Use of *Cutoff* and *SAS Code* node in SAS® Enterprise Miner to determine appropriate probability cutoff point for decision making with binary target models

Yogen Shah
Oklahoma State University, Stillwater, OK

ABSTRACT

This paper illustrates the effective use of *Cutoff* and *SAS Code* node in SAS® Enterprise Miner to change the default cut off value of predicted probability during decision making with binary target models. SAS®, by default uses cut-off value of 0.5 to predict a binary outcome from predicted probabilities i.e. the chance of primary outcome is same as that of a secondary outcome. A cut-off value of 0.5 is unacceptable because the observed proportion of a primary outcome in a given population can never be 50%. SAS® EM provides *Cutoff* node to adjust probability cut-off point based on model's ability to predict true positive, false positive & true negative. We need to add small snippets of SAS code under "score" section of *SAS Code* node to account for the new cut-off value in scoring dataset. This paper introduces a Technique to analyze probability cut-off using SAS® Enterprise Guide as well.

INTRODUCTION

In statistics, different kind of modeling techniques such as **Decision Tree** or **Logistic Regression** is used in situations wherein the target variable is binary. In most practical scenarios; however, it has been observed that the Primary Target proportion in a Population is never 50%; in fact the Proportion is, usually, much smaller. Such models usually predict the probability of a target to be equal to 1 or 0. SAS® then converts predicted probabilities to predicted binary responses (1's/0's) by choosing probability cut-off point i.e. if IP_1 is greater than 0.5 then we predict it as primary target.

This paper uses dataset from SAS® Data Mining shootout 2011 for illustration. Primary target variable "isAdmit" represents possibility of admit in the hospital following any storm.
IsAdmit = 1 → Admit in the hospital
IsAdmit = 0 →No admits in the hospital

Observations from the dataset reveal that the target variable shows a primary event occurring with a probability of 26.28%. In such situations, regardless of the model you build for predicting the primary event of target occurrence, the Primary Target Proportion should not be 50%. In other words (referring to the earlier example), the probabilities of admits and no admits are not identical. Hence, you would need to find out a more accurate **probability cutoff value,** different from default predicted probability cutoff value of 0.5.

SAS® Enterprise Miner provides "*Cutoff*" node to analyze the effect of various cut-off probabilities on true positive, false positive and true negative predictions. You can change the property of this node to change cutoff probability and run the flow again to get better binary predictions. You would still have to write small snippets of SAS code to apply the new cut-off value while scoring.

OVERVIEW OF THE CUTOFF NODE

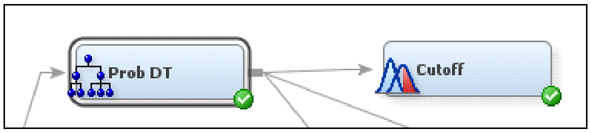
You can find Cutoff node under the Assess category in the SAS® data mining process of Sample, Explore, Modify, Model, and Assess (SEMMA).

The node provides tabular and graphical information to assist users in determining appropriate probability cutoff point(s) for decision making with binary target models. An appropriate use of the *Cutoff* node can help minimize the risk of generating false positives and false negatives.

