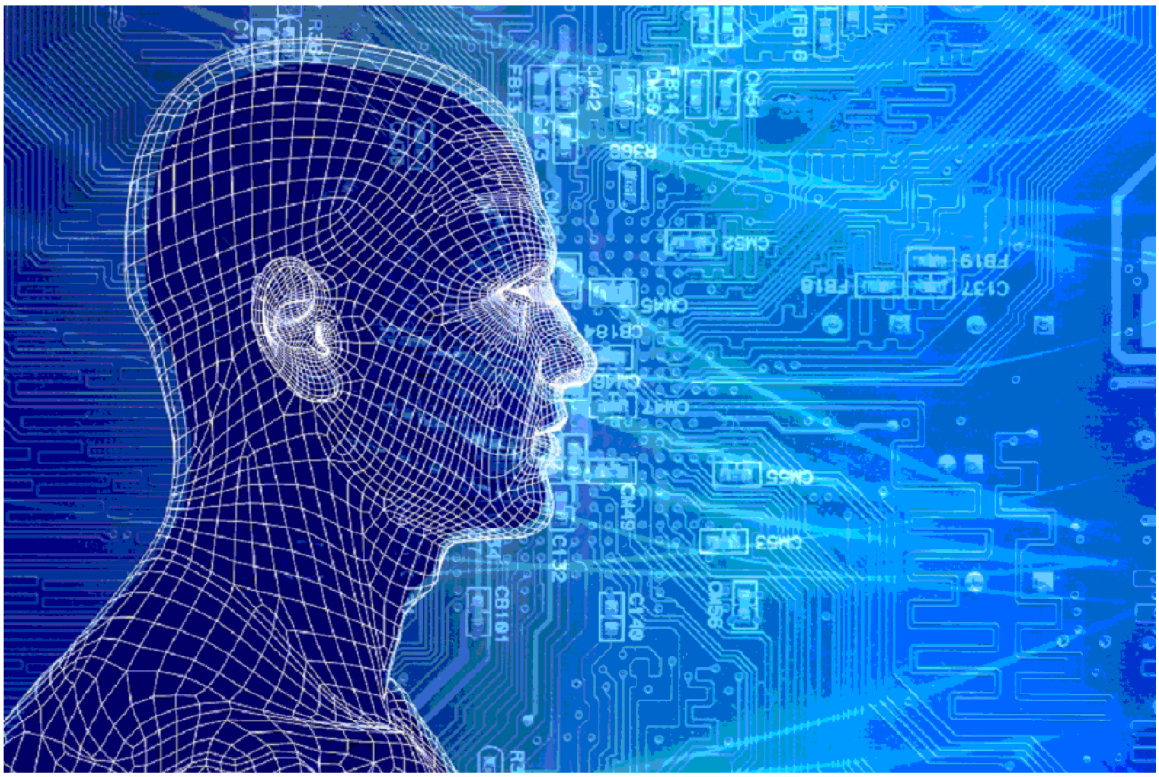

HT_Tijdreizen

Versies van het model

Versie nummer	Datum	Initialen	Belangrijkste wijziging
0.01	20-11-2013	10.2.e	Initieel document
0.02	24-01-2014		Nieuwe maand toegevoegd: 2014-01
0.03	27-01-2014		Opmerkingen MS verwerkt
0.04	10-02-2014		Aanpassing toewijzen cases aan run
1.0	20-12-2015		Wijziging nav overzetten naar AWS en bevriezen toeslagjaar 2013.
1.1	07-03-2015		Toevoegen Stappenplan (H6)
1.2	30-11-2018		Nieuwe opzet, 2013 apart

1. INHOUDSOPGAVE

2. DOEL

Doel van het tijdreizen is trainingscases 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_HT' 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 HT_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.

3.2. FLOWS

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

Autoexec: toewijzen van libnames

A. Initialize: Initialisatie van scripts en input Trainingscases_HT
B. Steekproef cases 2013: trekken van steekproef uit bestand 2013
C. Create Testcases dd-yy: 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 HT_Tijdreizen_2013), 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).

4. BESCHRIJVING HALFPRODUCT

4.1. 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).

4.1.1. A. INITIALIZE

In het script A02_SetMonth 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 2013.
- Er zijn gescoorde mutaties beschikbaar voor 31 mei, 2 augustus en 30 september 2013.
- De testcase wordt toegewezen aan maand augustus 2013.

En een tweede voorbeeld:

- BSN is beoordeeld op 20 december 2013.
- Er zijn gescoorde mutaties beschikbaar voor 2 augustus, 30 september en 13 december 2013.
- De testcase wordt toegewezen aan maand december 2013.

4.1.2. 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/HT_training_2013_selectie.

4.1.3. C. CREATE TESTCASES 2014

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

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

Deze flows kennen maar 1 stap: 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 HT_score tabel voor die risicoselectierun.

