



Gemeente Rotterdam

UITKERINGSFRAUDE VOORSPELLING

Draaiboek voor data scientist



Opdrachtgever	W&I Toetsing en Toezicht / BCO Onderzoek en BI
Auteurs	Projectgroep analytics uitkeringsfraude: XXXXXXXXXX
Datum	30-09-2020
Status	Versie 1.1

INTRODUCTIE

Dit document is bestemd voor data scientists die betrokken zijn bij de uitkeringsfraude voorspelling van W&I. Dit is een doorlopend product waarbij meermaals per jaar updates worden gedaan, dus OBI blijft ook doorlopend betrokken.

PROCES

In het proces van data tot aan onrechtmatigheid onderzoek wordt een deel uitgevoerd door data scientists binnen OBI, en een deel door W&I. Dit is grof geïllustreerd in onderstaand figuur.



Er zijn twee draaiboeken gemaakt:

1. Meermaals per jaar opleveren van een **nieuwe lijst** met risicoscores aan W&I.
2. Wanneer nodig het **volledig opnieuw trainen** van het voorspellende model.

Draaiboek 1 – het produceren van een nieuwe lijst – kost minder doorlooptijd (1 à 3 werkdagen) en is van toepassing wanneer W&I enkel een recentere lijst risicoscores nodig heeft. Bij draaiboek 2 doorlopen we een volledige train/testcyclus met modelselectie. Dit passen we toe als het bij draaiboek 1 gebruikte model gedateerd raakt (vuistregel: na 1 jaar). Mogelijk zijn er veranderingen in de brondata en wijzigingen in features. De doorlooptijd is daardoor langer (1-4 weken) en er is intensieve afstemming met W&I nodig over de gebruikte data, tussentijdse analyses en modelkeuzes.

NUMMERING

In de twee draaiboeken wordt een nummering gebruikt die refereert naar twee mappen die in de projectmap op de WOB server staan. In deze mappen staan de uit te voeren code en data bestanden, en worden ook de modellen en resultaten opgeslagen.

De map nummering ziet er als volgt uit:

1. CODE

- 10. Functies
- 11. Extractie
- 12. Laden
- 14. ABT
- 15. Featureselectie
- 16. Modelleren
- 18. Scoren
- 99. Rapporten

2. DATA

- 20. Stamdata
- 21. Brontabellen
- 22. Rdata
- 23. Ref
- 24. Features
- 25. ABT
- 26. Featureselecties
- 27. Model
- 28. Resultaten
- 29. Metadata








1 | PRODUCEREN NIEUWE LIJST

Dit deel van het draaiboek wordt gebruikt om op verzoek van W&I de complete lijst van uitkeringsontvangers een nieuwe onrechtmatigheids-risicoscore te geven. Nadat de lijst gegenereerd is worden de anonieme ID's door een data warehouse specialist omgezet tot BSN's. Die lijst wordt vervolgens beschikbaar gemaakt voor W&I. Zij halen zelf de top 500 BSN's (of meer) uit de lijst wanneer daar vraag naar is vanuit tHO.








Voordat de scripts kunnen worden uitgevoerd is het belangrijk dat een paar zaken worden gecontroleerd om de kwaliteit en structuur van de resultaten te waarborgen.

- 1) ESSENTIEEL: In het init.r (1.Code/10.Functies) bestand moet **dtlaad** worden aangepast naar de datum waarop de extractie gedaan gaat worden. In hetzelfde bestand moet **label** een beschrijving krijgen relevant voor deze run en moet **dtscopemodel** worden gecheckt (voor nu is dat 2014-01-01).
- 2) Overleg met de W&I process owner of de project- en afdelingsnamen - die zijn gebruikt voor het splitten van de data – nog steeds van toepassing zijn. Deze selecties zijn geïmplementeerd in het script ref_basis.r (in 1.Code/14.ABT).
- 3) Check voor voldoende schijfruimte. Breng oudere extracties (behalve de meest recente) onder in een zip-file.








0. EXTRACTIE

	BESCHRIJVING	Dit script extraheert de tabellen uit de Oracle DWH database en slaat ze als *.feather bestanden op de WOB server.
	SCRIPT	11.Extractie/extract_x_weifra_tabellen_masterscript.r <i>Let op: controleer of conn_db en conn_host naar de productie database verwijzen</i>
	DATA INPUT	Oracle DWH database
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	Onder het kopje "Set variables" voor controle_datum de datum van de vorige extractie invullen.
	OUTPUT	21.Brontabellen/*.feather
	CONTROLE	Bekijk de output van testcases.r (in de cosole). Hierin staan opvallende afwijkingen t.o.v. de vorige extractie. Controleer de oorzaak en onderneem actie indien nodig.