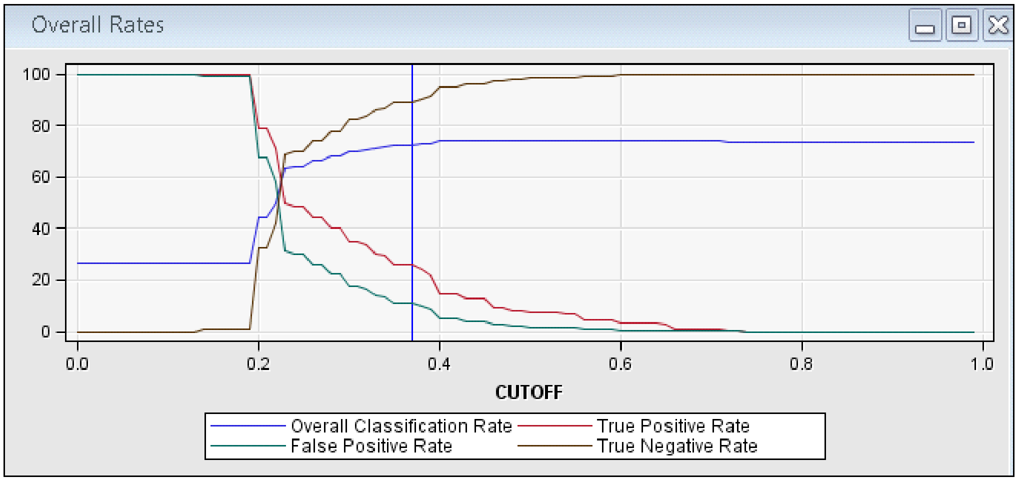
You need to run the node at least twice. In the first run, you would obtain all the plots and tables to narrow down to the best probability cutoff point. In subsequent runs, you would change the values of the Cutoff Method and Cutoff User Input properties, while customizing the plots, until an optimal cutoff value is obtained. The goal is to choose such a cutoff value that can provide the best balance between True Positive and False Positive predictions.

PROCEDURE FOR ENTERPRISE MINER

- Connect Cutoff node to your trained model.

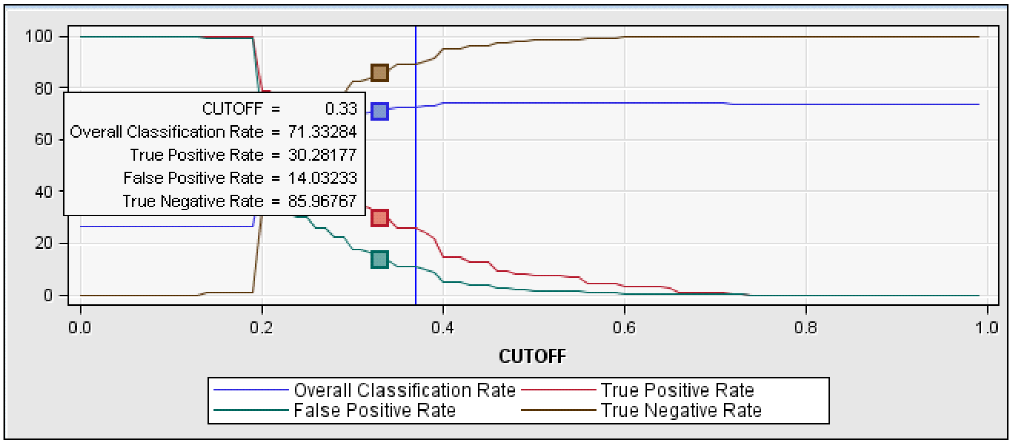


- Run the node. Go to Results → Overall Rates chart.



This plot shows that how overall classification, sensitivity, specificity and prediction accuracy changes with change in probability cut-off point

- You can click on any point on graph to find value of TP, FP, TN rate at particular cut-off point and choose best value based on problem requirement.



  - You will have to decide on the cutoff based on the basis of your business objective, level of impact and the trade-off between sensitivity, specificity and false positivity values.
  - You should select cutoff value such that you can improve sensitivity of the model by restricting the false positive rate to the lowest minimum value
  - You need to be very careful in selecting cut-off value because selecting a very low cut-off may give you a high True positive rate but the low cut-off will also increase the false positive rate, impacting your model badly.

- You may even use table instead of graph to get better insight on the effect of the various probability cutoff. Click view →table in the result window.