4.1.4. Z. APPEND CASES

Z01: In deze flow worden alle testcases zoals samengesteld in de voorgaande flows samengevoegd tot één tabel: HT_TRAININGSCASES_FINAL. 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. In de volgende situatie wordt een case uitgesloten:

1. Er is niet bekend of de case frauduleus was of niet (Y ne .); OF
2. De kale huur is onbekend (Kale_huur ne .);

In het verleden werden posten ook uitgesloten als de volgende variabelen missing waren:

3. Het oppervlakte per persoon is onbekend (oppervlakte_per_persoon ne .); OF
4. De oppervlakteprijs ten opzicht van de gemiddelde huurprijs in de Postcode 5-omgeving is onbekend (perc_oppervlakteprijs_omg ne .); OF
5. De huurprijs ten opzichte van de gemiddelde huurprijs in de Postcode 5-omgeving is onbekend (perc_huurprijs_huurprijs_omg ne .).

In gevallen 2-5 gaat het om indicatoren die in het verleden voor een kleine subset niet te reconstrueren bleek. Omdat deze cases de training van het model sterk beïnvloedden (want bij waarde missing was eventrate 100%) is besloten deze cases uit te sluiten. Echter op dit moment is de verdeling goed/fout bij de cases met missings op deze variabelen 50/50. Besloten is dan ok om deze wel in de trainingsset te laten.

Het bestand met trainingscases uit 2013 (zie hierboven) wordt vervolgens afzonderlijk toegevoegd.

Vervolgens wordt een steekproef getrokken op het totale bestand omdat het bestand onevenwichtig is opgebouwd voor wat betreft herkomst. We hebben geëxperimenteerd met de SAS Enterprise Miner en op basis daarvan gekozen voor een steekproef van:

- 4.000 posten op de DTCheck van 2017 (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. Wel altijd draaien!

5. Update HT_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 HT_Score tabel uit de betreffende risicoselectierun wordt tijdens het maken van de back-up van de desbtereffende risicoselectierun in de HT_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.

6. 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.

6.1. PROGRAMMA TIJDREIZEN

[AWS/files/AD10/data/RisicoSelectie/tijdreizen/ht_training/SAS/HT_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.HT_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 LIB0118.HT_TRAINING_201801_final AS
SELECT A.*,
  B.Type,
  B.Datum_beoordeeld,
  B.Rundatum,
  B.Herkomst,
  B.Omschrijving AS werkopdracht
FROM LIB0118.HT_score_201801 AS A
INNER JOIN HT_Train.Trainingscases_HT_final AS B
ON A.BSN = B.BSN AND a.toeslagJaar = B.toeslagJaar
WHERE B.runnr = 201801 ;
QUIT;
```

4.) In de process flow 'Z. Append' wordt tot slot in het programma 'Z01_append' de libname ook toegevoegd, voorbeeld: LIB0316.HT_TRAINING_201603_FINAL

6.2. INPUT DATA TOEVOEGEN

De data van elke risicoselectie run worden automatisch tijdens het maken van de back-up na elke run toegevoegd in:
AWS/ files/AD10/data/RisicoSelectie/tijdreizen/ht_training/Input/"betreffende jaar"/ht_score_'jaar''runnr'.

6.3. BACK-UP MAKEN

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. Nieuwe opzet

Met ingang van april 2019 zijn ook de jaren 2014 en 2015 bevroren. Deze bestanden zijn opgeslagen onder de namen: ht_trainingset_2014 en ht_trainingset_2015 in de map Results en deze worden vanaf april 2019 rechtstreeks in de append cases toegevoegd.

8. Resultaten per tijdreizen

Vanaf maart 2019 worden per keer ook de resultaten van het tijdreizen opgenomen

8.1. Maart 2019

		totaal			steekproef		
Herkomst	Toeslagjaar	GOED	FOUT	TOTAAL	GOED	FOUT	TOTAAL
DT Check	2017	12.252	nvt	12.252	4.000	nvt	4.000
Excel	2013	1.500	2.088	3.588	1.500	2.088	3.588
	2014	638	2.684	3.322	638	2.684	3.322
	2015	88	209	297	88	209	297
	totaal	2.226	4.981	7.207	2.226	4.981	7.207
Fraudeteams	2014	0	60	60	0	60	60
	2015	0	52	52	0	52	52
	2016	0	48	48	0	48	48
	2017	0	92	92	0	92	92
	2018	0	37	37	0	37	37
	2019	0					
	totaal	0	289	289	0	289	289
GreenLane	2016	7.555	0	7.555	2.000	0	2.000
	2017	659	127	786	659	127	786
	totaal	8.214	127	8.341	2.659	127	2.786
Zaak	2014	787	2.225	3.012	148	737	885
	2015	4.206	3.151	7.357	784	1.045	1.829
	2016	6.126	2.469	8.595	1.126	844	1.970
	2017	7.420	2.947	10.367	1.432	1.029	2.461
	2018	2.357	997	3.354	491	327	818
	2019	70	49	119	19	18	37
	totaal	20.966	11.838	32.804	4.000	4.000	8.000
Totaal		43.658	17.235	60.893	12.885	9.397	22.282