ULADZISLAU DARHEVICH
Nr indeksu 108519
Ud108519@student.sgh.waw.pl

Zadanie 1. Analiza braków danych, ich udziałów w czasie (czyli stabilności w czasie) i porównywania udziałów pomiędzy zbiorami **train** i **valid** (czyli stabilności na zbiorach) w postaci szczegółowego raportu tabelarycznego.

Celem projektu jest zapoznanie studenta z budową w pełni automatycznych procesów raportowych wykorzystując zarówno metody przetwarzania danych jak i analizy statystyczne. Projekt oparty jest na podstawowych sposobach raportowania i analizowania portfela kredytów detalicznych przydatnych w zarządzaniu ryzykiem kredytowym. W tym wypadku analizy koncentrują się na etapie weryfikacji danych wejściowych, analizie jakości i pierwszych identyfikacjach predyktorów ryzyka kredytowego

Stabilność na zbiorach

Pierwszym krokiem w analizie danych jest podjęcie decyzji o tym, jak obsłużyć brakujące wartości. Decyzja ta może obejmować ustalenie obserwacji i / lub zmiennych z nadmiarem brakujących danych, zastąpienie przypisanych wartości pominiętymi wartościami lub niepodejmowanie żadnych działań, jeżeli ilość brakujących danych jest nieznaczna i prawdopodobnie nie wpłynie na analizę.

Z każdego zbioru **train/ valid** zostało wybrane kilkanaście kolumn. To mogą być dowolne kolumny lub cały zbiór. Funkcja ***miss_report*** będzie tworzyła raport, gdzie będą przedstawiona ilość braków danych dla każdej zmiennej. Funkcja ***miss_report*** realizowana za pomocą procedury PROC FREQ. Funkcja przyjmuje dwa argumenty (nazwa zbioru i ścieżka do pliku) i zwraca tabele przedstawione poniżej w formacie .html.

TRAIN

VARIABLE	N_MISSING	%_MISSING	N_OK	%_OK
period	0	0.0	53,070	100.0
app_income	0	0.0	53,070	100.0
app_number_of_children	0	0.0	53,070	100.0
app_spendings	0	0.0	53,070	100.0
act_cus_seniority	0	0.0	53,070	100.0
act_cus_n_loans_hist	0	0.0	53,070	100.0
act_cus_n_statC	0	0.0	53,070	100.0
act_cus_n_statB	0	0.0	53,070	100.0
act_cus_n_loans_act	0	0.0	53,070	100.0
act_cus_pins	0	0.0	53,070	100.0
act_cus_utl	0	0.0	53,070	100.0
act_cus_dueutl	0	0.0	53,070	100.0
act_cus_cc	0	0.0	53,070	100.0
act_state_6_CMax_Days	15,935	30.0	37,135	70.0
act_state_6_CMax_Due	12,543	23.6	40,527	76.4
act_state_6_CMin_Days	15,935	30.0	37,135	70.0
act_state_6_CMin_Due	12,543	23.6	40,527	76.4
act_state_6_Cncr	50,339	94.9	2,731	5.1
act_cus_loan_number	0	0.0	53,070	100.0
app_char_job_code	0	0.0	53,070	100.0
app_char_marital_status	0	0.0	53,070	100.0
app_char_city	0	0.0	53,070	100.0
app_char_home_status	0	0.0	53,070	100.0
app_char_cars	0	0.0	53,070	100.0
year	0	0.0	53,070	100.0

