SCORECARD VARIABLE GROUPING AND SELECTION

Dr. Aric LaBarr

Institute for Advanced Analytics

Bin Variables cuz bye bye odds ratio. Essentially compare 2 categories. easiest interpretation out there. Mkaes it easy to compare. Can model non linearity

Optbinning package in Python

SMbinning package in R

Proc binning in SAS

Process Flow

Data Collection
- Variable Selection
- Sample Size
- Sample / Performance Window

Data Cleaning
- Eliminate Duplicates
- Examine / Remove Outliers

Variable Grouping and Selection
- Weights of Evidence (WOE)
- Information Value (IV)
- Gini Criterion

Initial Scorecard Creation
- Logistic Regression
- Accuracy
- Threshold
- Assessment

Reject Inference
- Remove bias resulting from exclusion of rejects

Final Scorecard Creation
- Final Model Assessment

VARIABLE GROUPING

Variable Grouping and Selection

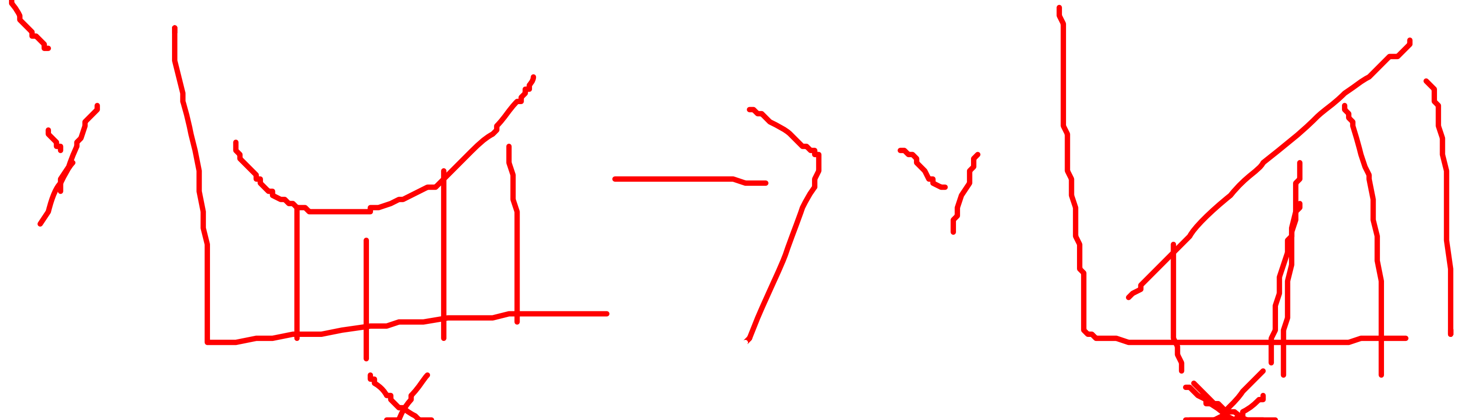
- Scorecards end up with only just groups within a variable.
- Objectives:
  1. Eliminate weak characteristics (variables) or those that do not conform to good business logic.
  2. Group the strongest characteristics' attribute levels in order to produce a model in scorecard format.
- Function/package "smbinning" in R.
- Package "scorecard" or "OptBinning" in Python.
- PROC BINNING in SAS VIYA.

Variable	Level
MISS	$x < 24$
MISS	$24 \le x < 36$
MISS	$36 \le x < 48$
MISS	$x \ge 48$
HOME	OWN
HOME	RENT

Why Grouping (Binning)?

- Goal is to help simplify analysis through grouping:
  - Useful for understanding relationships – no worries about explaining coefficients.
  - Modeling nonlinearities – similar to decision trees.
    (NO MORE LOGISTIC REGRESSION LINEARITY ASSUMPTION!)
  - Dealing with outliers – contained in the smallest / largest group.
  - Missing values typically in own group.

get their own category of missing



Initial Characteristic Analysis – SAS

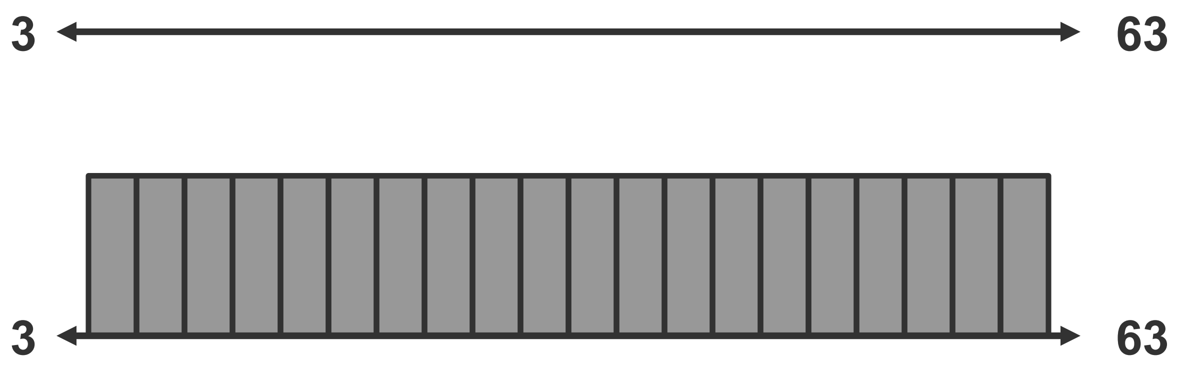
- Need a starting point for the grouping / binning.
  - Quantiles are most popular technique.
- Pre-bin the interval variables into a number of user-specified quantiles / buckets for fine detailed groupings.
- Aggregate the fine detailed groupings into a smaller number to produce coarse groupings.
  - Chi-squared tests to combine groups.

technique is pre bin and combine method. SAS first one to do this. Break into equal groups 20 to 100 groups. then they use chi sq test to combine together. predicting a binary target.

Initial Characteristic Analysis – SAS

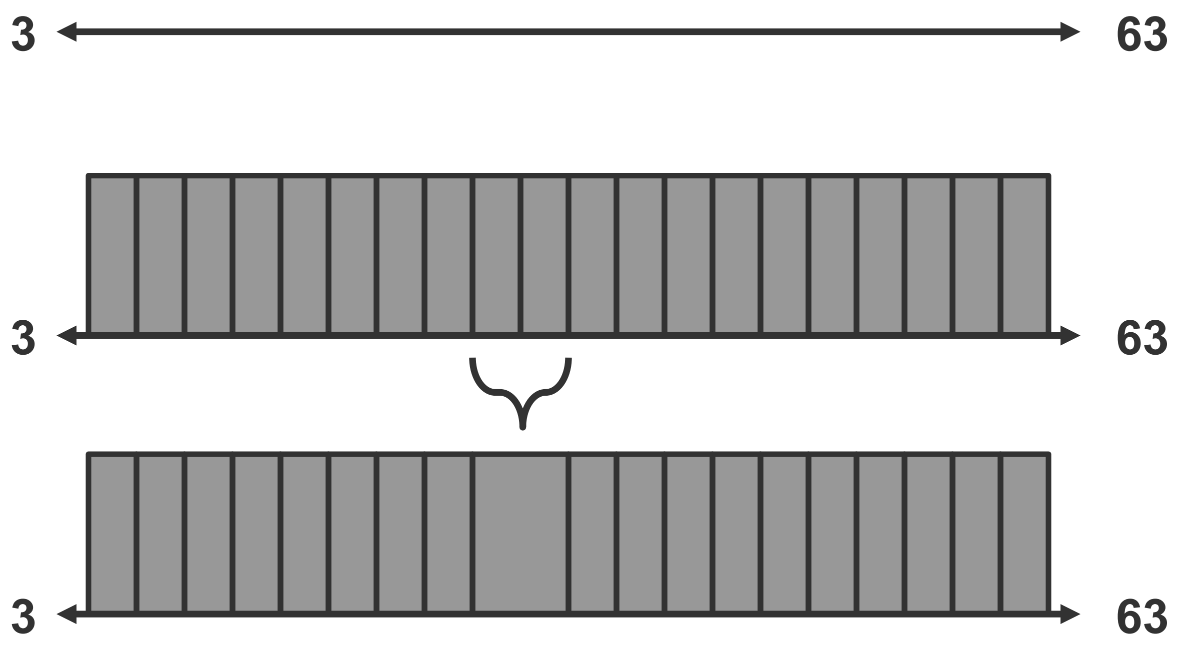
3 ⟷ **63**

Initial Characteristic Analysis – SAS

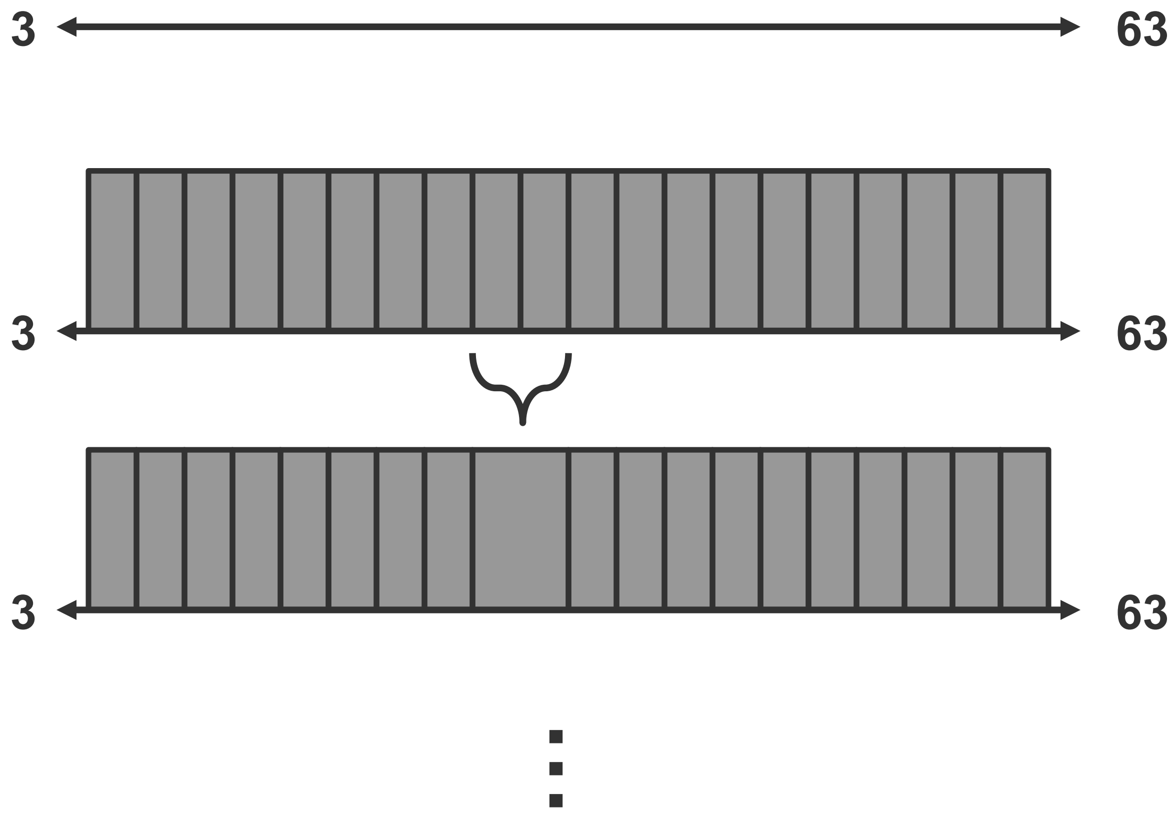


Initial Characteristic Analysis – SAS

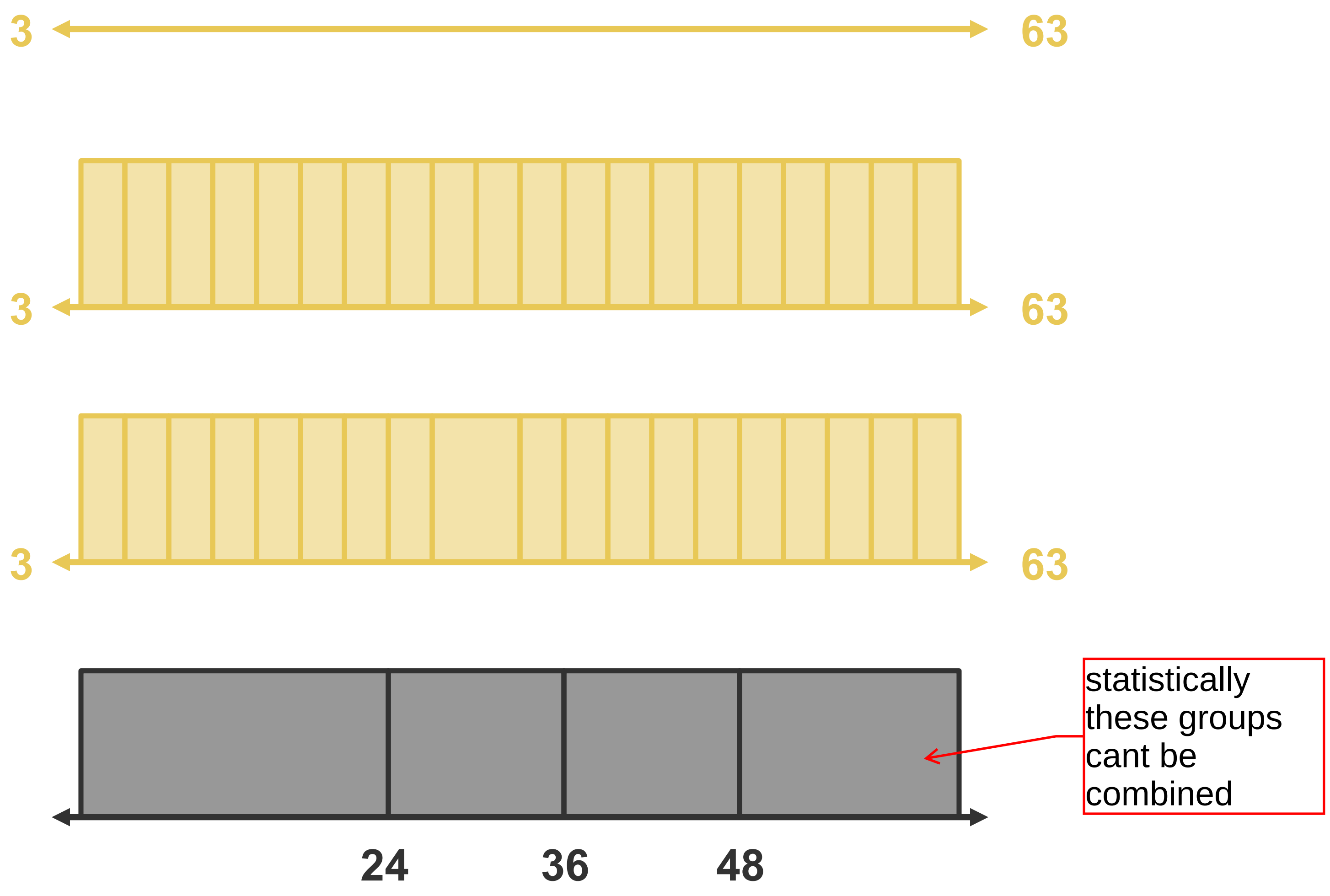
MH test - see if groups can be combined. of those pairs it combines pairs that are most significantly similar.



