms toegepast zal worden om het recidiverisico van andere populaties in te schatten dan alle volwassen daders, wordt er ook een zogenaamde substeekproefvalidatie gedaan. Hierin wordt gekeken voor niet-willekeurige selecties van data of de voorspellende waarde voldoende is.

De StatRec-schaal blijkt goed te generaliseren over de tijd. Het percentage juist geclassificeerd en de mate van discriminatie blijkt niet te veranderen als er meer recente data gebruikt wordt. Enkel de calibratie (de correspondentie tussen de geobserveerde en geschatte kans) blijkt gemiddeld een lichte afwijking (onderpredictie) te vertonen. De schaal blijkt nauwelijks variatie te tonen tussen arrondissementen. In substeekproeven doet de schaal het ook goed, met uitzondering van de groepen die extreem op het aantal eerdere justitiecontacten scoort: de first offenders en daders met 11 of meer justitiecontacten.

Qua performance is de schaal sterk vergelijkbaar met de OGRS (Offender Group Reconviction Scale) van het Verenigd Koninkrijk. Het toevoegen van dynamische (d.i. veranderlijke) data aan de StatRec-schaal, zoals burgerlijke staat, woonsituatie en dagbesteding blijken maar zeer beperkt de voorspelbaarheid van recidive te verbeteren, net als etniciteit en verblijfsstatus. Omdat de Statrec-schaal een relatief goede voorspelling genereert, over de tijd, plaats en subgroepen, kan de schaal gebruikt worden om schalen die ontwikkeld worden in het forensische veld te benchmarken.

Het verbeteren van voorspelkracht met moderne technieken
In dit hoofdstuk wordt getracht om de voorspelkracht van de StatRec-schaal te verbeteren door gebruik te maken van technieken uit het veld van machinaal leren (*machine learning*), predictieve data mining en moderne statistiek. Daarnaast wordt gekeken of het mogelijk is om Statrec-schalen te ontwikkelen voor gewelddadige en seksuele recidive. Het standaardmodel, logistische regressie, wordt vergeleken op een brede selectie van indicatoren die de mate van voorspelkracht kwantificeren met lineaire discriminantanalyse en meer geavanceerde technieken. De lijst geavanceerde modellen omvat multivariate adaptieve regressiesplines (MARS), flexibele discriminantanalyse, beslisbomen, neurale netwerken, adaptieve boosting, logitBoost, lineaire support vector machines, *k*-nearest neigbors en partiële kleinste kwadraten.

Voor algemene en gewelddadige recidivevoorspelling blijkt er geen verbetering mogelijk op een vooraf gespecificeerd logistisch regressiemodel. Bij seksuele recidive blijkt lineaire discriminantanalyse meer geschikt te zijn. Uit de studie blijkt dat, als er vooraf met nonlineariteit rekening wordt gehouden door continue variabelen te transformeren, automatische moderne methoden op deze data geen substantiële verbetering van voorspelkracht opleveren.

Het verbeteren voorspelkracht in data met binaire en survivaluitkomst
In dit hoofdstuk wordt geprobeerd op data met een binaire recidiveuitkomst (recidive ja/nee na 4 jaar) verbetering te vinden van voorspelkracht door de set technieken gebruikt in hoofdstuk 4 uit te breiden met random forests, L_1- en L_2-penalized logistische regressie, L_2-penalized discriminantanalyse, stochastische gradient boosting en Bayesiaanse additieve regressiebomen. Daarnaast wordt er gekeken welk soort klassiek/modern model het beste toegepast kan worden voor survivalanalyse. Er wordt een breed scala aan statistische modellen getest op gecensureerde recidivedata. Dit laatste houdt in dat niet van iedere observatie er een complete duur gemeten is. Omdat er weinig bekend is over de voorspelkracht van machinaal leren op survivaldata, worden generalisaties van technieken voor binaire uitkomstdata naar survivaldata vergeleken met de statistische modellen. Om de generaliseerbaarheid van de resultaten van de studie te verhogen, wordt er ook een historische Amerikaanse dataset gebruikt, de North Carolina Prison data. Wat betreft de binaire uitkomstdata blijkt ook met de uitgebreidere set methoden dat er geen verbetering kan worden gevonden ten opzichte van klassiek statistische methoden, zowel op de OBJD-data als de North Carolina Prison data, met uitzondering van de seksuele-recidivedata, waarbij L_1-penalized logistische regressie het hoogste percentage juist geclassificeerd behaalde.

