MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA

22 JUL 2020

Dear

On 30 June 2020, you emailed the Ministry of Social Development (the Ministry) requesting, under the Official Information Act 1982, the following information regarding how the Ministry uses statistical methods to help automate decisions:

- *On what decisions is stats being used?*
- *What sort of model is being used? (linear, k-NN, decision tree, neural network, etc...)*
- *What data is used to train these models? what are examples of the inputs and outputs?*
- *How much data is used? and where is it from?*
- *How does MSD evaluate the performance of models used for each decision?*
- *What is the performance of these models?*

You specified that you were only interested in machine learning that the Ministry has implemented or commissioned, and not statistics used within products such as Microsoft Office.

For the sake of clarity, each of your questions are addressed in turn.

- *On what decisions is statistical methods to help automate decisions being used?*

To date, the Ministry is currently using the Youth Service NEET (Not in education, employment or training) Model to help automate decisions.

The Youth Service is a contracted service, established in 2012, under which community-based providers work with disengaged, at risk or unemployed 16 to 19 year olds.

The Youth Service NEET model was developed to assist the business in identifying young people at risk of long-term benefit receipt to whom they can offer the Youth Service. The Youth Service NEET model primarily uses education information from the Ministry of Education (MoE) to identify young people who are identified as NEET or are at risk of becoming NEET.

The information about these young people allows the Youth Service NEET model to create a risk score. This risk score is then converted into a risk rating (Very Low, Low, Medium, High). The risk rating then helps determine if the young person qualifies for the Youth Service or not.

The Aurora Centre / 56-66 The Terrace / Wellington 6011
PO Box 1556 / Wellington 6140 / New Zealand
Phone: 04 916 3300 / **Fax:** 04 918 0099 / **www.msd.govt.nz**

A young person accessing the support of the Youth Service is entirely voluntary, and identification by the Youth Service NEET model is just one way a young person may access this support. Young people not identified by the Youth Service NEET model may also be referred directly to the service to be assessed, for example, by themselves, a parent or teacher.

The Ministry acknowledges that no model is perfect. The risk rating from this Youth Service NEET model is just one input for the Ministry's work with youth. This work to get the best outcome for youth comes from an on-going relationship with them, face-to-face engagement, listening to what they say, and offering the right support that is tailored to their individual circumstances.

Although it has been used in the past, Client Service Matching is not being used operationally, but you may be interested to read more about it at the following link: www.msd.govt.nz/documents/about-msd-and-our-work/work-programmes/initiatives/phrae/client-service-matching.pdf.

- *What sort of model is being used? (linear, k-NN, decision tree, neural network, etc...)*
- *What data is used to train these models? what are examples of the inputs and outputs?*
- *How much data is used? and where is it from?*
- *What is the performance of these models?*

Please find attached, a copy of the document *Youth Services – Plan A 'School Leavers Model' Update – Technical Report,* which is the most recent, finalised technical report. You will find the answers to your questions above, in detail, within the report.

The name of the author of this document is withheld under section 9(2)(a) of the Official Information Act in order to protect the privacy of natural persons. The need to protect the privacy of this individual outweighs any public interest in this information.

- *How does MSD evaluate the performance of models used for each decision?*

Please see the Treasury's Youth Service NEET report which describes the way the Ministry evaluates the performance of the Youth Service NEET Model in detail at the following link:
www.treasury.govt.nz/publications/wp/evaluation-impact-youth-service-neet-programme-html.

The principles and purposes of the Official Information Act 1982 under which you made your request are:

- to create greater openness and transparency about the plans, work and activities of the Government,
- to increase the ability of the public to participate in the making and administration of our laws and policies and
- to lead to greater accountability in the conduct of public affairs.

This Ministry fully supports those principles and purposes. The Ministry therefore intends to make the information contained in this letter and any attached documents available to the wider public. The Ministry will do this by publishing this letter on the

Ministry of Social Development's website. Your personal details will be deleted, and the Ministry will not publish any information that would identify you as the person who requested the information.

If you wish to discuss this response with us, please feel free to contact OIA_Requests@msd.govt.nz.

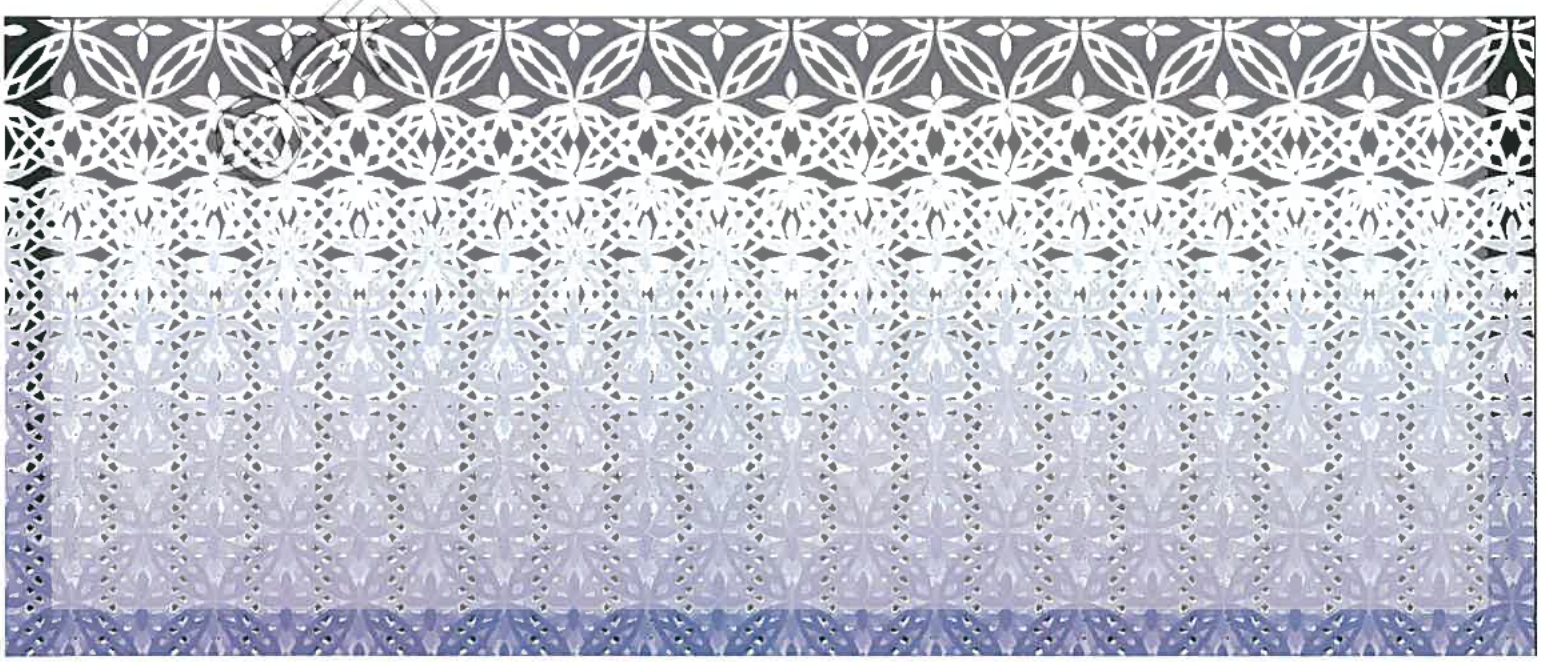
If you are not satisfied with this response regarding statistical methods used by the Ministry, you have the right to seek an investigation and review by the Ombudsman. Information about how to make a complaint is available at www.ombudsman.parliament.nz or 0800 802 602.

Yours sincerely

Daniel Lensen
General Manager
Client Business Intelligence

MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA

Youth Services – Plan A 'School Leavers Model' Update – Technical Report



Authors

s 9(2)(a) - Privacy of natural persons

Acknowledgements

Disclaimer

Date

ISBN

Contents

Table of figures and tables

Summary

This document describes the latest version of the Plan A 'School leavers' model for the Youth Service that has been trained against the most recent MoE extract (October 2015). It aims to replace the current version running into production since 2012. The main changes are the use of a bigger and more recent cohort of School leavers for the model training, the use of the CBI core for the MoE/WAI/CYF profile building and the use of the datamatch 2 for scoring. These changes result in a significant performance improvement: the AUR score is now 0.79 compared with 0.73 for the original 2012 version.

Here are summarized the main evolutions between the 2012 model and the 2016 update:

1. Project environment

The JIRA reference for the project is CBI-518 (link). The associated SVN folder is cbi-301_Youth_Services_Extension.

The EM project is cbi-301_Youth_Services_Extension, diagram YSX_AX06_01 for the last model built.

Note on libraries

The MOU signed between MSD and MOE specifies that the data provided for training should not appear in production environment –it can only be used for training and validation of the model. To deal with this restriction different libraries have been defined to host the 1992-2000 birth cohort data in one place and the regular data feed from MOE in another. What's more, the data comes with a frozen data match index. Consequently, when building the dataset to train[1] the models, the following datasets must be defined in your 'LIBNAMES' dataset:

```
SSIMOE:
/lev1_11/dev/cbi/external/SSI_301_Ministry_Of_Education_Data/files_training2_DO
_NOT_DELETE

SSIIDMGT:
/lev1_11/dev/cbi/external/SSI_301_Ministry_Of_Education_Data/datamatch_trainin
g2_DO_NOT_DELETE
```

Additionally, the commands

```
%clcm_override_libs(ssimoe);
%clcm_override_libs(ssiidmgt);
```

must be added in the ysx_autocbi.sas programme in order to override the official SSI libraries in case of training in DEV environment.

The library for the YS project (to be added in the 'LIBNAMES' dataset) is

```
CBIMYSX: /lev1_11/dev/cbi/ysx/files
```

CBI MOE events:

For training, CBI events libraries have to be used in DEV environment. This is important in particular for CBI MOE events, which have to have been generated by sourcing the DEV SSIMOE library given here above. This ensures the consistency between the training cohort and the corresponding MOE data.

For scoring cohorts of school leavers from the PROD SSIMOE library, PROD CBI events have to be used.

[1] The library paths given below are used for training only and will differ when scoring.

2. Building the training dataset

Modelling framework

The modelling framework for the model is as follow:

- Cohort: Students from the 1992-2000 birth cohorts who left school aged 15-17.
- Target: Being at least 3 month on a benefit in the three year window following the leaving event (Binary target). Target benefit group includes unemployment-related benefits –with the exception of Training related ones (UBT)-, emergency- and sickness-related benefits and sole parent ones. Note that student support and youth (YP/YPP) benefits are excluded from the target group.
- Profile window: 17 years (lifetime) leading to the leaving event (trigger for scoring).
- Forecast window: 3 years from the leaving event.
- Data sources: W&I, CYF, MOE (Enrolments, qualification, interventions and student identifiables). Data match based on the kiwid match.

The Figure 1 below depicts schematically the windows and events used for the creation of the training set.

Figure 1 - Depiction of windows and event for the training set

Training Cohort

As stated above, the aim of the risk rating model is to estimate the risk of long term benefit receipt for students aged 15, 16 or 17 when they leave school.

To build the model, data from Work & Income (WAI), Child, Youth and Family (CYF) and the Ministry of Education (MOE) related to every student from the 1992-2000 birth cohorts is considered (leaving year from 2007 to 2015). To ensure a 3 years forecast period to build the target variable at the time the model was built, only the students who left school between 2007 and 2012 at age 15 to 17 were used.