Initial Characteristic Analysis – SAS



Initial Characteristic Analysis – SAS



Initial Characteristic Analysis – R

Splitting not based on Gini. Based on Chi Sq test.

- The package (and function) “smbinning” uses a different approach than SAS.

- Conditional Inference Trees: CIT
  - CART methods have inherent bias – variables with more levels → more likely to be split on if split on Gini and Entropy.
  - CIT method adds extra statistical step before splits occur – statistical tests of significance.
  - What is **MOST** significant variable? → What is the best split (Chi-square) on **THIS** variable? → **REPEAT**.

SAS splits first then try and combine. in R, keep everythign combined then split. its basically a decision tree with one variable. split based on small p value. First do global test are there any splits.

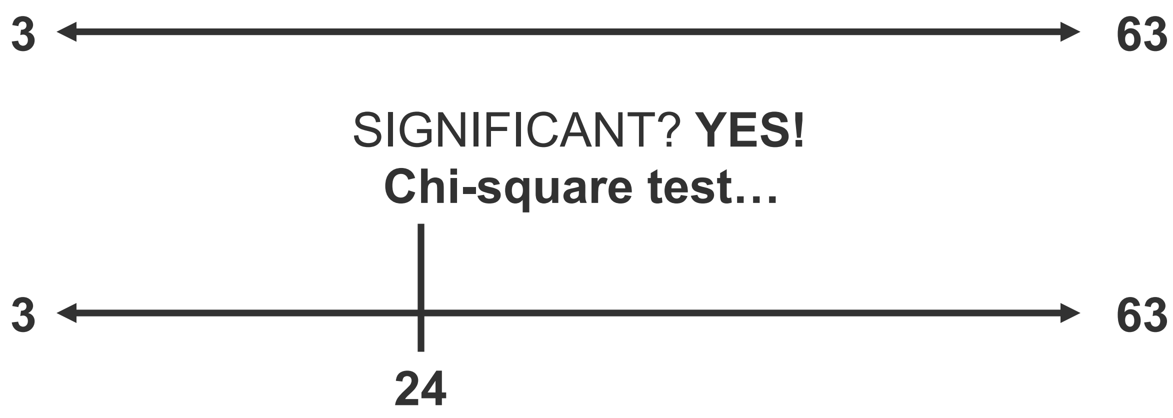
Opt binning package in Python does both methods - you pick technique A or B. Other package in Python does SAS way

Initial Characteristic Analysis – R

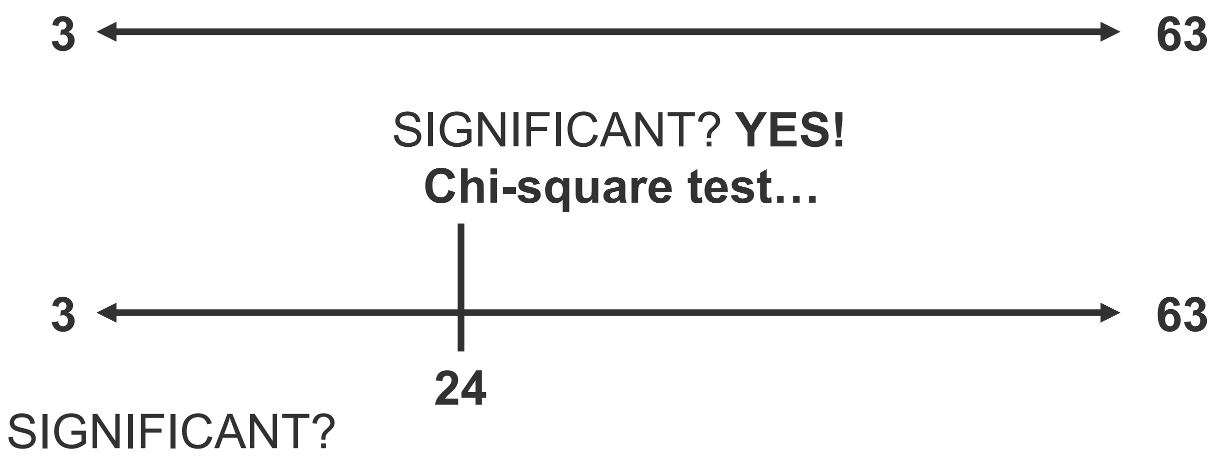
3 ⟵⟶ **63**

SIGNIFICANT?

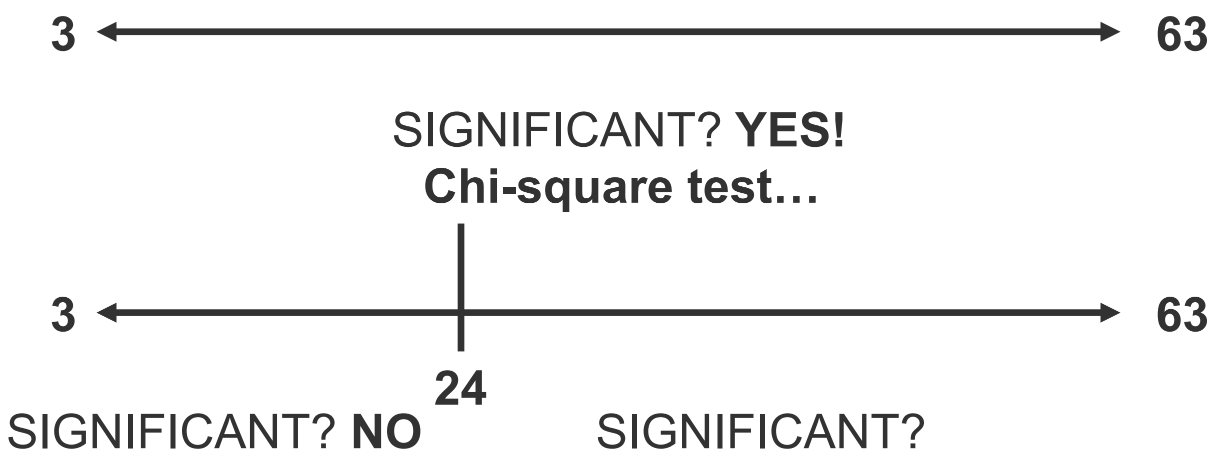
Initial Characteristic Analysis – R



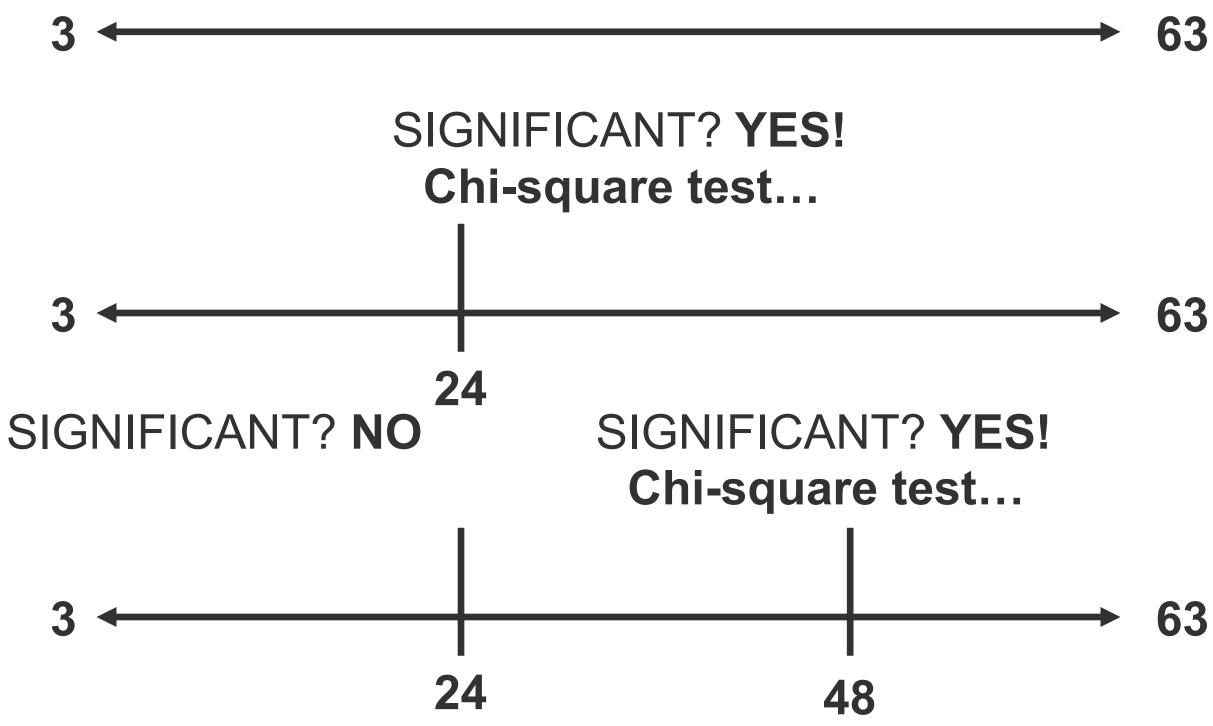
Initial Characteristic Analysis – R



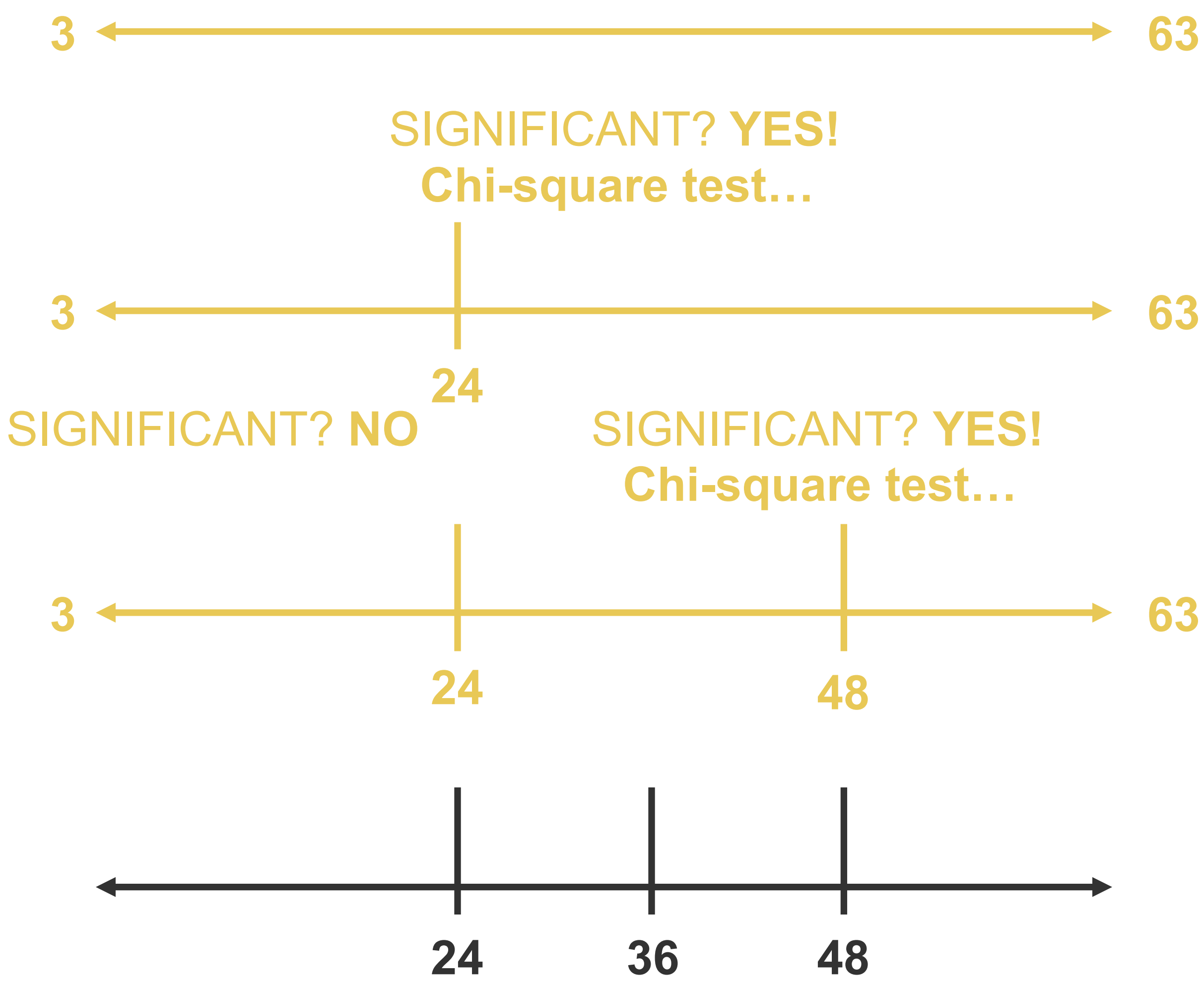
Initial Characteristic Analysis – R



Initial Characteristic Analysis – R



Initial Characteristic Analysis – R



Initial Characteristic Analysis

- Cut-offs may be rough from decision tree combining.
- **Optional to override** automatically generated groups to conform to business rules.
- Overrides may make groups suboptimal.

