Studentenuitval in bachelor 1 bij Erasmus School of Economics

Rapport Project Data & Business Analytics

- Pieter Vreeburg (193315pv)
- Roeland van der Molen (259280rm)

Introductie

Studentenuitval in bachelor 1 is een bekend probleem voor veel universiteiten. Nieuwe studenten zijn niet altijd even goed voorbereid op de eisen die de universiteit qua kennis en studievaardigheden aan hen gaat stellen, hebben moeite om lopende hun studie hun kennis en vaardigheden voldoende bij te spijkeren en lopen een verhoogd risico op uitval. Vanuit Erasmus School of Economics (ESE) is in de afgelopen jaren onderzoek gedaan naar de oorzaken van studentenuitval (Arnold (2014), Arnold (2012)) en zijn beleidsmaatregelen getroffen om uitval in het 1e jaar zoveel mogelijk te voorkomen.

Uit onderzoek naar studentenuitval binnen het economisch onderwijs komen een aantal variabelen naar voren die van invloed blijken te zijn op het risico van studie-uitval: genoten vooronderwijs (m.n. wiskunde onderwijs), intrinsieke motivatie (m.n. bij onvoldoende wiskunde onderwijs) (beiden Arnold & Straten, 2012) en geslacht (in mindere mate, Arnold & Rowaan, 2014).

Binnen Erasmus Universiteit Rotterdam is voor onderzoek naar studierendement de 'Onderzoeksdatabase Onderwijskwaliteit en Studiesucces' beschikbaar. Deze dataset is door leverancier RISBO opgebouwd voor onderzoek naar de mogelijke effecten van de 'nominaal = normaal' regeling (2013) en is sinds die tijd in actief beheer. De dataset bevat voor meerdere cohorten studenten persoonlijke gegevens en gegevens over de studievoortgang. Eén van de geregistreerde persoonlijke gegevens betreft de school waar de student zijn / haar vooropleiding heeft genoten. Dit gegeven stelt ons in staat om de dataset te combineren met een aanvullende set gegevens over de 'toeleverende scholen' (o.a. oordelen van de onderwijsinspectie, grootte, etc.) om op deze manier meer inzicht te krijgen in de invloed van vooropleiding op studentenuitval in bachelor 1.

Onderzoeksvragen

In dit project willen wij door het inzetten van algoritmen uit het machine-learning domein onderstaande onderzoeksvragen beantwoorden:

a) Welke variabelen uit de uitgebreide dataset dragen volgens ieder algoritme het meest bij aan het voorspellen van studententuitval? In hoeverre verschillen deze verzamelingen variabelen van de uit de literatuur en praktijk bekende verzameling variabelen?
b) Hoe generaliseren de voor vraag A getrainde modellen naar ongeziene data? Welk model levert de beste voorspellingen?
c) Welke van de voor A getrainde modellen is het eenvoudigst interpreteerbaar? Welke van de modellen is niet of nauwelijks interpreteerbaar? Welke balans bestaat er tussen kwaliteit van de voorspellingen enerzijds en de mate van interpreteerbaarheid van het model anderzijds?
d) Welke maatregelen kan ESE op basis van de bij A, B en C gegeven antwoorden nemen om studentenuitval in het eerste jaar te voorkomen.

Data

De ‘Onderzoeksdatabase Onderwijskwaliteit en Studiesucces’ vormt de basis van dit onderzoek. Sommige variabelen worden tijdens het 1e studiejaar opgebouwd (bijv. cursusresultaten en rendement), aangezien wij uitval willen voorspellen op basis van bij de start van het collegejaar bekende variabelen, verwijderen wij het eerstgenoemde type variabele uit de dataset. Onvolledig geregistreerde variabelen (bijv. voor maar 1 cohort) worden ook uit de dataset verwijderd.

Als afhankelijke variabele gebruiken wij het aan het eind van het eerste collegejaar afgegeven bindend studieadvies (BSA). De waarden 'Positief', 'Aangehouden / voorlopig positief' en ‘Persoonlijke omstandigheden’ zijn daarbij gecodeerd als de positieve klasse, 'Negatief' en '1 Feb-staker' (zelfgekozen uitval voor 1 februari van het betreffende collegejaar) zijn gecodeerd als de negatieve klasse. Observaties die niet in één van deze klassen vallen zijn uit de dataset verwijderd. De privacy van de geregistreerde studenten wordt beschermd door het volledig anonimiseren van het geregistreerde unieke studentennummer (door middel van salting & hashing). De resulterende dataset (ESE dataset) bevat 5332 observaties (studenten) met 13 bijbehorende kenmerken.

Met behulp van het zogenaamde BRIN-nummer (Basisregistratie Instellingen, 4 posities alfanumeriek) kan deze dataset worden gekoppeld aan openbare data over de aan ESE leverende scholen. Hiervoor verzamelen wij de volgende aanvullende gegevens van de websites van DUO en de onderwijsinspectie:

- Gemiddelde examencijfers per vak (verplichte vakken binnen de 4 profielen & de moderne talen)
- Slagingspercentage per school
- Totaal aantal leerlingen per school
- Leerling / docent ratio per school
- Docent full-time / part-time ratio per school
- Oordeel onderwijsinspectie
- Type school (VWO, HAVO/VWO of VMBO/HAVO/VWO)

Alle informatie van DUO is te vinden op: https://www.duo.nl/open_onderwijsdata/databestanden/. Daar hebben we de volgende bestanden vandaan gehaald.

- Eindexamencijfers en Schoolonderzoekcijfers voor alle scholen (schooljaren 2012-2013 t/m 2016-2017). Ieder schooljaar is één CSV-bestand. In totaal vijf bestanden.
- Beoordelingen Schoolinspectie (schooljaren 2012-2013 t/m 2016-2017). Ieder schooljaar is één CSV-bestand. In totaal vijf bestanden.
- Aantal leerlingen per school (schooljaren 2012-2013 t/m 2016-2017). Ieder schooljaar is één CSV-bestand. In totaal vijf bestanden.
- Slagingspercentages, aantal kandidaten en aantal geslaagden (schooljaren 2012-2013 t/m 2016-2017). Alles stond in één CSV bestand.
- Aantal fulltime/part-time docenten. (schooljaren 2012-2013 t/m 2016-2017). Alles stond in één CSV bestand.

We hebben de informatie uit de bovenstaande DUO-bronnen bewerkt. Deze tijdrovende ‘datawrangling’ stap heeft veel tijd gekost. Uiteindelijk is hier één bestand uit gekomen waarbij iedere rij/observatie bestaat variabelen voor uit één BRIN voor één schooljaar. De meeste scholen komen dus vijf keer terug, voor ieder jaar één keer. Dit leverde een probleem op bij de koppeling met de studenten uit de ESE-data, omdat er maar weinig studenten in de ESE-set zitten die

eindexamen hebben gedaan in de schooljaren 2012-2013, 2013-2014 en 2014-2015. Zo verloren wij veel observaties om de onderzoeksvragen mee te beantwoorden.

Daarom hebben we de aanvullende gegevens geaggregeerd op alleen BRIN-nummer en gekoppeld aan de onderzoeksdatabase. Voor de deel van de studenten in de onderzoeksdatabase is geen BRIN-nummer geregistreerd. Deze studenten worden in dit onderzoek buiten beschouwing gelaten. Na het koppelen van de aanvullende gegevens en het verwijderen van studenten zonder geregistreerd BRIN-nummer bevat de verrijkte dataset 5058 observaties met 39 bijbehorende kenmerken.

Onderzoeksmethode

Om de hierboven beschreven onderzoeksvragen te beantwoorden hebben wij ons onderzoek in onderstaande stappen ingedeeld.

Eerste selectie onafhankelijke variabelen

Voor een eerste selectie van de onafhankelijke variabelen trainen wij een Random forest en een Lasso-regressie op zowel de ESE dataset als op de verrijkte dataset. In beide gevallen houden wij een deel van de dataset achter als test set. Wij selecteren de dataset om mee verder te gaan op basis van de kwaliteit van de door de modellen gemaakte voorspellingen van de testdata. Onze gedachtegang hierbij is dat het toevoegen van (mogelijk relevante) onafhankelijke variabelen ook betere voorspellingen moet opleveren, als de kwaliteit van de voorspellingen afneemt heeft het verrijken van de ESE dataset geen informatie toegevoegd en is het verstandiger om alleen met de ESE dataset door te gaan.

Binnen de gekozen dataset selecteren wij de belangrijkste onafhankelijke variabelen met behulp van een 'Variable Importance Plot' (VIP) van het Random forest algoritme en de door de Lasso-regressie gerapporteerde coëfficiënten.

Trainen

Na de eerste selectie van de onafhankelijke variabelen trainen wij de volgende voorspellende algoritmen op een deel van de gekozen dataset:

- K-nearest neighbours (KNN): dit eenvoudige algoritme vormt de benchmark voor de volgende algoritmes. Een nadeel is dat dit algoritme geen interpreteerbaar model oplevert. Desalniettemin vinden wij (met het zuinigheidsprincipe in ons achterhoofd) dat de meer complexe algoritmen KNN overtuigend moeten verslaan om de 'kosten' van hun hogere complexiteit waard te zijn. Bij het trainen kiezen wij de K hyperparameter van dit algoritme voor een zo optimaal mogelijk foutpercentage (m.b.v. 10-fold crossvalidatie).
- Random forest: wij verwachten dat dit algoritme betere voorspellingen op zal leveren dan KNN. Het VIP maakt enige interpretatie van het resulterende model mogelijke. Bij het trainen controleren we of het standaard aantal bomen (500) voldoende is en kiezen het optimale aantal onafhankelijke variabelen per boom (de MTRY hyperparameter) aan de hand van het zogenaamde 'out of bag' foutpercentage dat voor verschillende waarden van deze parameter (+/- 2 rondom de standaard waarde) wordt gerapporteerd.
- Lasso regressie: dit algoritme (GLM-net met een alpha-parameter van 1 en 10-fold cross validation) willen wij gebruiken omdat dit individueel interpreteerbare coëfficiënten genereert, terwijl het ook een goede selector is de coëfficiënten die de meeste invloed hebben op de afhankelijke variabele. We verwachten dat de Lasso niet beter zal voorspellen dan de Random Forest. We trainen het model op 70% van de dataset (trainingset) en voorspellen daarmee op de overige 30% (testset).
- Decision tree: wij verwachten dat de voorspellingen van dit algoritme minder goed zullen zijn dan de door Random forest en Lasso-regressie gemaakte voorspellingen. De eenvoud waarmee het resulterende model kan worden geïnterpreteerd vinden wij echter een belangrijk voordeel van dit algoritme. Een nadeel van dit algoritme is de hoge mate van variabiliteit, een andere splitsing in een train- en testset kan een andere beslisboom opleveren. Wij proberen dit (deels) te ondervangen door de beslisboom alleen te trainen op de variabelen die door de Random forest en Lasso-regressie algoritmes als belangrijk zijn

aangemerkt. Bij het trainen gebruiken wij de ‘1 standaard error’ regel om de boom te snoeien tot de optimale omvang.