MOE has provided data reporting on the status of these students as of the 30th October 2015, including some personal characteristics (student ID and name, date of birth, gender and ethnicity), their history in the secondary education system (in terms of enrolments with -possibly several- schools, the start and end dates of each enrolments as well as the reason for leaving a given school) and the detail of interventions by the Ministry towards a given student (such as stand downs and suspensions, special education services, tests of English for speakers of other languages...).

Data matching algorithms are used to link the identity of students as recorded by MOE to the ones recorded by MSD in both the CYF and WAI space. This allows to get information on benefit history as well as on interaction with social services such as CNP and YJ to build a profile that "draws a picture" of the student as at the time of his leaving the education system.

The complete 1992-2000 birth cohorts based on the list of identities provided by MOE comprises 354,947 individuals. Of these, 173,098 have a 3 years forecast period before the model building (needed to build the target variable). Of these and after the data match process, 120,114 left school while they were aged 15-17.

These 120,114 individuals will constitute our training cohort.

The Table 1 below gives the distribution of the training cohort in term of age at leaving and target variable. Overall, 20.9% of the individuals of the training cohort have a positive outcome.

Table 1 - Leaving age distribution by target

Age at leaving school	Binary target variable: Over 3 months on benefit in outcome window		Total
	0	1	
15	8977 *78.57%*	2448 *21.43%*	**11425** ***9.51%***
16	30026 *72.49*	11393 *27.51*	**41419** ***34.48%***
17	55980 *83.22%*	11290 *16.78%*	**67270** ***56.01%***
Total	**94983** ***79.08%***	**25131** ***20.92%***	**120114**

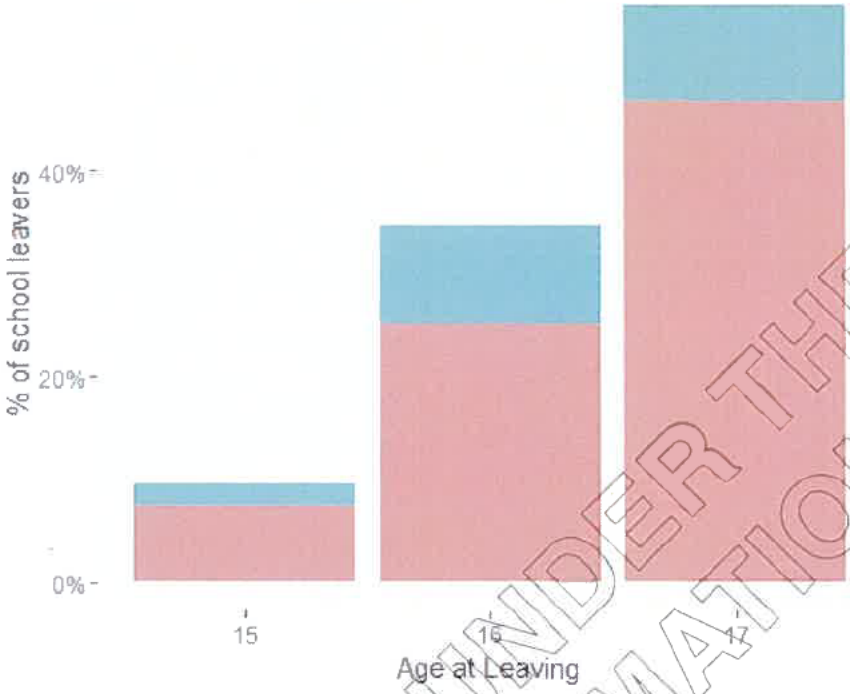


Figure 2- Leaving age distribution by target

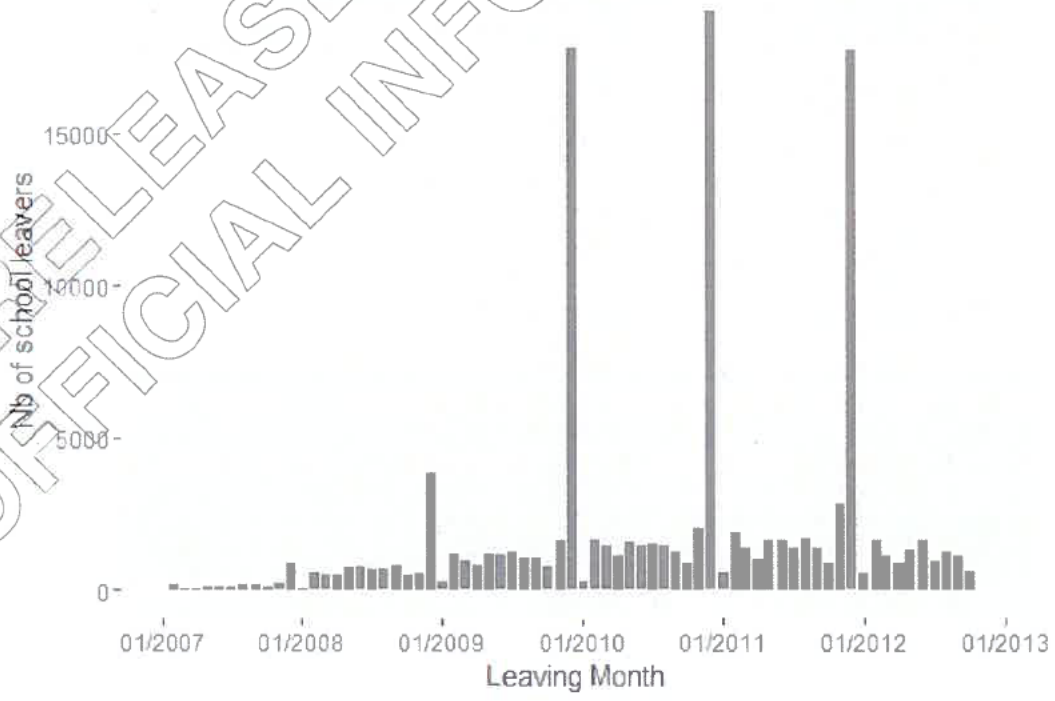


Figure 3 - Month of leaving distribution (training cohort)

Training Variables – predictors

The profile of the students used to predict to outcome is computed on a 17 years window leading to the leaving event, which represents the lifetime of the student. Classically, we compute for this window an extensive list of measures summarising the history of interaction with MSD (both in terms of benefit and CYF) in this window. The initial list of input variables (that is, the candidate predictors) include variables such as: the total time spent supported on a benefit (or more likely, associated to a caregiver's benefit) and corresponding number of spells, as well as breakdowns per type of benefit; the number of CYF (both CNP and YJ) events, again including breakdown per type of event.

Additionally, from the MOE data we compute similar summary variables. These include the number of NCEA level 1, 2 and 3 passes, the number of awards of merit or excellence, the count of all interventions (as detailed above), the number of enrolments as well as the reason for leaving school.

The Table 2 below summarizes the list the candidate predictors. The list of the 40 selected significant variables to be used by the model is given in Appendix 1 – List of selected predictors

Table 2 – List of candidate predictors for training

Source	Variables (predictors)	Target variables
MSD - WAI	Total time spent on benefit, number of spells, days since last spell, days to first spell (from start of the profile window), and status (Past or Current) at time of the profile date; the above is computed for all benefit types as well as per benefit type.	Binary indicator for over 3 months on benefit during the 3 years forecast period
MSD - CYF	Count, duration and costs of all events related with CNP, YJ and reports of concerns –including investigations for and findings of abuse (overall and per type). Indicator of level of involvement with CYF (None, investigations, findings, intake).	

MOE	Count and duration of enrolments and of interventions (overall and per type, e.g. per reason for ending an enrolment or per intervention type); number of NCEA L1, L2 and L3 passes; number of awards for merit or excellence. Reason for leaving school. Indicator of leaving school before the end of the school year. Characteristics of the student: gender, age at time of leaving.

Some predictor's distributions for the training cohort are plotted on the following figures.