Group Definition
Missing
< $35,200
$35,200 - $60,000
$60,000 - $85,000
$85,000 - $110,000
$110,000 - $142,530
> $142,530

Initial Characteristic Analysis

- Cut-offs may be rough from decision tree combining.
- **Optional to override** automatically generated groups to conform to business rules.
- Overrides may make groups suboptimal.

Group Definition	Override
Missing	Missing
< $35,200	< $35,000
$35,200 - $60,000	$35,000 - $60,000
$60,000 - $85,000	$60,000 - $85,000
$85,000 - $110,000	$85,000 - $110,000
$110,000 - $142,530	$110,000 - $140,000
> $142,530	> $140,000

Initial Characteristic Analysis

- Cut-offs may be rough from decision tree combining.
- **Optional to override** automatically generated groups to conform to business rules.
- Overrides may make groups suboptimal.

Group Definition	Override
Missing	Missing
< $35,200	**< $35,000**
$35,200 - $60,000	**$35,000** - $60,000
$60,000 - $85,000	$60,000 - $85,000
$85,000 - $110,000	$85,000 - $110,000
$110,000 - $142,530	$110,000 - **$140,000**
> $142,530	**> $140,000**

Initial Characteristic Analysis

- Calculate and examine the key assessment metrics:
  - **Weight of Evidence** (WOE) – how well attributes discriminate for each given characteristic
  - **Information Value** (IV) – evaluate a characteristic's overall predictive power
  - **Gini Statistic** – alternate to IV for selecting characteristics for final model.

how well is each bin separating 1 or 1.
IV is how well variable as a whole is
Gini is rarely used

how good we
at separating
1s and 0s
power

WEIGHT OF EVIDENCE

Weight of Evidence (WOE)

bins

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.
- WOE is based on comparing the proportion of goods to bads at each attribute level (levels of the predictor variable).

non defaulters

bankers came up with that.

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.
- WOE is based on comparing the proportion of goods to bads at each attribute level (levels of the predictor variable).

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

$$Dist.Good_i = \frac{Number\ Good\ in\ group\ i}{Total\ Number\ Good}$$

divided by total good in all bins

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.
- WOE is based on comparing the proportion of goods to bads at each attribute level (levels of the predictor variable).

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

$$Dist.Bad_i = \frac{Number\ Bad\ in\ group\ i}{Total\ Number\ Bad}$$

Weight of Evidence (WOE)

- What are we looking for?
  - Looking for “big” differences in WOE between groups.
  - Monotonic changes within an attribute for **interval variables** (not always required).

ppl like to see monotonic changes. As variable gets bigger, ..

- Why monotonic increases?
  - Oscillation back and forth of positive to negative values of WOE typically sign of variable that has trouble separating good vs. bad.
  - Not always required **if makes business sense** – credit card utilization for example.

WOE – Example

not a target variable, it is a perdictor variable

Good group

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 603	111	112	
2	604 – 662	378	678	
3	663 – 699	185	754	
4	700 – 717	74	440	
5	718 – 765	75	824	
6	> 765	15	498	
7	MISSING	80	153	
Total		**918**	**3,459**	

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 603	111	112	
2	604 – 662	378	678	
3	663 – 699	185	754	
4	700 – 717	74	440	
5	718 – 765	75	824	
6	> 765	15	498	
7	MISSING	80	153	
Total		**918**	**3,459**	

$$Dist.Good_1 = \frac{112}{3459}$$

$$= 0.032$$

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 603	111	112	
2	604 – 662	378	678	
3	663 – 699	185	754	
4	700 – 717	74	440	
5	718 – 765	75	824	
6	> 765	15	498	
7	MISSING	80	153	
Total		**918**	**3,459**	

$$Dist.Good_1 = \frac{112}{3459}$$

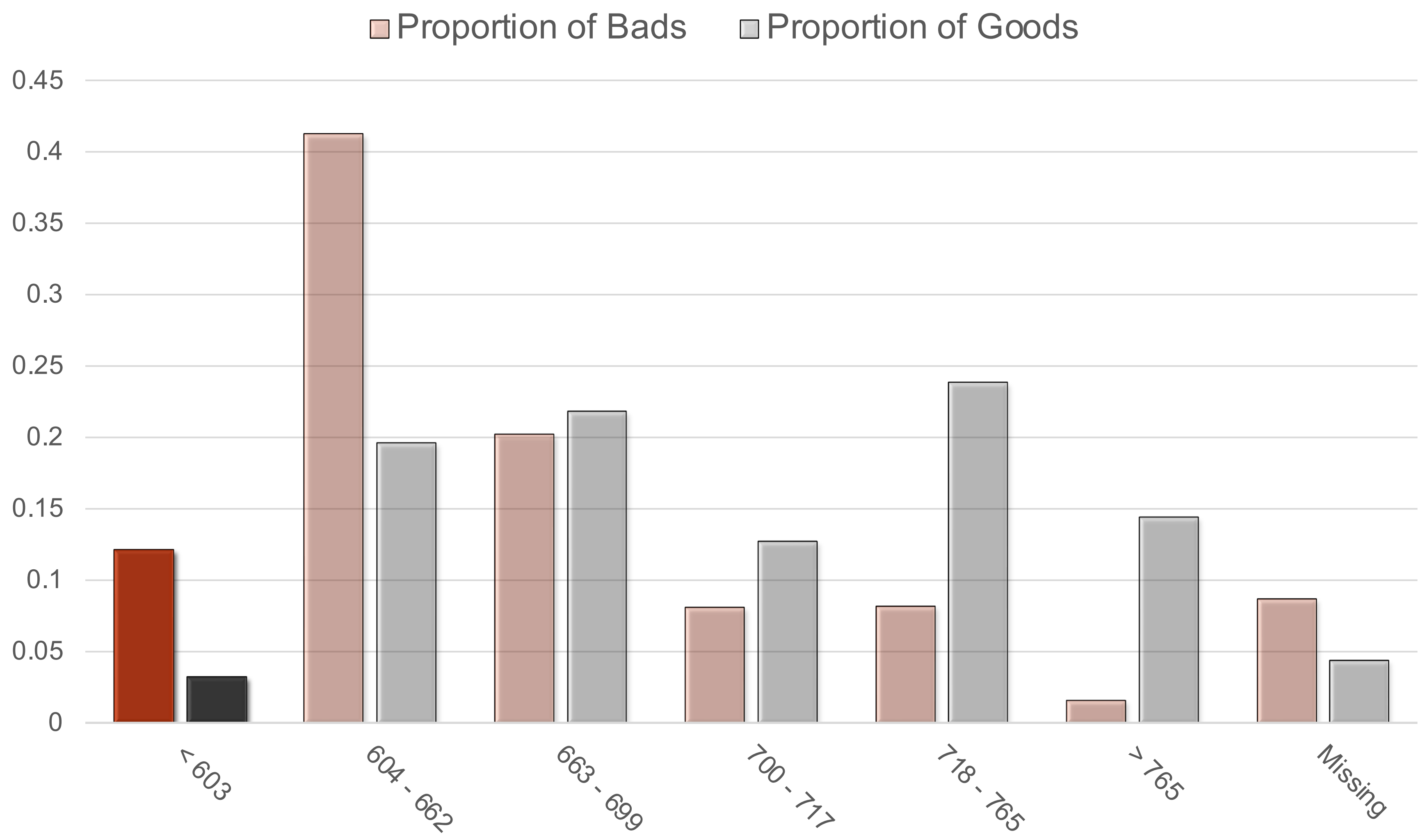
$$= 0.032$$

$$Dist.Bad_1 = \frac{111}{918}$$

$$= 0.121$$

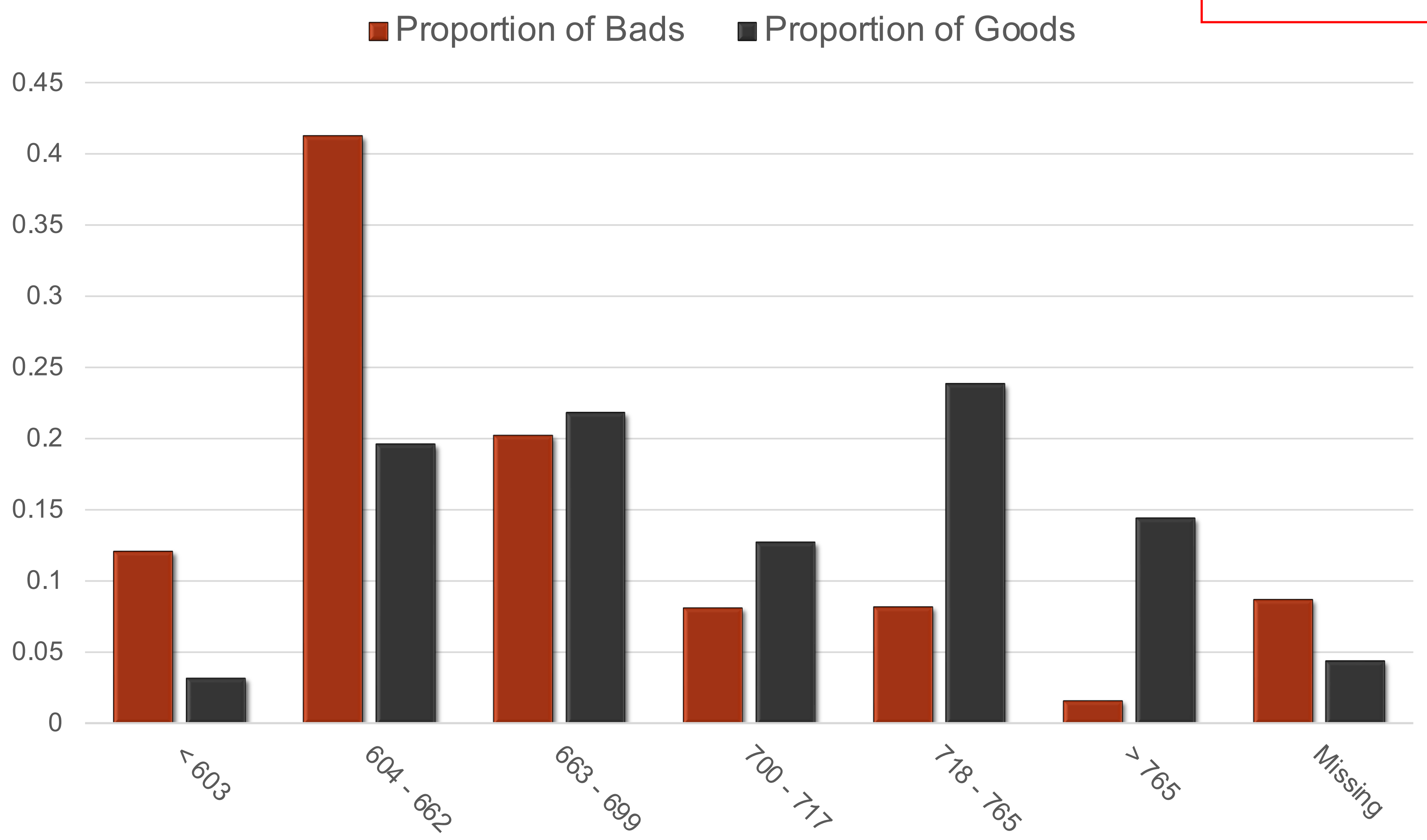
this category
leans towards
BAD 12 % vs 3%

WOE – Example

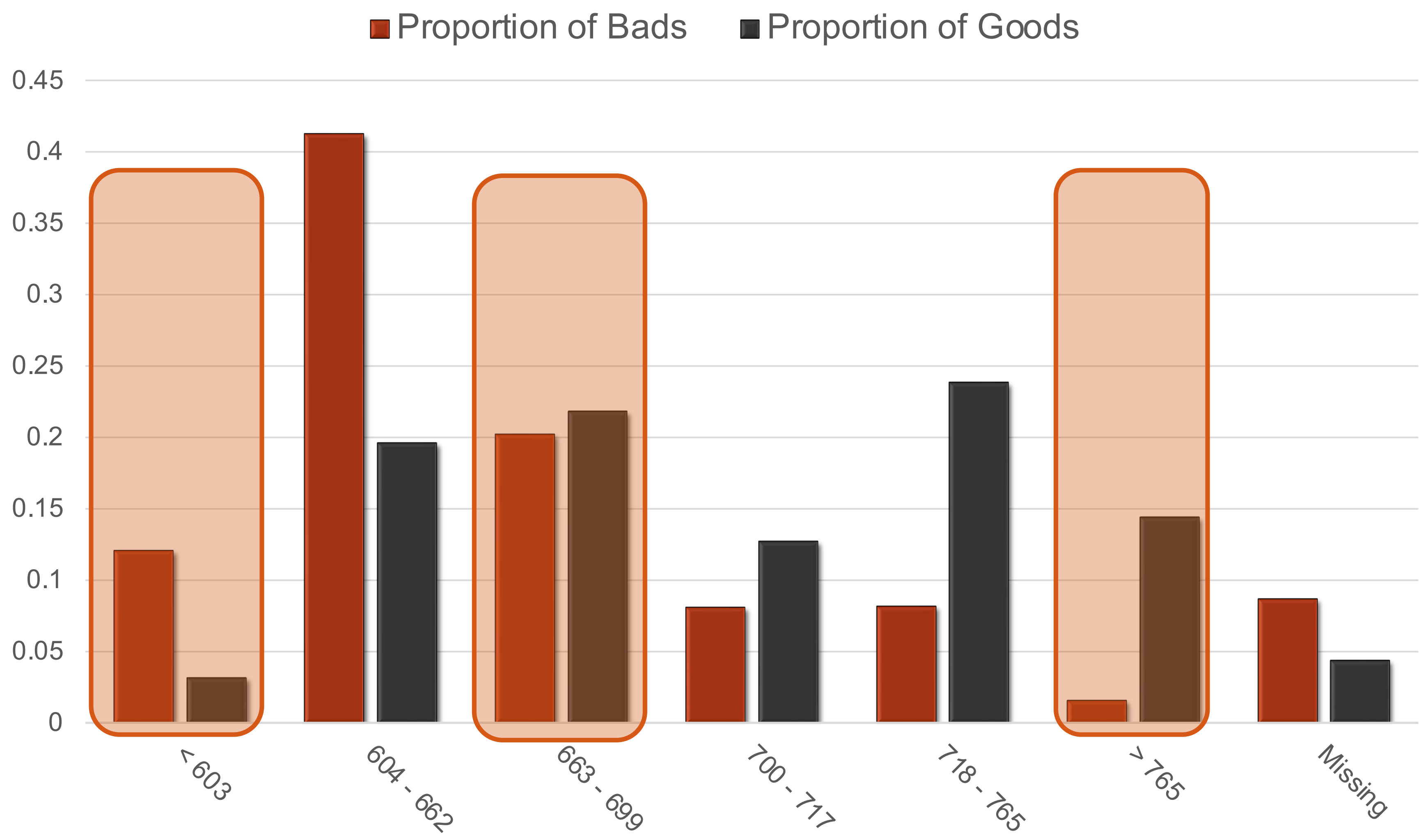

Proportion of Bads
Proportion of Goods
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
< 603
604 - 662
663 - 699
700 - 717
718 - 765
> 765
Missing

WOE – Example

All the left hand groups "Good" will sum to 1. ALL the right hand groups will sum to 1.