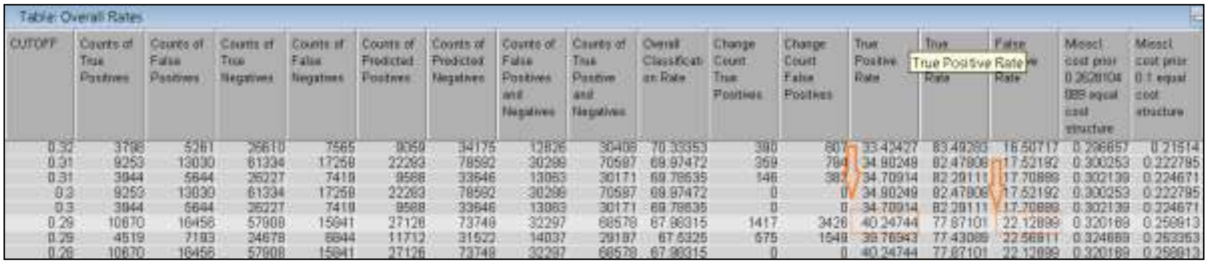
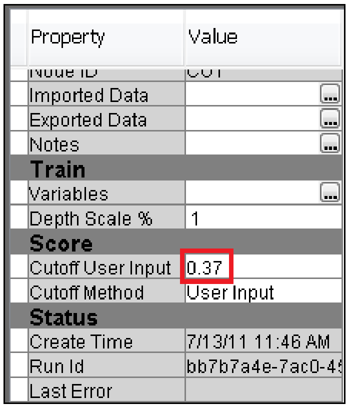


Table: Overall Rates

CUTOFF	Counts of True Positives	Counts of False Positives	Counts of True Negatives	Counts of False Negatives	Counts of Predicted Positives	Counts of Predicted Negatives	Counts of False Positives and Negatives	Counts of True Positive and Negatives	Overall Classification Rate	Change Count True Positives	Change Count False Positives	True Positive Rate	True Negative Rate	False Positive Rate	Miscl cost prior 0.2628104 089 equal cost structure	Miscl cost prior 0.1 equal cost structure
0.32	3798	5261	26610	7565	9059	34175	12826	30408	70.33353	380	807	33.42427	83.49283	16.50717	0.298657	0.21514
0.31	9253	13030	61334	17258	22283	78592	30288	70587	69.97472	359	784	34.90249	82.47808	17.52192	0.300253	0.222795
0.31	3944	5644	26227	7419	9588	33646	13063	30171	69.78635	146	387	34.70914	82.29111	17.70889	0.302139	0.224671
0.3	9253	13030	61334	17258	22283	78592	30288	70587	69.97472	0	0	34.90249	82.47808	17.52192	0.300253	0.222795
0.3	3944	5644	26227	7419	9588	33646	13063	30171	69.78635	0	0	34.70914	82.29111	17.70889	0.302139	0.224671
0.29	10670	16456	57908	15841	27126	73748	32297	68578	67.98315	1417	3426	40.24744	77.87101	22.12899	0.320189	0.258913
0.29	4519	7193	24678	6844	11712	31522	14037	29197	67.6325	675	1549	39.76943	77.43089	22.56911	0.324689	0.253353
0.28	10670	16456	57908	15841	27126	73748	32297	68578	67.98315	0	0	40.24744	77.87101	22.12899	0.320189	0.258913

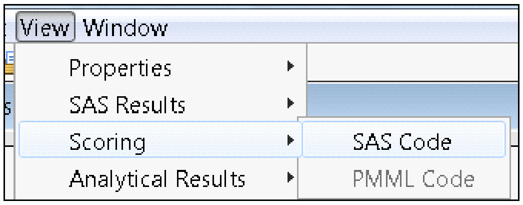
  - The tabular view will allow you to analyze minute change in probability cutoff value and select value up to two decimal places (for e.g. 0.29). Notice that, as you try to increase true positive rate, false positive rate also increases besides the decrease in true negative rate.

- As you decide on changing your cutoff value, click on the cutoff node and change the Cutoff User Input value to the value you desire in the property panel. As an example, I changed the cutoff to 0.37 from 0.5 after analyzing the plot and table generated above steps.

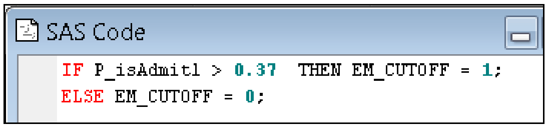


Property	Value
Node ID	CUT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Depth Scale %	1
Score	
Cutoff User Input	0.37
Cutoff Method	User Input
Status	
Create Time	7/13/11 11:46 AM
Run Id	bb7b7a4e-7ac0-4!
Last Error	

- You need to run the *Cutoff* node again to apply new cutoff value while predicting binary decision. Go to view →scoring → SAS Code under result window



- You can notice that how new cut-off value is applied & new variable for classification of predicted target “isAdmit” to 1 or 0 is generated.