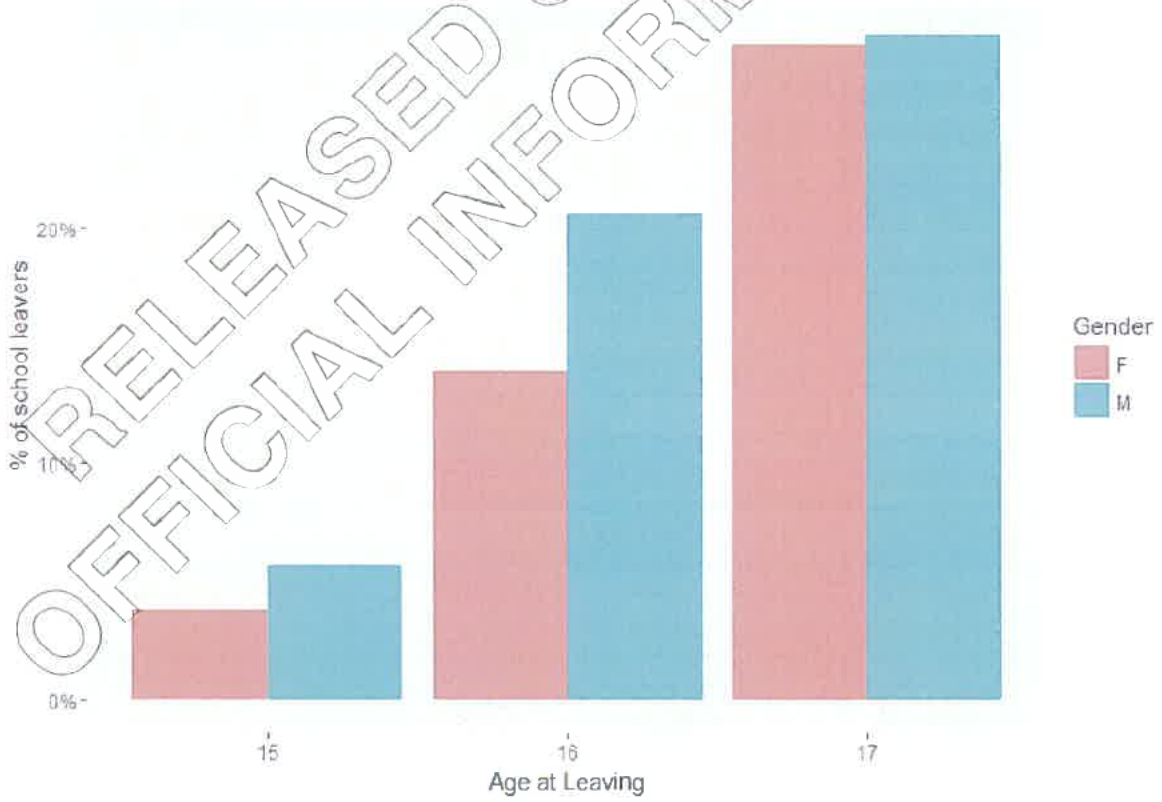


Figure 4 - Leaving age distribution by gender

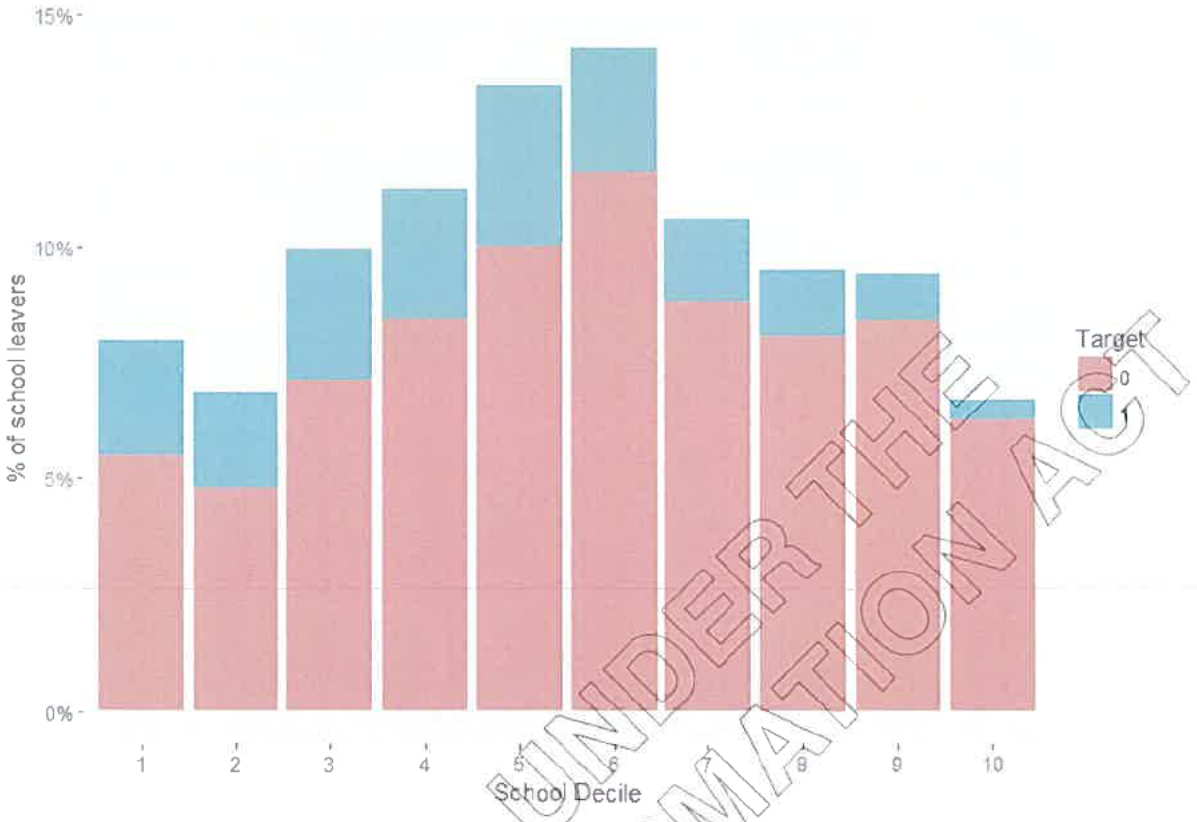


Figure 5 - Decile of last school enrolment distribution

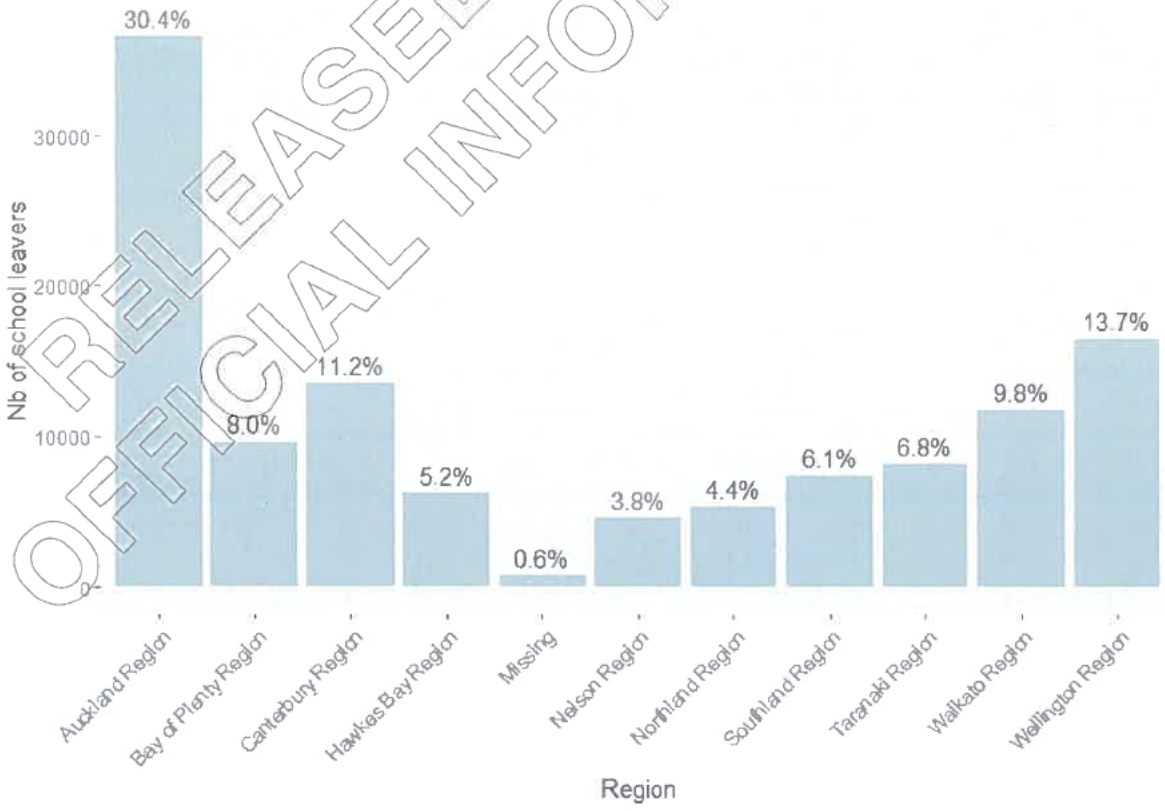


Figure 6 - School region distribution

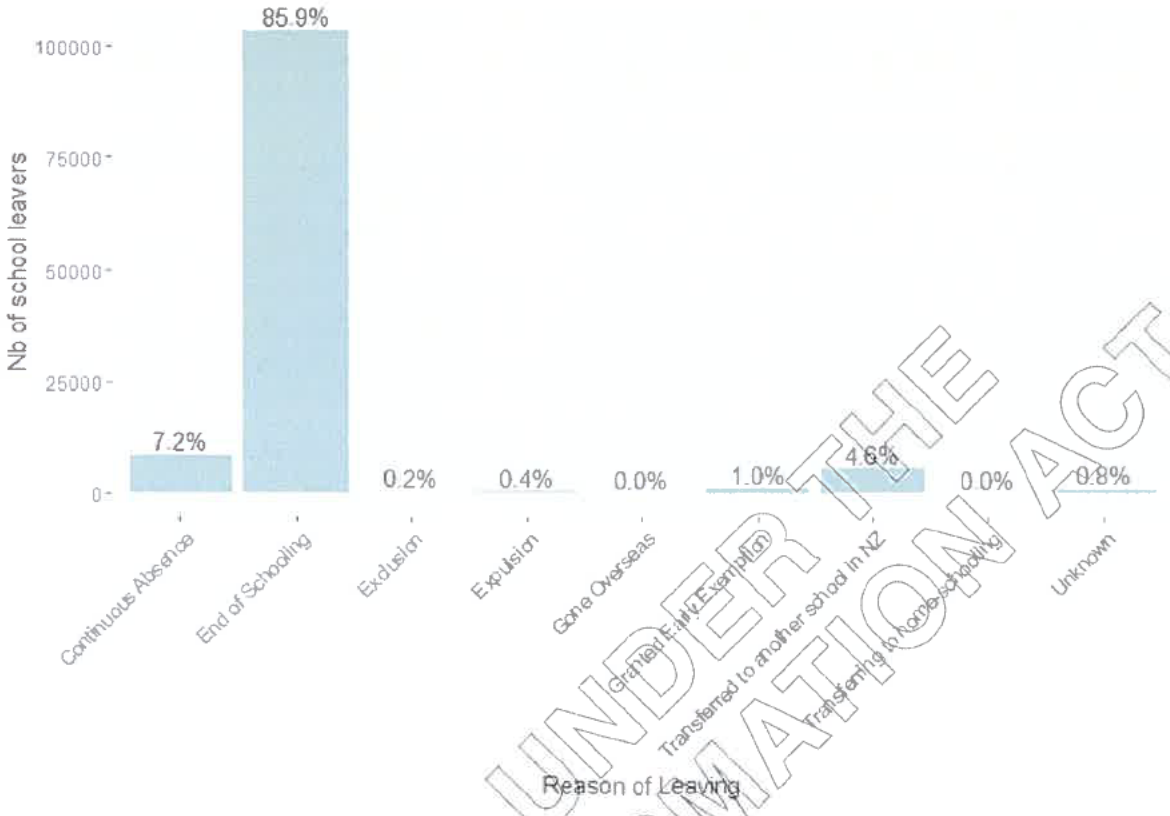


Figure 7 - Reason of leaving distribution

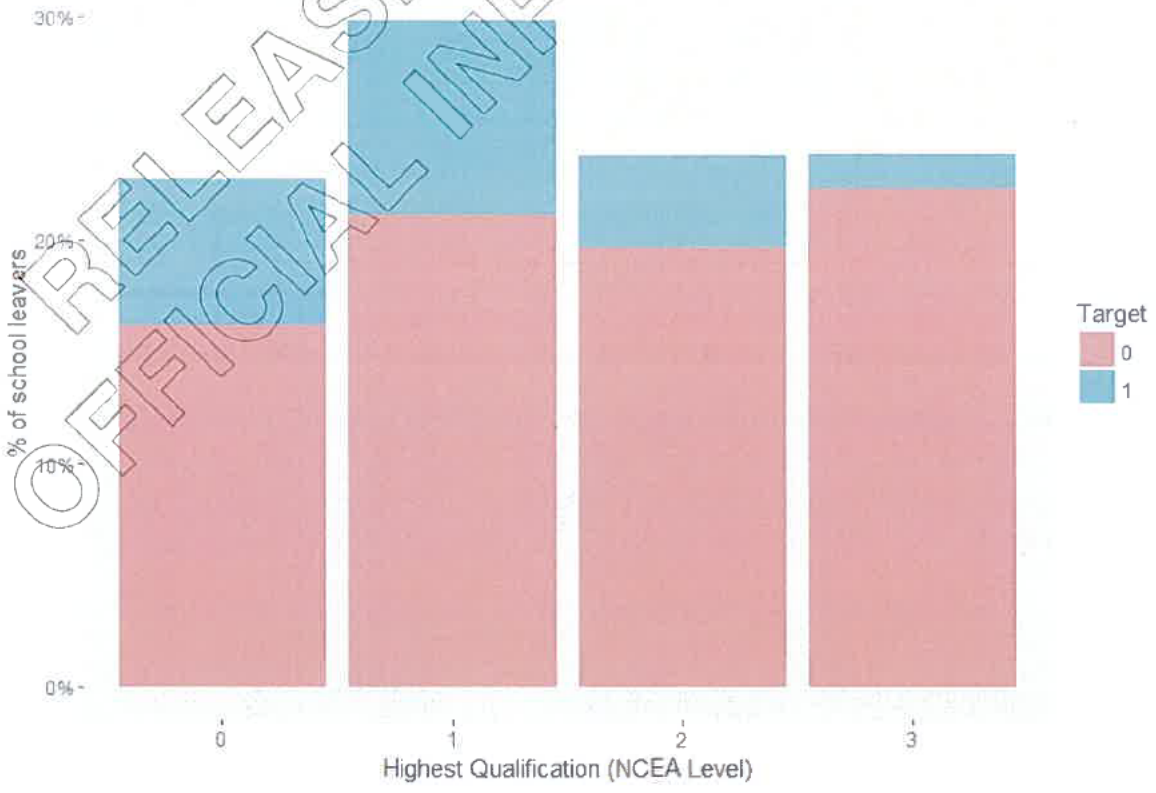


Figure 8 - Highest NCEA level at leaving

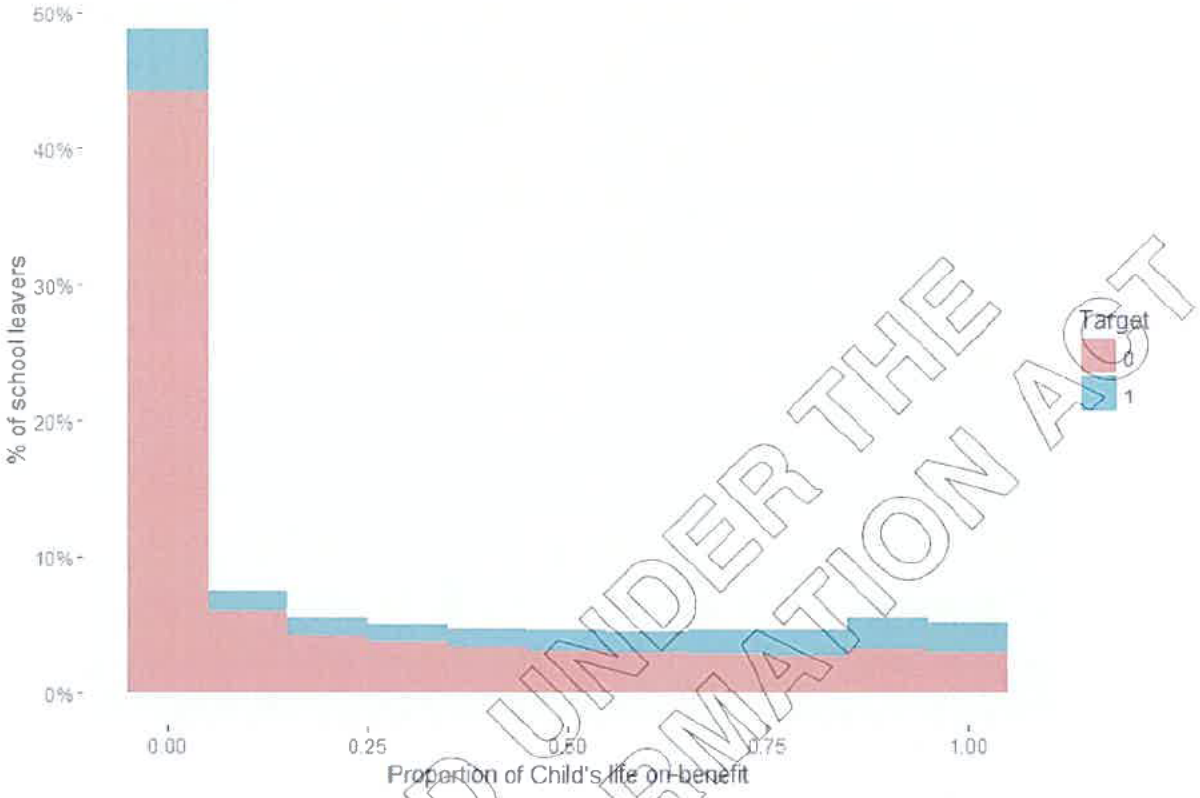


Figure 9 - Proportion of child's life on benefit distribution

Identity matching

Since the training set building process consists in gathering WAI, CYF and MOE profiles, a data match has to be used to match the different source id for each unique individual. A static master index table and specific to the training cohort is used (SSIIDMGT given in the first section). This master index contains ART, CYF, WAI and MOE source ids.

The used cluster id is the 'kiwid', which is the lowest matching level for youth services.

The Table 3 below provides the source ids distribution of this static master index used for the training cohort (354,943 Student MOE identities):

Table 3 - Training master index sources distribution

Source System Code				
SOURCE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ART	183809	13.57	183809	13.57
CYF	182355	13.46	366164	27.03
MOE	**354943**	26.20	721107	53.23
WAI	633654	46.77	1354761	100.00

Ambiguous datamatch

It may happen that several 'primary' id per source are found for a unique individual (unique kiwid) in the master index table. In that case, a flag 'Ambiguous datamatch' is set to 'Y' in the output dataset (training or scoring). These cases represent 2.55% of the training cohort.

SAS EG project

In cbi-301_Youth_Services_Extension.

The process flow for the project in SAS EG is as follows:

1. ysx_autocbi
2. ysx_generate_control
3. ysx_generate_cohort_training_plan_a
4. ysx_main

YSX_AUTOCBI

Simple setup of the environment; calls to `%clcm_override_libs(ssimoe)` and `%clcm_override_libs(ssiidmgt);`

YSX_GENERATE_CONTROL

Generates the control table.

The training cohort used to rebuild the Plan A 'School leavers' models is labelled YSX_AX06_TRAIN. As stated in the previous section, data from W&I, CYF and MOE is used with the 'kiwid' match and the CLUSTER option. No relationship is taken into account.

The scoring of children is triggered by the leaving event, so that profile dates are different for each student. Consequently, the 'USER' option is used for the specification of the cohort. The dataset 'cbimysx.ysx_ax_train_cohort' indicates the list of IDs as well as all the dates needed (history, profile, forecast).

