ASSIGNMENT

ASS04

DATA PREPARATION – STA6714

DONE BY:

SUSHMITHA MANI

**QUESTION:**

Write a simple program to perform smoothing of a given categorical variable when the target variable is “HasDetections”. The input of the function has the following fields: (1) Data Set Name; (2) the target variable; (3) the threshold for the minimum number of observations for any given category; (4) Categorical Variable to be smoothed. The output of the function is the smoothed variables added to the original input data set.

Use the following five variables to test your function.

1. AppVersion
2. Census\_ActivationChannel
3. Census\_ChassisTypeName
4. Census\_DeviceFamily
5. Census\_FirmwareManufacturerIdentifier

Data set used for this assignment is “Malware\_train\_s.csv”.

**SOLUTION:**

1. For the variable AppVersion

CODE:

proc freq data= work.import;

tables AppVersion / out=freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=0;

tables AppVersion / out=no\_freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=1;

tables AppVersion / out=yes\_freq nocum nopercent;

run;

proc sort data= freq(rename=(count = freq) drop=PERCENT);

by AppVersion;

proc sort data=yes\_freq(rename=(count = yes\_freq) drop=PERCENT);

by AppVersion;

proc sort data=no\_freq(rename=(count = no\_freq) drop=PERCENT);

by AppVersion;

run;

data merge\_out;

merge freq no\_freq yes\_freq;

by AppVersion;

run;

data prop;

set merge\_out;

yes\_prop = yes\_freq / freq;

no\_prop = no\_freq / freq;

run;

proc print data= prop;

var Appversion freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc print data= prop;

var Appversion freq yes\_freq no\_freq yes\_prop no\_prop;

where freq<100;

sum freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc sql;

insert into prop values ( 'Other', 1070, 530, 540, 20.7800, 24.2200);

Alter table prop

Add smoothinglogit DECIMAL(10,6);

update prop

%let smoothfactor=10;

set smoothinglogit= log((yes\_freq+(yes\_prop\*&smoothfactor))/(yes\_freq+(no\_prop\*&smoothfactor)));

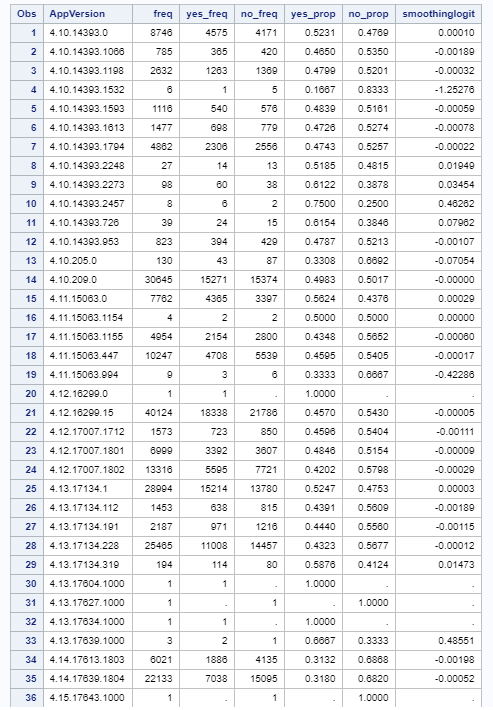
run;

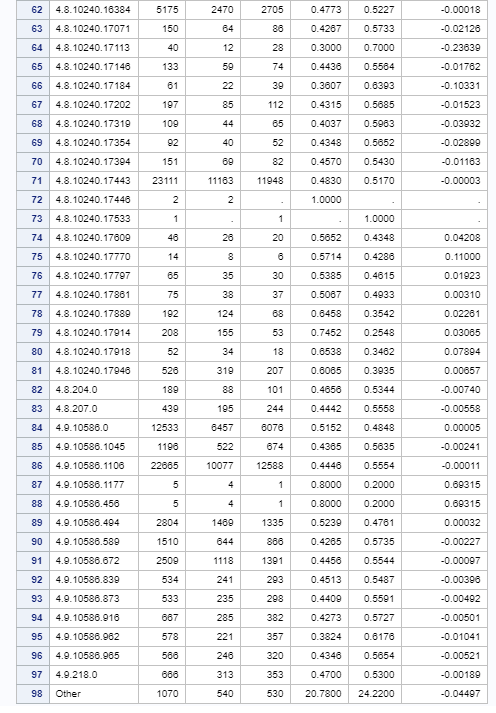
quit;

proc print data=prop;

var Appversion freq yes\_freq no\_freq yes\_prop no\_prop smoothinglogit;

run;