1. DATA LADEN




	BESCHRIJVING	Dit script roept de init en alle load scripts aan om de meest recente versies van alle data in te laden en klaar te zetten voor gebruik.
	SCRIPT	12.Laden/load_masterscript.r
	DATA INPUT	21.Brontabellen/*.feather 29.Metadata/*.csv
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	n.v.t.
	OUTPUT	22.Rdata/*.feather
	CONTROLE	n.v.t.

2. BOUW ABT








	BESCHRIJVING	<p>Inladen referentie tabel, en dan de features er bij joinen per tabel zoals die in vorige stap is ingeladen. Dit resulteert in een ABT feather bestand waar per ID alle features meegegeven zijn.</p> <p><i>Let op: in script ref_basis.R wordt de selectie voor de train-, test- en controleset gemaakt. Controleer dit en pas het aan indien nodig.</i></p>
	SCRIPT	14.ABT/abt.r
	DATA INPUT	22.Rdata/*.feather
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	n.v.t.
	OUTPUT	23.Ref/ref_basis_<dtlaad>_<label>.feather 24.Features/*.feather 25.ABT/abt_basis_<dtlaad>_<label>.feather
	CONTROLE	<p>In het <i>abt.r</i> script wordt 1) gecontroleerd op de afwezigheid van NA's, 2) dat het aantal rijen gelijk is aan de referentie tabel, 3) dat afwijkende tekens en spaties uit de featurenamen worden gehaald, en 4) character features worden gecontroleerd op gelijke levels tussen train en scoring set.</p> <p>Draai script 1.Code/99.Rapporten/Controlerapport_abt.r. Het geeft beschrijvende informatie over 1) de actualiteit van de</p>

bronbestanden, 2) verdwenen en/of nieuwe features, en 3) de variatie binnen de features. Pas hierbij **abt_oud_locatie** aan naar de abt waarmee je wil vergelijken.








3. FEATURE SELECTIE

	BESCHRIJVING	Feature selectie o.b.v. 1) handmatige opgave, 2) near zero variance, 3) multicollineariteit. De multicollineariteit wordt bepaald door VIF a.d.h.v. terugwaartse selectie van features.
	SCRIPT	15.Featureselectie/nzv.r 15.Featureselectie/multicol.r
	DATA INPUT	25.ABT/abt_basis_<dtlaad>_<label>.feather
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	In 15.Featureselectie/nzv.r: In de variabele drop_features kun je optioneel handmatig features opgeven om te verwijderen.
	OUTPUT	26.Featureselecties. Hier wordt een lijst (no_nzv_no_mc_<dtlaad>_<label>.csv) weggeschreven met namen van de features die zijn overgebleven na de feature selectie.
	CONTROLE	n.v.t.

4. HERTRAIN FINALE MODEL

	BESCHRIJVING	Hertrainen met alle gelabelde data om het finale model te maken. Er wordt niet getuned of getest. Van het al bestaande (referentie)model worden de tuning parameters gebruikt. Het model wordt na hertrainen gebruikt voor het scoren van populatie voor een nieuwe lijst.
	SCRIPT	16.Modelleren/train_model_nieuwelijst.r
	DATA INPUT	25.ABT/abt_basis_<dtlaad>_<label>.feather 26.Featureselecties/no_nzv_no_mc_<dtlaad>_<label>.feather
	MODEL INPUT	27.Model/<dtlaad>_<label>/finale_model.rds (NB: <i>dtlaad</i> en <i>label</i> van referentiemodel, zie label_refmodel)
	HANDMATIGE INPUT	Geef in label_refmodel de naam op van het model waar de hyperparametersettings van overgenomen moeten worden
	OUTPUT	27.Model/<dtlaad>_<label>/finale_model.rds (NB: hier <i>dtlaad</i> en <i>label</i> zoals in init gespecificeerd)
	CONTROLE	Draai script 16.Modelleren/train_model_controle.r. Hierin wordt een los model getraind om de performance op een (min of meer) aselecte controleset te bepalen. De output hiervan komt in de markdown <i>controle_model.html</i> terecht, in folder 27.Model/<dtlaad>_<label>

5. SCOREN

	BESCHRIJVING	In deze fase worden 1) risicoscores per ID berekend, 2) ID's gesorteerd op aflopende score, en 3) wordt de lijst als csv weggeschreven.
	SCRIPT	18.Scoren/scoring.r
	DATA INPUT	25.ABT/abt_basis_<dtlaad>_<label>.feather
	MODEL INPUT	27.Model/<dtlaad>_<label> (zoals nu in init gespecificeerd)
	HANDMATIGE INPUT	n.v.t.
	OUTPUT	28.Resultaten/lijs_t_alle_finalemodel_<dtlaad>_<label>.csv: CSV lijst met ID's gesorteerd op aflopende risicoscore, inclusief een kolom met rangnummer, een kolom per business rule(doelgroep scope) en een kolom met de berekende risicoscore. 28.Resultaten/abt_prob_finalemodel_<dtlaad>_<label>.feather voor naslag
	CONTROLE	Handmatig: bekijk de output en de scores.