Note that since the option is set to 'USER', the parameters `cles_prfl_period`, `cles_frcst_period`, `cles_days`, `cles_agemin` and `cles_agemax` are not used by the programme.

The corresponding complete line in the control table is given below.

```
,ysx_ax06,TRAIN,USER  ,ADULT, 17y,3y,kiwid          ,
FAST,CLUSTER,N,NONE,Y,cbimysx,Y,Y,Y,N,N,N,YSX_AX06 Training
set,365,15,17,cbimysx.ysx_ax06_train_cohort,
```

Note that the control table contains a line that defines the cohort considered in scoring mode, and will be detailed in a following section.

YSX_GENERATE_COHORT_TRAINING_PLAN_A

Creates the window dataset `YSX_AX06_TRAIN_COHORT` that is looked for in the CBIMYSX library by the programme (as indicated in the control table).

The dataset contains 173,098 MOE source ids from the 1992-2000 birth cohorts alongside all the dates needed to create the profiles and target. For each individual, the profile date is set as the recorded time of leaving school and the history date is set 17 years before that. The forecast date is set 3 years after the profile date and the student is kept in the cohort only if the forecast date is prior to the MOE data extract date (to have a full 3 years forecast period). The analysis date (not used) is set at the 30th October 2015 (date of the MOE data extract).

YSX_MAIN

The main programme building the training dataset.

The following steps are standard from the CBI core macros, calling parameters, creating the specific master index, building the standard profile and doing standard imputation and cleaning:

`%cles_setup()`, `%cles_get_master_index()`, `%cles_get_cl_window()`, `%cles_get_cl_profiles()`, `%cles_get_related_persons()`, `%cles_final_prep_01_relpers()`, `%cles_final_prep_02_expvars()`, `%cles_final_prep_03_targets()`, `%cles_final_prep_04_shapes()`, `%cles_final_prep_05_imputes()`, `%cles_summary_stats()` and `%cles_cleanup()`

Note that `%cles_get_related_persons()` does not do anything in the present case.

Project-specific programmes are called:

`%ysx_final_prep_02_expvars()`

In addition to the standard profile variables generated by the CBI core, a set of 'expert variables' are created to enrich the list of candidate predictors:

- Youth Service: count of qualifications achieved with merit or excellence:

```
ysx_exp_moe_awa_mer_exc =SUM(moe_yse_qal_2awm_cnt,moe_yse_qal_2awe_cnt)
```

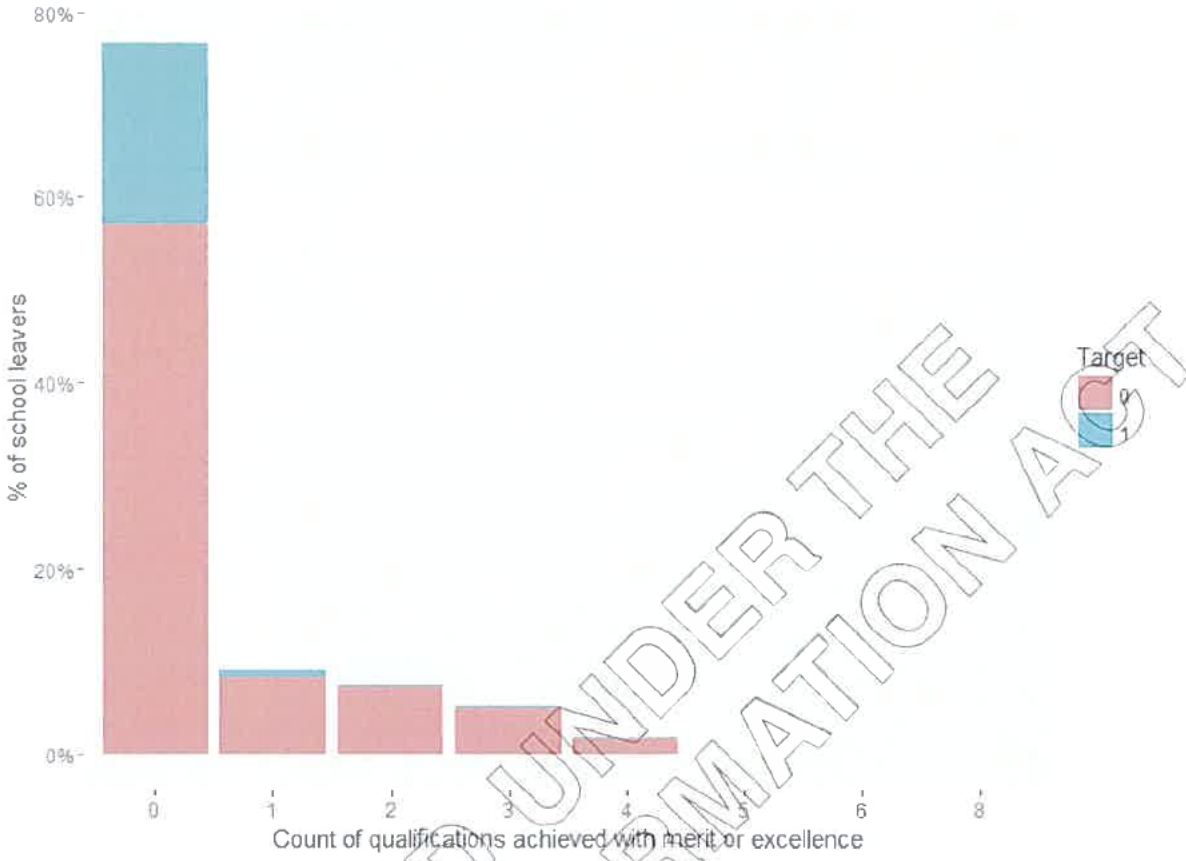


Figure 10 - Qualifications with merit or excellence

- Youth Service: indicator of CYF history in profile period

```
if cyf_cec_all_cnt>0 then ysx_exp_cyf = 1;  else ysx_exp_cyf = 0;
```

