TransUnion

PROC HPBIN

Solve your WOEs and more

Meera Ragunathan

NOVEMBER 2017

Agenda

Features

Binning Methods

Syntax

WOEs and IVs

Features

- “**H**igh **P**erformance **Bin**ning”
- Provides a mapping table
- Highly multithreaded during all phases
  (computational work organized into multiple tasks)
- Can calculate Weight of Evidence (WOE) and Information Value (IV) based on results

TransUnion

Binning

Refresher on independent variables

- **Nominal (*categorical*)**

take distinct values that assume no order or scale relationship between them

ex. Province

Refresher on independent variables

- **Nominal (*categorical*)**

take distinct values that assume no order or scale relationship between them

ex. Province

- **Ordinal (*rank*)**

similar to **nominal variable** but with order defined on categories

ex. Level of education

Refresher on independent variables

- **Nominal (*categorical*)**

take distinct values that assume no order or scale relationship between them

ex. Province

- **Ordinal (*rank*)**

similar to **nominal variable** but with order defined on categories

ex. Level of education

- **Continuous**

take values represented on real number scale

ex. Income

Binning

- Common step in model building
- Data preparation stage

- Different binning methods
  - **Bucket**
  - Quantile
  - Pseudo-Quantile
  - Winsorized

Bucket Binning

- Bucket lengths of equal length are calculated by:

$$L = \frac{max(x) - min(x)}{n}$$

- Split points of *k*th variable is $s_k = min(x) + (L)(k)$

  - L= length of each bucket
  - n = number of observations

Syntax for PROC HPBIN

```
PROC HPBIN
        DATA = SAS-data-set <options>;
        INPUT variables-list;
RUN;
```

Specifies data set being used

Specifies interval variable(s) to be binned

PROC HPBIN Statement Options

	Option	Description
Basic	**DATA=**	Specifies input data set
Basic	**OUTPUT=**	Specifies output data set
Binning level	**NUMBIN=**	Specifies global number of bins for all binning variables
Binning method	**BUCKET**	Specifies bucket binning method
Binning method	**QUANTILE**	Specifies quantile binning method
	WOE	Computes WOE and IVs ; must specify target variable

ODS Tables produced by PROC HPBIN

Table Name	Description	Options
BinInfo	Basic binning information and parameters	Default
Mapping	Level mapping information	Default
PerformanceInfo	Information about high-performance computing environment	Default
WOE	Weight of Evidence for each bin	WOE
InfoValue	Information value for each variable	WOE
NObs	Number of observations read and used	WOE

Further detail on ODS tables produced

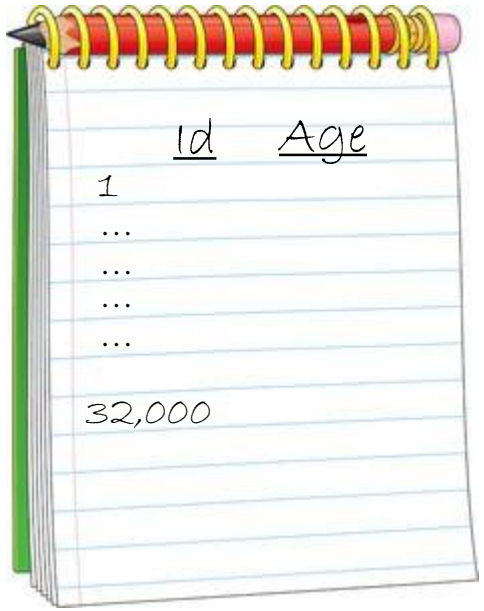
- ***BinInfo***
  - Includes information about binning method, number of bins, number of variables
- ***Mapping***
  - Level starts at 1 and increases to value of *NUMBINS*
  - Bin level can be less than *NUMBINS* if input data are small
- ***PerformanceInfo***
  - Information about execution mode
    - Single-machine
    - Distributed-mode

Further detail on ODS tables produced

- ***WOE***
  - Provides level mapping information, binning information, weight of evidence and information value for each bin
  - When WOE table is printed, mapping table is not printed
  - Other values: event/non-event counts and event/non-event rates for each bin
- ***InfoValue***
  - Generates Information Value for each variable

Example 1a

Bucket Binning using PROC HPBIN




```
/* The following code bins the Age variable into 5 bins */

proc hpbin data=data.sample
     output=data.binned
     numbin=5
     bucket;
     input age;
run;
```

Example 1a (cont'd)