2 | VOLLEDIG OPNIEUW TRAINEN MODEL

Dit deel van het draaiboek wordt gebruikt wanneer het bestaande model gedateerd raakt (vuistregel: na 1 jaar), of wanneer blijkt dat de prestatie tot een ongewenst niveau is gedaald. Aan de hand van dit draaiboek wordt het model volledig opnieuw gebouwd, met nieuwe data, parameters, technieken, etc.. Aan het eind van dit proces is er een nieuw gevalideerd model, en een nieuwe risicoscore lijst. Nadat de lijst gegenereerd is worden de anonieme ID's door een data warehouse specialist omgezet tot BSN's. Die lijst wordt vervolgens beschikbaar gemaakt voor W&I. Zij halen zelf de top 500 BSN's uit de lijst wanneer daar vraag naar is vanuit tHO.








Voordat de scripts kunnen worden uitgevoerd is het belangrijk dat een paar zaken worden gecontroleerd om de kwaliteit en structuur van de resultaten te waarborgen.

- 1) ESSENTIEEL: In het init.r (1.Code/10.Functies) bestand moet **dtlaad** worden aangepast naar de datum waarop de extractie gedaan gaat worden. In hetzelfde bestand moet **label** een beschrijving krijgen relevant voor deze run en moet **dtscopemodel** worden gecheckt (voor nu is dat 2014-01-01).
- 2) Overleg met de W&I process owner of de project- en afdelingsnamen - die zijn gebruikt voor het splitsen van de data – nog steeds van toepassing zijn. Deze selecties zijn geïmplementeerd in het script ref_basis.r (in 1.Code/14.ABT).
- 3) Check voor voldoende schijfruimte. Breng oudere extracties (behalve de meest recente) onder in een zip-file.








0. EXTRACTIE

	BESCHRIJVING	Dit script extraheert de tabellen uit de Oracle DWH database en slaat ze als *.feather bestanden op op de WOB server.
	SCRIPT	11.Extractie/extract_x_weifra_tabellen_masterscript.r <i>Let op: controleer of conn_db en conn_host naar de productie database verwijzen</i>
	DATA INPUT	Oracle DWH database
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	Onder het kopje "Set variables" voor controle_datum de datum van de vorige extractie invullen.
	OUTPUT	21.Brontabellen/*.feather
	CONTROLE	Bekijk de output van testcases.r (in de cosole). Hierin staan opvallende afwijkingen t.o.v. de vorige extractie. Controleer de oorzaak en onderneem actie indien nodig.

1. DATA LADEN





	BESCHRIJVING	Dit script roept de init en alle load scripts aan om de meest recente versies van alle data in te laden en klaar te zetten voor gebruik.
	SCRIPT	12.Laden/load_masterscript.r
	DATA INPUT	21.Brontabellen/*.feather 29.Metadata/*.csv
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	n.v.t.
	OUTPUT	22.Rdata/*.feather
	CONTROLE	n.v.t.

2. BOUW ABT







	BESCHRIJVING	<p>Inladen referentie tabel, en dan de features er bij joinen per tabel zoals die in vorige stap is ingeladen. Dit resulteert in een ABT feather bestand waar per ID alle features meegegeven zijn.</p> <p><i>Let op: in script ref_basis.R wordt de selectie voor de train-, test- en controleset gemaakt. Controleer dit en pas het aan indien nodig.</i></p>
	SCRIPT	14.ABT/abt.r
	DATA INPUT	22.Rdata/*.feather
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	n.v.t.
	OUTPUT	23.Ref/ref_basis_<dtlaad>_<label>.feather 24.Features/*.feather 25.ABT/abt_basis_<dtlaad>_<label>.feather
	CONTROLE	<p>In het <i>abt.r</i> script wordt 1) gecontroleerd op de afwezigheid van NA's, 2) dat het aantal rijen gelijk is aan de referentie tabel, 3) dat afwijkende tekens en spaties uit de featurenamen worden gehaald, en 4) character features worden gecontroleerd op gelijke levels tussen train en scoring set.</p> <p>Draai script 1.Code/99.Rapporten/Controlerapport_abt.r. Het geeft beschrijvende informatie over 1) de actualiteit van de</p>

bronbestanden, 2) verdwenen en/of nieuwe features, en 3) de variatie binnen de features. Pas hierbij **abt_oud_locatie** aan naar de abt waarmee je wil vergelijken.