Voorspellen

Na de trainingsfase evalueren wij de modellen op de volgende punten:

- Kwaliteit van de voorspellingen op de testset, onder andere op basis van Cohen’s Kappa, accuracy, sensitivity en specificity.
- Over- of underfit door het vergelijken van de kwaliteit van de voorspellingen op de train- en testsets.

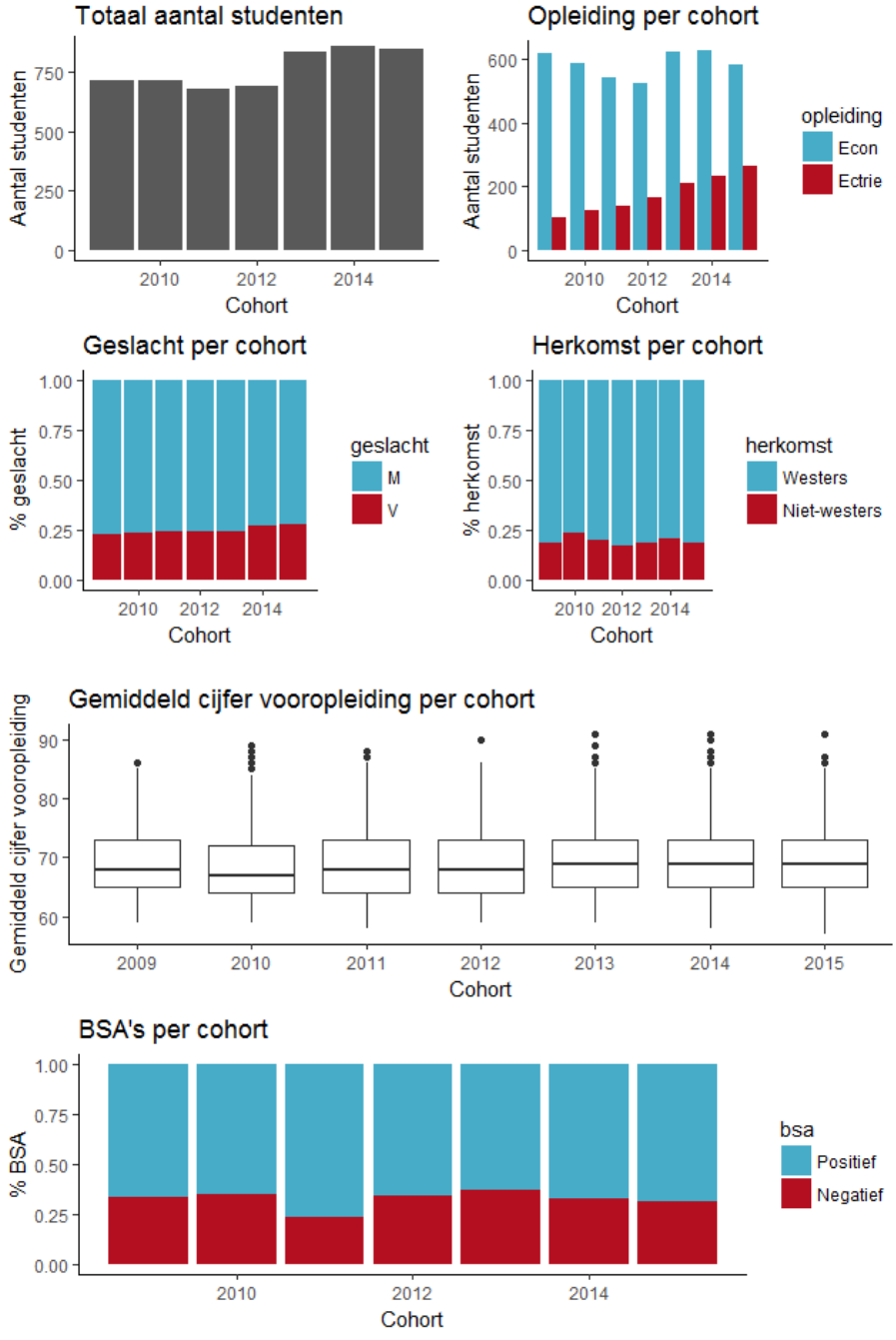
Interpreteren

Wij sluiten ons onderzoek af met een antwoord op de eerder gestelde onderzoeksvragen. Hierbij proberen wij een balans te vinden tussen de enerzijds de voorspelkracht en anderzijds de interpretabiliteit van de gemaakte modellen. Mogelijk complementeren de verschillende modellen elkaar, het ene model levert een bijdrage aan het ‘conceptuele model’ van het probleem in het hoofd van de beleidsmaker of bestuurder, het andere model komt op basis van dezelfde data tot een serie goede voorspellingen.

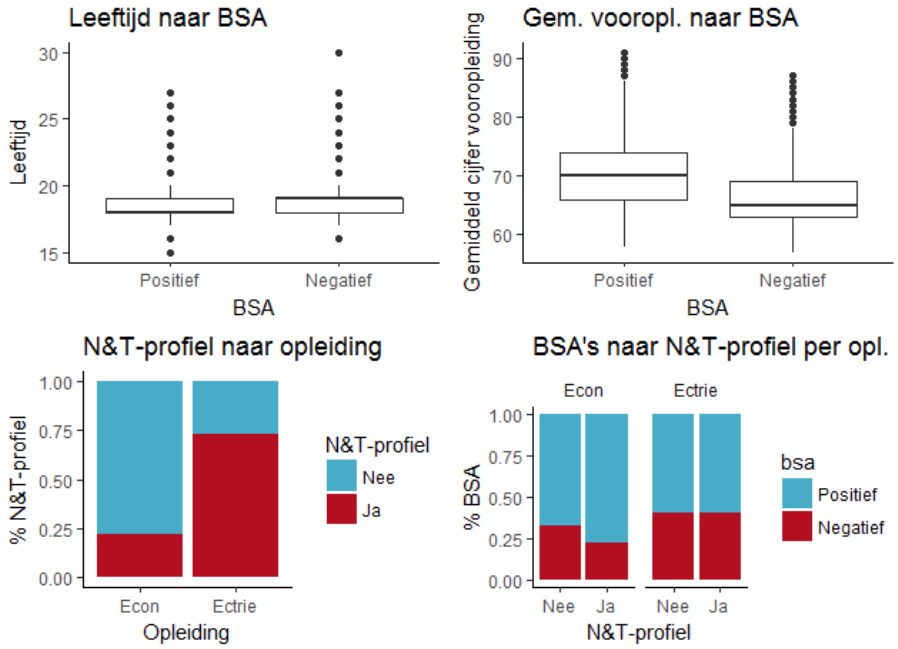
Resultaten

Beschrijving

De ESE dataset bevat 5058 observaties verdeeld over 7 cohorten (2009 – 2015). Tot 2013 was het aantal eerstejaarsstudenten redelijk stabiel op 700. In 2013 is dit aantal gegroeid tot 850 studenten en daarna gestabiliseerd. Het aantal studenten bij Econometrie vertoont een gestaag stijgende lijn, het aantal studenten bij Economie is redelijk stabiel. Het aandeel vrouwen per cohort is zeer licht gestegen. Het aandeel niet-westerse allochtonen per cohort is redelijk stabiel. Het grootste deel van de studenten heeft een gemiddeld (VWO-)eindexamencijfer tussen 6.5 en 7.3. Het aandeel negatieve BSA's aan het eind van jaar 1 (incl. voortijdige gestopte studenten) schommelt rond 33% per cohort.



Uit bestaand onderzoek is bekend dat het gemiddelde vooropleidingscijfer, de leeftijd en het gevolgde wiskunde B onderwijs van invloed zijn op de kans op uitval in het eerste jaar. Onderstaande grafieken tonen deze verbanden in een bivariate context. In deze dataset wordt het gevolgde Wiskunde B onderwijs (bij gebrek aan een directe variabele) gecodeerd door middel van de variabele 'Natuur & techniek profiel'. Wiskunde B is een verplicht onderdeel van dit middelbare school profiel.

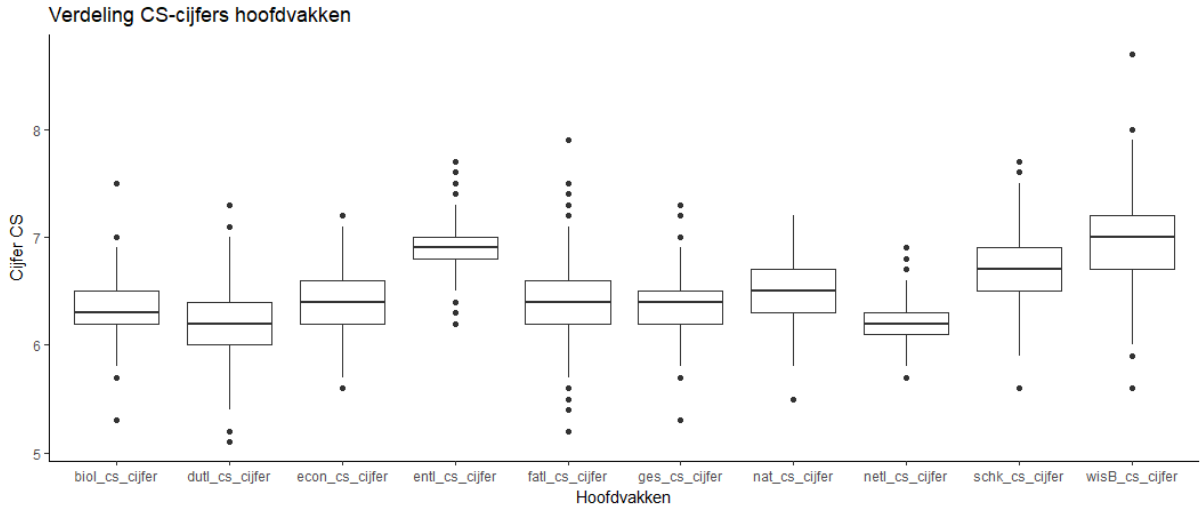


De verdeling van de N&T-profiel variabele en het mogelijke effect van deze variabele op het uitgebrachte BSA zijn interessant. Qua verdeling op deze variabele lijken de twee groepen studenten van positie te wisselen, waar bij de ene groep ongeveer 75% geen N&T-profiel blijkt hebben, heeft bij de andere groep ongeveer 75% juist wel een N&T profiel. Bij de economiestudenten lijken de studenten met een N&T-profiel vaker een positief advies te krijgen dan de studenten zonder N&T-profiel, maar bij de econometriestudenten speelt deze variabele geen rol van betekenis. Een mogelijke verklaring voor het laatste is dat Econometriestudenten zonder N&T-profiel wel Wiskunde B als keuzevak in hun middelbare school pakket hadden opgenomen.