1. For the Variable: Census\_ActivationChannel

CODE:

proc freq data= work.import;

tables Census\_ActivationChannel / out=freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=0;

tables Census\_ActivationChannel / out=no\_freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=1;

tables Census\_ActivationChannel / out=yes\_freq nocum nopercent;

run;

proc sort data= freq(rename=(count = freq) drop=PERCENT);

by Census\_ActivationChannel;

proc sort data=yes\_freq(rename=(count = yes\_freq) drop=PERCENT);

by Census\_ActivationChannel;

proc sort data=no\_freq(rename=(count = no\_freq) drop=PERCENT);

by Census\_ActivationChannel;

run;

data merge\_out;

merge freq no\_freq yes\_freq;

by Census\_ActivationChannel;

run;

data prop;

set merge\_out;

yes\_prop = yes\_freq / freq;

no\_prop = no\_freq / freq;

run;

proc print data= prop;

var Census\_ActivationChannel freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc sql;

Alter table prop

Add smoothinglogit DECIMAL(10,6);

update prop

%let smoothfactor=10;

Set smoothinglogit= log((yes\_freq+(yes\_prop\*&smoothfactor))/(yes\_freq+(no\_prop\*&smoothfactor)));

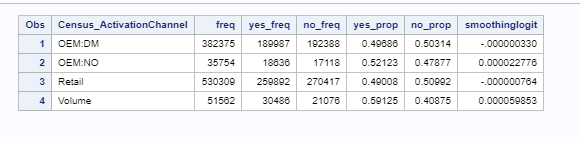
run;

quit;

proc print data=prop;

var Census\_ActivationChannel freq yes\_freq no\_freq yes\_prop no\_prop smoothinglogit;

run;



1. For the Variable : Census\_ChasisTypeName

CODE:

proc freq data= work.import;

tables Census\_ChassisTypeName / out=freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=0;

tables Census\_ChassisTypeName/ out=no\_freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=1;

tables Census\_ChassisTypeName / out=yes\_freq nocum nopercent;

run;

proc sort data= freq(rename=(count = freq) drop=PERCENT);

by Census\_ChassisTypeName;

proc sort data=yes\_freq(rename=(count = yes\_freq) drop=PERCENT);

by Census\_ChassisTypeName;

proc sort data=no\_freq(rename=(count = no\_freq) drop=PERCENT);

by Census\_ChassisTypeName;

run;

data merge\_out;

merge freq no\_freq yes\_freq;

by Census\_ChassisTypeName;

run;

data prop;

set merge\_out;

yes\_prop = yes\_freq / freq;

no\_prop = no\_freq / freq;

run;

proc print data= prop;

var Census\_ChassisTypeName freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc print data= prop;

var Census\_ChassisTypeName freq yes\_freq no\_freq yes\_prop no\_prop;

where freq<100;

sum freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc sql;

insert into prop values ( 'Other',263, 113, 150, 7.80249, 9.19751);

Alter table prop

Add smoothinglogit DECIMAL(10,6);

update prop

%let smoothfactor=10;

set smoothinglogit= log((yes\_freq+(yes\_prop\*&smoothfactor))/(yes\_freq+(no\_prop\*&smoothfactor)));

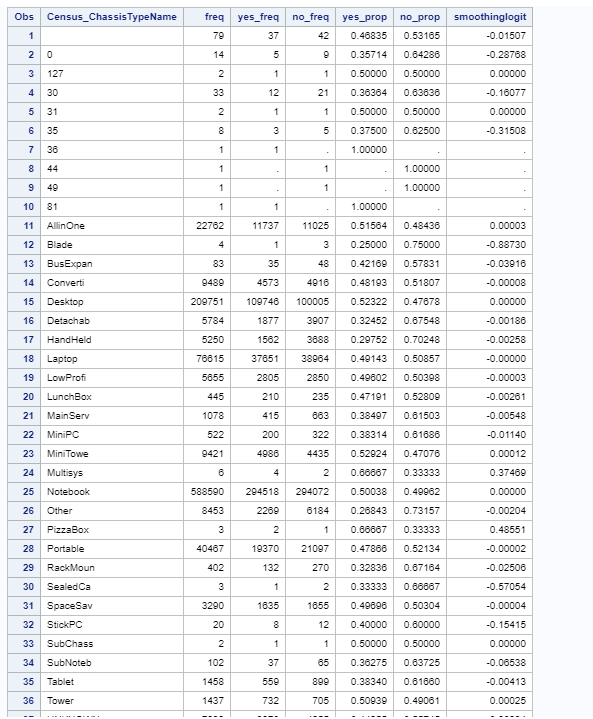
run;

quit;

proc print data=prop;

var Census\_ChassisTypeName freq yes\_freq no\_freq yes\_prop no\_prop smoothinglogit;

run;



