Chapter 5
Predicting High-Risk-Bin Memberships (1,445 Families)

General Purpose

Optimal bins describe continuous predictor variables in the form of best fit categories for making predictions, e.g., about families at high risk of bank loan defaults. In addition, it can be used for, e.g., predicting health risk cut-offs about individual future families, based on their characteristics (Chap. 56).

Specific Scientific Question

Can optimal binning also be applied for other medical purposes, e.g., for finding high risk cut-offs for overweight children in particular families?

Example

A data file of 1,445 families was assessed for learning the best fit cut-off values of unhealthy lifestyle estimators to maximize the difference between low and high risk of overweight children. These cut-off values were, subsequently, used to determine the risk profiles (the characteristics) in individual future families.

This chapter was previously published in "Machine learning in medicine-cookbook 2" as Chap. 2, 2014.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_5

Var 1	Var 2	Var 3	Var 4	Var 5
0	11	1	8	0
0	7	1	9	0
1	25	7	0	1
0	11	4	5	0
1	5	1	8	1
0	10	2	8	0
0	11	1	6	0
0	7	1	8	0
0	7	0	9	0
0	15	3	0	0

Var = variable
Var 1 fruitvegetables (times per week)
Var 2 unhealthysnacks (times per week)
Var 3 fastfoodmeal (times per week)
Var 4 physicalactivities (times per week)
Var 5 overweightchildren (0 = no, 1 = yes)

Only the first 10 families of the original learning data file are given, the entire data file is entitled “optimalbinning1” and is in extras.springer.com.

Optimal Binning

SPSS 19.0 is used for analysis. Start by opening the data file.

Command:

Transform….Optimal Binning….Variables into Bins: enter fruitvegetables, unhealthysnacks, fastfoodmeal, physicalactivities….Optimize Bins with Respect to: enter "overweightchildren"….click Output….Display: mark Endpoints….mark Descriptive statistics….mark Model Entropy….click Save: mark Create variables that contain binned data….Save Binning Rules in a Syntax file: click Browse…. open appropriate folder….File name: enter, e.g., "exportoptimalbinning"….click Save….click OK.

fruitvegetables/wk

	End point		Number of cases by level of overweight children		
Bin	Lower	Upper	No	Yes	Total
1	[a]	14	802	340	1142
2	14	[a]	274	29	303
Total			1076	369	1445

unhealthysnacks/wk

	End point		Number of cases by level of overweight children		
Bin	Lower	Upper	No	Yes	Total
1	[a]	12	830	143	973
2	12	19	188	126	314
3	19	[a]	58	100	158
Total			1076	369	1445

fastfoodmeal/wk

	End point		Number of cases by level of overweight children		
Bin	Lower	Upper	No	Yes	Total
1	[a]	2	896	229	1125
2	2	[a]	180	140	320
Total			1076	369	1445

physicalactivities/wk

	End point		Number of cases by level of overweight children		
Bin	Lower	Upper	No	Yes	Total
1	[a]	8	469	221	690
2	8	[a]	607	148	755
Total			1076	369	1445

Each bin is computed as Lower <= physicalactivities/wk <Upper
a. Unbounded

In the output sheets the above table is given. It shows the high risk cut-offs for overweight children of the four predicting factors. E.g., in 1,142 families scoring under 14 units of (1) fruit/vegetable per week, are put into bin 1 and 303 scoring over 14 units per week, are put into bin 2. The proportion of overweight children in bin 1 is much larger than it is in bin 2: 340/1,142 = 0.298 (30 %) and 29/303 = 0.096 (10 %). Similarly high risk cut-offs are found for (2) unhealthy snacks less than 12, 12–19, and over 19 per week, (3) fastfood meals less than 2, and over 2 per week, (4) physical activities less than 8 and over 8 per week. These cut-offs will be used as meaningful recommendation limits to 11 future families.

fruit	snacks	fastfood	physical
13	11	4	5
2	5	3	9
12	23	9	0
17	9	6	5
2	3	3	3
10	8	4	3
15	9	3	6
9	5	3	8
2	5	2	7
9	13	5	0
28	3	3	9

Var 1 fruitvegetables (times per week)
Var 2 unhealthysnacks (times per week)
Var 3 fastfoodmeal (times per week)
Var 4 physicalactivities (times per week)

The saved syntax file entitled "exportoptimalbinning.sps" will now be used to compute the predicted bins of some future families. Enter the above values in a new data file, entitled, e.g., "optimalbinning2", and save in the appropriate folder in your computer. Then open up the data file "exportoptimalbinning.sps"….subsequently click File….click Open….click Data….Find the data file entitled "optimalbinning2"….click Open….click "exportoptimalbinning.sps" from the file palette at the bottom of the screen….click Run….click All.

When returning to the Data View of "optimalbinning2", we will find the underneath overview of all of the bins selected for our 11 future families.

fruit	snacks	fastfood	physical	fruit _bin	snacks _bin	fastfood _bin	physical _bin
13	11	4	5	1	1	2	1
2	5	3	9	1	1	2	2
12	23	9	0	1	3	2	1
17	9	6	5	2	1	2	1
2	3	3	3	1	1	2	1
10	8	4	3	1	1	2	1
15	9	3	6	2	1	2	1
9	5	3	8	1	1	2	2
2	5	2	7	1	1	2	1
9	13	5	0	1	2	2	1
28	3	3	9	2	1	2	2

This overview is relevant, since families in high risk bins would particularly qualify for counseling.

Conclusion

Optimal bins describe continuous predictor variables in the form of best fit categories for making predictions, and SPSS statistical software can be used to generate a syntax file, called SPS file, for predicting risk cut-offs in future families. In this way families highly at risk for overweight can be readily identified. The nodes of decision trees can be used for similar purposes (Machine learning in medicine Cookbook One, Chap. 16, Decision trees for decision analysis, pp 97–104, Springer Heidelberg Germany, 2014), but it has subgroups of cases, rather than multiple bins for a single case.

Note

More background, theoretical and mathematical information of optimal binning is given in Machine Learning in Medicine Part Three, Chap. 5, Optimal binning, pp 37–48, Springer Heidelberg Germany 2013, and Machine learning in medicine Cookbook One, Optimal binning, Chap. 19, pp 101–106, Springer Heidelberg Germany, 2014, both from the same authors.