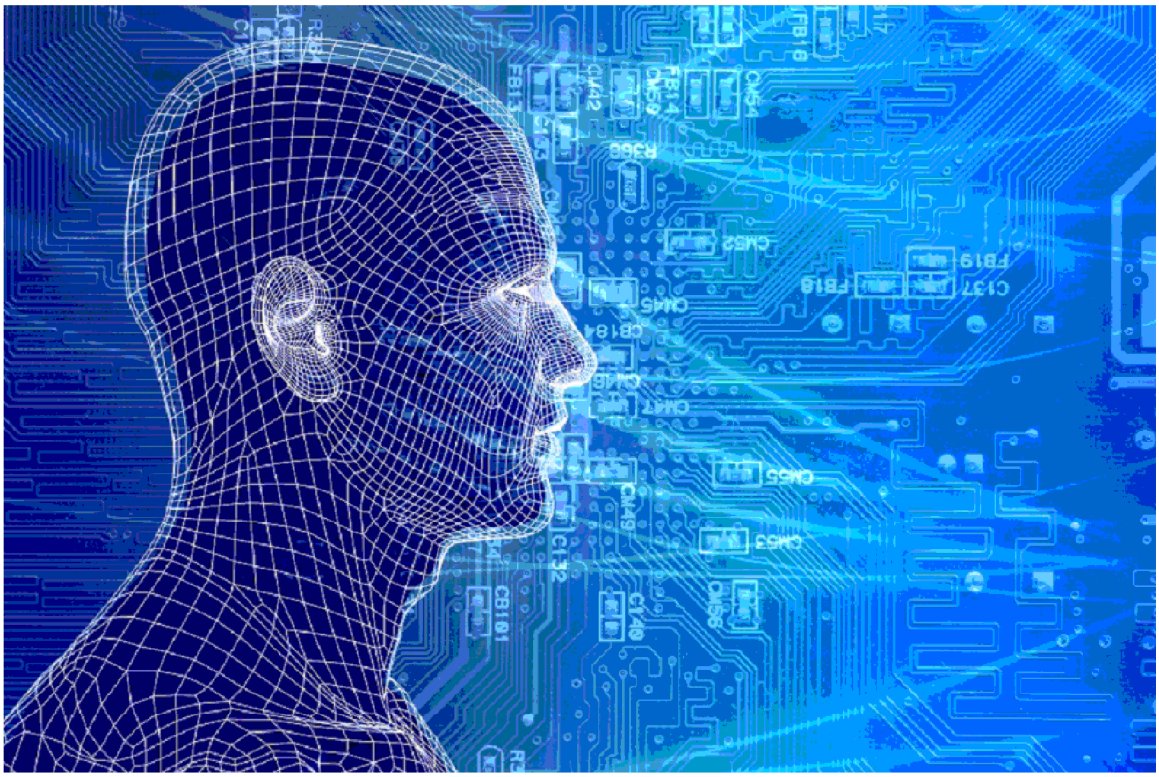

KOT_Tijdreizen

Versies van de documentatie

Versie nummer	Datum	Initialen	Belangrijkste wijziging
0.01	28-11-2013	10.2.e	Initieel document
0.02	10-02-2014		Nieuwe maand toegevoegd: 2014-01
1.0	20-12-2015		Wijziging nav overzetten naar AWS en bevriezen toeslagjaar 2013.
1.1	10-02-2016		Stappenplan toegevoegd
1.2	30-11-2018		Nieuwe opzet, 2013 apart

1. INHOUDSOPGAVE

2. DOEL

Doel van tijdreizen is testcases voorzien van de informatie zoals deze op het moment van beoordeling (als goed of fout) actueel was.
Dit heeft tevens tot gevolg dat als er nieuwe indicatoren worden toegevoegd aan het model, deze ook met terugwerkende kracht voor alle trainingscases toegevoegd moeten kunnen worden.