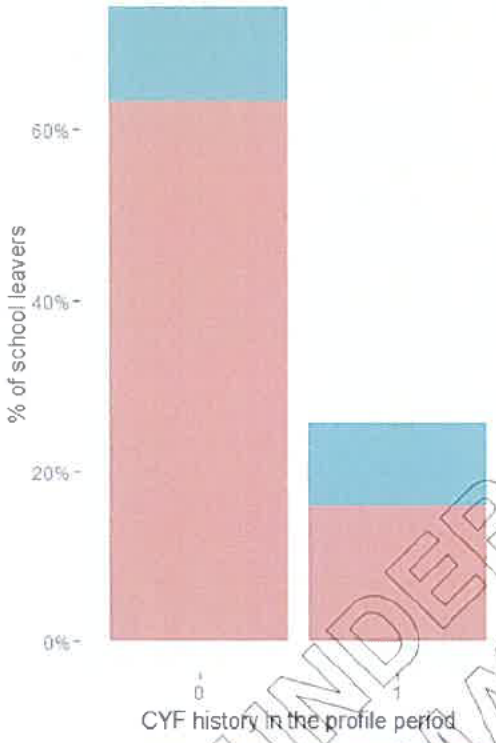


Figure 11 - CYF history flag

- Youth Service: indicator of WI history in profile period

```
if win_bdd_mbs_dur then ysx_exp_wi = 1;    else ysx_exp_wi = 0;
```

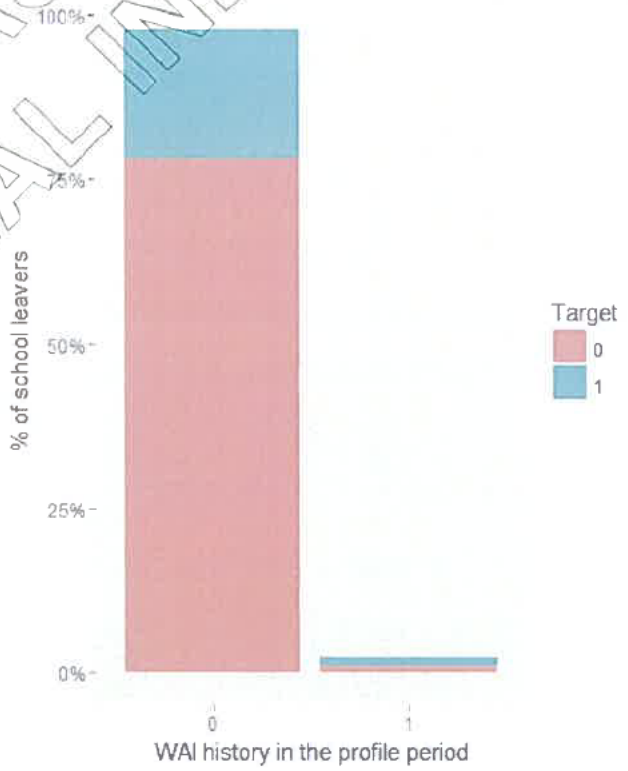


Figure 12 - WAI history flag

- Youth Service: proportion of child's life on benefit

```
ysx_exp_win_bdd_chd_life_prop = min(1,max(0,
win_bdd_chd_dur/(moe_yse_pch_lst_sch_day - moe_yse_pch_dob)));
```

- Youth Service: year that left school

```
ysx_exp_moe_yse_pch_lst_sch_year = year(moe_yse_pch_lst_sch_day);
```

Figure 13 - Year of leaving

- Youth Service: month that left school

```
ysx_exp_moe_yse_pch_lst_sch_mth = month(moe_yse_pch_lst_sch_day);
```

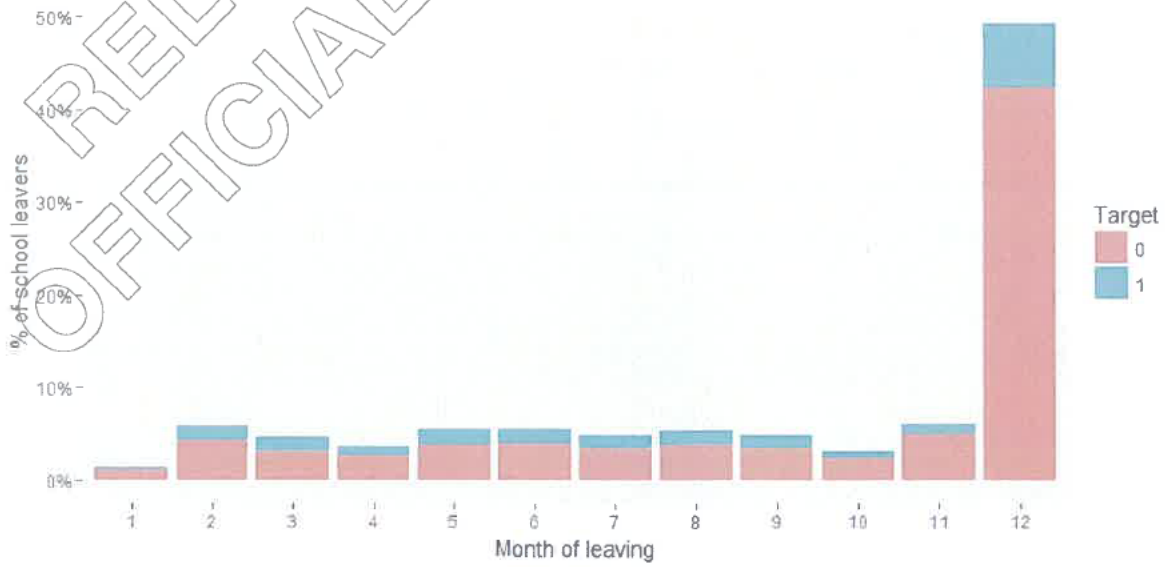


Figure 14 - Month of leaving

- Youth Service: highest NCEA qualification

```
ysx exp moe yse qal highest = 0;
if moe_yse_qal_1lv1_cnt then ysx_exp_moe_yse_qal_highest = 1;
if moe_yse_qal_1lv2_cnt then ysx_exp_moe_yse_qal_highest = 2;
if moe_yse_qal_1lv3_cnt then ysx_exp_moe_yse_qal_highest = 3;
```

- Youth Service: MOE high need intervention

```
if moe_yse_int_1net_cnt or moe_yse_int_1s78_cnt or moe_yse_int_1sd6_cnt
or moe_yse_int_1mp8_dsf or moe_yse_int_1019_dsf or moe_yse_int_1021_dsf
or moe yse int 1022 dsf or moe yse int 1ae5_cnt or moe_yse_int_1ses_dsf
then ysx_exp_moe_yse_int_high_need = 1;
     else ysx_exp_moe_yse_int_high_need = 0;
```

Figure 15 - MOE intervention flag

`%ysx_final_prep_03_targets()`

Computes the defined target, based on the total, cumulative duration spent on benefit in the forecast window. The duration is computed as follows:

```
tgt_ysx_ben_dur =
/* unemployment related benefit -except '608': UBT-related*/
f_win_bdd_mbs_1unm_dur - f_win_bdd_mbs_1unm_2608_dur
/* emergency related benefit */
+ f_win_bdd_mbs_1emo_dur
/* sickness related benefit */
+ f_win_bdd_mbs_1sic_dur
/* sole-parent related benefit */
```

```
+ f_win_bdd_mbs_lsop_dur;
```

And the binary target variable `tgtb_ysx_ind_3mth` is then computed as:

```
if tgt_yse_ben_dur > 91 then tgtb_yse_ind_3mth = 1;
     else tgtb_yse_ind_3mth = 0;
```

`%ysx_final_prep_04_shapes()`

- Drops some useless variables (dates) and *_dod and cluster_* variables which may contain future information.
- Only keeps students between 15 and 17 years old at profile date, non-deceased.

`%ysx_final_prep_05_imputes()`

- Imputes some missing values

`%ysx_final_prep_06_labels()`

- Merge back the sourceid in the output dataset
- Labels the output dataset

`%ysx_score_all()`

- Does nothing in TRAIN mode, the SCORE mode is detailed in a following section.

DATASETS FLOW

In the Table 4 below is given the details of the created intermediate and final datasets:

Table 4 - Datasets flow for training set building

Description	Datasource	Count
Number of StudentIDs in the cohort. (year of birth 1992-2000, leaving year 2007-2015)	SSIMOE.student_identifiable (extract_date=30Oct2015)	354,947
Number of StudentIDs in cohort dataset passed to modelling program. **Only students with a full forecast period of 3 years from the last day to the extract date**	CBIMYSX.ysx_ax06_train_cohort	173,098
Number of distinct ClusterIDs in matching table after rejection of bad identities (WIN&CYF)	CBIMYSX.ysx_ax06_master_index	354,136
Number of distinct ClusterIDs in cluster table after rejection of bad clusters	CBIMYSX.ysx_ax06_master_clusters	353,886

Number of distinct ClusterIDs in client windowset	CBIMYSX.ysx_ax06_window_cl	172,491
Number of distinct ClusterIDs in MOE client profile	CBIMYSX.ysx_ax06_clpr_cl_moe	172,491
Number of distinct ClusterIDs in WIN client profile	CBIMYSX.ysx_ax06_clpr_cl_win	136,650
Number of distinct ClusterIDs in CYF client profile	CBIMYSX.ysx_ax06_clpr_cl_cyf	47,321
Number of distinct ClusterIDs in merged client profile (MOE + CYF + WIN)	CBIMYSX.ysx_ax06_final_cl	172,491
Number of distinct ClusterIDs in merged client profile (MOE + CYF + WIN + Related persons)	CBIMYSX.ysx_ax06_merged01	172,491
As above but after addition of expert variables	CBIMYSX.ysx_ax06_merged02	172,491
As above but after adding targets	CBIMYSX.ysx_ax06_merged03	172,491
As above but after shaping: include only students who are 15-17 years old at last school-day, do not die on or before the forecast date, not registered as deceased with MoE and have at least one enrolment in the profile period	CBIMYSX.ysx_ax06_merged04	120,114
As above but after imputation	CBIMYSX.ysx_ax06_merged05	120,114
As above but after final sort and labelling	CBIMYSX.ysx_ax06_train	120,114

3. Building the model

SAS EM project

The EM project is cbi-301_Youth_Services_Extension:

- Data source: CBIMYSX.YSX_AX06_TRAIN
- Diagram YSX_AX06_01 for the last model built, adapted from the CBI model template cbi-283_EM_template, given in appendix...

The candidate models are built using the classic flow:

- Random sample of 50,000 observation extracted from YSX_AX06_TRAIN
- A data partition node does a 70/30 split to create the training and validation datasets.
- Variable selection nodes
- Candidate models nodes: Logistic regression (forward and stepwise), decision trees (entropy and gini), ensemble trees, gradient boosting, random forest. SVM and neural networks were tested but failed to converge
- Model comparison node for the model selection.

An iteration loop was done to select manually the 40 most significant predictors. The list of considered variables is given in Appendix 1 – List of selected predictors.