WOE – Example


Proportion of Bads
Proportion of Goods
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
< 603
604 - 662
663 - 699
700 - 717
718 - 765
> 765
Missing

WOE – Example

Event is Deafaulted so " Bad"

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 603	111	112	**-1.32**
2	604 – 662	378	678	
3	663 – 699	185	754	
4	700 – 717	74	440	
5	718 – 765	75	824	
6	> 765	15	498	
7	MISSING	80	153	
Total		**918**	**3,459**	

$$Dist.Good_1 = \frac{112}{3459}$$

$$= 0.032$$

$$Dist.Bad_1 = \frac{111}{918}$$

$$= 0.121$$

NOT BASE 10, it is base e

$$WOE_1 = \log\left(\frac{0.032}{0.121}\right)$$

$$= -1.32$$

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 603	111	112	**-1.32**
2	604 – 662	378	678	**-0.74**
3	663 – 699	185	754	**0.08**
4	700 – 717	74	440	**0.46**
5	718 – 765	75	824	**1.07**
6	> 765	15	498	**2.18**
7	MISSING	80	153	**-0.68**
Total		**918**	**3,459**	

Higher the number. higher the evidence. 0 means both proporiton equal, ratio is 1.

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- WOE approximately zero implies what?

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- WOE approximately zero implies % good approximately equal to % bad so group doesn't separate good vs. bad well.

WOE 0 means no evidence to separate Good from Bad.

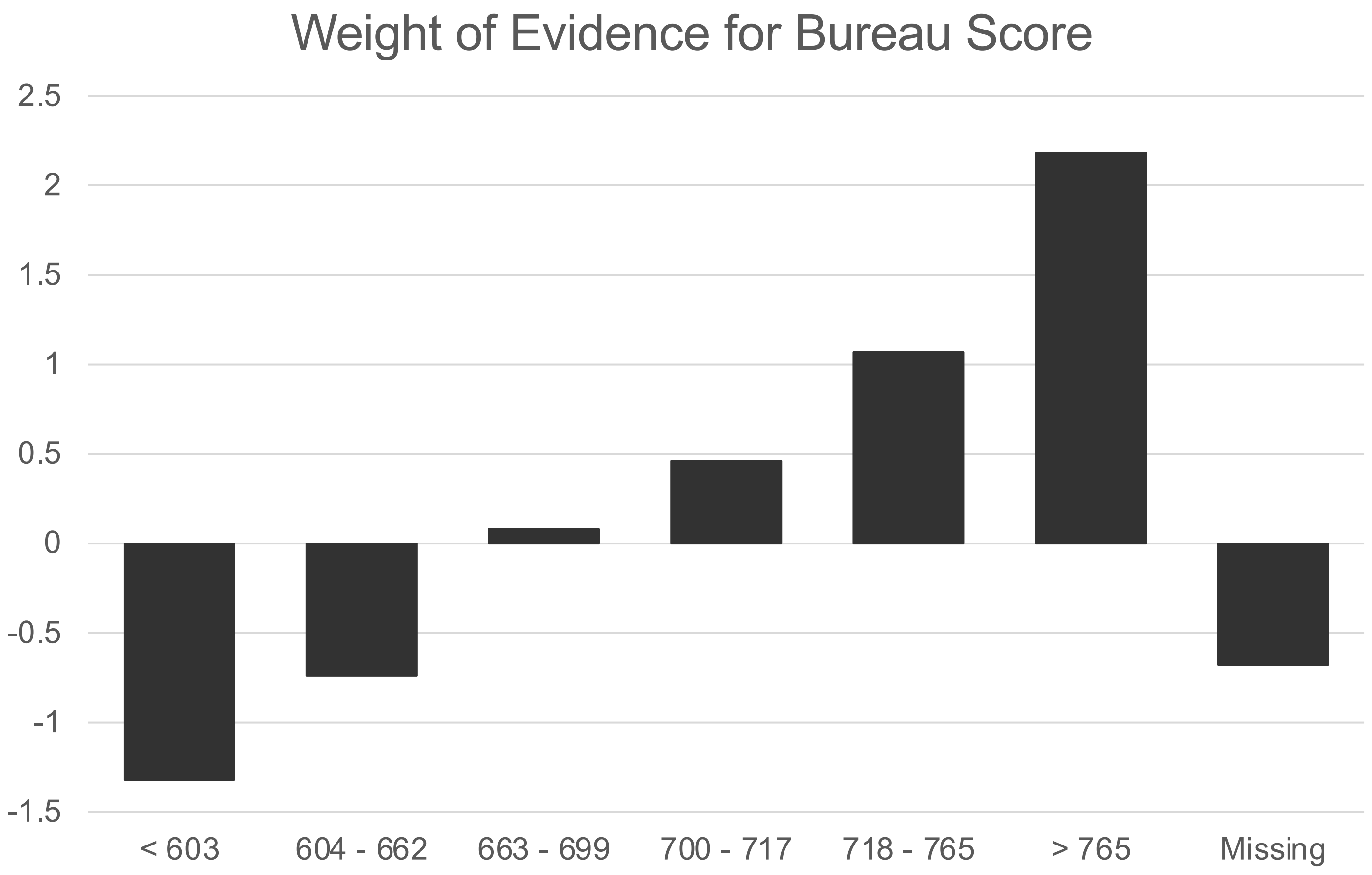
Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.

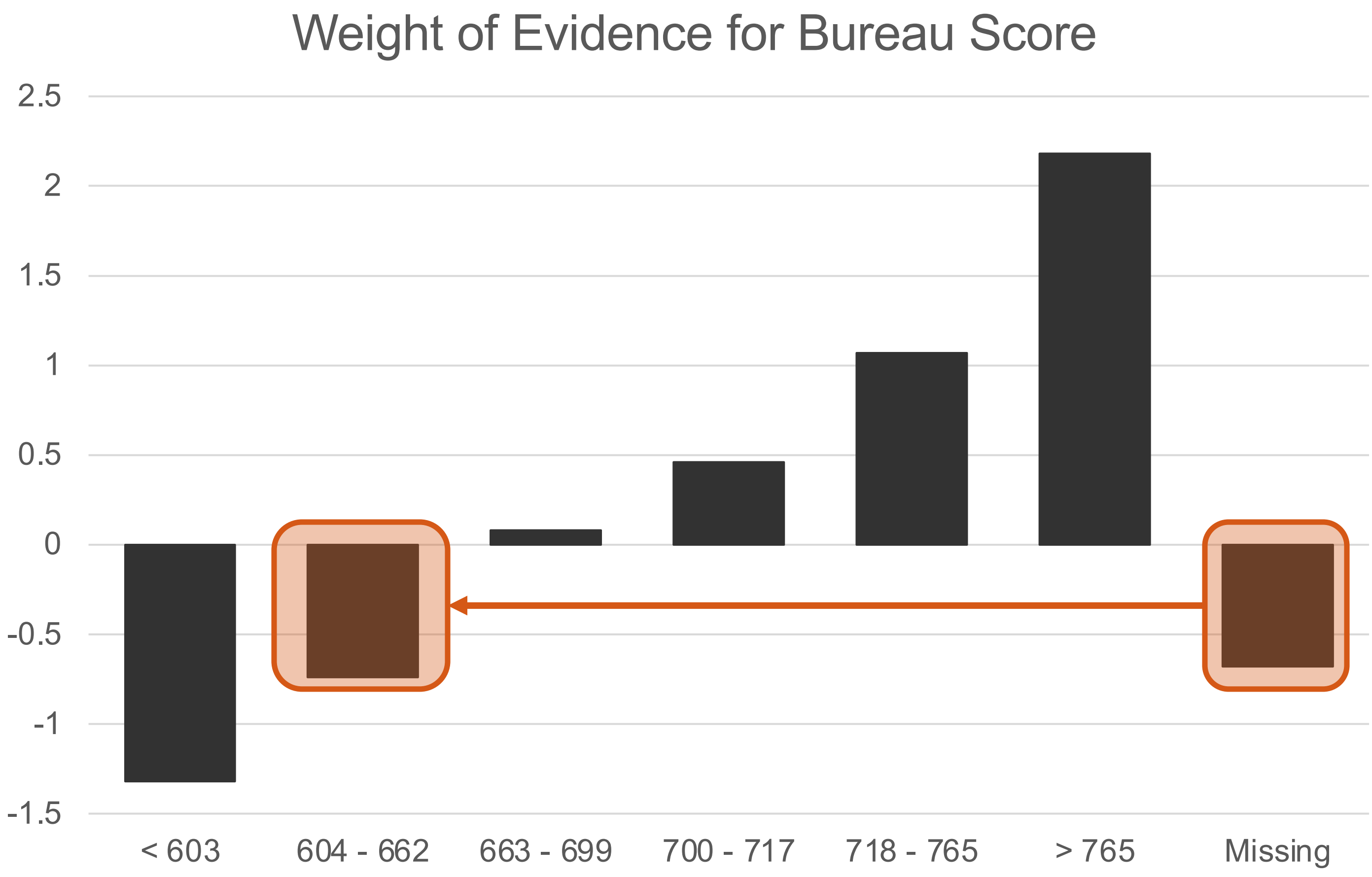
$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- WOE approximately zero implies % good approximately equal to % bad so group doesn't separate good vs. bad well.
- WOE positive implies group identifies people who are good.
- WOE negative implies group identifies people who are bad.

WOE – Example


Weight of Evidence for Bureau Score
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
< 603
604 - 662
663 - 699
700 - 717
718 - 765
> 765
Missing

WOE – Example


Weight of Evidence for Bureau Score
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
< 603
604 - 662
663 - 699
700 - 717
718 - 765
> 765
Missing

df

WOE – R

smbinning is a function and a package. Takes df, y=target variable. One downside is notice how variable is called good 1 on top of weight of eveidence calc and 0 on bottom (1 is numerator). I need variable that flags 1 as bad, 0s as good. Notice it is variable name in quotes as y and x. Then it finds cuts for you and also

numerator column name is good

```
result <- smbinning(df = train, y = "good", x = "bureau_score")
result$ivtable
```

name of column

```
##    Cutpoint CntRec CntGood CntBad CntCumRec CntCumGood CntCumBad PctRec
## 1    <= 603    223     112    111       223        112       111 0.0509
## 2    <= 662   1056     678    378      1279        790       489 0.2413
## 3    <= 699    939     754    185      2218       1544       674 0.2145
## 4    <= 717    514     440     74      2732       1984       748 0.1174
## 5    <= 765    899     824     75      3631       2808       823 0.2054
## 6     > 765    513     498     15      4144       3306       838 0.1172
## 7   Missing    233     153     80      4377       3459       918 0.0532
## 8     Total   4377    3459    918        NA         NA        NA 1.0000
```

```
##    GoodRate BadRate    Odds LnOdds     WoE     IV
## 1    0.5022  0.4978  1.0090 0.0090 -1.3176 0.1167
## 2    0.6420  0.3580  1.7937 0.5843 -0.7423 0.1602
## 3    0.8030  0.1970  4.0757 1.4050  0.0785 0.0013
## 4    0.8560  0.1440  5.9459 1.7827  0.4562 0.0213
## 5    0.9166  0.0834 10.9867 2.3967  1.0701 0.1675
## 6    0.9708  0.0292 33.2000 3.5025  2.1760 0.2777
## 7    0.6567  0.3433  1.9125 0.6484 -0.6781 0.0291
## 8    0.7903  0.2097  3.7680 1.3265  0.0000 0.7738
```

only thing you care about is cut points and weight of evidence.