VALID

VARIABLE	N_MISSING	%_MISSING	N_OK	%_OK
period	0	0.0	52,841	100.0
app_income	0	0.0	52,841	100.0
app_number_of_children	0	0.0	52,841	100.0
app_spendings	0	0.0	52,841	100.0
act_cus_seniority	0	0.0	52,841	100.0
act_cus_n_loans_hist	0	0.0	52,841	100.0
act_cus_n_statC	0	0.0	52,841	100.0
act_cus_n_statB	0	0.0	52,841	100.0
act_cus_n_loans_act	0	0.0	52,841	100.0
act_cus_pins	0	0.0	52,841	100.0
act_cus_utl	0	0.0	52,841	100.0
act_cus_dueutl	0	0.0	52,841	100.0
act_cus_cc	0	0.0	52,841	100.0
act_state_6_CMax_Days	15,876	30.0	36,965	70.0
act_state_6_CMax_Due	12,503	23.7	40,338	76.3
act_state_6_CMin_Days	15,876	30.0	36,965	70.0
act_state_6_CMin_Due	12,503	23.7	40,338	76.3
act_state_6_Cncr	50,053	94.7	2,788	5.3
act_cus_loan_number	0	0.0	52,841	100.0
app_char_job_code	0	0.0	52,841	100.0
app_char_marital_status	0	0.0	52,841	100.0
app_char_city	0	0.0	52,841	100.0
app_char_home_status	0	0.0	52,841	100.0
app_char_cars	0	0.0	52,841	100.0
year	0	0.0	52,841	100.0

Z tablic widać, że ilość brakujących danych w zbiorach **train** i **valid** jest podobna.

Na podstawie raportu użytkownik może zdecydować, jakie zmienne będą upuszczone. Na przykład możemy opuścić te zmienne, które mają wszystkie brakujące wartości lub określony procent obserwacji z brakującymi wartościami (50%) lub obserwacje ze zbyt dużą liczbą brakujących wartości (10%). Za pomocą procedury PROC SQL można stworzyć listę upuszczonych zmiennych.

Deleted variables

Next variables will be deleted bacause the % of missing values >= 50

VARIABLE
act_state_6_Cncr

W celu dalszej pracy z danymi może być warto wypełnić brakujące wartości. Zrobić to można za pomocą procedury PROC MI. Po działaniu tej procedury będziemy mieli zbiór *ready* już bez brakujących danych.

Stabilność w czasie

Dla oceny stabilności zbiorów w czasie stwórzmy tabele w których będzie podana ilość brakujących danych zgrupowanych po latach. Funkcja ***time_stab*** tworzy takie tabele za pomocą procedury PROC MEANS.

Na podstawie tych tabel można wywnioskować, że ilość brakujących danych na zbiorach **train** i **valid** jest podobna. W roku 2018 mamy mniej brakujących danych, bo ostatnia obserwacja zrobiona w lipcu tego roku.

TRAIN

NAME OF FORMER VARIABLE	_2004	_2005	_2006	_2007	_2008	_2009	_2010	_2011	_2012	_2013	_2014	_2015	_2016	_2017	_2018
app_income_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
app_number_of_children_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
app_spendings_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_seniority_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_loans_hist_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_statC_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_statB_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_loans_act_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_pins_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_utl_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_dueutl_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_cc_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_state_6_CMax_Days_NMiss	910	1047	1032	1057	1053	1132	1111	1081	1110	1100	1144	1188	1096	1171	644
act_state_6_CMax_Due_NMiss	721	829	777	835	854	885	870	841	861	885	915	935	859	917	519
act_state_6_CMin_Days_NMiss	910	1047	1032	1057	1053	1132	1111	1081	1110	1100	1144	1188	1096	1171	644
act_state_6_CMin_Due_NMiss	721	829	777	835	854	885	870	841	861	885	915	935	859	917	519
act_state_6_Cncr_NMiss	3049	3312	3294	3435	3397	3419	3552	3352	3545	3561	3527	3497	3435	3618	2060
act_cus_loan_number_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
total_number	6311	7064	6912	7219	7211	7453	7514	7196	7487	7531	7645	7743	7345	7794	4386

VALID

NAME OF FORMER VARIABLE	_2004	_2005	_2006	_2007	_2008	_2009	_2010	_2011	_2012	_2013	_2014	_2015	_2016	_2017	_2018
app_income_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
app_number_of_children_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
app_spendings_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_seniority_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_loans_hist_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_statC_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_statB_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_n_loans_act_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_pins_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_utl_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_dueutl_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_cus_cc_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act_state_6_CMax_Days_NMiss	979	1038	1045	1031	1051	1121	1082	1132	1123	1098	1136	1120	1149	1145	685
act_state_6_CMax_Due_NMiss	752	809	825	812	838	897	861	895	888	841	899	878	918	896	534
act_state_6_CMin_Days_NMiss	979	1038	1045	1031	1051	1121	1082	1132	1123	1098	1136	1120	1149	1145	685
act_state_6_CMin_Due_NMiss	752	809	825	812	838	897	861	895	888	841	899	878	918	896	534
act_state_6_Cncr_NMiss	2986	3369	3349	3378	3388	3521	3485	3526	3619	3469	3492	3469	3607	3569	2112
act_cus_loan_number_NMiss	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
total_number	6448	7063	7089	7064	7166	7557	7371	7580	7641	7347	7562	7465	7741	7651	4550