De schooldataset bestaat uit 460 observaties. Iedere observatie heeft het unieke BRIN-kenmerk. Iedere BRIN vertegenwoordigt een school of scholengemeenschap. De hele set is een gewogen aggregatie van metingen uit vijf schooljaren: 2012-2013 tot en met 2016-2017. In het hoofdstuk 'data' hebben we al toegelicht waarom we dit hebben gedaan.

Resultaten Centraal Schriftelijk: tien vaste vakken

We hebben alleen de vakken overgehouden die verplicht zijn bij één van de mogelijke profielen die er zijn. Daardoor blijven er dus tien hoofdvakken over. Het valt op dat alle klassieke exacte vakken (natuurkunde, scheikunde en wiskunde B) gemiddeld beter worden gemaakt dan bijvoorbeeld de talen (met uitzondering van Engels), maar tegelijkertijd de meeste spreiding kennen.

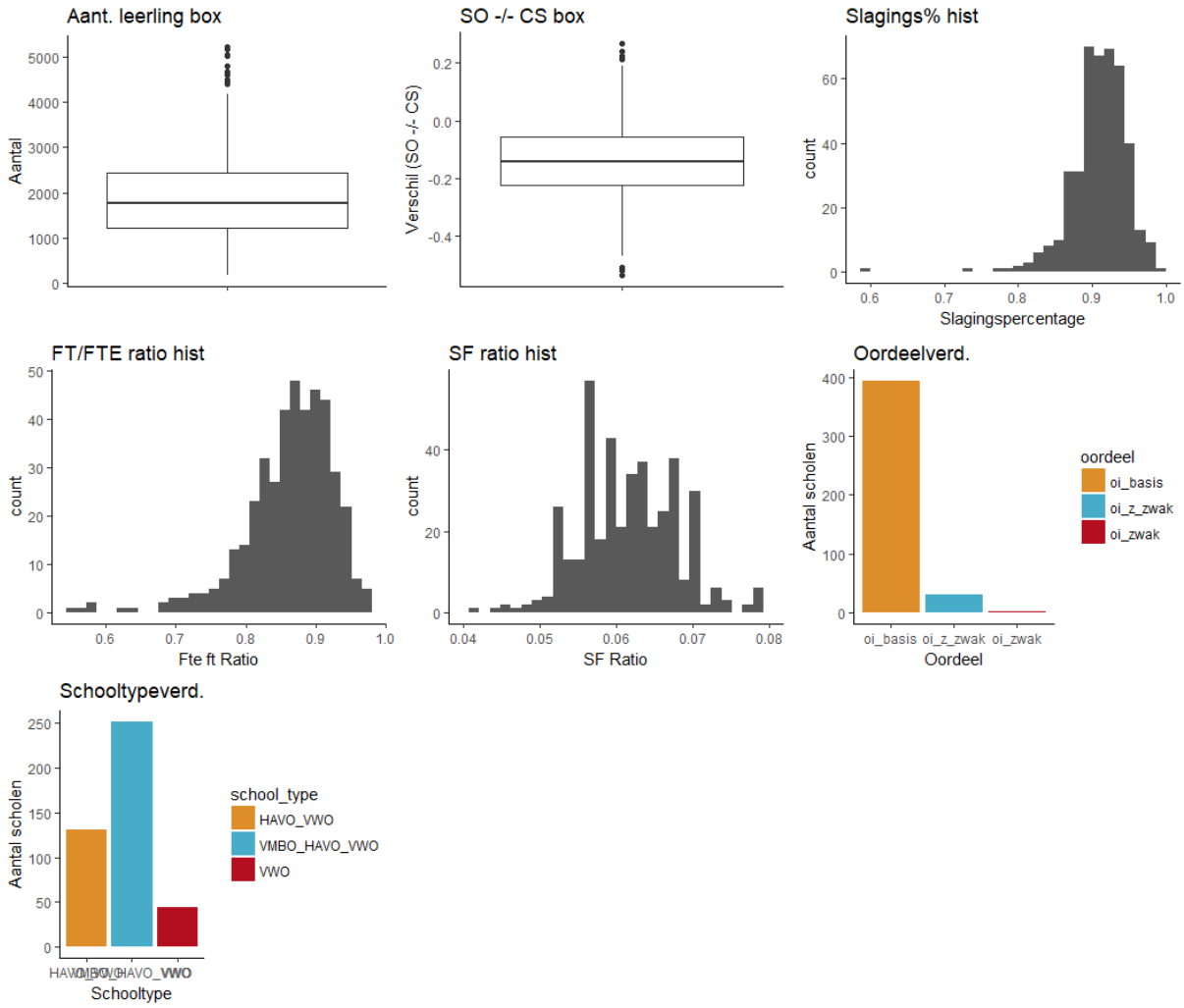


Boxplots

- Veel scholen(gemeenschappen) zijn vrij groot: het overgrote deel heeft in tussen de 1200 en 2400 leerlingen over alle schooljaren.
- Verder valt op dat veruit de meeste scholen over alle vakken gemiddeld slechter scoren op het Centraal Schriftelijk dan voor de schoolonderzoeken. Pas in het vierde kwantiel zijn de CS-resultaten beter dan het SO, maar dan zitten we al in de staart van de verdeling.

Histograms

- De slagingspercentages concentreren zich rondom de bandbreedte van 89% en 93%. Daar past het landelijke gemiddelde van 2017 (91%) prima in.
- Een iets andere verdeling zien we bij de verhouding van full time docenten en part time docenten: 83% tot 90% Dat betekent dat de meeste scholen vooral veel full time docenten voor de klas hebben staan.
- De verhouding docenten en leerlingen concentreert zich vooral tussen de 0.05 en 0.07 procent. Dat is een kleine waarde en laat zien dat voor de meeste scholen ‘evenveel’ aandacht is voor de leerlingen.



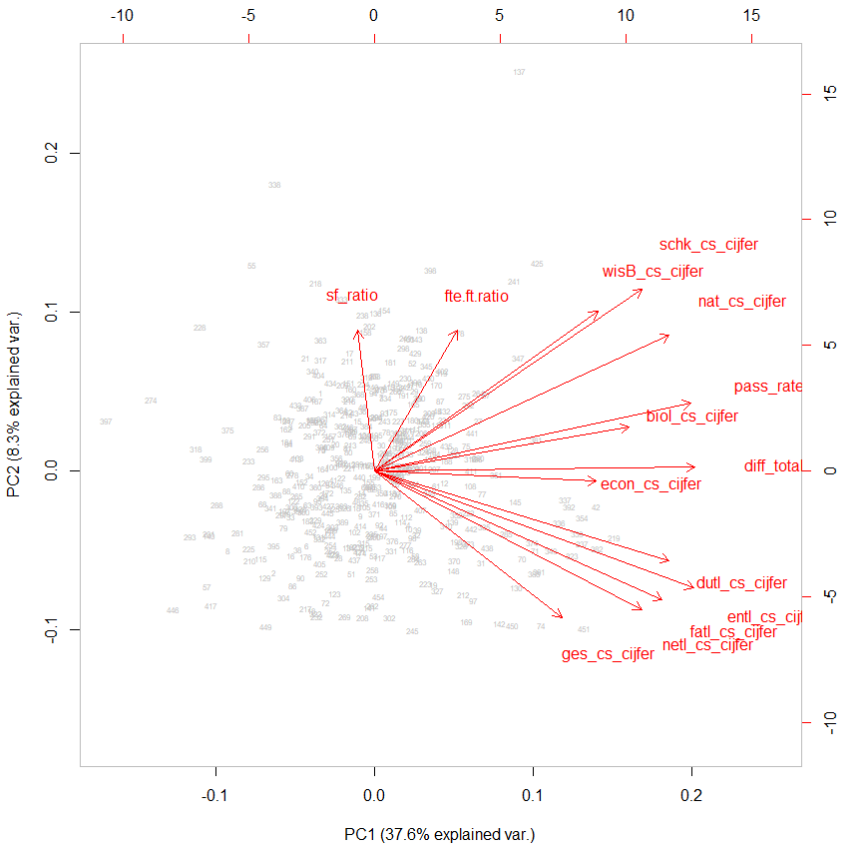
Barcharts

- De onderwijsinspectie deelt (gelukkig) weinig oordelen ‘zwak’ of ‘zeer zwak’ uit. Het nadeel hiervan is dat de verdeling tamelijk scheef is. De vraag is hoeveel invloed deze variabele zal hebben op het geheel, met de ‘basiswaarde’ zo dominant aanwezig.
- De verdeling ziet er iets beter uit voor het schooltype. Veruit de meeste scholen(gemeenschappen) in Nederland hebben zowel een VMBO-, als HAVO- en VWO-afdeling. *VWO-only scholen* komen het minste vaak voor.

Biplot: samenhang variabelen schooldata

Een biplot op de eerste twee dimensies van een PCA.out-matrix levert een logische samenhang tussen de variantie van de bovenstaande variabelen op. Beide dimensie verklaren ruim 45 procent van alle variantie; de onderstaande bevindingen zijn dus niet 100 procent zeker.

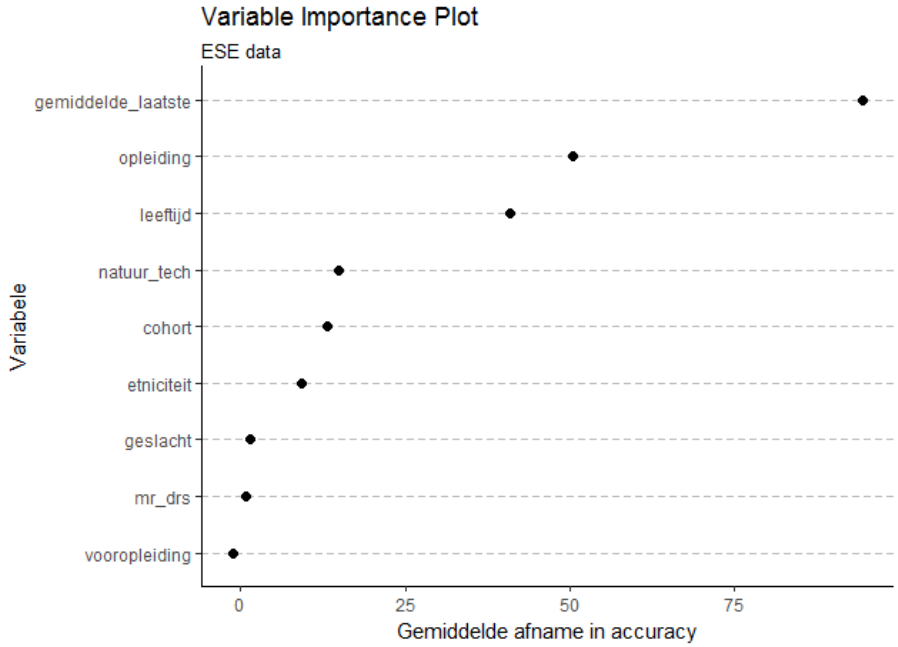
- Typische ‘staffing’ variabelen’ zoals fulltime/partime-percentages en leerling-docent ratio zitten in een andere dimensie dan ‘resultaatvariabelen’ (cijfers, verschil met so-resultaten en slagingspercentage).
- Het is opvallend dat het slagingspercentage en het verschil tussen het centraal schriftelijk en het schoolonderzoek tussen scholen eenzelfde variantie kennen en elkaar flink versterken.