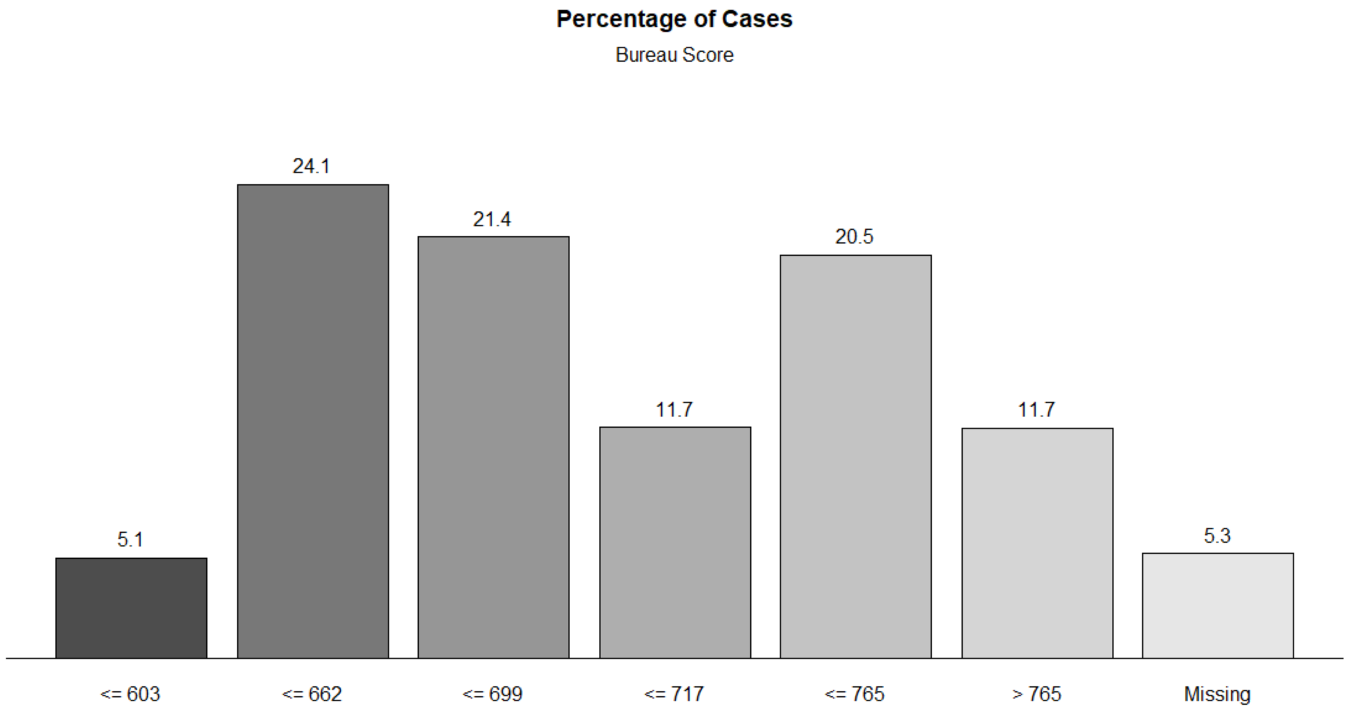
WOE – R

```
result$cut

## [1] 603 662 699 717 765

result$iv

## [1] 0.7738
```

WOE – R

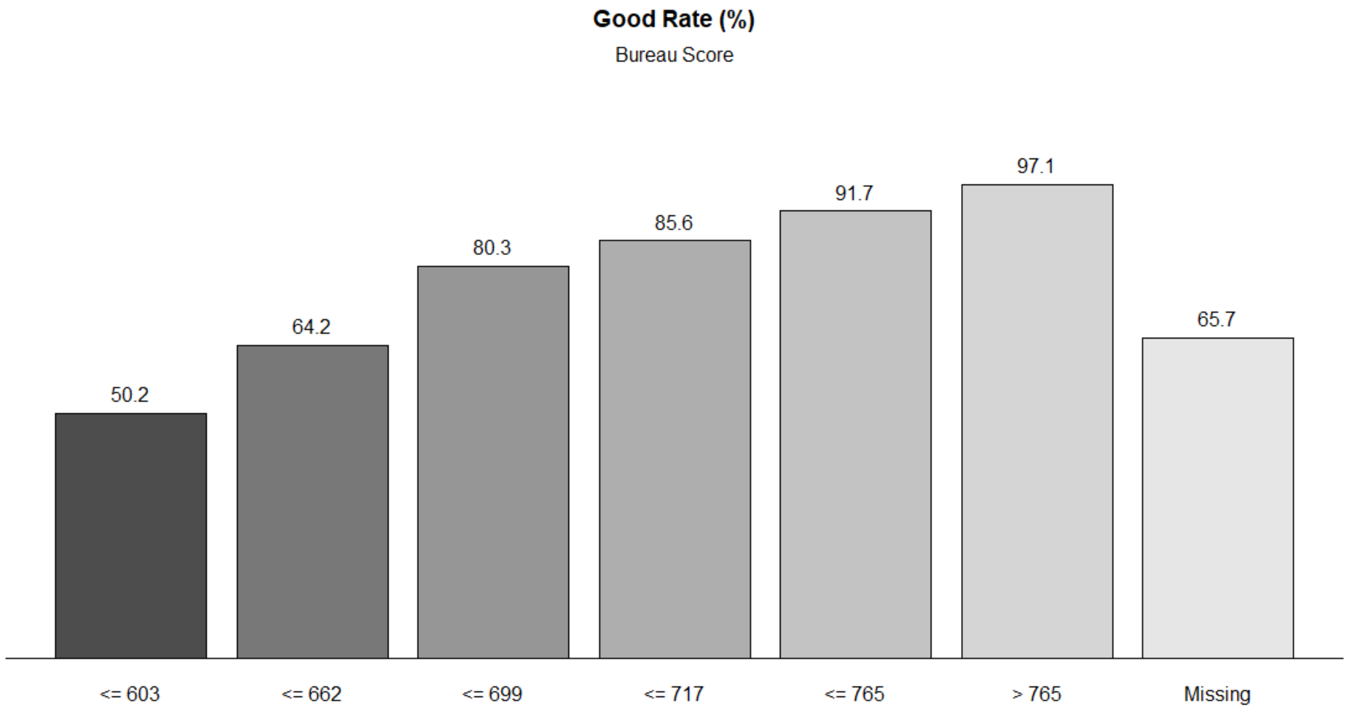
distributions

```
smbinning.plot(result, option = "dist", sub = "Bureau Score")
```



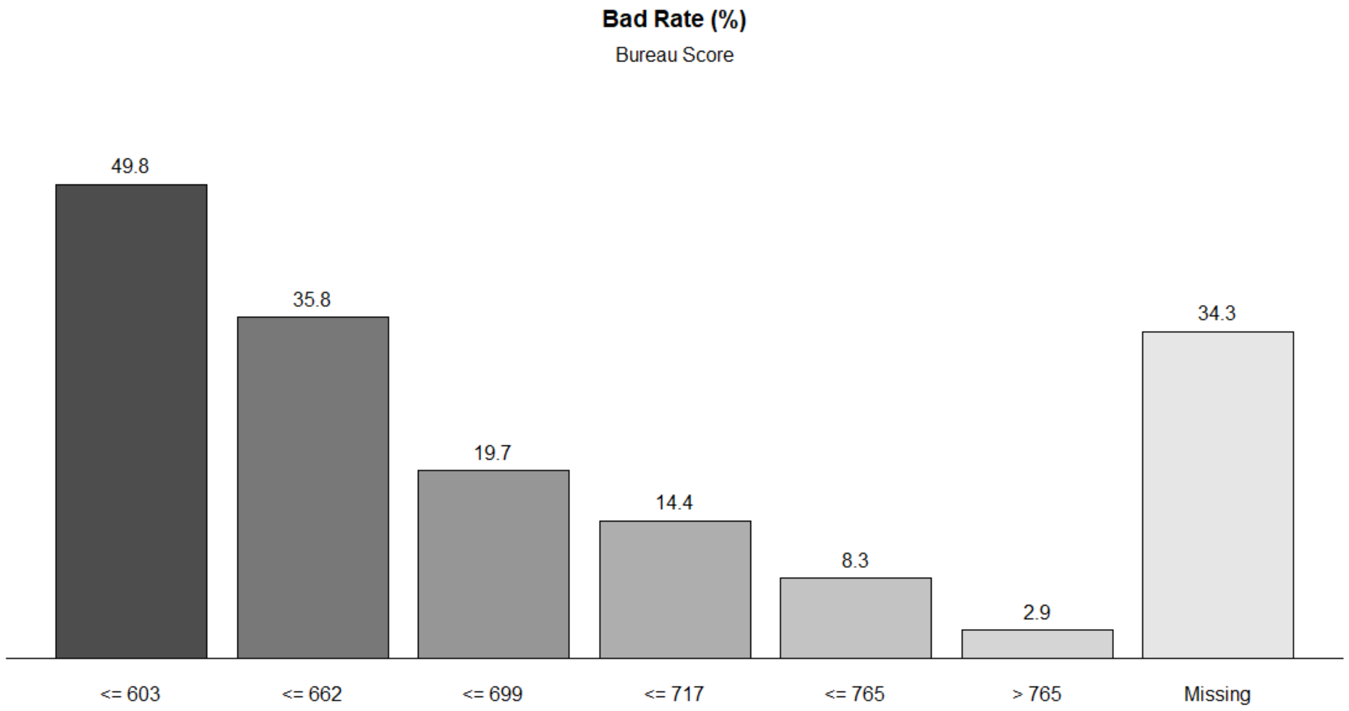
WOE – R

```
smbinning.plot(result, option = "goodrate", sub = "Bureau Score")
```



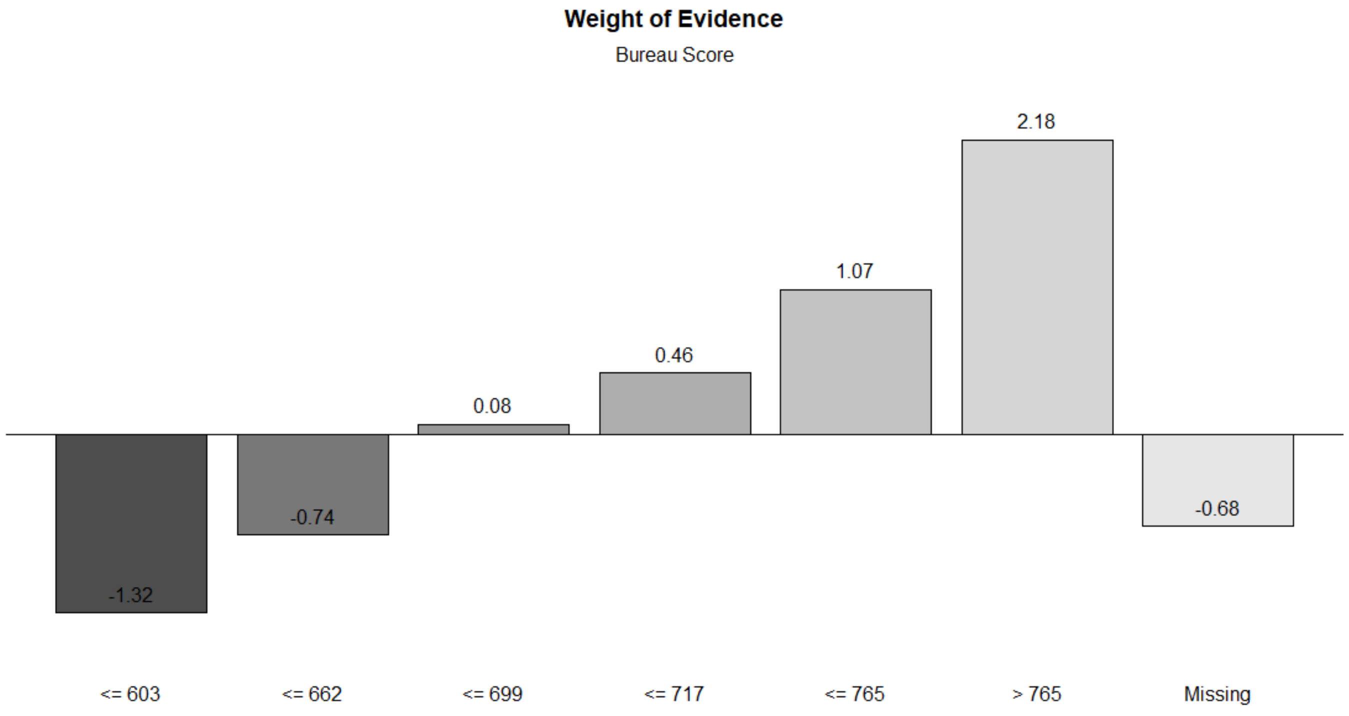
WOE – R

```
smbinning.plot(result, option = "badrate", sub = "Bureau Score")
```



WOE – R

```
smbinning.plot(result, option = "WoE", sub = "Bureau Score")
```



WOE – R

sm binning will not ocmbine existing categories inside a variable. Like it thinks not to touch categorical variable that already has bins. It is always looking for missing.

```
result <- smbinning.factor(df = train, y = "good", x = "purpose")
result$ivtable
```

numerator, good is column name

col name

```
##     Cutpoint CntRec CntGood CntBad CntCumRec CntCumGood CntCumBad PctRec
## 1 = 'LEASE'    1466    1149    317      1466       1149       317 0.3349
## 2  = 'LOAN'    2911    2310    601      4377       3459       918 0.6651
## 3    Missing      0       0      0      4377       3459       918 0.0000
## 4      Total   4377    3459    918        NA         NA        NA 1.0000

##   GoodRate BadRate   Odds LnOdds     WoE     IV
## 1   0.7838  0.2162 3.6246 1.2877 -0.0388 0.0005
## 2   0.7935  0.2065 3.8436 1.3464  0.0199 0.0003
## 3      NaN     NaN    NaN    NaN     NaN    NaN
## 4   0.7903  0.2097 3.7680 1.3265  0.0000 0.0008
```

Separation Issues Remain

This should be found in exploration phase

- Quasi-complete separation still a problem:

	Non-Event	Event	WOE
A	28	7	**-0.032**
B	16	0	∞
C	94	11	**0.728**
D	23	21	**-1.327**
Total	**161**	**39**	

Adjusted WOE

- Adjust the WOE calculation to account for possible quasi-complete separation:

$$Adjusted\ WOE_i = \log\left(\frac{Dist.Good_i + \eta_1}{Dist.Bad_i + \eta_2}\right)$$

- The η_1 and η_2 parameters are smoothing parameters that correct for potential overfitting and also protect against quasi-complete separation.
- Most software just sets $\eta_1 = \eta_2$ and has one parameter.

Adjusted WOE ($\eta_1 = \eta_2 = 0.005$)

- Quasi-complete separation no longer a problem:

	Non-Event	Event	WOE
A	28	7	**-0.031**
B	16	0	**3.039**
C	94	11	**0.719**
D	23	21	**-1.302**
Total	**161**	**39**	

Smoothed WOE (SWOE)

- SAS has recently proposed a slightly different smoothed version of the WOE calculation to account for possible quasi-complete separation:

$$SWOE_i = \log\left(\frac{\#Bad_i + (Overall\ Prop.Bad)\times c}{\#Good_i + (Overall\ Prop.Good)\times c}\right)$$

- This is just a smoothing parameter put in a slightly different place in the WOE calculation based on more Bayesian inference techniques.
- Haven't seen it really used elsewhere.