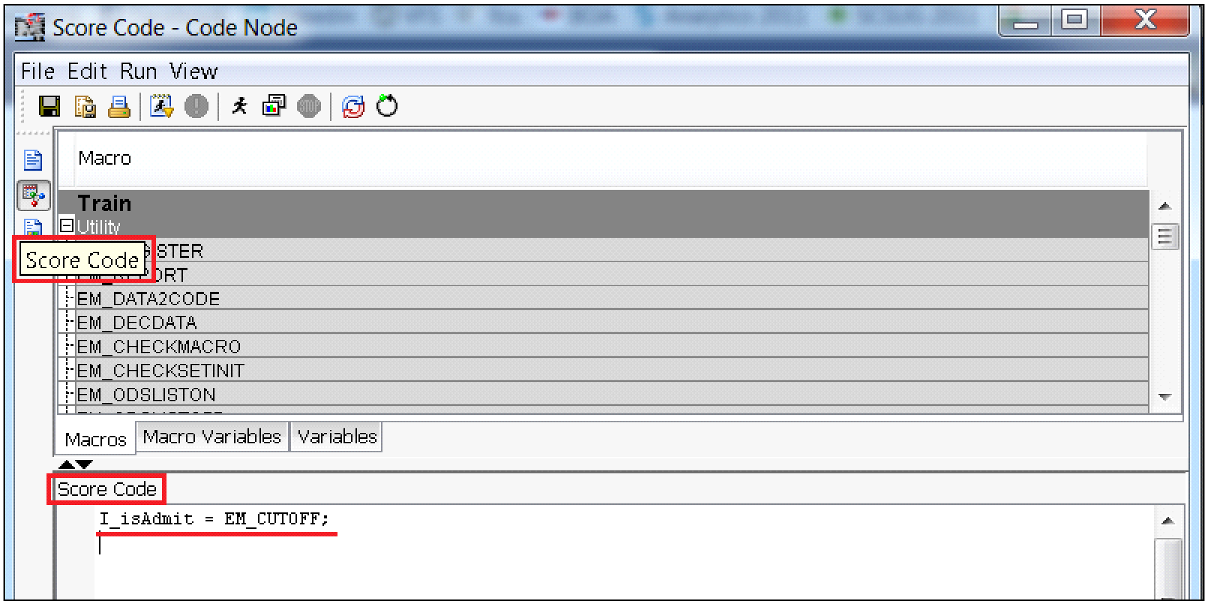


```
IF P_isAdmit1 > 0.37  THEN EM_CUTOFF = 1;
ELSE EM_CUTOFF = 0;
```

Note: P_isAdmit is equivalent to P_<*your target*>

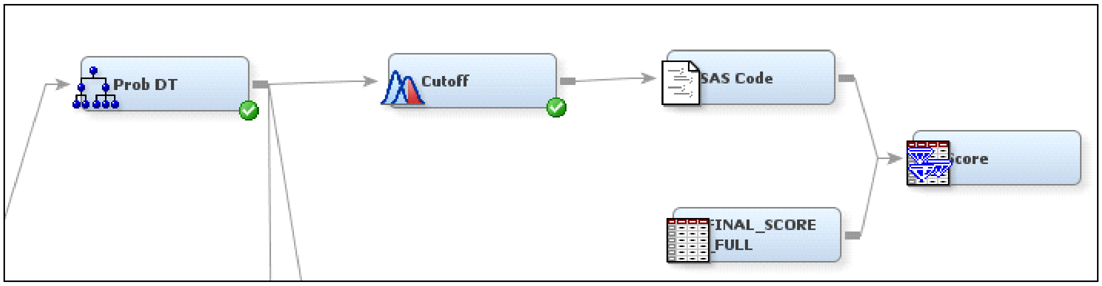
- Above updated cutoff will not be applied while scoring dataset. You need to connect *SAS Code* node to *Cutoff* node and write simple assignment statement in score code tab of *SAS Code* node