Podsumowanie

Analiza pokazała, że braki danych dla zbiorów **train** i **valid** są podobne. Stworzone funkcje ***miss_report*** i ***time_stab*** są uniwersalne i mogą być użyte do dowolnych zbiorów dowolnego rozmiaru

Bibliografia

Mike Zdeb "An Easy Route to a Missing Data Report with ODS+PROC FREQ+A Data Step" NESUG 2011 https://www.lexjansen.com/nesug/nesug11/ds/ds12.pdf

```
libname data '/folders/myfolders/sasuser/Projekt';

data data.test_train;
        set data.abt_sam_beh_train(keep=period act_state_6: app_: act_cus:);
        year = substr(period, 1, 4);
run;

data data.test_valid;
        set data.abt_sam_beh_valid(keep=period act_state_6: app_: act_cus:);
        year = substr(period, 1, 4);
run;

%macro miss_report(din, fout);
proc format;
value nm   .  = '0' other = '1';
value $ch ' ' = '0' other = '1';
run;

ods listing close;
ods output onewayfreqs=tables;
proc freq data=&din;
tables _all_ / missing;
format _numeric_ nm. _character_ $ch.;
run;
ods output close;
ods listing;

data report;
length var $32;
do until (last.table);
        set tables;
        by table notsorted;
        array names(*) f_: ;
        select (names(_n_));
                when ('0') do;
                        miss = frequency;
                        p_miss = percent;
                end;
                when ('1') do;
                        ok   = frequency;
                        p_ok   = percent;
                end;
        end;
end;
miss   = coalesce(miss,0);
ok     = coalesce(ok,0);
p_miss = coalesce(p_miss,0);
p_ok   = coalesce(p_ok,0);
var    = scan(table,-1);

keep var miss ok p_:;
format miss ok comma7. p_: 5.1;
label
miss   = 'N_MISSING'
ok     = 'N_OK'
p_miss = '%_MISSING'
p_ok   = '%_OK'
var    = 'VARIABLE';
run;

ods listing close;
ods html file="&fout" style=barrettsblue;
proc print data=report label noobs;
        id var;
        var miss p_miss ok p_ok;
run;
```

```
ods html close;
ods listing;
%mend;

%miss_report(data.test_train, /folders/myfolders/sasuser/-
missing_values_report_train.html);
%miss_report(data.test_valid, /folders/myfolders/sasuser/-
missing_values_report_valid.html);

ODS HTML body="/folders/myfolders/sasuser/Deleted_Variables.HTML" style=barrettsblue;
run;
title1 'Deleted variables';
title2 'Next variables will be deleted bacause the % of missing values >= 50';
proc sql;
select var into :droplist separated by ' '
from report where p_miss ge 50;
quit;
run;
ods html close;

data cleaned;
set data.test_valid(drop=&droplist);
if cmiss(of _all_) ge 10 then delete;
if missing(cats(of _all_))then delete;
run;

proc mi data=cleaned ROUND=1 NIMPUTE=1 out=ready;
ods select misspattern;
run;

%macro time_stab(din, fout);
proc means data=&din noprint;
by year;
output out=output_means(drop=_type_ _freq_) nmiss=/autoname;
run;

data time_report;
        set output_means;
        total_number = sum(of _numeric_);
run;

proc transpose data=time_report name=varName out=out_report;
id year;
run;

ods listing close;
ods html file="&fout" style=barrettsblue;
proc print data=out_report label noobs;
var _all_;
run;
ods html close;
ods listing;
%mend;

%time_stab(data.test_train, /folders/myfolders/sasuser/time_train.html);
%time_stab(data.test_valid, /folders/myfolders/sasuser/time_valid.html);
```