3. BESCHRIJVING PROJECT

3.1. BRON DATA

Er is generieke brondata en brondata per risicoselectierun:

- Generiek is het SAS-bestand 'Trainingscases_KOT' dat volgt uit het SAS project 'Trainingscases' (Q:\VEPROW63\TSL_DM_Handhavingsregie\Profiling 2013\Trainingscases).
- Daarnaast is er per risicoselectierun die in het verleden is uitgevoerd een tabel met alle indicatoren vereist. Dit is de tabel KOT_Score. Let op: de tabel Export_Temp3 kan ook, die bevat dezelfde indicatoren, en daar boven op een aantal andere variabelen die uit de risicoselectierun volgen, maar niet gebruikt worden tijdens het tijdreizen.
- Voor KOT is zijn na 2015 2 nieuwe indicatoren toegevoegd, namelijk of iemand bijstand heeft, en zo ja, het bedrag.
  De bijstandsgegevens zijn afkomstig uit het Halfproduct FLG, het resultaat hiervan wordt gekopieerd naar ….. dit bestand wordt per jaar opnieuw gedraaid.

3.2. FLOWS

Het SAS project wordt met ingang van december 2015 uitgevoerd op de AWS omgeving: AD010\data\RisicoSelectie\Tijdreizen\KOT\SAS\KOT_Tijdreizen.
Het project bestaat uit meerdere process flows:

Autoexec: toewijzen van libnames

A. Initialize: Initialisatie van scripts en input Trainingscases_KOT

B. Steekproef cases 2013: trekken van steekproef uit bestand 2013

C.-H. Create Testcases 2014-2019: het toevoegen van details vanuit een historische risicoselectierun die uitgevoerd is vlak voor de datum waarop de case als trainingscase beoordeeld is.

Z. Append: voeg alle testcases samen in 1 tabel

In december 2015 is besloten om de trainingscases voor toeslagjaar 2013 te bevriezen. De process flows die dit regelden waren omvangrijk en er werden nauwelijks meer nieuwe trainingscases voor dit toeslagjaar toegevoegd.
De bevriezing houdt in dat het tijdreizen voor deze toeslagjaren eenmalig is gerund (zie aparte documentatie KOT_Tijdreizen_2013 voor een toelichting op deze process flows), waarna het bestand met cases inclusief alle indicatoren opgeslagen is. Dit bestand wordt aan de gehele set met trainingscases toegevoegd in de process flow Z. Append. De implicatie is dat geen nieuwe trainingscases voor 2013 meer toegevoegd kunnen worden. Omdat het een omvangrijk en niet representatief bestand is, wordt een steekproef van 2013 aan het totaal toegevoegd (zie par. 3.3.3).
Met ingang van maart 2019 zijn de cases van 2014 en 2015 bevroren, ook hieraan kunnen geen nieuwe cases toegevoegd worden. (wat te doen met fraudecases?)

3.3. BESCHRIJVING HALFPRODUCT

3.3.1. AUTOEXEC

Deze process flow wordt gebruikt voor het aanmaken om:

- Verschillende libraries aan te maken, 1 generieke, en 1 per maand;
- De generieke brondata (BSN's van alle testcases in te lezen).

3.3.2. A. INITIALIZE

In het script A02_SetMonth vindt de bepaling van de meest historische data per testcase plaats. Hiervoor wordt eerst per testcase bepaald wat de datum van beoordeling was, op basis van de diverse beschikbare datumkolommen.

Vervolgens worden aan de testcases alle beschikkingen uit de risicoselectieruns gelijk of voor de beoordeeldatum gekoppeld. Principe hierachter is dat de beoordeling van de BSN waarschijnlijk plaatsvindt naar aanleiding van een conceptbeschikking, en dat die conceptbeschikking terug te vinden zou moeten zijn in de run die op of voor die datum plaats heeft gevonden. Als de BSN in meerdere risicoselectieruns terugkomt, wordt alleen de laatste mutatie behouden.

Dus bijvoorbeeld:

- BSN is beoordeeld op 2 augustus 2016.
- Er zijn gescoorde mutaties beschikbaar voor 31 mei, 2 augustus en 30 september 2016.
- De testcase wordt toegewezen aan maand augustus 2016.

En een tweede voorbeeld:

- BSN is beoordeeld op 20 december 2016.
- Er zijn gescoorde mutaties beschikbaar voor 2 augustus, 30 september en 13 december 2016.
- De testcase wordt toegewezen aan maand december 2016.

3.3.3. B. STEEKPROEF CASES 2013

In deze process flow wordt het steekproefbestand voor 2013 aangemaakt. Het programma is in deze process flow opgenomen zodat het mogelijk is om een andere steekproef te trekken indien dat wenselijk is. Het resultaat van de steekproef staat in de map Results/KOT_training_2013_selectie.