Bucket Binning using PROC HPBIN

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Data Access Information			
Data	Engine	Role	Path
DATA.SAMPLE	V9	Input	On Client
DATA.BINNED	V9	Output	On Client

Binning Information	
Method	Bucket Binning
Number of Bins Specified	5
Number of Variables	1

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Age	BIN_Age	Age < 31.6	3372	0.13015787
		31.6 <= Age < 47.2	9412	0.36329949
		47.2 <= Age < 62.8	9117	0.35191261
		62.8 <= Age < 78.4	3525	0.13606361
		78.4 <= Age	481	0.01856641

Example 1b

Bucket Binning using PROC HPBIN

```
122  /* This time, the number of bins specified in line 125 overwrites
123  the global number of binning levels (in this case 5)*/
124
125  proc hpbin data=data.sample
126       output=data.binned_2
127       numbin=5
128       bucket;
129       input age;
130       input income/numbin=4;
131  run;
```

Example 1b (cont'd)

Bucket Binning using PROC HPBIN

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Data Access Information			
Data	Engine	Role	Path
DATA.SAMPLE	V9	Input	On Client
DATA.BINNED_2	V9	Output	On Client

Binning Information	
Method	Bucket Binning
Number of Bins Specified	Between 4 and 5
Number of Variables	2

Mapping				
Variable	Binned Variable	Range	Frequency	Proportion
Income	BIN_Income	Income < 58.25	20679	0.78087002
		58.25 <= Income < 116.5	5239	0.19783249
		116.5 <= Income < 174.75	516	0.01948493
		174.75 <= Income	48	0.00181255
Age	BIN_Age	Age < 31.6	3372	0.13015787
		31.6 <= Age < 47.2	9412	0.36329949
		47.2 <= Age < 62.8	9117	0.35191261
		62.8 <= Age < 78.4	3525	0.13606361
		78.4 <= Age	481	0.01856641

TransUnion

WOEs and IVs

Weight of Evidence and Information Values

WOEs and Information Values (IVs)

- **Weight of Evidence (WOE)**
  - Measure of predictive power of independent variable on dependent variable
  - Strength of a grouping used to separate good and bad risk

- Applying WOE transformation is called *Coarse Classing*

- WOEs are later used in place of predictors in Logistic Regression for subsequent modelling

How do you calculate WOE?

$$\text{WOE} = \ln\left(\frac{\textbf{Bad}\ \text{Distribution}_i}{\textbf{Good}\ \text{Distribution}_i}\right)$$

- **Nominal variable**
  i represents category

- **Continuous variable**
  i represents bin level

$$\textbf{Bad}\ \text{Distribution}_i = \frac{\text{Number of}\ \textbf{Bad}_i}{\text{Total number of Bad}}$$

$$\textbf{Good}\ \text{Distribution}_i = \frac{\text{Number of}\ \textbf{Good}_i}{\text{Total number of Good}}$$

Logarithm of a modified form of odds ratio

Example 2

Solve WOEs for nominal variables manually

Row Labels	Bad	Good	Bad Dist.	Good Dist.	WOE
British Columbia	1,188,423	2,945,712			
Ontario	3,384,277	8,224,103			
Quebec	1,708,456	4,976,089			
Total	**6,281,156**	**16,145,904**			

Defintion of events:

Bad
consumers with score less than 700

Good
consumers with score greater than 700

Example 2

Solve WOEs for nominal variables manually

Row Labels	Bad	Good	Bad Dist.	Good Dist.	WOE
British Columbia	1,188,423	2,945,712	0.189		
Ontario	3,384,277	8,224,103			
Quebec	1,708,456	4,976,089			
Total	**6,281,156**	**16,145,904**			

$$\text{Bad Dist} = \frac{1{,}188{,}423}{6{,}281{,}156} = 0.189$$

Example 2

Solve WOEs for nominal variables manually

Row Labels	Bad	Good	Bad Dist.	Good Dist.	WOE
British Columbia	1,188,423	2,945,712	0.189	0.182	0.0364
Ontario	3,384,277	8,224,103			
Quebec	1,708,456	4,976,089			
Total	**6,281,156**	**16,145,904**			

$$\text{Bad Dist} = \frac{1{,}188{,}423}{6{,}281{,}156} = 0.189$$

$$\text{WOE} = \ln\left(\frac{0.189}{0.182}\right) = 0.0364$$

Example 2

Solve WOEs for nominal variables manually

Row Labels	Bad	Good	Bad Dist.	Good Dist.	WOE
British Columbia	1,188,423	2,945,712	0.189	0.182	0.0364
Ontario	3,384,277	8,224,103	0.539	0.509	0.0562
Quebec	1,708,456	4,976,089	0.272	0.308	-0.1249
Total	**6,281,156**	**16,145,904**			