Eerste selectie onafhankelijke variabelen

ESE dataset

Een aantal variabelen in de ESE dataset zijn mogelijk niet erg relevant om mee te nomen door een bijzonder scheve verdeling (vooropleiding, deelname Mr.Drs programma) of een zeer sterke samenhang met andere variabelen in de dataset (universitaire opleidings- en middelbare school profielkeuze van de studenten). Om een gevoel te krijgen voor het effect van deze scheve verdelingen en onderlinge samenhangen in de dataset trainen wij een Random forest (met alle hyperparameters op de standaard instellingen) op de gehele ESE dataset en beslissen op basis van het 'Variable Importance Plot' welke variabelen verder worden meegenomen.

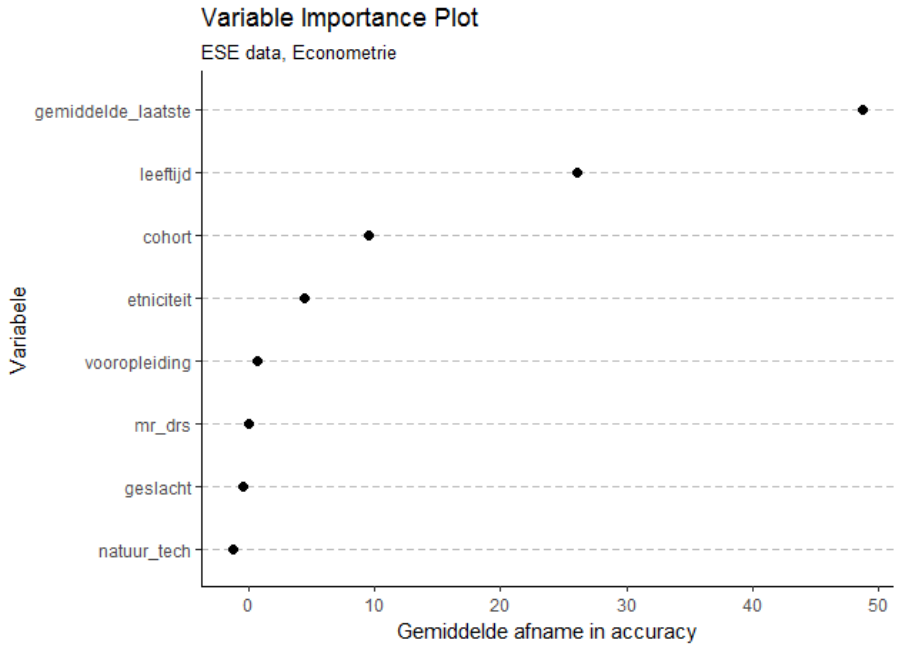


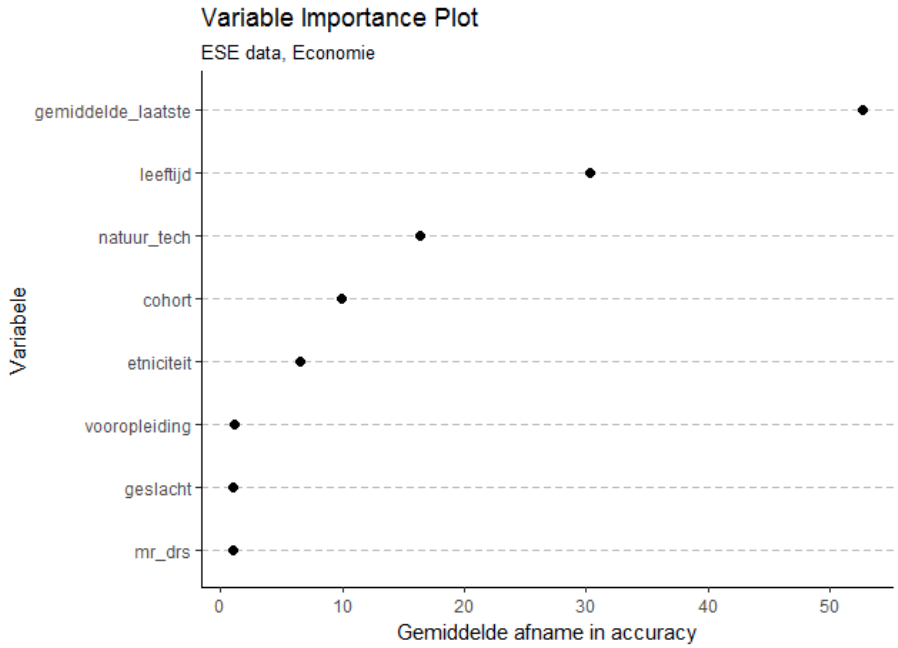
Op basis van dit plot kiezen wij ervoor om zowel de vooropleiding als de deelname aan het Mr.Drs programma te verwijderen. Deze variabelen voegen weinig informatie toe aan het model en maken door hun scheve verdeling het splitsen van de dataset in een train- en testset lastiger. Dit laatste is op te lossen door het trekken van een over meerdere variabelen gestratificeerde steekproef, maar gelet op de beperkte hoeveelheid informatie die deze variabelen opleveren vinden wij dat niet de moeite waard.

De variabele cohort voegt informatie toe aan het model. Er zijn blijkbaar verschillen in de prestaties van studenten die samenhangen met hun cohort. Hierbij kan gedacht worden aan de inzet van specifieke docenten in bepaalde jaren of aan groeps-effecten binnen het cohort (bijv. onderlinge negatieve of positieve beïnvloeding door studenten). Zonder aanvullende informatie zijn deze cohort-effecten echter lastig te interpreteren. Omdat wij streven naar interpreteerbare modellen kiezen wij ervoor deze variabele uit de dataset te verwijderen.

Het effect van de sterke samenhang tussen opleidingskeuze en het al dan niet hebben van een N&T profiel op het model is interessant. Als beide variabelen aan het model worden toegevoegd blijkt opleidingskeuze veruit de grootste voorspellende waarde te hebben, terwijl het hebben van een N&T profiel in belang afneemt. Uit bestaand onderzoek is echter bekend dat het hebben van een N&T profiel wel degelijk van waarde is bij het voorspellen van studentenuitval (in ieder geval voor economiestudenten).

Om meer inzicht te krijgen in dit effect hebben wij de dataset gesplitst op opleiding en op iedere dataset een nieuw Random forest getraind (opnieuw met alle hyperparameters op de standaard instellingen). Uit de resulterende VIPs blijkt dat voor de econometriestudenten middelbare school profielkeuze geen informatie aan het model toevoegt, wat logisch is aangezien deze studenten grotendeels een N&T profiel hebben. Voor de economiestudenten voegt deze variabele echter wel bijzonder veel informatie toe gaan het model. Bij het controleren voor opleidingskeuze levert het hebben van een N&T profiel dus wel degelijk informatie op.





Wij kiezen ervoor om dit probleem op te lossen door het introduceren van een nieuwe variabele waarin wij de 2 opleidingen en het hebben van een N&T profiel combineren (Econometrist met N&T, Econometrist zonder N&T, Econoom met N&T, Econoom zonder N&T). De oorspronkelijke variabelen worden uit de dataset verwijderd.

De resterende variabelen voegen veel (cijfer vooropleiding, leeftijd) of weinig (geslacht, etniciteit) informatie toe, maar hebben verder geen problematische eigenschappen. Deze variabelen blijven behouden. De uiteindelijke ESE dataset bestaat uit de volgende variabelen:

ESE variabelen
Opleiding * NT Profiel
Cijfer laatste
Leeftijd
Geslacht
Etniciteit

Verrijkte dataset

Aan de uiteindelijke ESE dataset voegen wij de schooldata toe om te komen tot de verrijkte dataset. De verrijkte dataset bestaat naast de ESE variabelen uit de volgende variabelen:

School variabelen
Gem. resultaat CS vaste profielvakken
Verschil SO minus SO profielvakken samen
Slagingspercentage
Onderwijsoordeel
Schooltype
Percentage pull-time vs part-time docenten
Leerling / docent ratio

Om een keuze te maken tussen de uiteindelijke ESE dataset en de verrijkte dataset trekken wij uit beide datasets een gestratificeerde (op basis van de afhankelijke variabele) steekproef van 70% van het totaal aantal observaties. De overige 30% van de observaties houden wij achter als testset. Op beide trainingsets trainen wij een Random forest en Lasso-regressie (met tuning van de relevante hyperparameters – zie onderzoeksmethode voor details) en vergelijken de gerapporteerde fout op de training- en testsets en de voor de testsets gerapporteerde Cohen's Kappa scores.