I_<*your target*> = EM_CUTOFF;



```
I_isAdmit = EM_CUTOFF;
```

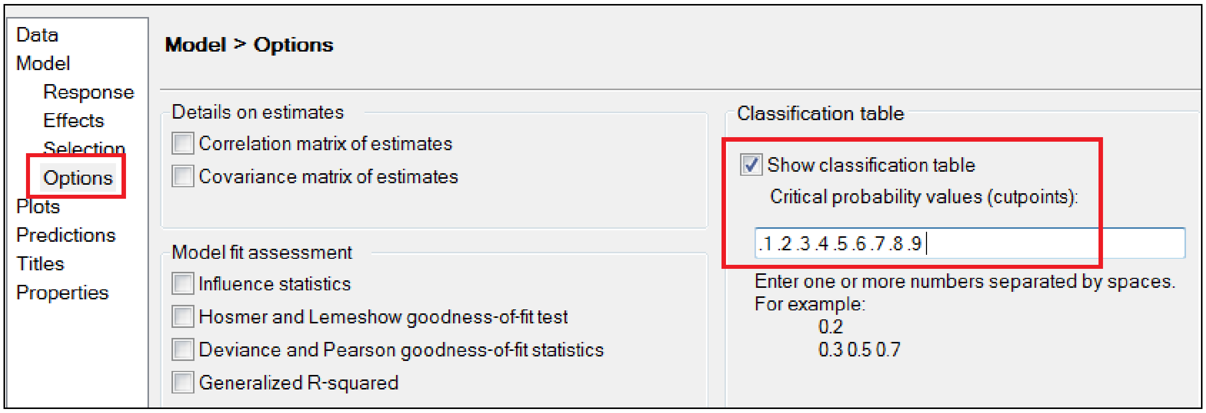
- Save and run the node. Connect *Score* node to *SAS Code* node & your scored data set will use modified probability cutoff value.



PROCEDURE FOR SAS® ENTERPRISE GUIDE

- SAS® Enterprise Guide automatically creates classification table for various possible probability cutoff value similar to EM Cutoff node table

- The following screenshot has been taken from a Logistic Regression model. Under Model Option, you can check the check box for “Show Classification table” to display table in output result. You can optionally specify custom cut points, if you are not interested in entire table to be displayed.



Data
Model
Response
Effects
Selection
Options
Plots
Predictions
Titles
Properties

Model > Options

Details on estimates

□ Correlation matrix of estimates
□ Covariance matrix of estimates

Model fit assessment

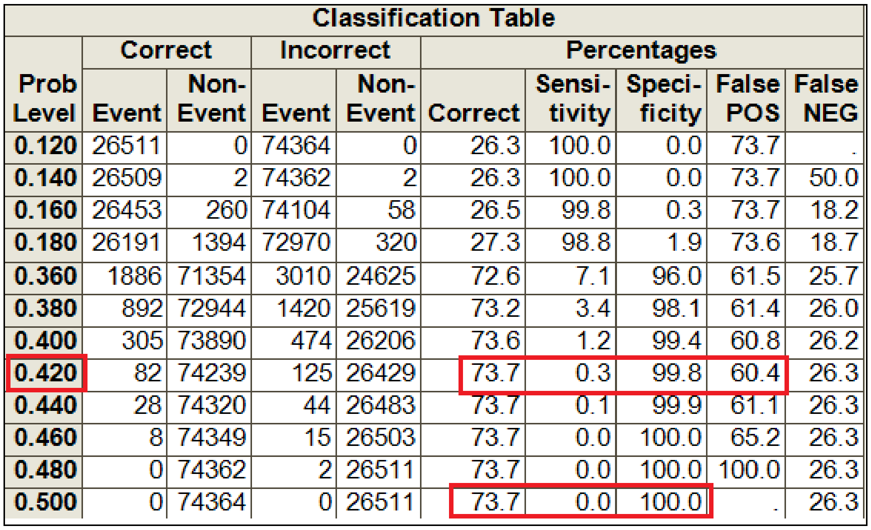
□ Influence statistics
□ Hosmer and Lemeshow goodness-of-fit test
□ Deviance and Pearson goodness-of-fit statistics
□ Generalized R-squared