The Figure 16 and Table 5 below give the ROC curves and the AUC values for the different tested models:

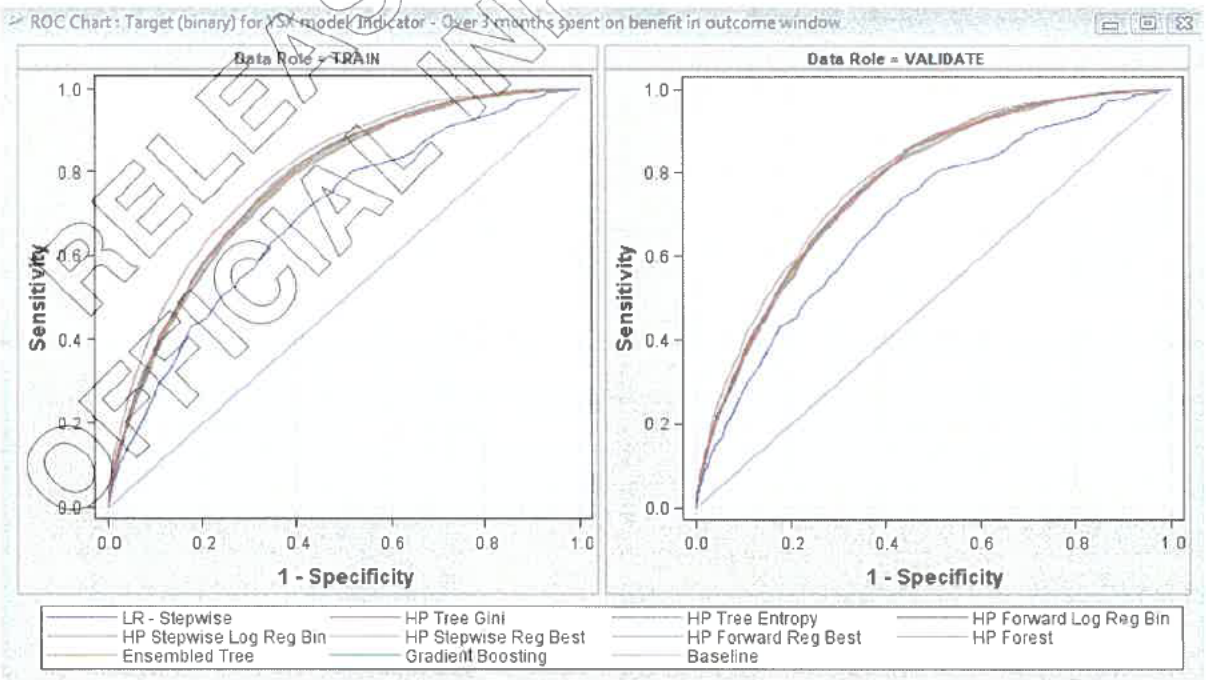


Figure 16 - ROC curves of candidate models

Table 5 - Models performance comparison

Selected Model	Model Node	Model Description	Target Variable	Train: Roc Index	Selection Criterion: Valid: Roc Index
Y	HPDMForest	HP Forest	tgtb_ysx_ind_3mth	0.798	0.785
	Ensmbl	Ensembled Tree	tgtb_ysx_ind_3mth	0.779	0.776
	HPTree	HP Tree Entropy	tgtb_ysx_ind_3mth	0.775	0.772
	HPTree2	HP Tree Gini	tgtb_ysx_ind_3mth	0.776	0.772
	Boost	Gradient Boosting	tgtb_ysx_ind_3mth	0.773	0.771
	HPReg	HP Forward Reg Best	tgtb_ysx_ind_3mth	0.768	0.768
	HPReg2	HP Stepwise Reg Best	tgtb_ysx_ind_3mth	0.768	0.768
	HPReg3	HP Stepwise Log Reg Bin	tgtb_ysx_ind_3mth	0.766	0.766
	HPReg4	HP Forward Log Reg Bin	tgtb_ysx_ind_3mth	0.766	0.766
	Reg	LR - Stepwise	tgtb_ysx_ind_3mth	0.693	0.695

The winning model is the random forest with AUC(train) = 0.798 and AUC(validate) = 0.785.

Scoring and performances

The scoring code is extracted from the SAS EM scoring node and integrated to the SAS EG project in the macros `%ysx_score_all()`, `%ysx_score_ysx_ax06()` and `%ysx_ax06_ise_3mth_ind()`.

The HP random forest model is not scored by a classic SAS code using data steps but thanks to the SAS proc 'HP4SCORE' and a binary score file generated by EM. This binary file has to be included in the deployment process and deployed in the considered environment.

Additionally to the risk score (probability between 0 and 1), a risk rating representing 4 levels of risk is generated: High (top 10% of the caseload), Medium (10-20%), Low (20-40%) and Very Low (40-100%).

The following Table 6 is the classification table for the scored training set, giving the lift, the True Positive and Negative Rates and the Classification rate relative to the caseload:

Table 6 - Final model classification table

Risk Rating	Caseload	Lift	TPR = Sensitivity	TNR = Specificity	Classification rate = Accuracy
High	5%	3.1	16%	98%	81%
Medium	**10%**	**2.8**	**28%**	**95%**	**81%**

	20%	**2.4**	**47%**	**87%**	**79%**
Low	30%	2.1	62%	78%	75%
	40%	**1.9**	**74%**	**69%**	**70%**
Very Low	50%	1.7	83%	58%	63%
	60%	1.5	90%	48%	56%
	70%	1.4	95%	36%	48%
	80%	1.2	98%	25%	39%
	90%	1.1	99%	12%	30%
	100%	1	100%	0%	20%

The Figure 17 below shows the score distribution and the risk rating thresholds for the scored training dataset.

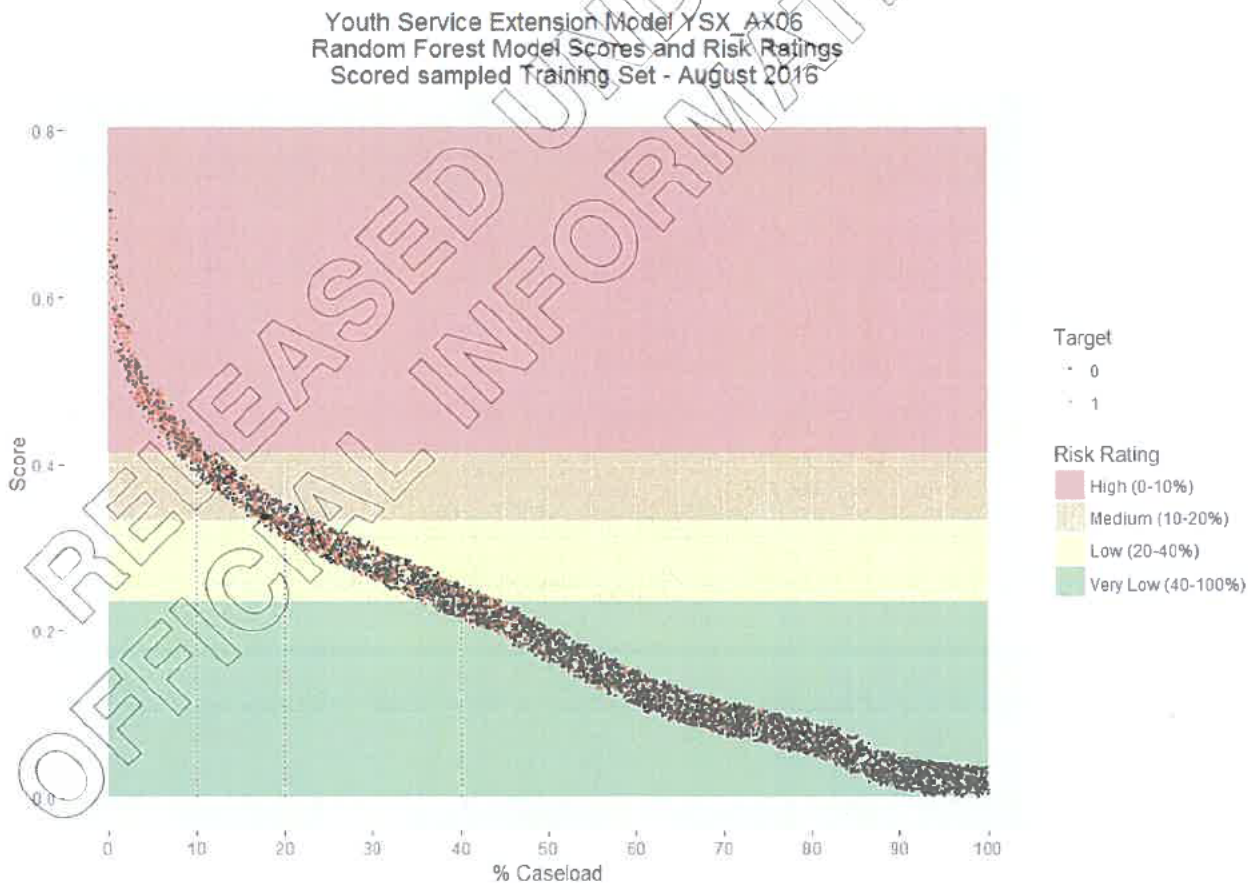


Figure 17 - Model scores distribution

The Table 7 below gives some characteristics of the scored training cohort.

Table 7 - Training cohort characteristics

	% by risk rating of the scored training set				
Characteristic	High (0-10%)	Med (10-20%)	Low (20-40%)	VeryLow (40-100%)	All
≥ 3 months on benefit in outcome window	54.99	37.59	24.79	7.89	20.92
Gender=Male	35	47.14	60.53	57.76	54.5
1+ Passes at Level 3	1.42	2.57	5.78	40.49	23.77
1+ Passes at Level 2	9.64	18.45	29.56	69.05	47.31
1+ Passes at Level 1	56.73	63.21	69.39	87.13	76.67
1+ Endorsements Achieved with merit	0.26	0.61	3.89	40.87	23.29
2+ Enrolments	82.99	69.94	63.25	39.75	53.94
1+ Interventions	70.28	55.16	45.54	14.7	33.2
Has CYF involvement in profile window	77.47	48.74	32.59	5.37	25.76
Has WIN involvement as a child	99.02	99.17	93.72	26.93	58.76
Left school before end of year	64.74	59.91	53.9	24.83	40.14

4. Scoring the production data

MOE school leavers data feed

The SSIMOE production library is updated fortnightly by a MOE data feed with the latest school leavers data from the previous two weeks. This data feed process triggers the model scoring process which scores the latest school leavers.

Until August 2016, 224,808 school leavers from age 15 to 17 have been scored in total by the previous 2012 model.

The Figure 18 below shows the school leavers distribution in SSIMOE:

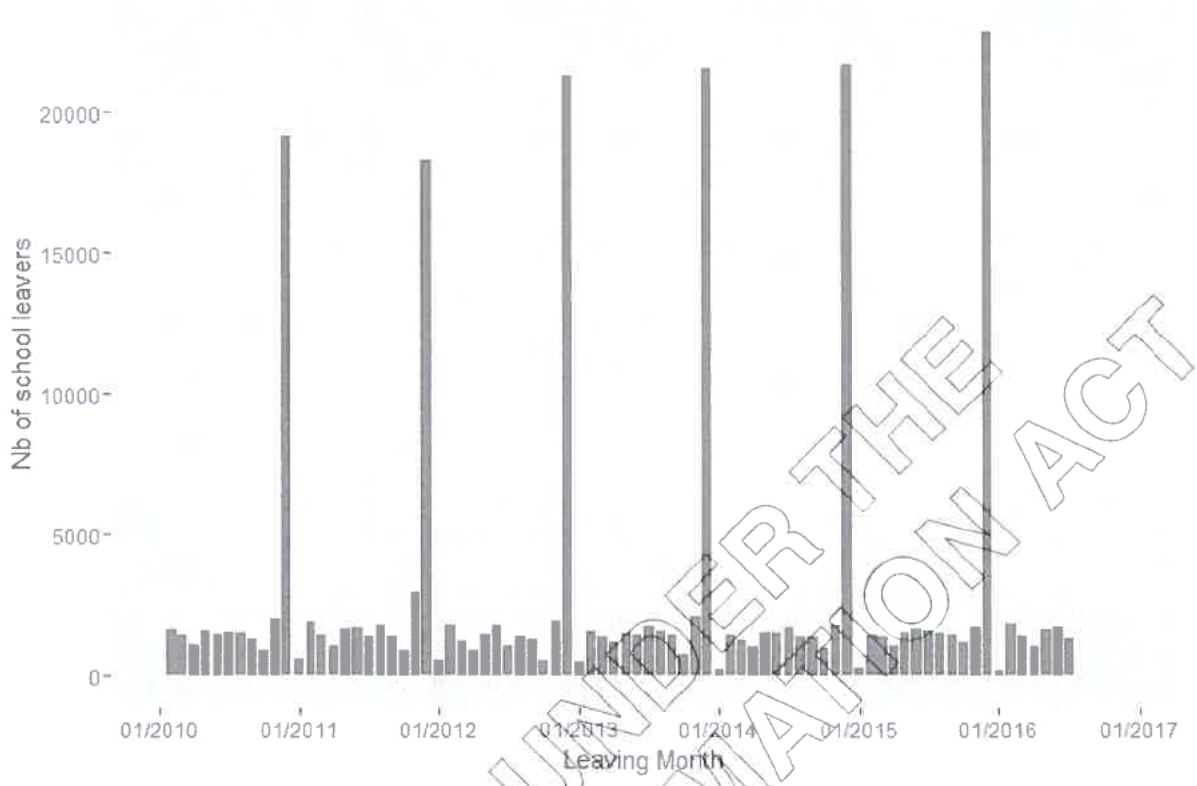


Figure 18 - School leavers' distribution (production)

Youth Services datamatch

As specific datamatch process has been coded by IAP, producing a daily master index table for identity matching, and derived from the official datamatch 2. The purpose of this specific datamatch is to exclude Corrections identities, not allowed to be used in the framework of this model.