Random forest

Dataset	**Fout train**	**Fout test**	**Cohen's Kappa**
ESE	0.28	0.29	0.28
Verrijkt	0.31	0.30	0.20

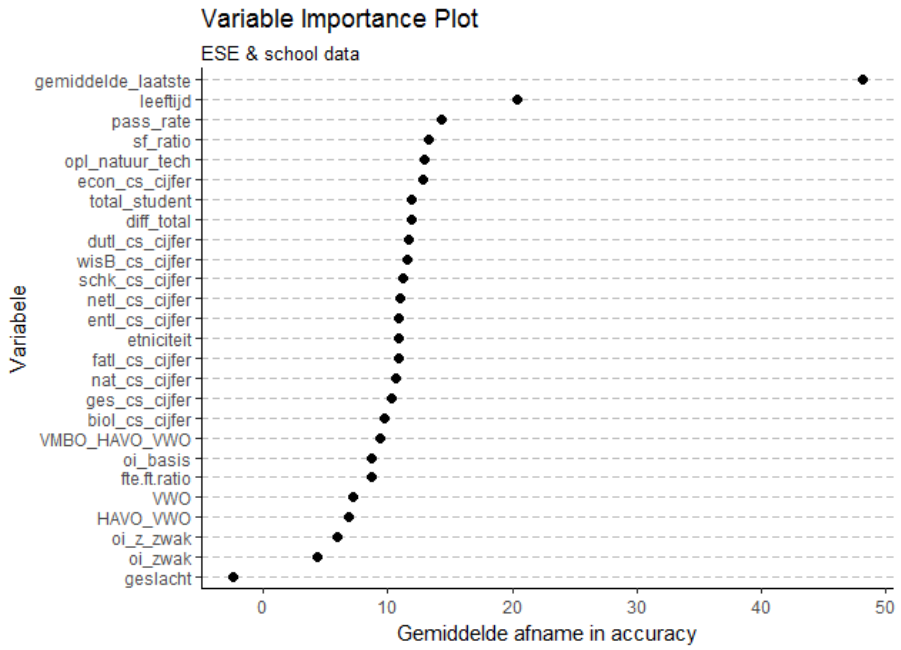
Lasso-regressie

Dataset	**Fout train**	**Fout test**	**Cohen's Kappa**
ESE	0.35	0.35	0.32
Verrijkt	0.35	0.36	0.29

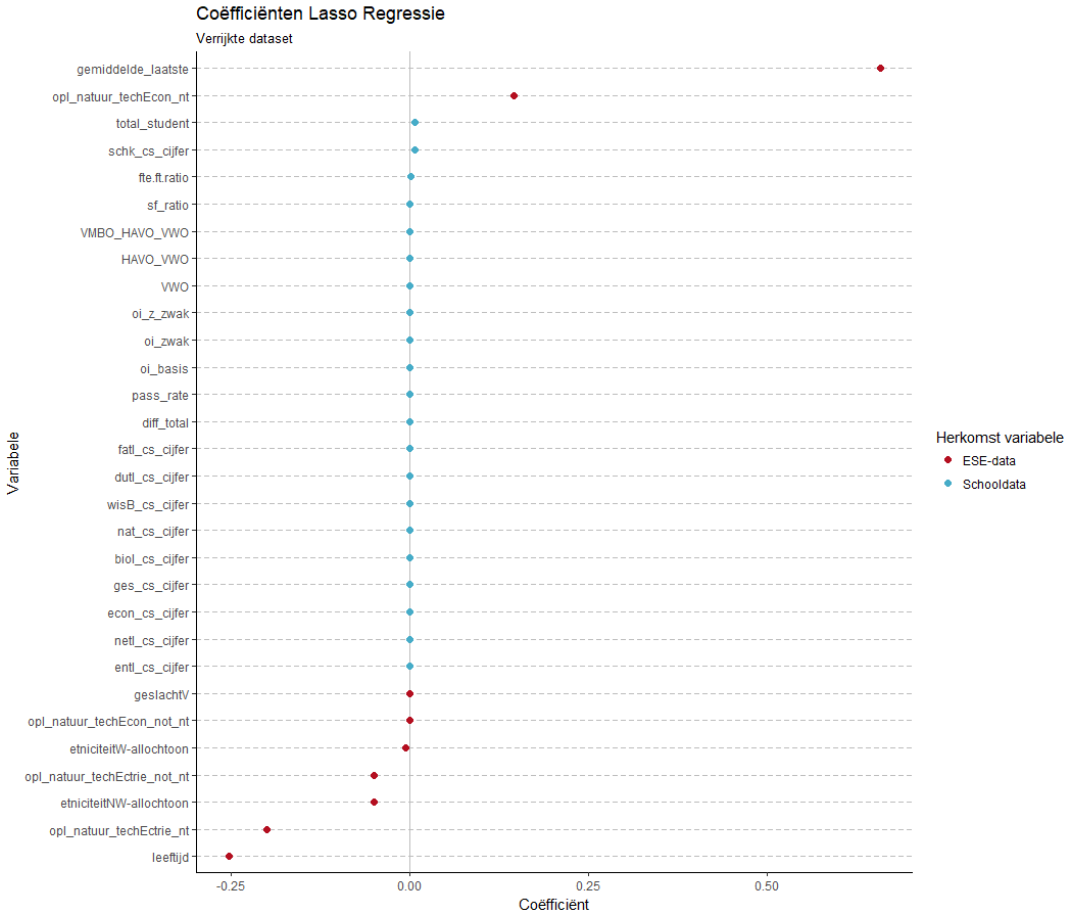
Voor beide datasets zijn de gerapporteerde waardes voor Cohen's Kappa laag en liggen dicht bij elkaar. Het VIP van het Random forest en de door de Lasso-regressie gerapporteerde coëfficiënten op de verrijkte data lijkt een zeker 'schooleffect' te suggereren. Variabelen als het school slagingspercentage en de school grootte voegen enige informatie toe aan het model. Het verrijken

van de dataset resulteert echter niet in betere voorspellingen. Wellicht draagt de meer granulaire data uit de ESE dataset toch meer informatie dan de toegevoegde geaggregeerde data.

Variable Importance Plot: verrijkte dataset



Lasso coëfficiënten: verrijkte dataset



Coefficiënten verrijkte dataset		Coefficiënten ESE dataset	
opl_natuur_techEcon_not_nt	.	opl_natuur_techEcon_not_nt	.
opl_natuur_techEcon_nt	0.145548470	opl_natuur_techEcon_nt	0.126125112
opl_natuur_techEctrie_not_nt	-0.050489477	opl_natuur_techEctrie_not_nt	-0.088516239
opl_natuur_techEctrie_nt	-0.200378752	opl_natuur_techEctrie_nt	-0.244811373
etniciteitW-allochtoon	-0.005543181	etniciteitW-allochtoon	-0.003133245
etniciteitNW-allochtoon	-0.050496390	etniciteitNW-allochtoon	-0.026866348
geslachtV	.	geslachtV	.
leeftijd	-0.253322477	leeftijd	-0.250980768
gemiddelde_laatste	0.659239322	gemiddelde_laatste	0.663355116
entl_cs_cijfer	.		
netl_cs_cijfer	.		
econ_cs_cijfer	.		
ges_cs_cijfer	.		
biol_cs_cijfer	.		
nat_cs_cijfer	.		
schk_cs_cijfer	0.007337781		
wisB_cs_cijfer	.		
dutl_cs_cijfer	.		
fatl_cs_cijfer	.		
diff_total	.		
pass_rate	.		
oi_basis	.		
oi_zwak	.		
oi_z_zwak	.		
VWO	.		
HAVO_VWO	.		
VMBO_HAVO_VWO	.		
total_student	0.007975776		
fte.ft.ratio	0.001394235		
sf_ratio	.		

De Kappa-scores en in het bijzonder de coëfficiënten van het Lasso-model laten zien dat de schoolvariabelen van de verrijkte dataset vooral ruis toevoegen. De ESE-variabelen veranderen bijna niet met de 'komst' van de schooldata variabelen. De correlaties die nog overblijven zijn verwaarloosbaar.

Omdat de toegevoegde data onze lakmoesproef van betere voorspellingen niet heeft kunnen doorstaan kiezen wij, ook vanuit het zuinigheidsprincipe, om alleen met de ESE dataset door te gaan.

Trainen

Naast de al eerder getrainde Random forest en Lasso-regressie modellen trainen wij een KNN model en een Decision tree model (CART algoritme) op de eerder willekeurig getrokken 70% van de ESE data (met tuning van de relevante hyperparameters – zie onderzoeksmethode voor details).

Voorspellen

Op basis van de voorspellingen van de op de ESE data getrainde Random Forest, Lasso regressie, KNN en Decision tree modellen komen wij tot de volgende kwaliteitsmaten voor de verschillende modellen.

Model	**Cohen's Kappa**	**Accuracy**	**Sensitivity**	**Specificity**	**Fout train**	**Fout test**
KNN	0.29	0.71	0.85	0.41	0.26	0.29
Random Forest	0.28	0.72	0.87	0.38	0.28	0.29
Lasso-regressie	0.32	0.65	0.60	0.77	0.35	0.35
Decision tree	0.29	0.72	0.90	0.36	0.28	0.28

Opvallend genoeg presteren alle modellen ongeveer even goed (of nauwkeuriger: slecht). Wij hadden verwacht dat de meer complexe Random forest en Lasso-regressie modellen betere voorspellingen zouden leveren dan het zeer simpele KNN model. Alleen het Lassomodel scoort echter met een kappa van 0.32 hoger dan het KNN-model. We denken dat dit te maken heeft dat dit model weliswaar minder scoort op de true positives, maar fors beter op de true negatives.

De scores voor Cohen's Kappa zijn over de gehele linie laag. Qua accuracy scoren de modellen tussen de 4 en 10 procentpunten beter dan een compleet naïeve voorspelling waarbij altijd de meerderheidsklasse wordt voorspeld (accuracy: 0.68). Blijkbaar kunnen ook de geavanceerde modellen weinig voorspellende informatie uit de aangeboden dataset persen. Alle modellen leveren bijzonder scheef verdeelde voorspellingen, het voorspellen van de positieve casus gaat vrij goed, het voorspellen van de negatieve casus gaat over het algemeen slecht.

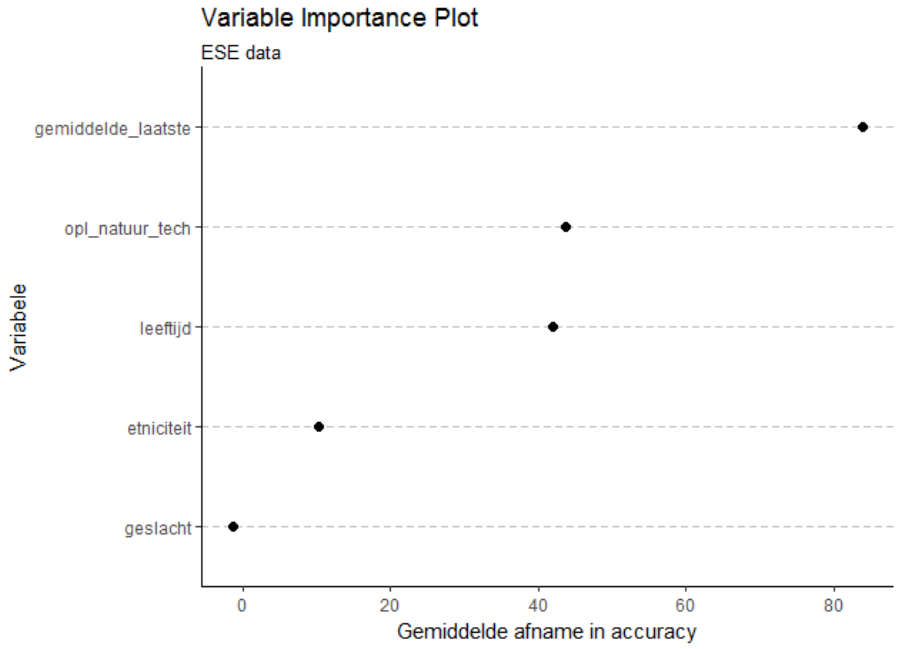
Op basis van de gerapporteerde kwaliteitsmaten alleen gaat onze voorkeur uit naar het KNN model. De voorspellingen zijn niet slechter dan die van de overige modellen, terwijl het onderliggende algoritme zo eenvoudig mogelijk is gehouden. Zowel Random forest, als KNN en Decision tree zijn (zonder tuning van gewichten of cutoff waarden voor de voorspellingen) erg ongebalanceerd in hun voorspellingen, ware positieven worden veelal gevonden, het voorspellen van ware negatieven gaat vaker fout dan goed. Het Lasso model daarentegen maakt meer gebalanceerde voorspellingen, ware negatieven worden in dit geval vaker goed voorspeld dan ware negatieven.