Classification table

☑ Show classification table
Critical probability values (cutpoints):
.1 .2 .3 .4 .5 .6 .7 .8 .9

Enter one or more numbers separated by spaces.
For example:
0.2
0.3 0.5 0.7

- The result window will provide following table for analysis. You will notice that probability cutoff of 0.5 is not suitable in given example. Looking at all other values, cut-off value of 0.42 seems to be perfect with lowest false positive & highest true positive.

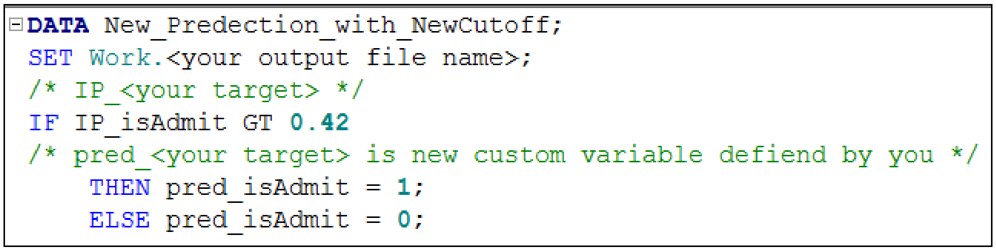


Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.120	26511	0	74364	0	26.3	100.0	0.0	73.7	.
0.140	26509	2	74362	2	26.3	100.0	0.0	73.7	50.0
0.160	26453	260	74104	58	26.5	99.8	0.3	73.7	18.2
0.180	26191	1394	72970	320	27.3	98.8	1.9	73.6	18.7
0.360	1886	71354	3010	24625	72.6	7.1	96.0	61.5	25.7
0.380	892	72944	1420	25619	73.2	3.4	98.1	61.4	26.0
0.400	305	73890	474	26206	73.6	1.2	99.4	60.8	26.2
0.420	82	74239	125	26429	73.7	0.3	99.8	60.4	26.3
0.440	28	74320	44	26483	73.7	0.1	99.9	61.1	26.3
0.460	8	74349	15	26503	73.7	0.0	100.0	65.2	26.3
0.480	0	74362	2	26511	73.7	0.0	100.0	100.0	26.3
0.500	0	74364	0	26511	73.7	0.0	100.0	.	26.3

Please note that above screen shot has been taken from an initially trained model. Further tuning should be required before selecting cutoff point.

- To change default probability cutoff , we need to run small SAS program on output data set



```
DATA New_Predection_with_NewCutoff;
 SET Work.<your output file name>;
 /* IP_<your target> */
 IF IP_isAdmit GT 0.42
 /* pred_<your target> is new custom variable defiend by you */
     THEN pred_isAdmit = 1;
     ELSE pred_isAdmit = 0;
```

CONCLUSION

SAS® Enterprise Miner and SAS® Enterprise guide allows enough flexibility to the users to change SAS® default probability cutoff value using *Cutoff* node and *SAS code* so as to obtain more accurate decision type predictions.

REFERENCES

[1] "Logistic regression", "http://en.wikipedia.org/wiki/Logistic_regression"

[2] SAS® Enterprise Miner Help

[3] Course material from Oklahoma State University's "SAS and OSU Data Mining Certificate Program"

ACKNOWLEDGEMENTS

Thanks for IICH and Analytics2011 team for allowing me to work on this dataset.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yogen Shah, Oklahoma State University, Stillwater, OK, Email: yogen.shah@live.com

Yogen Shah is a master's student in Management Information Systems at Oklahoma State University. He is SAS® Certified Base Programmer and Predictive Modeler with 4 years of industry experience in IT consulting, Database Marketing and Statistical Data Mining using BI tools & techniques. Yogen is recognized for strong technical competencies and, implementation of new technologies, for optimal results.