Bij de gecensureerde data is er bij de North Carolina prison data één model gevonden dat het beter doet dan de overige modellen, namelijk het stochastische gradient boosting survivalmodel. Anders dan bij het binaire uitkomstgeval lijkt er bij survivalmodellen nog wel winst te behalen door verbeterde algoritmen. Om te controleren of de performance van de modellen uit de machine learning niet het gevolg is van niet-optimale tuningparameters, is een volautomatisch algoritme gebruikt dat tegelijkertijd de tuning parameters en de performance optimaliseert. Dit algoritme was ook niet in staat om de performance te verbeteren ten opzichte van de met de hand getunede modellen. Deze automatische techniek was echter alleen mogelijk met de binaire-uitkomstdata.

In gevallen van gelijke predictieve performance zou dan voor (klassieke) statistische methoden gekozen moeten worden omdat zij op andere vlakken dan performance voordelen hebben. Het belangrijkste is dat zij meer transparant zijn in hoe ze met de data tot een voorspelling komen. Hierdoor is gemakkelijker te controleren wanneer een model voorspellingen oplevert die onzinnig zijn.

Het schatten van het effect van een interventie op observationele data met veel missende gegevens

Dit hoofdstuk gaat in op de praktische toepassing van effectschatting van de ISD-maatregel, een maximaal twee jaar durende maatregel voor veelplegers. De veelplegers zijn een groep gekarakteriseerd door een hoge pleegfrequentie en relatief lichte delicten. Het was nog onbekend hoe groot het effect van de ISD-maatregel zou zijn op recidive na ontslag uit de inrichting was. Ook was er vanwege hun hoge pleegfrequentie een incapacitatie-effect van de maatregel te verwachten van onbekende hoogte. Omdat er geen willekeurige toewijzing van veelplegers naar de maatregel plaatsvindt, is het lastig om een niet-vertekende effectschatting te verkrijgen. Een voor de hand liggende strategie is om aan iedere ISD'er een veelpleger te matchen op zoveel mogelijk kenmerken die samenhangen met de uitkomst aan een veelpleger die de standaardbehandeling heeft gekregen, zijnde een korte gevangenisstraf. Hierdoor wordt de vergelijkbaarheid van de groepen op voorhand gemaximaliseerd.

Het aantal ISD-maatregelen is relatief klein ten opzichte van de totale doelgroep, de populatie zeer actieve volwassen veelplegers. In dit geval is er een groot aantal potentiële kandidaten die aan de ISD-groep gematcht kunnen worden op achtergrondkenmerken. Echter, zelfs met een grote matchingsgroep is het lastig een perfecte match te vinden als het aantal kenmerken groot is (in dit onderzoek 20). In dit geval is *propensity score matching* (PSM) een voor de hand liggende techniek om tot een effectschatting te komen.

Een extra complicatie ontstaat wanneer er relatief veel missende waarden op achtergrondkenmerken zijn; PSM kan dan niet zomaar toegepast worden. In dit onderzoek wordt gekozen om multipele imputatie - een vorm van imputatie waarbij meerdere complete datasets worden gegenereerd uit voorspellingsmodellen voor iedere variabele met ontbrekende waarden - te combineren met PSM. Een zwakte van PSM is dat er, zelfs bij een groot aantal relevante matchingskenmerken, sprake kan zijn van *confounding*. Dit houdt in dat er verschillen tussen de behandelgroep en de controlegroep blijven bestaan op niet gemeten kenmerken, die het verschil in de uitkomstvariabelen kunnen verklaren, anders dan het behandeleffect.