Interpreteren

Eén van de doelstellingen van dit onderzoek is het verrijken van het 'conceptueel model' van de beleidsmaker of bestuurder die zich met het probleem van studentenuitval bezig houdt. Kunnen wij aan deze personen een heuristiek meegeven om in hun dagelijks werk snel goede beslissingen over dit onderwerp te kunnen nemen? Het behalen van deze doelstelling staat of valt met de mate waarin de verschillende modellen te interpreteren zijn. De voorkeursoptie uit de voorgaande paragraaf, KNN, heeft in dit opzicht slechte kaarten. Het algoritme levert geen interpreteerbaar model op. De andere algoritmen doen dit wel.

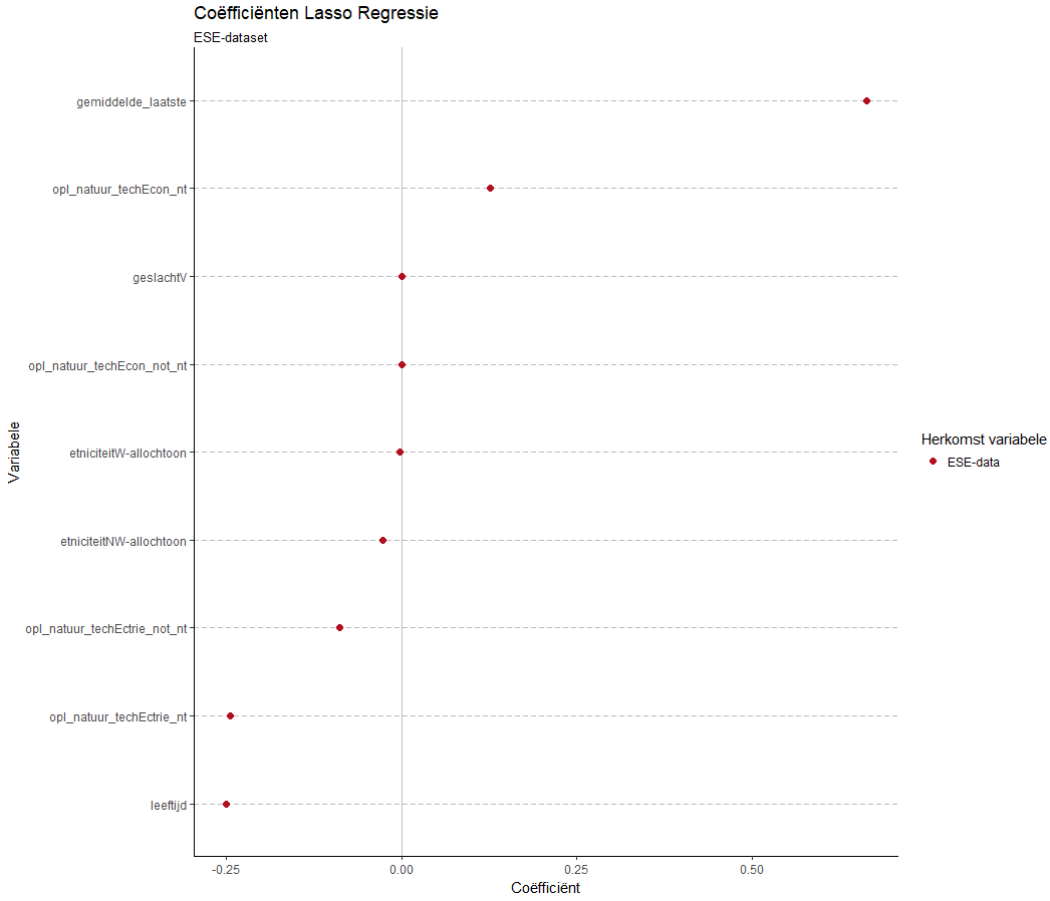
Random Forest

Het Random forest model is te interpreteren aan de hand van het VIP. Veruit de meeste informatie wordt geleverd door het cijfer van de vooropleiding, de combinatie van opleiding met N&T profiel en de leeftijd. Etniciteit speelt een zeer beperkte rol en geslacht lijkt geen rol te spelen. Een nadeel van deze manier van interpreteren is dat het alleen mogelijk is om iets te zeggen over het relatieve belang van de variabelen. Vanwege het niet-lineaire karakter van het algoritme is het echter niet goed mogelijk om met zekerheid iets te zeggen van de richting van de verschillende variabelen. Met het toevoegen van enige domeinkennis kan er natuurlijk wel een plausibele inschatting van de richting worden gemaakt, maar helemaal zeker kunnen wij hier niet over zijn. Een ander nadeel is dat er niet over de afzonderlijke niveaus van de combinatie opleiding met N&T profiel wordt gerapporteerd. Een meer diepgaande interpretatie van deze variabele is op basis van het VIP niet mogelijk. Een mogelijkheid die wij, vanwege de beperkt beschikbare tijd, in dit onderzoek niet verder verkennen is interpretatie van het model aan de hand van 'Partial dependence plots'.



Interpretatie Lasso-regressiemodel

Het Lasso-regressiemodel kan worden geïnterpreteerd aan de hand van de gerapporteerde coëfficiënten. Het grote voordeel van Lasso-regressie is dat het een *generalized linear model* is, wat er voor zorgt dat het in tegenstelling tot Random forest met zekerheid iets te zeggen is over het absolute belang van variabelen. Hierdoor is in één oogopslag te zien welke onafhankelijke variabelen de afhankelijke variabele positief of negatief beïnvloeden en hoe groot deze invloed is.

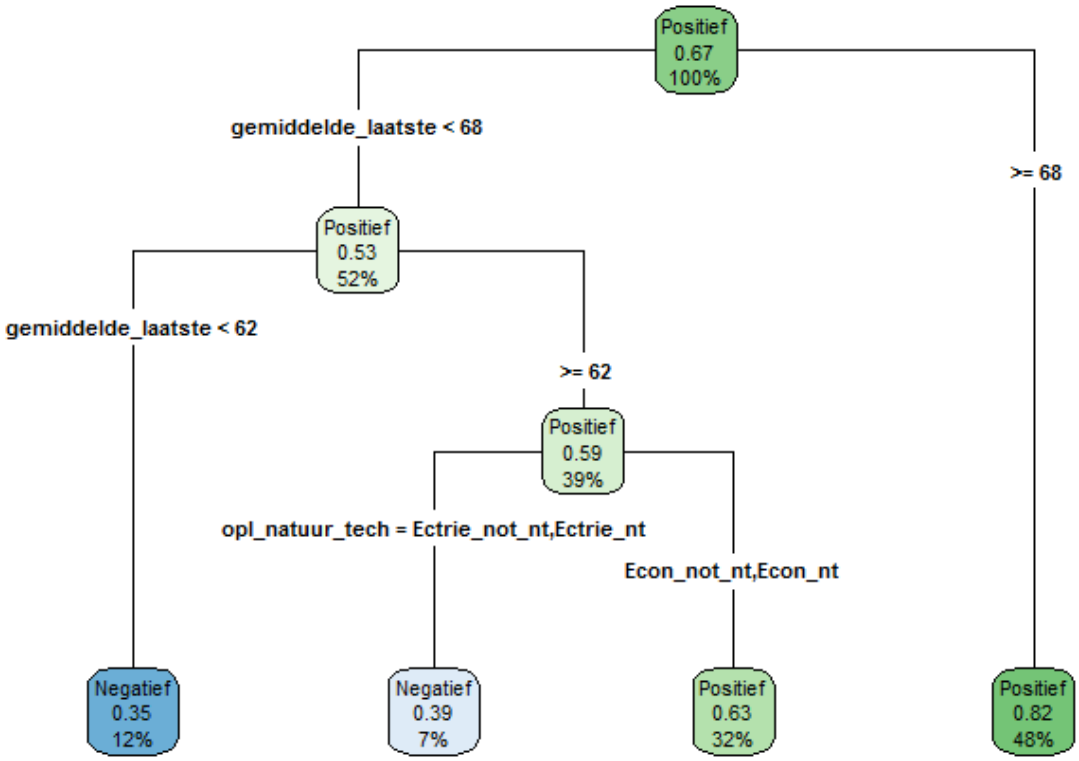


Coefficiënten ESE dataset	
opl_natuur_techEcon_not_nt	.
opl_natuur_techEcon_nt	0.126125112
opl_natuur_techEctrie_not_nt	-0.088516239
opl_natuur_techEctrie_nt	-0.244811373
etniciteitW-allochtoon	-0.003133245
etniciteitNW-allochtoon	-0.026866348
geslachtV	.
leeftijd	-0.250980768
gemiddelde_laatste	0.663355116

Uit bovenstaande coëfficiënten blijkt overduidelijk dat het vooropleidingscijfer veruit de grootste invloed heeft op een positief studieadvies. Op afstand volgt de kruisterm economiestudenten met een N&T-profiel. Negatief van invloed zijn vooral leeftijd (hoe hoger de leeftijd, hoe minder kans op

een positief studieadvies) en de kruisterm econometriestudenten met een N&T-profiel. Voor deze laatste observatie hebben we geen sluitende verklaring. Het is mogelijk dat een studiekeuze voor econometrie de kans verkleint op een positief advies omdat dit een 'moeilijke' studie is. Verder is het niet ondenkbaar dat de econometriestudenten uit de dataset met een niet-NT-profiel bepaalde eigenschappen bezitten waardoor zij een betere kans maken op een positief studieadvies dan hun medestudiegenoten met een NT-profiel.
De verdeling van positieve en negatieve adviezen tussen de N&T en niet-N&T groepen lijkt (zie beschrijving) niet op grote verschillen in BSA tussen deze groepen te wijzen. Geslacht, etniciteit en het volgen van Econometrie zonder NT-profiel lijken vrijwel geen rol te spelen.

Decision tree
Het decision tree model kan worden geïnterpreteerd aan de hand van een plot van de gesnoeide boom. Een groot voordeel van dit algoritme ten opzichte van het Random forest is dat de niveaus van de in het model opgenomen categorische variabelen afzonderlijk kunnen worden geïnterpreteerd (net als bij Lasso-regressie). Een nadeel van dit model is de hoge variabiliteit van de gerapporteerde boom, deze is nogal afhankelijk van de initiële splitsing van de dataset in train- en testsets. Een algoritme op maat om de gerapporteerde boom te stabiliseren of ten minste inzicht te krijgen in de stabiliteit van de gerapporteerde boom zou erg nuttig zijn, maar valt niet binnen de voor dit project beschikbare tijd te realiseren.[1] Dit model is strenger in haar variabelenselectie dan het Random forest en houdt alleen het cijfer van de vooropleiding en de combinatie van opleiding met N&T profiel over.