1. For the variable Census\_DeviceFamily

proc freq data= work.import;

tables Census\_DeviceFamily / out=freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=0;

tables Census\_DeviceFamily/ out=no\_freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=1;

tables Census\_DeviceFamily / out=yes\_freq nocum nopercent;

run;

proc sort data= freq(rename=(count = freq) drop=PERCENT);

by Census\_DeviceFamily;

proc sort data=yes\_freq(rename=(count = yes\_freq) drop=PERCENT);

by Census\_DeviceFamily;

proc sort data=no\_freq(rename=(count = no\_freq) drop=PERCENT);

by Census\_DeviceFamily;

run;

data merge\_out;

merge freq no\_freq yes\_freq;

by Census\_DeviceFamily;

run;

data prop;

set merge\_out;

yes\_prop = yes\_freq / freq;

no\_prop = no\_freq / freq;

run;

proc print data= prop;

var Census\_DeviceFamily freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc print data= prop;  
var Census\_DeviceFamily freq yes\_freq no\_freq yes\_prop no\_prop;  
where freq<100;   
sum freq yes\_freq no\_freq yes\_prop no\_prop;  
run;

proc sql;

insert into prop values ( 'Other',3, 2,1, 0.66667, 0.33333);

Alter table prop

Add smoothinglogit DECIMAL(10,6);

update prop

%let smoothfactor=10;

set smoothinglogit= log((yes\_freq+(yes\_prop\*&smoothfactor))/(yes\_freq+(no\_prop\*&smoothfactor)));

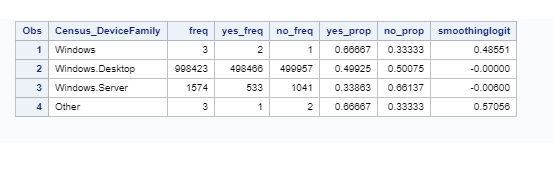
run;

quit;

proc print data=prop;

var Census\_DeviceFamily freq yes\_freq no\_freq yes\_prop no\_prop smoothinglogit;

run;



1. For The Variable: Census\_FirmwareManufacturerIdent

proc freq data= work.import;

tables Census\_FirmwareManufacturerIdent / out=freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=0;

tables Census\_FirmwareManufacturerIdent/ out=no\_freq nocum nopercent;

run;

proc freq data=work.import noprint;

where HasDetections=1;

tables Census\_FirmwareManufacturerIdent / out=yes\_freq nocum nopercent;

run;

proc sort data= freq(rename=(count = freq) drop=PERCENT);

by Census\_FirmwareManufacturerIdent;

proc sort data=yes\_freq(rename=(count = yes\_freq) drop=PERCENT);

by Census\_FirmwareManufacturerIdent;

proc sort data=no\_freq(rename=(count = no\_freq) drop=PERCENT);

by Census\_FirmwareManufacturerIdent;

run;

data merge\_out;

merge freq no\_freq yes\_freq;

by Census\_FirmwareManufacturerIdent;

run;

data prop;

set merge\_out;

yes\_prop = yes\_freq / freq;

no\_prop = no\_freq / freq;

run;

proc print data= prop;

var Census\_FirmwareManufacturerIdent freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc print data= prop;

var Census\_FirmwareManufacturerIdent freq yes\_freq no\_freq yes\_prop no\_prop;

where freq<100;

sum freq yes\_freq no\_freq yes\_prop no\_prop;

run;

proc sql;

insert into prop values (0, 3235, 1301, 1934, 133.008, 213.992);

Alter table prop

Add smoothinglogit DECIMAL(10,6);

update prop

%let smoothfactor=10;

set smoothinglogit= log((yes\_freq+(yes\_prop\*&smoothfactor))/(yes\_freq+(no\_prop\*&smoothfactor)));

run;

quit;

proc print data=prop;

var Census\_FirmwareManufacturerIdent freq yes\_freq no\_freq yes\_prop no\_prop smoothinglogit;

run;