INFORMATION VALUE

how good ALL variables are at predicting, used for ranking imp variable. Get 1 # for each varibale. then put them all in 1 table.

Higher the # better at predicting Good/bad

Information Value (IV)

Uses WOE

looks at all 20 variables

- How big is a “big” difference when looking across groups for WOE?
- IV measures the ability of the characteristic to separate goods vs. bads.

entire variable

$$IV = \sum_{i=1}^{L} (Dist.Good_i - Dist.Bad_i) \times \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

IC is the sum of all groups, diff bet Good/Bad distr times WOE

Weight of Evidence

multiplying to make it positive

Information Value (IV)

- How big is a “big” difference when looking across groups for WOE?
- IV measures the ability of the characteristic to separate goods vs. bads.

$$IV = \sum_{i=1}^{L} (Dist.Good_i - Dist.Bad_i) \times \boxed{\log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)}$$

Weight of Evidence!

Information Value (IV)

- How big is a “big” difference when looking across groups for WOE?
- IV measures the ability of the characteristic to separate goods vs. bads.

$$IV = \sum_{i=1}^{L} (Dist.Good_i - Dist.Bad_i) \times \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- Used to select characteristics with strong predictive value.

Information Value (IV)

banks use IV to include or exclude variables in model

- Characteristics of IV:
  - $IV \geq 0$
  - Bigger is Better!

- Rules of Thumb:
  - IV < 0.02 – Not predictive
  - 0.02 < IV < 0.1 – Weak predictor
  - 0.1 < IV < 0.25 – Medium predictor
  - 0.25 < IV – Strong predictor

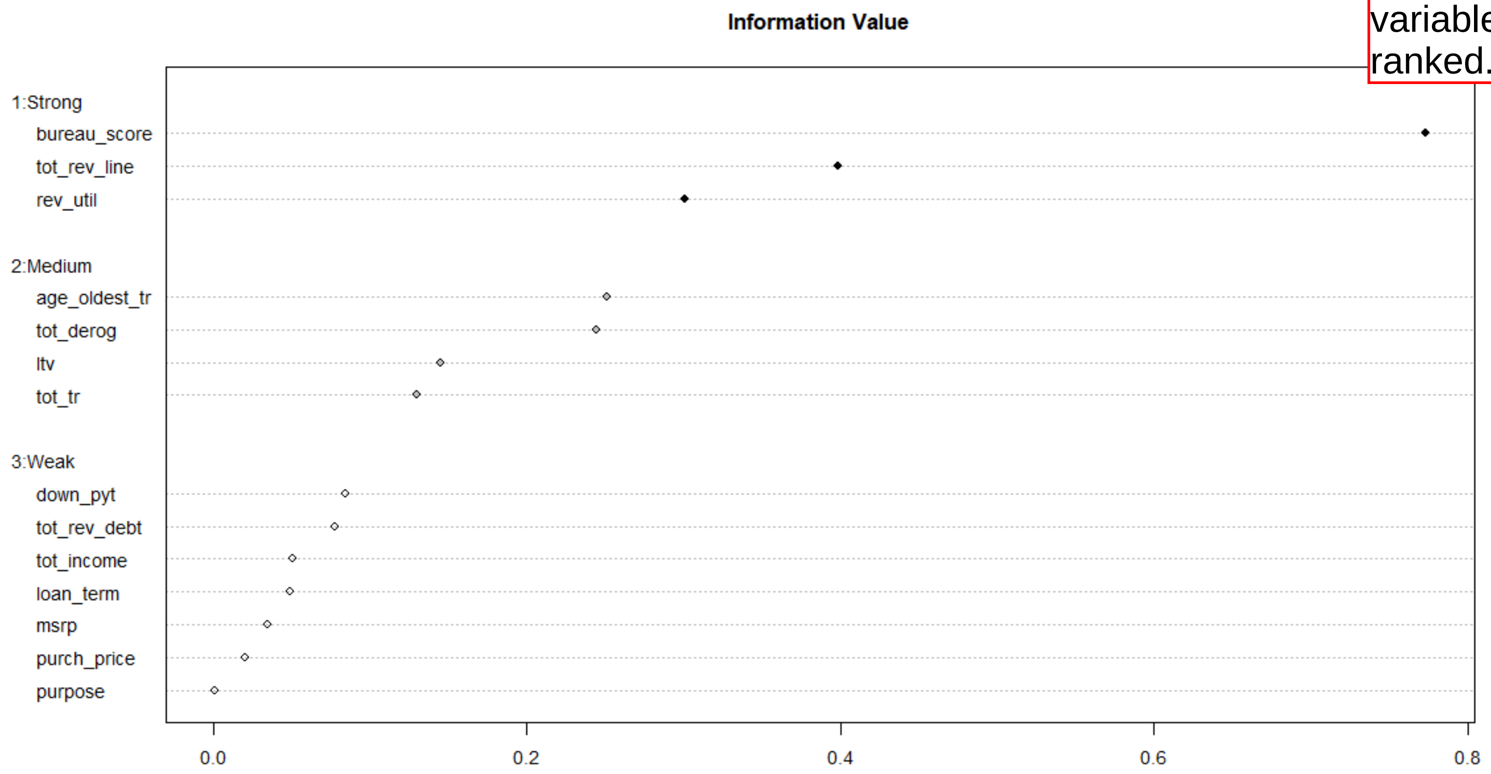
These are banking rules of thumbs, in banking dont like seeing above 0.5. it implies over predicting. Bureau scores already have greater than 0.5. I have already given them loans, how do you think i gave them loans? I used credit scores to give them loan so variable credit scores will be good already - looping kinda variable..HOW DID YOU MKAE THAT ORIGINAL DECISION?

IN scenarios like this, we build 2 models, 1 with bureau score and 1 without. Then we ensemble together. This is a banking construct.

Information Value (IV) – R

```
iv_summary <- smbinning.sumiv(df = train, y = "good")

smbinning.sumiv.plot(iv_summary)
```

everything else it assumes predictor variables. Give you imp variables ranked.



Information Value (IV) – R

dont like >0.5 cuz over corelate bureau score

smbinning by default looks at all variables , anything numeric it tries to split with CIT ie chi square test to split/bin numeric variable. So ensure variables coded 0 and 1 are factors, that way doesnt cause issues.

```
iv_summary

##              Char     IV              Process
## 12  bureau_score 0.7738    Numeric binning OK
## 10  tot_rev_line 0.3987    Numeric binning OK
## 11      rev_util 0.3007    Numeric binning OK
## 6  age_oldest_tr 0.2512    Numeric binning OK
## 4      tot_derog 0.2443    Numeric binning OK
## 19           ltv 0.1454    Numeric binning OK
## 5         tot_tr 0.1304    Numeric binning OK
## 15      down_pyt 0.0848    Numeric binning OK
## 9   tot_rev_debt 0.0782    Numeric binning OK
## 20    tot_income 0.0512    Numeric binning OK
## 17     loan_term 0.0496    Numeric binning OK
## 14          msrp 0.0353    Numeric binning OK
## 13   purch_price 0.0204    Numeric binning OK
## 16       purpose 0.0008     Factor binning OK
## 1     bankruptcy     NA    Uniques values < 5
## 2            bad     NA    Uniques values < 5
## 3         app_id     NA No significant splits
## 7    tot_open_tr     NA No significant splits
## 8     tot_rev_tr     NA No significant splits
## 18      loan_amt     NA No significant splits
## 21      used_ind     NA    Uniques values < 5
## 22        weight     NA    Uniques values < 5
```

no probs with this variable.

will not bin if numeric variables if less than 5 unique value.s

dont worru about this one <5.

no sig splits means numeric variable that it could not find statistic relationship with your target variable statistically. Loan amt had 0 predictive power.

Information Value (IV) – R

```
iv_summary

##               Char     IV                Process
## 12    bureau_score 0.7738      Numeric binning OK
## 10    tot_rev_line 0.3987      Numeric binning OK
## 11        rev_util 0.3007      Numeric binning OK
## 6    age_oldest_tr 0.2512      Numeric binning OK
## 4        tot_derog 0.2443      Numeric binning OK
## 19             ltv 0.1454      Numeric binning OK
## 5           tot_tr 0.1304      Numeric binning OK
## 15        down_pyt 0.0848      Numeric binning OK
## 9     tot_rev_debt 0.0782      Numeric binning OK
## 20      tot_income 0.0512      Numeric binning OK
## 17       loan_term 0.0496      Numeric binning OK
## 14            msrp 0.0353      Numeric binning OK
## 13     purch_price 0.0204      Numeric binning OK
## 16         purpose 0.0008       Factor binning OK
## 1       bankruptcy     NA      Uniques values < 5
## 2              bad     NA      Uniques values < 5
## 3           app_id     NA No significant splits
## 7      tot_open_tr     NA No significant splits
## 8       tot_rev_tr     NA No significant splits
## 18        loan_amt     NA No significant splits
## 21        used_ind     NA      Uniques values < 5
## 22          weight     NA      Uniques values < 5
```

Information Value (IV) – R

```
iv_summary

##                Char     IV                 Process
## 12  bureau_score 0.7738    Numeric binning OK
## 10  tot_rev_line 0.3987    Numeric binning OK
## 11      rev_util 0.3007    Numeric binning OK
## 6  age_oldest_tr 0.2512    Numeric binning OK
## 4      tot_derog 0.2443    Numeric binning OK
## 19           ltv 0.1454    Numeric binning OK
## 5         tot_tr 0.1304    Numeric binning OK
## 15      down_pyt 0.0848    Numeric binning OK
## 9   tot_rev_debt 0.0782    Numeric binning OK
## 20    tot_income 0.0512    Numeric binning OK
## 17     loan_term 0.0496    Numeric binning OK
## 14          msrp 0.0353    Numeric binning OK
## 13   purch_price 0.0204    Numeric binning OK
## 16       purpose 0.0008     Factor binning OK
## 1     bankruptcy     NA    Uniques values < 5
## 2            bad     NA    Uniques values < 5
## 3         app_id     NA No significant splits
## 7    tot_open_tr     NA No significant splits
## 8     tot_rev_tr     NA No significant splits
## 18      loan_amt     NA No significant splits
## 21      used_ind     NA    Uniques values < 5
## 22        weight     NA    Uniques values < 5
```

sm binning will not bin numerical variable if it has less than 5 variables. less than 5 is basically a categorical variable

Information Value (IV) – R

```
iv_summary

##                  Char     IV                 Process
## 12  bureau_score 0.7738    Numeric binning OK
## 10  tot_rev_line 0.3987    Numeric binning OK
## 11      rev_util 0.3007    Numeric binning OK
## 6  age_oldest_tr 0.2512    Numeric binning OK
## 4      tot_derog 0.2443    Numeric binning OK
## 19           ltv 0.1454    Numeric binning OK
## 5         tot_tr 0.1304    Numeric binning OK
## 15      down_pyt 0.0848    Numeric binning OK
## 9   tot_rev_debt 0.0782    Numeric binning OK
## 20    tot_income 0.0512    Numeric binning OK
## 17     loan_term 0.0496    Numeric binning OK
## 14          msrp 0.0353    Numeric binning OK
## 13   purch_price 0.0204    Numeric binning OK
## 16       purpose 0.0008     Factor binning OK
## 1     bankruptcy     NA    Uniques values < 5
## 2            bad     NA    Uniques values < 5
## 3         app_id     NA No significant splits
## 7    tot_open_tr     NA No significant splits
## 8     tot_rev_tr     NA No significant splits
## 18      loan_amt     NA No significant splits
## 21      used_ind     NA    Uniques values < 5
## 22        weight     NA    Uniques values < 5
```

this means this is numeric variable that could not find any statistical relationship with target variable.

sm binning have a shot at variable selection if you have cont variables

Information Value (IV)

- Characteristics of IV:
  - $IV \geq 0$
  - Bigger is Better!

- Rules of Thumb:
  - IV < 0.02 – Not predictive
  - 0.02 < IV < 0.1 – Weak predictor
  - 0.1 < IV < 0.25 – Medium predictor
  - 0.25 < IV < 0.5 – Strong predictor
  - IV > 0.5 – Over-predicting?

Information Value (IV)

- Rules of Thumb:
  - IV < 0.02 – Not predictive
  - 0.02 < IV < 0.1 – Weak predictor
  - 0.1 < IV < 0.25 – Medium predictor
  - 0.25 < IV < 0.5 – Strong predictor
  - IV > 0.5 – Over-predicting?
- Over-predicting Example:
  - All previous mortgage decisions have been made only on bureau score so of course bureau score is highly predictive – becomes only significant variable!
  - Create two models – one with bureau score, one without bureau score and **ensemble**.

GINI STATISTIC

Gini Statistic

- **Gini statistic** is optional technique that tries to answer the same question as Information Value – which variables are strong enough to enter the scorecard model?
- IV is more in line with WOE calculation and used more often.
- Characteristics:
  - Range is 0 to 100.
  - Bigger is Better.

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E}\times\sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E}\times n_{i,NE}\right)\right)}{N_E\times N_{NE}}\right)\times 100$$

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Number of events in group *i*

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Number of non-events in group *i*

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Total number of events and non-events

?