[1] Er zijn packages beschikbaar (stablelearner, dtree) die dit proberen te doen. Helaas is er niet genoeg tijd beschikbaar om goed naar deze packages te kijken.

Voor studenten met een vooropleidingscijfer van gemiddeld 6.8 of hoger voorspelt het model een positief studieadvies, voor studenten met een gemiddeld cijfer van 6.2 of lager voorspelt het model een negatief studieadvies. Voor alle studenten met een cijfer daartussen voorspelt het model een positief advies bij de economen en een negatief advies bij de econometristen. Opvallend genoeg zet het model de combinatie van opleiding met N&T profiel in als plaatsvervanger voor opleiding alleen, blijkbaar domineert in deze dataset de opleidingsvariabele de N&T profiel variabele wat informatiewaarde betreft.

Qua interpretatiewaarde gaat onze voorkeur uit naar de Lasso-regressie en Decision tree modellen. De coëfficiënten van de Lasso-regressie bieden duidelijke mogelijkheden voor interpretatie, net als de grafische weergave van het Decision tree model. Het Lasso-regressie model is waarschijnlijk wel aanzienlijk robuuster voor wijzigingen ten gevolge van een andere splitsing in train- en testset dan het Decision tree model. Het Random forest model biedt minder mogelijkheden voor interpretatie en heeft wat dit betreft niet onze voorkeur. Dit geldt in nog sterkere mate voor KNN.

Modelkeuze

Op basis van de machine learning theorie hadden wij verwacht een balans te moeten zoeken tussen voorspelkracht enerzijds en interpretabiliteit anderzijds. In dit onderzoek blijkt voorspelkracht geen factor van onderscheidend belang. Alle modellen presteren ongeveer gelijk wat voorspellen betreft. Op basis van het zuinigheidsprincipe geven wij bij het voorspellen het KNN model de voorkeur. Deze keuze is echter strijdig met één van de andere doelstelling van dit onderzoek: het zoeken naar een interpreteerbaar model. Qua interpretabiliteit geven wij de voorkeur aan de Lasso-regressie en Decision tree modellen.

Wij vatten onze bevindingen samen in onderstaande tabel.

Model	**Rangorde voorspellen**	**Opmerking**	**Rangorde interpretatie**	**Opmerking**
KNN	1	Zuinig model dat min of meer even goed voorspelt als de rest	4	Niet interpreteerbaar
Lasso-regressie	2	Hoogste Kappa score	1	Goed interpreteerbaar
Random forest	3 (gedeeld)	Lagere Kappa score dan Lasso	3	Beperkt interpreteerbaar
Decision tree	3 (gedeeld)	Lagere Kappa score dan Lasso	2	Goed interpreteerbaar, maar hoge variabiliteit

Wanneer zowel voorspelkracht, zuinigheid en interpretabiliteit in een afweging worden meegenomen komt het Lasso-model als winnaar uit de bus. Dit model biedt min of meer dezelfde kwaliteit voorspellingen als de rest (met overigens een groot verschil in sensitivity en specificity ten opzichte van de andere modellen) en is goed interpreteerbaar.

Lessons learned

- 80% van de tijd heeft gezeten in *data wrangling* werkzaamheden. In het eerste blok van deze opleiding zijn wij hier voor gewaarschuwd. Wij waren dus op onze hoede. Desondanks was onze teleurstelling toch aanzienlijk, toen bleek dat de schooldata vooral ruis toevoegde.
- Kwantiteit is niet altijd kwaliteit. Het toevoegen van (zelfs goed doordachte) variabelen levert niet altijd extra informatie op. De kans is simpelweg groot dat we gewoon ruis toevoegen. In het vervolg zullen we zorgvuldiger uitdenken of een variabele nu echt informatieve waarde gaat toevoegen. Helaas is trial-and-error niet altijd te voorkomen. In ons geval hadden we echter vooraf kunnen weten dat schoolinformatie minder informatief zou zijn dan individuele studentinformatie.
- Het loont om in een vroeg stadium om met een aantal verschillende modellen te benchmarken of extra variabelen het model beter, slechter of vooral ingewikkelder maken. Zowel de VIPs van Random Forest als de coëfficiënten van de Lasso bleken ijzersterke scherprechters in variabelenselectie.

Conclusie

In de introductie van dit onderzoek stelden wij ons een serie vragen over studentenuitval in bachelor 1 bij Erasmus School of Economics.

Welke variabelen uit de uitgebreide dataset dragen volgens ieder algoritme het meest bij aan het voorspellen van studententuitval? In hoeverre verschillen deze verzamelingen variabelen van de uit de literatuur en praktijk bekende verzameling variabelen?

Alle modellen selecteren ongeveer dezelfde belangrijkste variabelen uit de aangeboden datasets: cijfer vooropleiding, leeftijd en de combinatie opleiding * N&T-profiel. Dit onderschrijft de in eerder onderzoek gevonden resultaten. De overige variabelen in de ESE dataset (etniciteit en geslacht) spelen geen rol van betekenis in de gemaakte modellen.

Het toevoegen van de schooldata aan de ESE data resulteert niet in betere voorspellingen. Integendeel, de verrijkte data levert slechtere voorspellingen dan de niet-verrijkte data. Eén van de geleerde lessen is dan ook dat het loont om a priori stil te staan bij de verwachte voorspellende waarde van nieuw toe te voegen variabelen alvorens veel werk te steken in het verzamelen en bewerken van nieuwe data.

Hoe generaliseren de voor vraag A getrainde modellen naar ongeziene data? Welk model levert de beste voorspellingen?

Over het algemeen zijn de door de modellen geleverde voorspellingen van matige kwaliteit. Vooral het correct voorspellen van de negatieve casus (negatief BSA) is voor de meeste modellen (met uitzondering van Lasso-regressie) moeilijk. Op basis van de kwaliteit van de voorspellingen (in combinatie met het zuinigheidscriterium) kiezen wij KNN als het best voorspellende model.

In de epiloog van dit rapport (zie hierna) experimenteren wij met het toevoegen van cutoff waarden aan het Random forest model om de voorspellingen van dit model meer in balans te brengen.

Welke van de voor A getrainde modellen is het eenvoudigst interpreteerbaar? Welke van de modellen is niet of nauwelijks interpreteerbaar? Welke balans bestaat er tussen kwaliteit van de voorspellingen enerzijds en de mate van interpreteerbaarheid van het model anderzijds?

De gemaakte modellen verschillen aanzienlijk qua interpretabiliteit. KNN en Random forest zijn niet / beperkt interpreteerbaar. Lasso-regressie en Decision tree zijn zeer goed te interpreteren. Op dit punt kiezen wij Lasso-regressie als het best interpreteerbare model.

Welke maatregelen kan ESE op basis van de bij gegeven antwoorden nemen om studentenuitval in het eerste jaar te voorkomen?

Het blijkt erg lastig om een goede voorspelling van het uiteindelijke BSA te doen op basis van de bij aanvang van de studie bekende variabelen. Waarschijnlijk is dit maar goed ook; een student is niet, op basis de combinatie van zijn (haar) studiekeuze en gedrag in het verleden, voorbestemd om een positief of negatief BSA te krijgen. In dit onderzoek niet meegenomen variabelen als intelligentie, talent, studiediscipline en motivatie dragen blijkbaar voor een groot deel bij aan het studiesucces in jaar 1 (of het uitblijven daarvan).

Op basis van de resultaten van dit onderzoek adviseren wij ESE om bij studenten met een grotere kans op een negatief BSA (oudere studenten met een lager gemiddeld eindcijfer die econometrie studeren of die economie studeren zonder op de middelbare school Wiskunde B te hebben gevolgd) tijdens het eerste jaar goed te monitoren en bij dreigende uitval (bijvoorbeeld slechte resultaten in

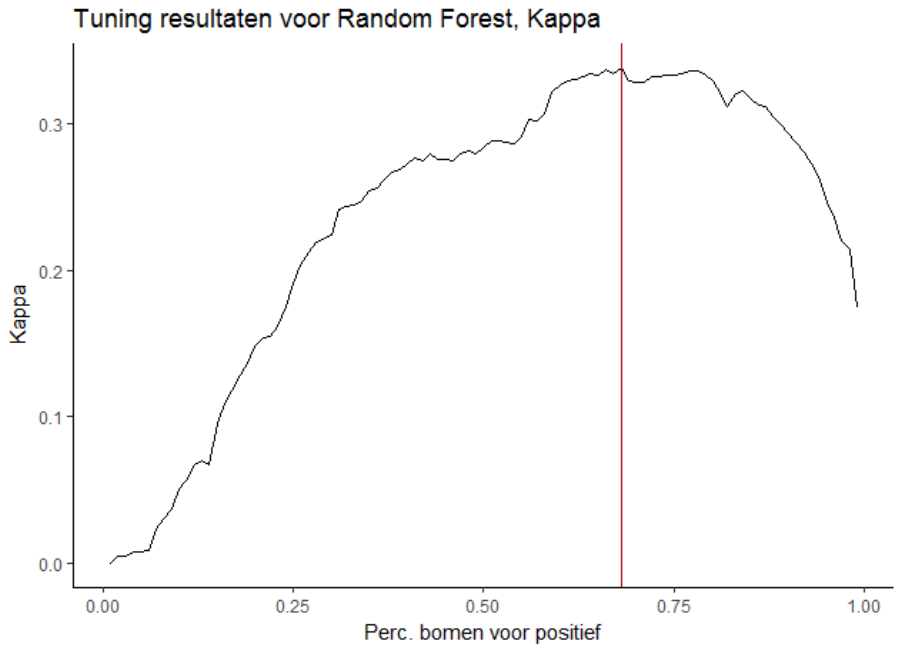
blok 1) in te grijpen. De ‘Studiekeuzeactiviteit’ die op dit moment door ESE wordt aangeboden aan alle 1e jaars studenten (met in sommige gevallen een verplichte online wiskundecursus) past goed in dit kader.

Epiloog

Tijdens ons werk aan dit onderzoek hebben wij geëxperimenteerd met een aantal manieren om de gevonden voorspellingen te verbeteren of om ons vertrouwen in de gevonden resultaten te verbeteren. Om een duidelijke lijn in het rapport te houden kiezen wij ervoor om de resultaten van deze experimenten niet in de hoofdtekst van het rapport, maar in deze epiloog op te nemen.