The details of this process are given in Appendix 3 – Youth Services Datamatch process.

Scores distributions

The Figure 19 below shows the risk rating distribution of the scored production MOE cohorts since 2012 by the previous model (logistic regression).

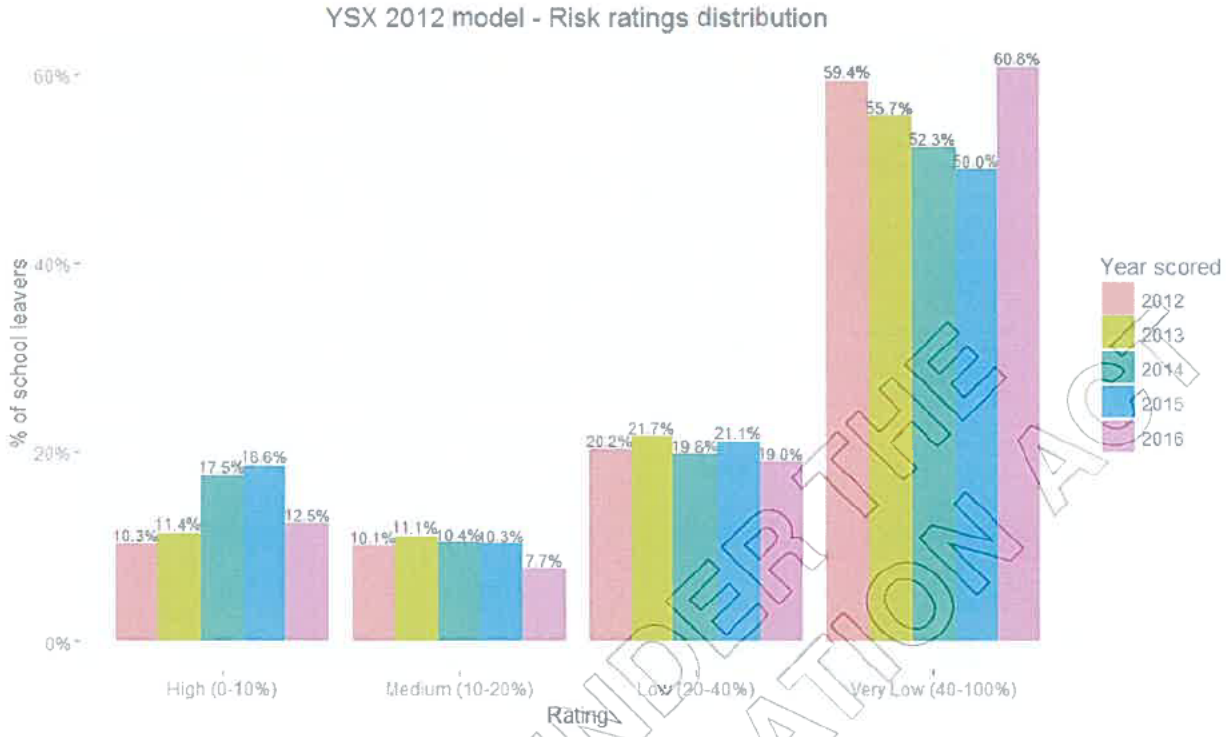


Figure 19 - 2012 model risk rating distribution

The risk rating thresholds between High/Medium/Low/Very_Low in term of risk score have been initially established with the first model in 2012. A significant drift can be noticed since 2012 in terms of the size of the risk rating groups, especially for the High and Very_Low categories. The High risk group increases from 10.3% in 2012 to 18.6% in 2015, while the Very_Low group drops from 59.4% in 2012 to 50% in 2015. This means that the proportion (and the number) of school leavers reported "at risk" has increased during the last years. Note that the 2016 distribution does not reflect a full year distribution and cannot be compared to the previous years.

These thresholds have been updated in July 2016 for the 2012 model, based on 2015 scores, in order to counter this drift and adjust the risk rating sizes to the target 10%-10%-20%-60%.

The Figure 20 shows the risk rating distribution for the 2016 updated model, with thresholds based on 2015 risk scores. The random forest model demonstrates a better stability in time in terms of the size of risk ratings groups. Again, the 2016 distribution does not reflect a full year distribution because does not include lower risk population of students leaving school at the end of the year.

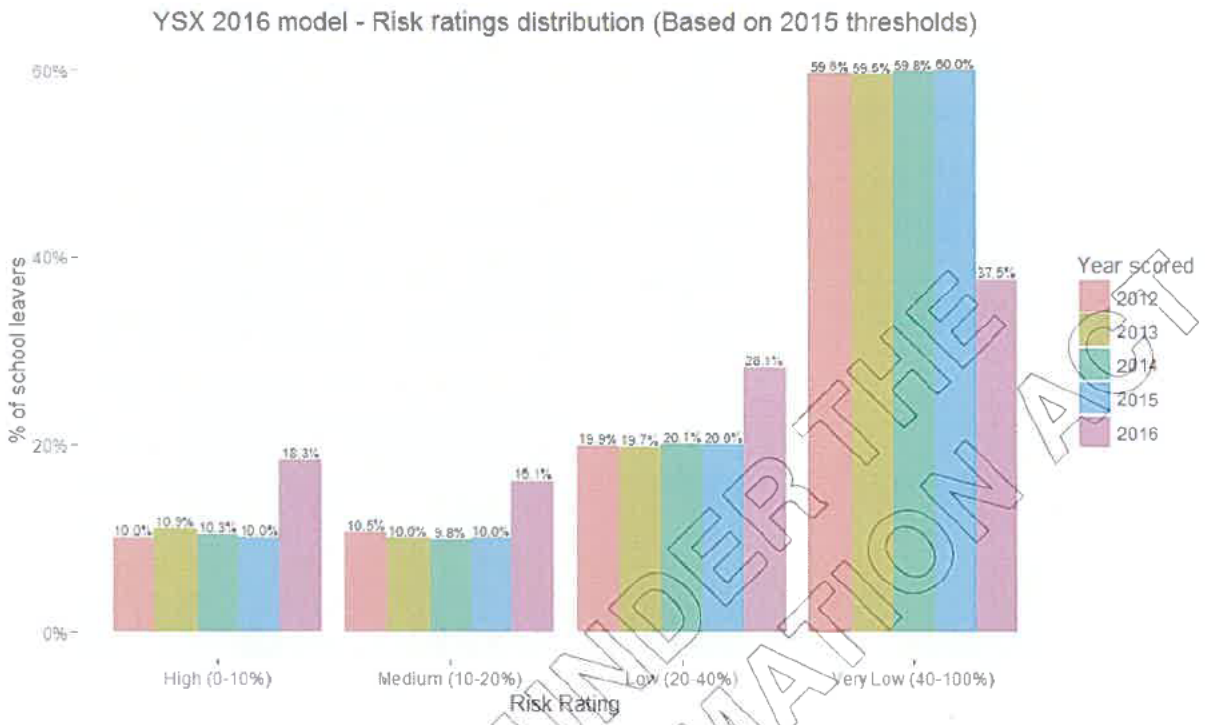


Figure 20 - Updated 2016 model risk rating distribution

Risk ratings thresholds

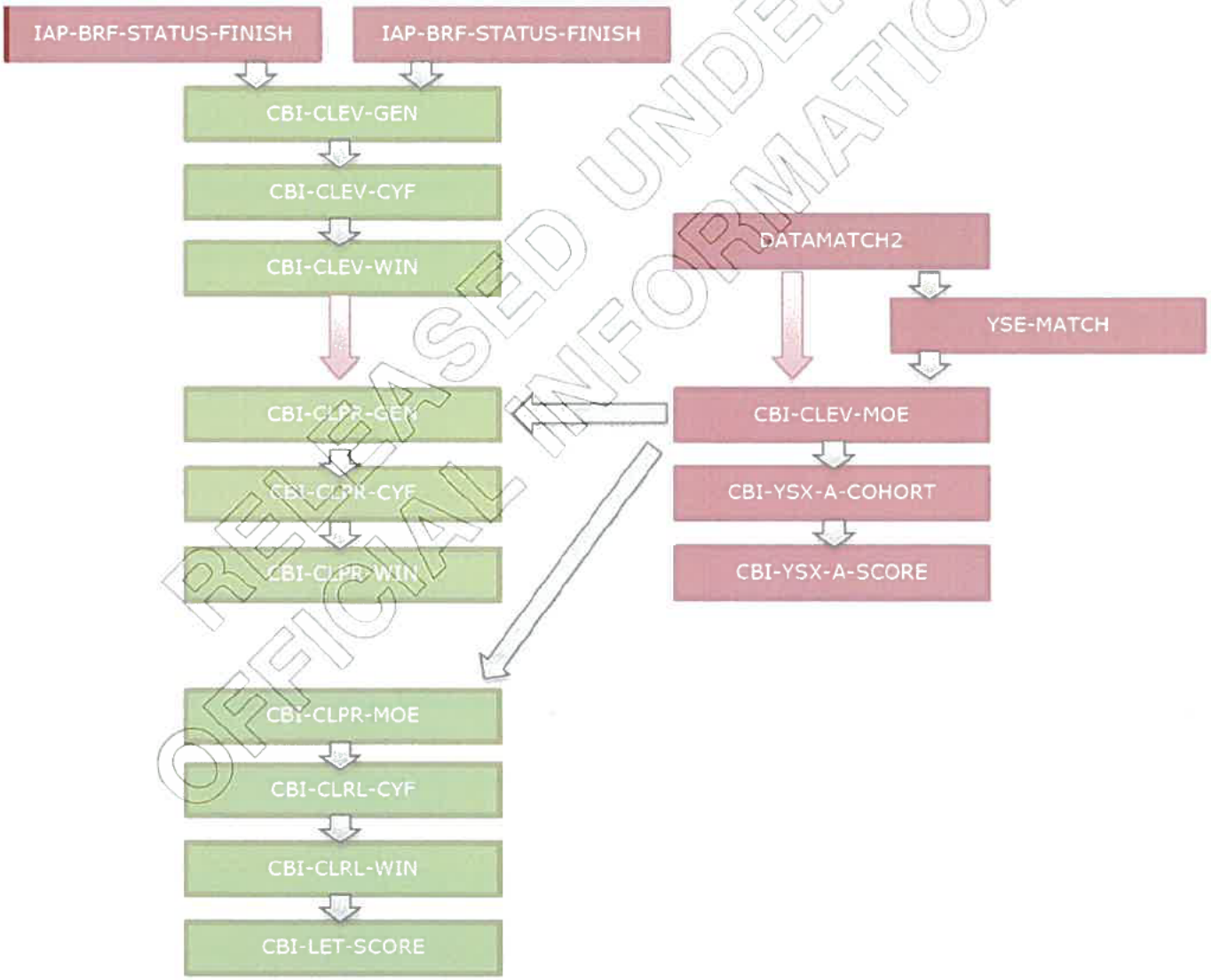
The Table 8 below gives the calculated thresholds for the mapping of risk rating categories. These thresholds have been established considering the distribution of scores for the school leavers in 2015, the latest full school year. As seen in the previous part, the risk rating distribution is quite constant between 2012 and 2015, which means that we can be relatively confident about their consistency in the future.