PROC BINNING IN SAS VIYA

WOE – SAS Viya

```
proc binning data = public.train method = tree(initbin = 100
                 maxnbins = 10);
    target bad / level = int;
    input bureau_score / level = int;
    ods output BinDetails = bincuts VarTransInfo = bincount;
run;
```

WOE – SAS Viya

Bin Details										
Variable	Bin ID	Lower Bound	Upper Bound	Bin Width	N Levels	Number of Observations	Mean	Standard Deviation	Minimum	Maximum
bureau_score										

WOE – SAS Viya

Transformation Information					
Variable	N	N Miss	N Bins	Importance	Relative Importance
bureau_score					

WOE – SAS Viya

```
data _null_;
   set bincount;
   call symput('numbin', Nbins - 1);
run;

proc sql;
   select Max
      into :cuts separated by ' '
      from bincuts(firstobs = 2 obs = &numbin);
quit;

proc binning data = public.train numbin = &numbin
             method=cutpts(&cuts) woe;
   target bad / event = '1';
   input bureau_score / level = int;
run;
```

WOE – SAS Viya

Bin Details							
Variable	Bin ID	Lower Bound	Upper Bound	Bin Width	Number of Observations	Mean	Standard Deviation
bureau_score							

WOE – SAS Viya

Bin Details						
Variable	**Bin ID**	**Minimum**	**Maximum**	**Event Count**	**Weight of Evidence**	**Information Value**
bureau_score						

WOE – SAS Viya

Variable Information Value	
Variable	**Information Value**
bureau_score	

WOE – SAS Viya

```
proc tabulate data=public.train out=facwoe;
   class bad purpose;
   table purpose, bad*colpctn / rts=10;
run;

proc transpose data = facwoe out = facwoe2(rename=
                                   (col1 = bad0 col2 = bad1));
   var PctN_10;
   by purpose;
run;

data facwoe2;
   set facwoe2;
   WOE = log(bad1/bad0);
run;
```

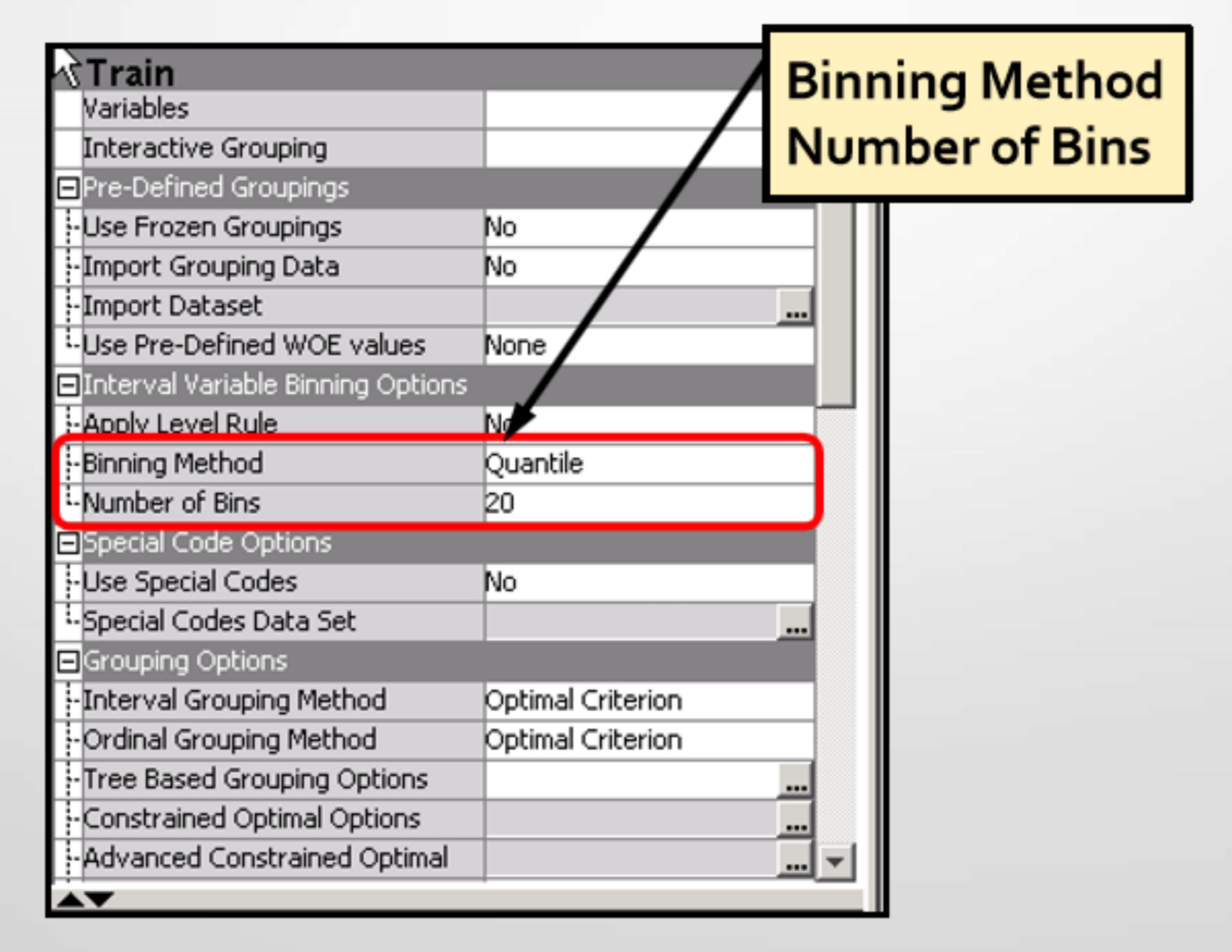
WOE – SAS Viya

	bad	
	0	1
	ColPctN	ColPctN
purpose		
LEASE		
LOAN		

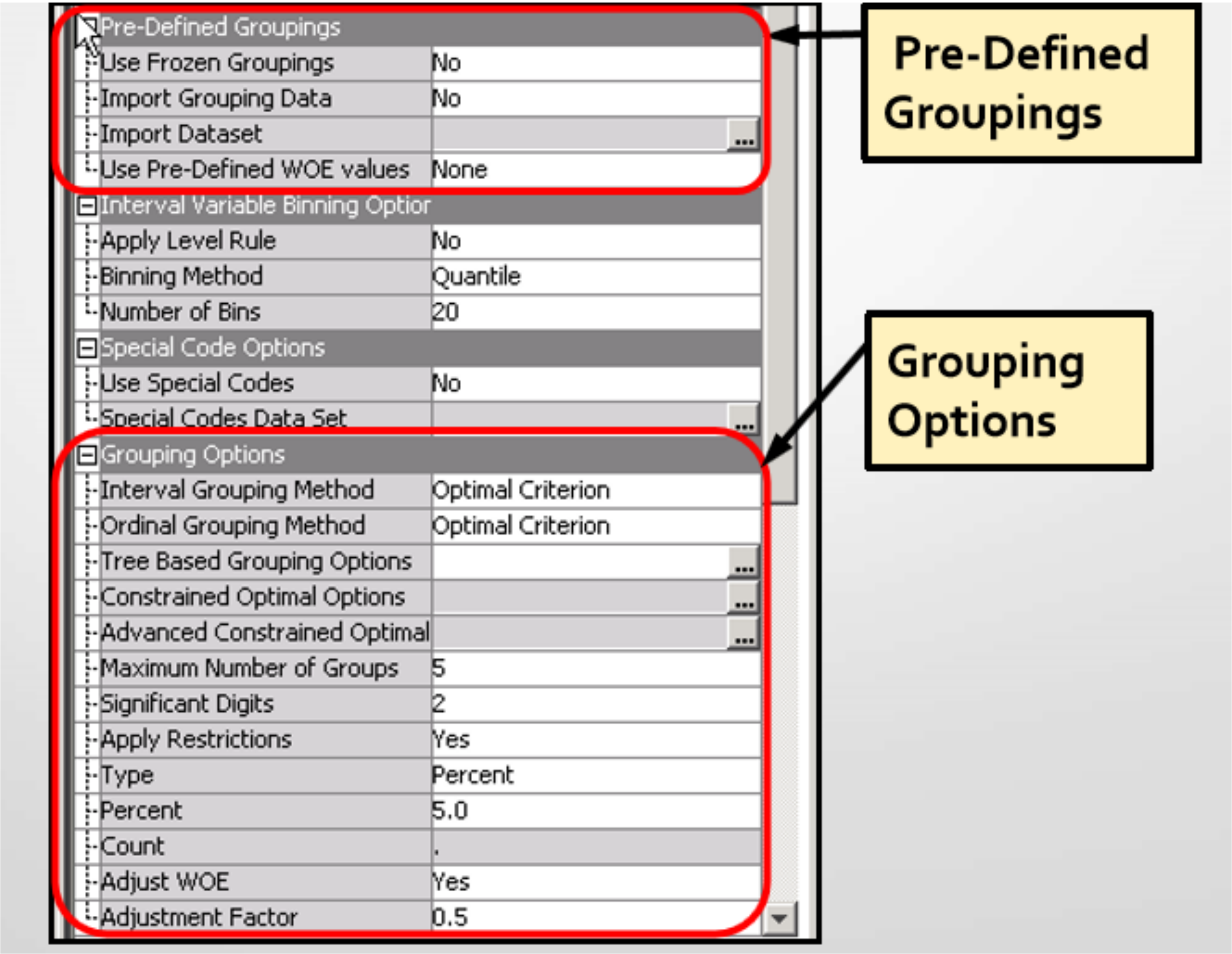
Obs	purpose	_NAME_	bad0	bad1	WOE
1					
2					