Gewichten

Voor het Random forest voegen wij gewichten toe aan het algoritme om te onderzoeken in hoeverre de voorspellingen op deze manier verbeterd kunnen worden. Hiertoe lopen wij een verzameling cutoff-waardes af waar het aandeel benodigde bomen voor een positieve of negatieve voorspelling mee kan worden beïnvloed. De kwaliteit van de met iedere cutoff-waarde gemaakte voorspellingen beoordelen wij opnieuw aan de hand van de gerapporteerde score voor Cohen's Kappa. Onderstaande grafiek toont het verloop van deze score voor de verschillende (positieve) cutoff waardes. De rode verticale lijn geeft de positieve cutoff waarde voor de hoogst gevonden Kappa score.



Het schuiven met de cutoff-waarde heeft een duidelijk effect op de kwaliteit van de gemaakte voorspellingen (tot onze verbazing, aangezien wij a priori hadden verwacht dat dit voornamelijk zou neerkomen op het inwisselen van het ene type fout voor een ander type fout). De algehele kwaliteit van de voorspellingen neemt toe. De sensitivity neemt enigszins af, maar de specificity neemt sterker toe. Onderstaande tabel toont de scores voor zowel het originele model als voor het geoptimaliseerde model.

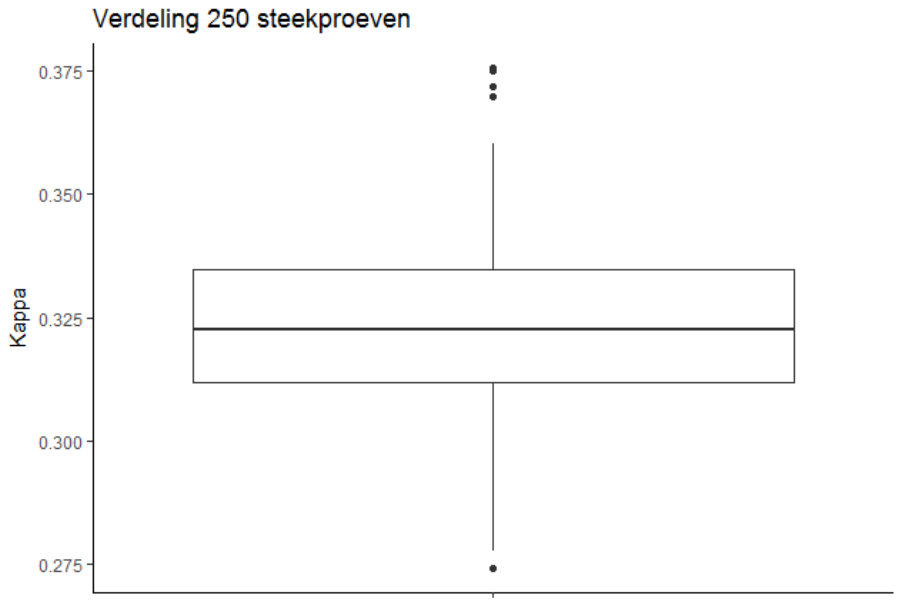
Model	**Cohen's Kappa**	**Accuracy**	**Sensitivity**	**Specificity**	**Cutoff (pos / neg)**
RF origineel	0.28	0.72	0.87	0.39	50 / 50
RF met cutoff	0.34	0.71	0.79	0.54	68 / 32

Kruisvalidatie train- / testset

Een mogelijk risico van het splitsen in een train- en testset is dat een toevallige splitsing aan het uiteinde van de kansverdeling een te grote onbalans tussen train- en testset veroorzaakt, waardoor

over- of underfit optreedt. Wij controleren hiervoor door voor ieder model de fout op de train- en testset te vergelijken. Bij min of meer gelijke fouten nemen wij aan dat er geen over- of underfit is opgetreden.

Om meer gevoel te krijgen voor de mogelijke effecten van een andere splitsing in train- en testset passen wij een kruisvalidatie variant toe op de Lasso-regressie. In deze variant trekken wij 250 op de afhankelijke variabele gestratificeerde train- en testsets uit de ESE dataset. Voor iedere combinatie van train- en testset voeren wij een Lasso-regressie uit. De kwaliteit van de door de verschillende modellen geleverde voorspellingen beoordelen wij aan de hand van de score Cohen's Kappa. Onderstaand boxplot toont de verdeling van deze score.



Hoewel de minimale en de maximale Kappa score zich op ongeveer 10 punten afstand van elkaar bevinden, is de afstand tussen de ondergrens van het eerste kwartiel en de bovengrens van het derde kwartiel klein. Wij concluderen hieruit dat, in meer algemene zin, het risico op het optreden van over- of underfit door een 'slechte' split in deze dataset beperkt is.

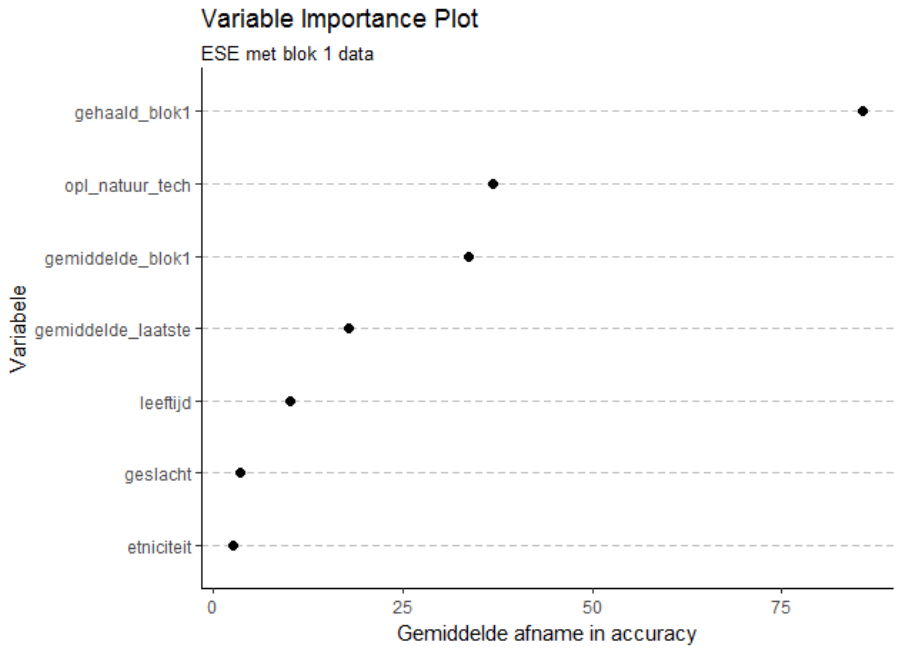
Toevoegen data blok 1

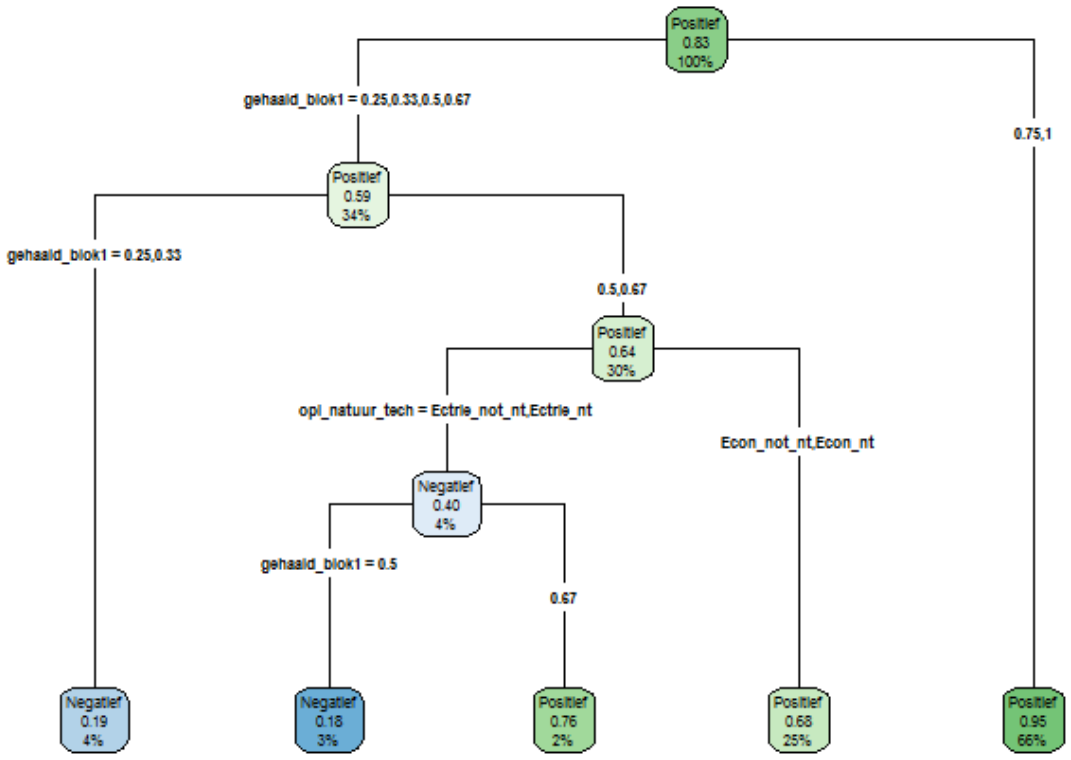
Om de kwaliteit van de voorspellingen te verbeteren kijken wij ook naar het effect van het toevoegen van informatie over het verloop van het 1e blok in bachelor 1. Hiervoor voegen wij 2 nieuwe variabelen aan de ESE dataset toe:

- Succespercentage blok 1: het percentage vakken in blok 1 waarvoor de student een voldoende of hoger dan een 5.5 heeft gehaald.
- Gemiddeld cijfer blok 1: ongewogen gemiddelde van de in blok 1 gegeven cijfers (> 5.5)

Op basis van deze nieuwe dataset trainen wij een Random forest en een Decision tree. In beide gevallen neemt de kwaliteit van de voorspellingen flink toe ten opzichte van de eerdere modellen zonder blok 1 data. Daarnaast domineren in beide gevallen de nieuw toegevoegde variabelen de oorspronkelijke variabelen. In het Random forest model nemen het gemiddelde vooropleidingscijfer en de leeftijd in belang af en in de Decision tree komen de oorspronkelijke variabelen bijna niet meer terug.

Model	Cohen's Kappa	Accuracy	Sensitivity	Specificity	Fout train	Fout test
Random Forest	0.46	0.86	0.97	0.42	0.13	0.13
Decision tree	0.42	0.88	0.97	0.34	0.13	0.12





Geraadpleegde literatuur

- Arnold, Ivo, J.M., Rowaan, Wietske, (2014), First-year study success in economics and econometrics: the role of gender, motivation and math skills, The Journal of Economic Education, 45, 1, 25-35.
- Arnold, Ivo, J.M., Straten, Jerry, T., (2012), Motivation and math skills as determinants of first-year performance in economics, The journals of Economic Education, 43, 1, 33-47.