3. FEATURE SELECTIE

	BESCHRIJVING	Feature selectie o.b.v. 1) handmatige opgave, 2) near zero variance, 3) multicollineariteit. De multicollineariteit wordt bepaald door VIF a.d.h.v. terugwaartse selectie van features.
	SCRIPT	15.Featureselectie/nzv.r 15.Featureselectie/multicol.r
	DATA INPUT	25.ABT/abt_basis_<dtlaad>_<label>.feather
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	In 15.Featureselectie/nzv.r: In de variabele drop_features kun je optioneel handmatig features opgeven om te verwijderen.
	OUTPUT	26.Featureselecties. Hier wordt een lijst (no_nzv_no_mc_<dtlaad>_<label>.csv) weggeschreven met namen van de features die zijn overgebleven na de feature selectie.
	CONTROLE	n.v.t.

4. TRAIN KANDIDAAT MODELLEN

	BESCHRIJVING	In deze stap worden alle eerste modellen ontwikkeld op basis van de ABT en de featureselectie. Er worden meerdere methodes en performance metrics aangeroepen. De seed en het aantal folds & repetities staan vast. Van elk model worden de 1) hitrates, 2) variable importances, 3) probabilities, en een overzicht van alle modellen opgeslagen in feather en CSV bestanden. Het modelleren is in deze fase een iteratief proces. Doorloop het proces een paar keer en pas tuning parameter, data selectie etc. aan ter verbetering van de modellen.
	SCRIPT	16.Modelleren/train_model_kandidaat.r
	DATA INPUT	25.ABT/abt_basis_<dtlaad>_<label>.feather 26.Featureselecties/no_nzv_no_mc_<dtlaad>_<label>.feather
	MODEL INPUT	n.v.t.
	HANDMATIGE INPUT	Traincontrol, grids
	OUTPUT	27.Model/<dtlaad>_<label> (zoals in init gespecificeerd) - model file (kandidaat_modellen.rds) - modelinformatie notebook (kandidaat_modellen.html)



CONTROLE

In het notebook worden de modellen met elkaar vergeleken door verschillende metrics en doorsnedes te gebruiken. Uiteindelijk is het aan de data scientist om de model prestaties te interpreteren en de beste modellen te selecteren om verder mee te werken.

De volgende analyses worden gedaan: 1) descriptieve tabel met diverse performance metrics, 2) hitrate per positie plot, 3) ROC plot, 4) kalibratie plot, 5) variable importance, 6) model stabiliteit tabel.

5. HERTRAIN FINALE MODEL



BESCHRIJVING

Hertrainen met alle gelabelde data om het finale model te maken. Er wordt niet meer getuned of getest. Van het best presterende model uit de vorige stap worden de tuning parameters gebruikt. Het model wordt na hertrainen gebruikt voor het scoren van populatie voor een nieuwe lijst.



SCRIPT

16.Modelleren/train_model_finale.r



DATA INPUT

25.ABT/abt_basis_<dtlaad>_<label>.feather
26.Featureselecties/no_nzv_no_mc_<dtlaad>_<label>.feather



MODEL INPUT

27.Model/<dtlaad>_<label>/kandidaat_modellen.rds



HANDMATIGE INPUT

Geef bij "VUL HIER DE CRITERIA VOOR HET BESTE MODEL IN" hoe je het best presterende model uit de kandidaatmodellen selecteert



OUTPUT

27.Model/<dtlaad>_<label>/finale_model.rds



CONTROLE

Draai script 16.Modelleren/train_model_controle.r. Hierin wordt een los model getraind om de performance op een (min of meer) aselecte controleset te bepalen. De output hiervan komt in de markdown *controle_model.html* terecht, in folder 27.Model/<dtlaad>_<label>

6. SCOREN



BESCHRIJVING

In deze fase worden 1) risicoscores per ID berekend, 2) ID's gesorteerd op aflopende score, en 3) wordt de lijst als csv weggeschreven.



SCRIPT

18.Scoren/scoring.r



DATA INPUT

25.ABT/abt_basis_<dtlaad>_<label>.feather



MODEL INPUT

27.Model/<dtlaad>_<label>



HANDMATIGE INPUT

n.v.t.



OUTPUT

28.Resultaten/lijt_alle_finalemodel_<dtlaad>_<label>.csv: CSV lijst met ID's gesorteerd op aflopende risicoscore, inclusief een kolom met rangnummer, een kolom per business rule(doelgroep scope) en een kolom met de berekende risicoscore.
28.Resultaten/abt_prob_finalemodel_<dtlaad>_<label>.feather voor naslag



CONTROLE

Handmatig: bekijk de output en de scores.