INTERACTIVE GROUPING NODE IN SAS EM

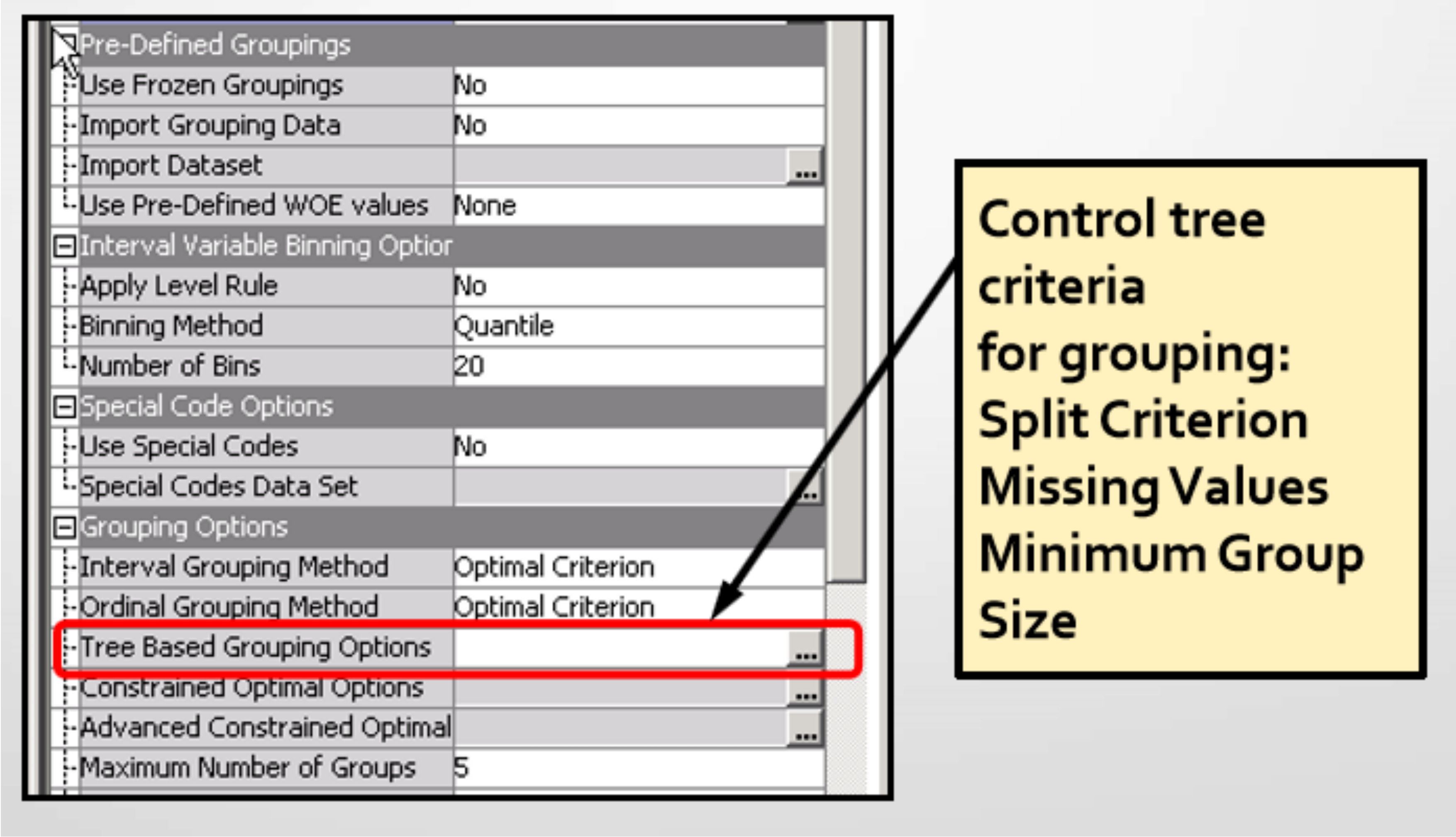
Pre-Binning of the Interval Variables



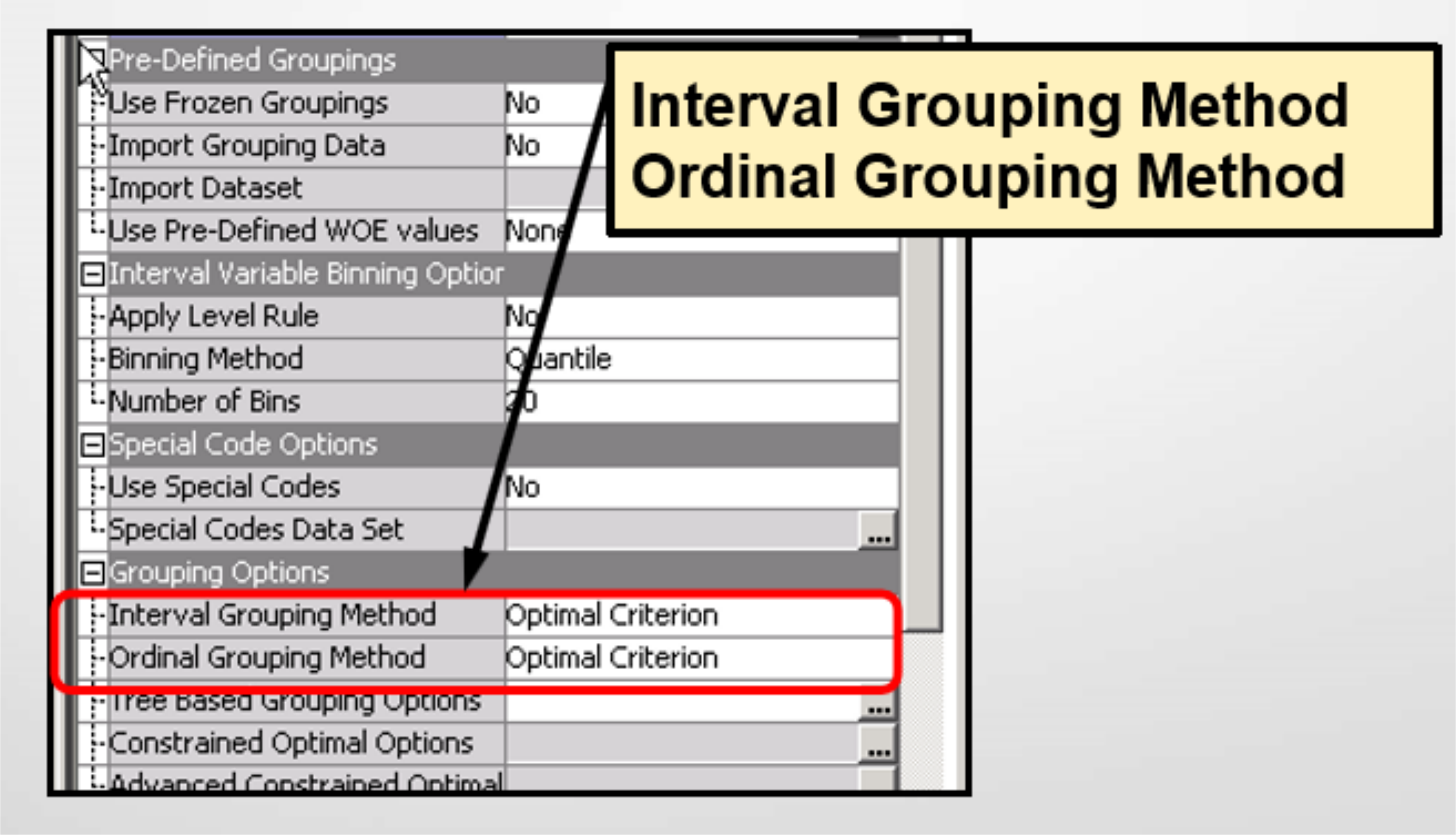
Grouping Options



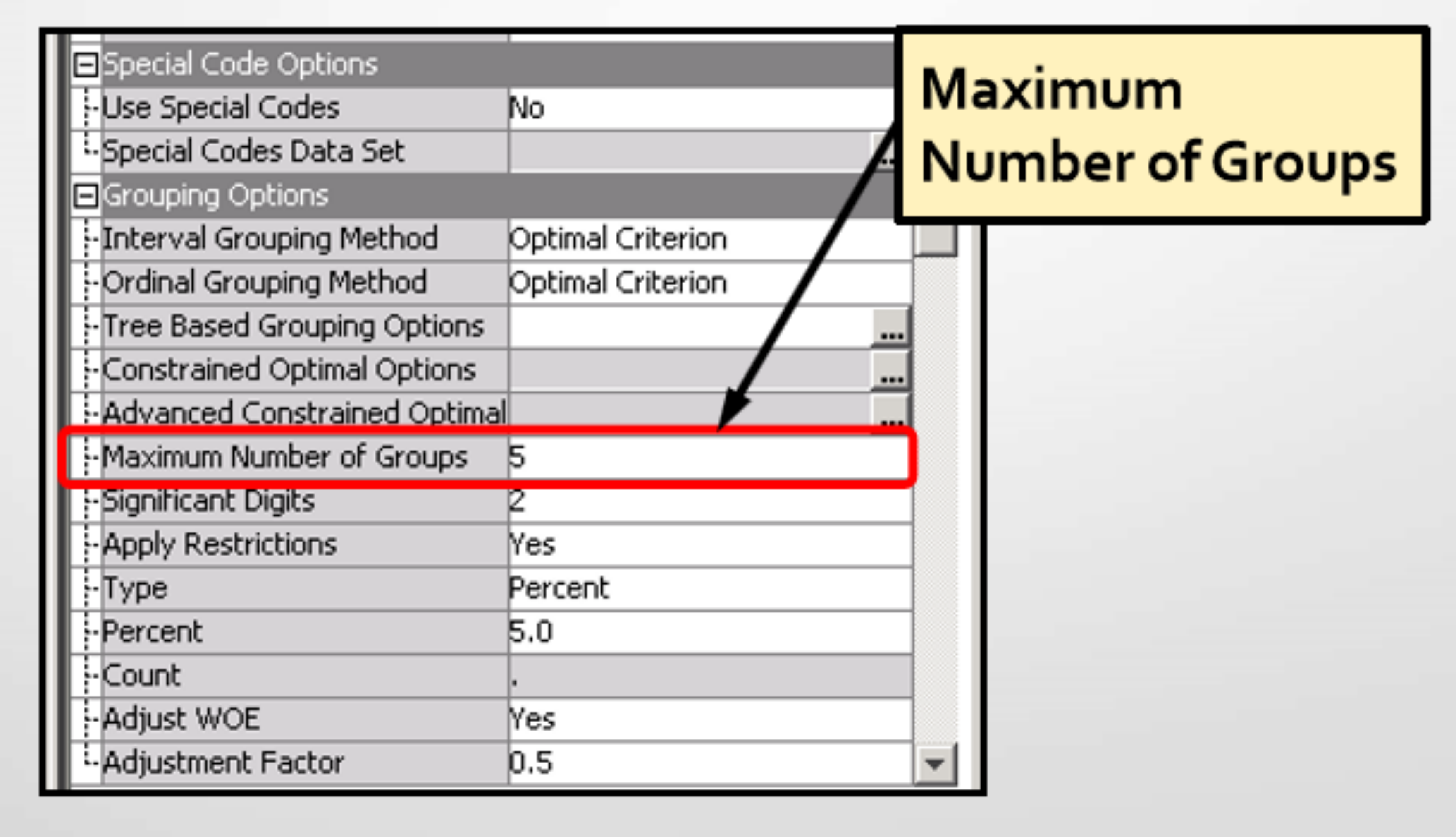
Grouping Options: Tree Criteria



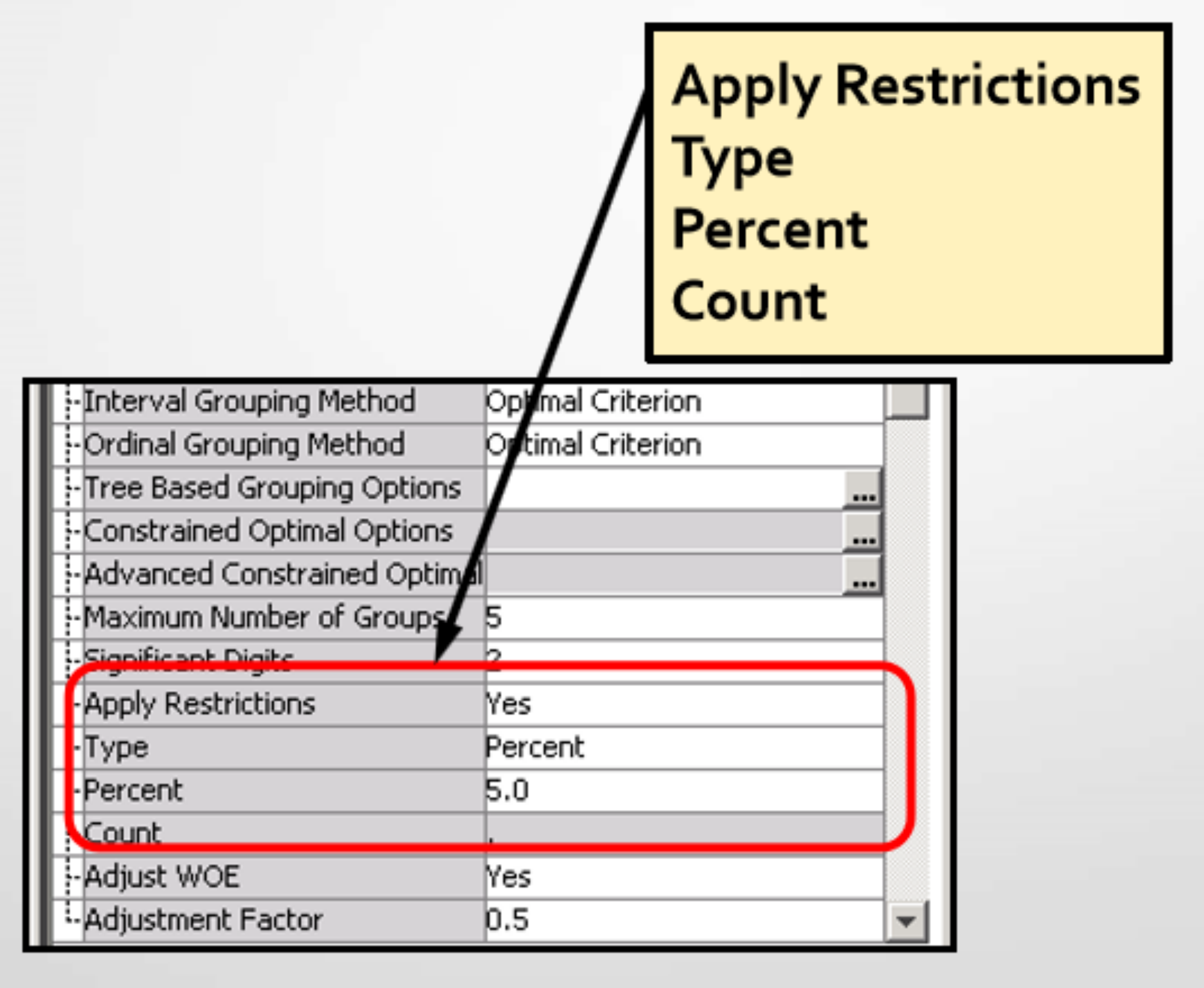
Grouping Options: Interval vs. Ordinal



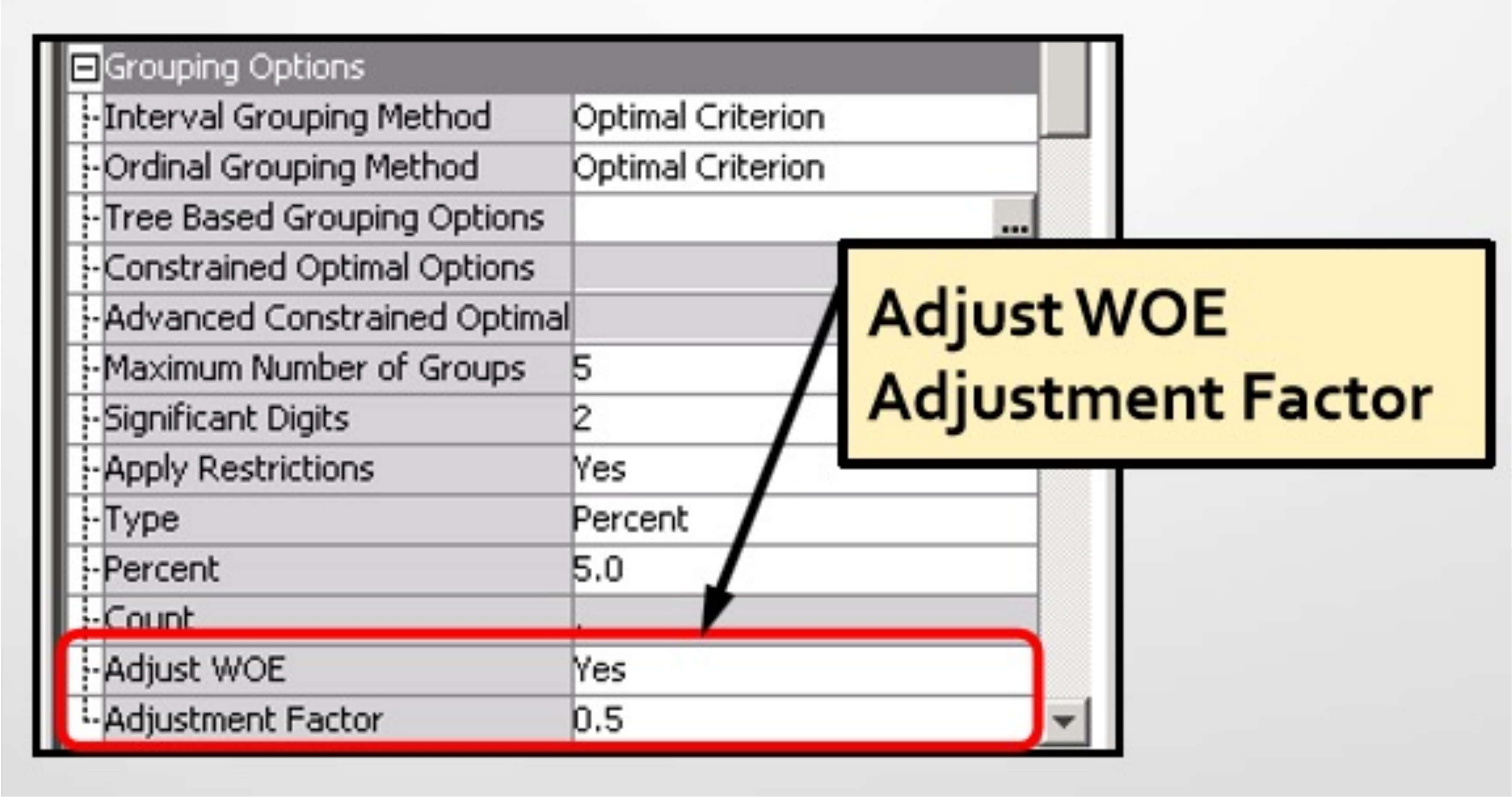
Grouping Options: Number of Groups



Grouping Options: Stopping Rules



Grouping Options: WOE Adjustments



?