Table 8 - Risk ratings thresholds

Model score	NEET model rating	Comment
.	Missing	
0.00000000 -< 0.20253685	Very Low	0th - 70th percentile
0.20253685 -< 0.32347140	Low	70th - 80th percentile
0.32347140 -< 0.41431427	Medium	80th - 90th percentile
0.41431427 - 1	High	90th - 100th percentile
-100	Age out of Range	Age <15 or >17
-101	No MoE record	No MoE data
OTHER	Other	Out of training range

These values are included in the SAS format `ysx_ax06_rating_original` which have to be generated once or after any update, after the deployment process, by the macro `%ysx_formats()`.

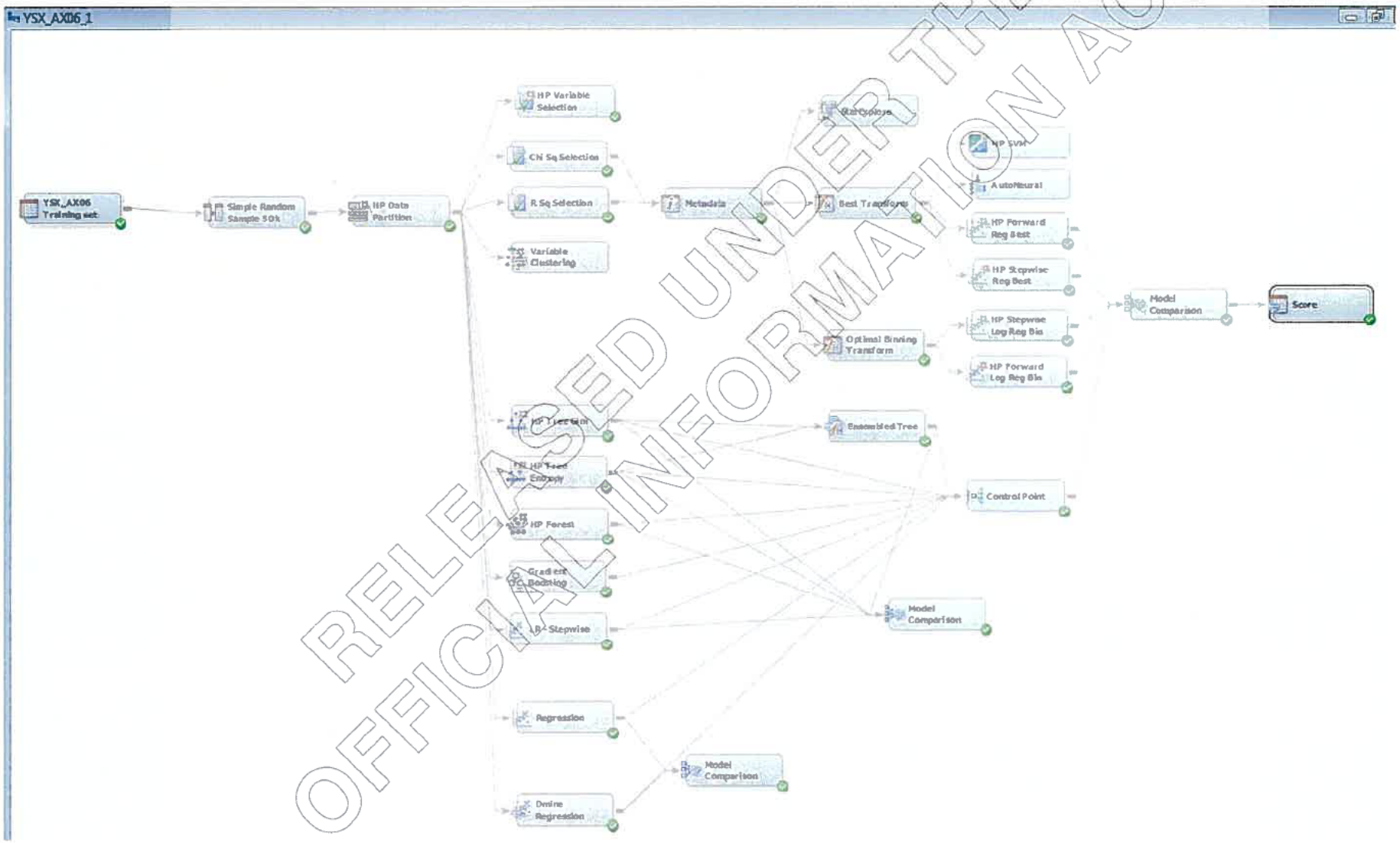
5. Model deployment and scheduled jobs flow



Appendix 1 – List of selected predictors

Variable Name	Number	Label	Train: Gini Reduction
moe_yse_pch_gender	397	MOE: Youth Service: Characteristics: Gender	0.002652
cyf_cec_all_del	311	P_cyf_cec_all_del CYF: Client Event Cost: All spells Window = Profile Metric= Days to end of window since last event(s) of all types	0.004114
ysx_exp_moe_yse_int_high_need	257	Youth Service: MOE high need intervention	0.000867
moe_yse_erl_1eos_cnt	239	P_moe_yse_erl_1eos_cnt Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Count of event(s) of type Evatt1: Leaving reason = End of Schooling	0.000885
ysx_exp_moe_yse_qal_highest	234	Youth Service: highest NCEA qualification	0.001822
win_bdd_chd_dur	211	P_win_bdd_chd_dur Work and Income: BDD: Child spells Window = Profile Metric= Duration (days) of event(s) of all types	0.004350
moe_yse_pch_pdate_age	198	Age at Profile date	0.000961
win_bdd_chd_del	172	P_win_bdd_chd_del Work and Income: BDD: Child spells Window = Profile Metric= Days to end of window since last event(s) of all types	0.007600
moe_yse_pch_decile	169	MOE: Youth Service: Decile of school of last enrolment	0.000485
moe_yse_pch_psc_act	164	MOE: Youth Service: Post-school activity	0.000531
moe_yse_erl_1eos_dsf	162	P_moe_yse_erl_1eos_dsf Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Days since start of window to first event(s) of type Evatt1: Leaving...	0.001152
win_bdd_chd_dsf	162	P_win_bdd_chd_dsf Work and Income: BDD: Child spells Window = Profile Metric= Days since start of window to first event(s) of all types	0.004773
ysx_exp_win_bdd_chd_life_prop	162	Youth Service: proportion of child's life on benefit	0.007120
moe_yse_qal_cnt	152	P_moe_yse_qal_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of all types	0.000831
win_bdd_chd_cnt	144	P_win_bdd_chd_cnt Work and Income: BDD: Child spells Window = Profile Metric= Count of event(s) of all types	0.002685
ysx_exp_moe_awa_mer_exc	139	Youth Service: count of quals achieved with merit or excellence	0.000817
moe_yse_cnt	137	MOE events in profile window (Total)	0.000390
moe_yse_qal_2awm_cnt	137	P_moe_yse_qal_2awm_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt2: Endorsement = Achieved with ...	0.002270
moe_yse_qal_1lv2_cnt	134	P_moe_yse_qal_1lv2_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt1: NQF level = Level2	0.001191
cles_expert_age_group	133	CLES expert variable: age-group at profile date	0.000304
moe_yse_pch_left_bef_eoy	132	MOE: Youth Service: Indicator: left before end of School-year	0.000274
ysx_exp_wi	131	Youth Service: indicator of WI history in profile period	0.000746
moe_yse_erl_2u30_dsf	129	P_moe_yse_erl_2u30_dsf Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Days since start of window to first event(s) of type Evatt2: Post sc...	0.000622
win_bdd_mbs_dur	127	P_win_bdd_mbs_dur Work and Income: BDD: Main benefit spells Window = Profile Metric= Duration (days) of event(s) of all types	0.000485
moe_yse_qal_1lv2_2ne6_cnt	119	P_moe_yse_qal_1lv2_2ne6_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt1: NQF level = Level2 Evatt2:...	0.000242
moe_yse_erl_2e30_swe	118	P_moe_yse_erl_2e30_swe Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Status at window-end event(s) of type Evatt2: Post school activity...	0.000191
win_bdd_chd_sws	117	P_win_bdd_chd_sws Work and Income: BDD: Child spells Window = Profile Metric= Status at window-start event(s) of all types	0.002737
moe_yse_qal_1lv1_2ne6_cnt	112	P_moe_yse_qal_1lv1_2ne6_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt1: NQF level = Level1 Evatt2:...	0.000197
moe_yse_erl_2u30_cnt	110	P_moe_yse_erl_2u30_cnt Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Count of event(s) of type Evatt2: Post school activity = Unknown	0.000232
ysx_exp_moe_yse_pch_lst_sch_mth	109	Youth Service: month that left school	0.000232
moe_yse_qal_1lv3_cnt	107	P_moe_yse_qal_1lv3_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt1: NQF level = Level3	0.001074
moe_yse_qal_2ne6_cnt	104	P_moe_yse_qal_2ne6_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt2: Endorsement = No Endorsement	0.000229
win_involvement	104	WIN involvement in profile window [-|Low|Medium|High]	0.000711
moe_yse_pch_dob	101	MOE: Youth Service: Characteristics: Date of birth	0.000320
win_bdd_chd_swe	98	P_win_bdd_chd_swe Work and Income: BDD: Child spells Window = Profile Metric= Status at window-end event(s) of all types	0.001607
moe_yse_int_dsf	92	P_moe_yse_int_dsf Ministry of Education: Youth Service: Student intervention spell Window = Profile Metric= Days since start of window to first event(s) of all types	0.000275
win_bdd_mbs_sws	88	P_win_bdd_mbs_sws Work and Income: BDD: Main benefit spells Window = Profile Metric= Status at window-start event(s) of all types	0.000542
moe_yse_erl_2u30_swe	81	P_moe_yse_erl_2u30_swe Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Status at window-end event(s) of type Evatt2: Post school activity...	0.000157
moe_yse_erl_2feo_dsf	70	P_moe_yse_erl_2feo_dsf Ministry of Education: Youth Service: Student Enrolment spells Window = Profile Metric= Days since start of window to first event(s) of type Evatt2: Post sch...	0.000146
moe_yse_qal_1lv1_2awm_cnt	67	P_moe_yse_qal_1lv1_2awm_cnt Ministry of Education: Youth Service: Student Qualification spel Window = Profile Metric= Count of event(s) of type Evatt1: NQF level = Level1 Evatt2:...	0.000239

Appendix 2 – SAS EM diagram



Appendix 3 – Youth Services Datamatch process