3.3.4. C. CREATE TESTCASES 2014

Per jaar is er momenteel 1 process flow.
Brondata specifiek voor deze flow:

KOT_score_yyyymm	Output van het model zoals gedraaid tijdens een risicoselectierun (bijvoorbeeld januari 2014 over toeslagjaar 2014).

Deze flows kennen 2 stappen:

1. selecteer in tabel met alle trainingscases de cases die beoordeeld zijn op het moment van de betreffende risicoselectierun run, en gebruik hiervoor alle indicatoren beschikbaar in de KOT_score tabel voor die risicoselectierun.
2. Voeg toe of aanvragers een bijstandsuitkering hebben en koppel dat aan het bestand. Deze indicator is later toegevoegd, vandaar dat het toevoegen van deze gegevens in een apart programma gaat.

3.3.5. Z. APPEND CASES

Z01: In deze flow worden alle testcases zoals samengesteld in de voorgaande flows samengevoegd tot één tabel: KOT_TRAININGSCASES_yyyymmdd. Alleen velden die nodig zijn om te modelleren in Enterprise Miner worden meegenomen, de rest wordt gedropt.
Bovendien wordt een laatste filter gezet op de cases die uiteindelijk als trainingscase naar Enterprise Miner gaan.
Het bestand met trainingscases uit 2013 (zie hierboven) wordt vervolgens afzonderlijk toegevoegd. Vervolgens wordt een steekproef getrokken op het hele bestand omdat het bestand onevenwichtig is opgebouwd voor wat betreft herkomst (voor een beschrijving van 'herkomst' zie de documentatie van Trainingscases). We hebben geëxperimenteerd met de SAS Enetrprise Miner en op basis daarvan gekozen voor een steekproef van:

- 4.000 posten op de DTCheck van 2017 (dit betreft alleen goede posten)
- 1.000 posten uit GreenLane 2015 (dit betreft alleen goede posten)
- 1.000 posten uit GreenLane 2016 (dit betreft alleen goede posten)
- 4.000 goede posten uit het bestand van eerder gecontroleerde posten
- 4.000 foute posten uit het bestand van eerder gecontroleerde posten

Het programma is zodanig opgezet dat de steekproef eenvoudig aangepast kan worden.
Z02: check op alle numerieke velden of missende waarden voorkomen. Deze output wordt verder niet gebruikt, is bedoeld voor controledoeleinden.

4. UPDATE KOT_TIJDREIZEN

Het project is per risicoselectierun opgezet. Dat betekent dat:

1. Er diverse scripts zijn waarin hard-coded de betreffende maanden aangeroepen worden;
2. Er 1 process flow per risicoselectierun is, waarin de relevante testcases geselecteerd worden en de indicatoren vanuit die risicoselectierun toegevoegd worden aan de cases die vlak na die risicoselectierun zijn beoordeeld.

Bij het toevoegen van een nieuwe risicoselectierun moet dus:

1. Nieuwe (bron)data die relevant is voor deze risicoselectierun ontsloten worden. De KOT_Score tabel uit de betreffende risicoselectierun wordt tijdens het maken van de back-up van de desbtereffende risicoselectierun in de KOT_Tijdreizen folder op de AWS omgeving gezet.
2. Een tab ingevoegd worden waarin testcases voor de betreffende risicoselectierun geselecteerd en indien nodig aangevuld worden.
3. Diverse algemene scripts aangepast worden zodat ook de nieuwe risicoselectierun meeloopt. Het gaat om:
   Autoexec: A01: nieuwe library
   A02: nieuwe risicoselectierun hardcoded toevoegen aan stap 3 (2x).
   Z01: finaal testcases bestand voor de nieuwe risicoselectierun toevoegen aan set statement in stap 1.
4. Bepaald worden welke indicatoren nog niet beschikbaar zijn, en (indien van toepassing) hoe deze op basis van de tijdens de run voor die risicoselectierun gebruikte gegevens toegevoegd kunnen worden. Dit komt momenteel niet voor.

5. STAPPENPLAN

In dit deel staan kort de verschillende stappen die ondernomen moeten worden bij het aanvullen van het tijdreizen. In het stappenplan moet de input data worden aangevuld, en het programma moet worden aangepast.