Information Values

- Weighted sum of WOE of categories
- Useful concept for variable selection in a predictive model

$$IV = \sum (DistributionGood_i - DistributionBad_i) \times ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

$$IV = \sum (DistributionGood_i - DistributionBad_i) \times WOE_i$$

Information Value	Variable Predictiveness
< 0.02	Not useful for prediction
0.02 to 0.1	Weak
0.1 to 0.3	Medium
0.3 to 0.5	Strong
>0.5	Suspiciously good

Example 3

Solve WOEs for continuous variables using PROC HPBIN

Age	Income	Ins

Age
of customer in years

Income
in thousands of dollars

Ins
Did customer buy insurance product?
(Binary target variable)

1 = Yes **0 = No**

Example 3

Solve WOEs for continuous variables using PROC HPBIN

```
proc hpbin data=data.sample
     output=data.woe_example
     numbin=5;
     input age income;
     ods output Mapping=Mapping;
run;
```

```
proc hpbin data=data.sample WOE BINS_META=Mapping;
   target Ins/level=nominal;
run;
```

When bins_meta dataset is specified, PROC HPBIN does not do binning. It takes binning results from bins_meta and calculates values from bins_meta data set.

Target variable must be specified when calculating WOE

Example 3

Solve WOEs for continuous variables using PROC HPBIN

The HPBIN Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	4

Data Access Information			
Data	Engine	Role	Path
DATA.SAMPLE	V9	Input	On Client

Binning Information	
Method	BinsMeta
Number of Bins Specified	See BinsMeta
Number of Variables	2

Number of Observations Read	32264
Number of Observations Used	25907

Variable Information Value	
Variable	Information Value
Income	0.00084011
Age	0.00114314

Weight of Evidence								
Variable	Binned Variable	Range	Non-event Count	Non-event Rate	Event Count	Event Rate	Weight of Evidence	Information Value
Income	BIN_Income		3770	0.65202352	2012	0.34797648	-0.0071267	9.11191E-6
		Income < 46.6	11643	0.65891341	6027	0.34108659	0.02338333	0.00029837
		46.6 <= Income < 93.2	4718	0.64321745	2617	0.35678255	-0.0457161	0.00047842
		93.2 <= Income < 139.8	820	0.64976228	442	0.35023772	-0.0170780	0.00001144
		139.8 <= Income < 186.4	121	0.63684211	69	0.36315789	-0.0733884	0.00003206
		186.4 <= Income	17	0.68000000	8	0.32000000	0.11869937	0.00001071
Age	BIN_Age		4124	0.64873368	2233	0.35126632	-0.0215948	0.00009219
		Age < 31.6	2242	0.66488731	1130	0.33511269	0.05007825	0.00026005
		31.6 <= Age < 47.2	6205	0.65926477	3207	0.34073523	0.02494706	0.00018085
		47.2 <= Age < 62.8	5933	0.65076231	3184	0.34923769	-0.0126807	0.00004553
		62.8 <= Age < 78.4	2258	0.64056738	1267	0.35943262	-0.0572449	0.00036110
		78.4 <= Age	327	0.67983368	154	0.32016632	0.11793513	0.00020343

Summary

Powerful binning tool which saves the guesswork and adds options

Customized binning levels across variables

Weight of Evidence

Information Values

Thank you.

References

- “SAS® Enterprise Miner™ High-Performance Data Mining Procedures and Macro Reference for SAS® 9.3”

  https://support.sas.com/documentation/onlinedoc/miner/emhp71/emhpprcref.pdf

- Base SAS® 9.4 Procedures Guide High-Performance Procedures

  http://support.sas.com/documentation/cdl/en/prochp/67530/HTML/default/viewer.htm#prochp_hpbin_overview.htm

- Upadhyay, Roopam. “Information Value (IV) and Weight of Evidence (WOE) – A Case Study from Banking (Part 4)”

  http://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/

Contact Information

Meera Ragunathan
Data Scientist, TransUnion Canada

Work:
mraguna@transunion.com

Personal:
https://www.linkedin.com/in/meeraragunathan/
m.ragunathan@icloud.com