Daarom is in dit onderzoek ook geprobeerd om de bruikbaarheid van de combinatie van PSM en difference-in-difference (DD) te beoordelen. Bij DD mag er overgebleven vertekening overblijven in de voormeting van de uitkomst, onder de assumptie dat de vertekening constant over de tijd is en de hellingshoek van de lijn tussen de voor- en nameting gelijk zou zijn geweest als er geen behandeling was geweest. Uit het onderzoek blijkt dat het mogelijk is om een effectschatting te maken van het effect van de ISD-maatregel en van haar incapacitatie-effect. Echter, voor het schatten van het effect op recidivefrequentie blijkt dat men rekening moet houden met een regressie naar het gemiddelde. Omdat de (op de voormeting) meest extreem frequent plegende veelplegers geselecteerd worden voor de ISD, is het op grond van toevalsfluctuatie te verwachten dat ze lager zullen scoren op de nameting. Bij PSM is dat waarschijnlijk geen probleem, omdat er gematcht werd op de hoogte van voormeting. Bij PSD-DD is dat wel een probleem omdat verschillen op de voormeting mogen blijven bestaan; daardoor wordt waarschijnlijk een artefact gemeten. Daarom is de conclusie dat er geen effect is op recidivefrequentie.

Conclusie

Het is mogelijk om op basis van strafzaakregistraties een redelijk goede voorspelling te doen van de recidivekans, zoals de StatRec-schaal laat zien. Deze voorspelling is ook redelijk robuust over tijd en plaats. Wel is er voorzichtigheid geboden bij daders die niet of nauwelijks een strafrechtelijke voorgeschiedenis hebben, zoals first offenders. Omdat het aantal eerdere justitiecontacten en strafzaakdichtheid de sterkste voorspellers zijn, ontstaat bij deze groep een informatiegat. Ook bij daders die juist veel recidiveren is het lastig om recidivisten en niet-recidivisten uit elkaar te houden.

Het gebruik van computerintensieve methoden om de voorspelling te verbeteren van voorspellingsmodellen op Nederlandse strafzaakdata leverde weinig tot geen winst op. De voorspelling is meestal nagenoeg gelijk aan die van een handmatig gespecificeerd statistisch model. Daarom lijkt het vinden van meer en betere predictoren een meer vruchtbare weg om de voorspelling van recidive te verbeteren.

Bij gelijke voorspelkracht zou een statistisch model de voorkeur moeten krijgen boven een computerintensief model omwille van de transparantie. Een transparant model kan eenvoudiger gecheckt worden op vreemde resultaten en het is duidelijk hoe een model tot een specifieke voorspelling komt. Ook zal er dan minder weerstand zijn in de praktijk, omdat een model eerder vertrouwd zal worden. Wel zou een automatisch computerintensief model routinematig toegepast moeten worden om te zien of de modelspecificatie adequaat genoeg is. Voor kleine steekproeven, die geregeld voorkomen in het forensische risicotaxatieveld, zou standaard lineaire discriminantanalyse ook uitgeprobeerd moeten worden, ook al voldoen de data duidelijk niet aan de statistische aannames. Dit model bleef namelijk overeind in de seksuele-recidivedata, terwijl logistische regressie de data ging overfitten. Dit is een relevant resultaat voor het risicotaxatieveld, waarin bijna enkel logistische regressie bijna exclusief wordt toegepast. Een ander alternatief voor deze data is L_1-penalized logistische regressie. Deze zorgt ervoor dat de coefficienten van de covariaten met de minste informatie over de voorspelling naar nul worden gekrompen. Dit heeft echter wel sterke gevolgen voor de calibratie.

Ook al is de voorspelkracht van verschillende modellen op geaggregeerd niveau hetzelfde, ze kunnen op voorspellingen voor individuen substantieel uiteenlopen. Ook op Nederlandse strafzaakdata kan dit erop wijzen dat verschillende modellen optimaal zijn voor verschillende (groepen) individuen. Het maken van een ensemble van verschillende modellen zou in dat geval verbeterde voorspellingen op kunnen leveren. Echter, dan wordt er wel transparantie opgeofferd en het is de vraag of de praktijk de voorspellingen van een dergelijk gemengd model wel zal vertrouwen.

Het is mogelijk gebleken om een effectschatting van een justitiële maatregel verkrijgen met behulp van observationele data. Het gebruik van gekoppelde bestanden heeft ervoor gezorgd dat er een groot aantal matchingskenmerken beschikbaar was, waardoor er minder zorg is dat er vertekening in de effectschatting plaatvindt. Omdat de betreffende interventie, de ISD-maatregel, wordt toegepast op de meest extreem 'scorende' veelplegers, moet er echter wel worden gecontroleerd voor de uitkomstmaat op de voormeting. Dit om eventuele vertekening door regressie naar het gemiddelde tegen te gaan.