5.1.1. PROGRAMMA TIJDREIZEN:

[AWS/files/AD10/data/RisicoSelectie/tijdreizen/kot_training/SAS/KOT_tijdreizen.epg]

1.) In process flow 'A0.Initialize' in programma 'A01_SetParameters' onderaan een libname toevoegen met de nieuwe run. Voorbeeld: LIBNAME LIB0316 "&PATH./Input/2016";
2.) Zelfde process flow in programma 'A02_Set month' de libnames toevoegen en onderaan de nieuwe risicoselectie run met betreffende datum en het jaar waarop de run betrekking heeft toevoegen. Voorbeeld: LIB0316.KOT_score_201603 (IN = IN032016 KEEP = BSN)

```
En:  IF IN032016 THEN DO;
        runnr = 201603;
        Toeslagjaar = 2016;
        Rundatum = INPUT('29/01/2016', ddmmyy10.);
END;
```

3.) Vervolgens wordt er in de process flow van het huidige jaar een nieuw stukje aan het programma geplakt waarin nieuwe runs worden toegevoegd. Voorbeeld van een programma:

```
PROC SQL;
CREATE TABLE LIB0316.KOT_training_201603_Final AS
SELECT A.*,
          B.Type,
          B.Datum_beoordeeld,
          B.Rundatum,
          B.Herkomst
FROM LIB0316.KOT_score_201603 AS A
INNER JOIN KOTTrain.Trainingscases_KOT_FINAL AS B
WHERE B.runnr = 201603;
QUIT;
```

4.) In de Process Flow 'ZZ.Append' wordt tot slot in het programma 'Z01_append' de libname ook toegevoegd, voorbeeld: LIB0316.KOT_TRAINING_201603_FINAL

5.1.2. INPUT DATA TOEVOEGEN:

Vervolgens moet de data van elke risicoselectie run worden toegevoegd in: AWS/ files/AD10/data/RisicoSelectie/tijdreizen/kot_training/Input/"betreffende jaar"/kot_score_'jaar''runnr'.
Met ingang van 2017 gebeurt dit automatisch tijdens het maken van de back-up van de risicoselectie.

6. NIEUWE OPZET

Met ingang van april 2019 zijn ook de jaren 2014 en 2015 bevroren. Deze bestanden zijn opgeslagen in xxxx en worden rechtstreeks in de append cases toegevoegd.

Als het proces is afgerond het programma en de datasets opslaan op de q-schijf:
Q:\VEPROW63\TSL_DM_Handhavingsregie\Profiling 2013\Trainingscases\Backup
Het project in de map sas_epg en de bestanden in de map: bestanden.
Telkens opslaan met bestandsnaam_jjjjmmdd.

7. RESULTATEN PER TIJDREIZEN

Vanaf maart 2019 wordt per keer dat we tijdreizen draaien een overzicht toegevoegd van het totale bestand en de steekproef naar type, herkomst en toeslagjaar.

7.1. MAART 2019

		Totaal			Steekproef		
		goed	fout	totaal	goed	fout	totaal
Herkomst	Toeslagjaar						
DT Check	2017	12.567	0	12.567	4.000	0	4.000
Excel	2013	1.500	1.938	3.438	1.500	1.938	3.438
	2014	622	859	1.481	622	859	1.481
	2015	15	61	76	15	61	76
	totaal	2.137	2.858	4.995	2.137	2.858	4.995
Fraudeteams	2013	0	562	562	0	562	562
	2014	0	657	657	0	657	657
	2015	0	262	262	0	262	262
	2016	0	263	263	0	263	263
	2017	0	167	167	0	167	167
	2018	0	99	99	0	99	99
	2019	0	4	4	0	4	4
	totaal	0	2.014	2.014	0	2.014	2.014
GreenLane	2015	3.382	0	3.382	1.000	0	1.000
	2016	4.740	0	4.740	1.000	0	1.000
	2017	410	30	440	410	30	440
	totaal	8.532	30	8.562	2.410	30	2.440
Zaak	2013	131	123	254	39	58	97
	2014	1.324	1.538	2.862	361	758	1.119
	2015	3.774	2.673	6.447	987	1.329	2.316
	2016	2.956	1.196	4.152	778	623	1.401
	2017	3.575	1.491	5.066	981	734	1.715
	2018	2.975	978	3.953	839	485	1.324
	2019	47	21	68	15	13	28
	totaal	14.782	8.020	22.802	4.000	4.000	8.000
Totaal		38.018	12.922	50.940	12.547	8.902	21.449