Paper 8920-2016

Reducing and screening redundant variables in logistic regression models using SAS/STAT® software and SAS ENTERPRISE MINER®

Xinghe Lu, AmeriHealth Caritas Family of Companies, Philadelphia, PA

ABSTRACT

Logistic regression is one of the popular regression models in statistics. It measures the relationship between categorical depend variable and independent variable(s) and predicts the likelihood of having the event associated with outcome variable. Variable reduction and screening are the techniques that reduce the redundant independent variables and filter out the variable(s) with less predictive power. Variable reduction and screening are important and critical for building a reasonable statistical model, especially when dealing with hundreds or thousands of independent variables. This is because they can 1) decrease the model estimation time, 2) avoid occurrence of non-convergence in the model, 3) stabilize the parameter estimate, and 4) decrease the chance of over-fitting.

This paper will explore three variable reduction approaches: Effect-selection method, single logistic regressions method and variable clustering method using SAS/STAT® software. It also will review three nodes from SAS ENTERPRISE MINER®: variable selection node, variable clustering node and decision tree node.

Key words: logistic regression, CHAID, empirical logit plot, hoeffding correlation, Spearman/Pearson correlation. R squares

INTRODUCTION

Reducing and screening redundant variables is a necessary step before building a formal statistical model, especially when fitting a logistic regression model with hundreds or even thousands independent variables on thousands of millions observations. Too many variables can reduce model efficiency. It also can destabilize the parameter estimates when input variables are highly correlated with each other. When input variable has weak association with the target variable, it will decrease the predictive power. Finally, it is much more difficult to have explainable model when having so many variables and their interactions effect in the model. This paper will explore various variable reduction approaches including single logistic regressions, effect-selection method and variable clustering using SAS/STAT®. It also will present hands-on examples with three nodes from SAS ENTERPRISE MINER®: variable selection node, variable clustering node and decision tree node.

EXAMPE DATA

The data set used in paper is called “campaign” (N=5,000), which is used to determine who is likely to be the donors in a non-profit organization campaign and to target them for donation solicitation. The data contains 1 binary target variable and 33 predictors, including 10 dummy variables converted from 5 categorical variables originally.

The example data in the paper is a simulated data and it does not refer to any organization/company.

SAS/STAT®

EFFECT-SELECTION METHOD

Effect-selection method is the one of most common used methods for variable selection. It can be implemented by specifying the SELECTION option in the MODEL statement of PROC LOGISTIC procedure. For example,

```
proc logistic data= campaign;
     model donated (event='1')= &var_list /selection=stepwise slentry=0.3 slstay=0.35;
run;
```

&var_list includes all input variables. Selection=stepwise indicates that stepwise selection method was used in the variable selection. Slentry=0.3 means a significant level of 0.3 is required to allow a variable enter into the model; while slstay=0.35 means a significant level of 0.35 is required for a variable to stay in the model. There are 4 other options available for selection statement-Backward, Forward, None and score.

SINGLE LOGISTIC REGRESSION METHOD

Effect-selection method doesn't work well in the model with thousands of input variables, as it is not very efficient and requires a lot of computational time. However, running single logistic regression between each predictor and target variable before applying effect-selection method can fix the problem. The reason is that if a predictor is not significant at a threshold of P-value, such as 0.25, in the single logistic regression model, it is almost impossible to be significant at the level of 0.05 when interacting with other effects in the multiple logistic regression model. These non-significant predictors will be screened out and the rest of the predictors can be applied into the logistic regression with or without effect-selection method. For instance,

```
proc logistic data= campaign;
     model donated (event='1')= Pct_attribute_1;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Pct_attribute_1	1	-0.0051	0.00791	0.4154	0.5193

In the single logistic regression model, P-value of Pct_attribute_1 is 0.5193 which is greater than 0.25. So this predictor should not be shown in the logistic regression model.

VARIABLE CLUSTERING METHOD

There are two methods which can be used for segmenting variables: principal components analysis (PCA) and variable clustering analysis (VCA).

Principal components analysis (PCA) is used to transform redundant variables into a set of principal components (PC), which are linear combinations of variables, to explain the total variability among the original variables. The numbers of PC are determined by total proportion explained by total variation from covariance matrix and these PC can be used to substitute the original variable in the analysis. PCA can be implemented by PROC PRINCOMP or PROC FACTOR. PROC PRINCOMP emphasizes more of the linear combinations of the variables to form the components, while PROC FACTOR expresses variables as linear combinations of the components in the output.

Variable clustering analysis is another technique for reducing dimension of variables. Unlike PCA, variable clustering analysis groups correlated subsets of the variables; selects variables with minimal resulting collinearity and chooses a "best" variable or multiple "best" variables from each cluster as the new input variables. PROC VARCLUS is the procedure used for variable clustering analysis.

In this paper, only variable clustering analysis will be focused on.

```
proc varclus data = campaign maxeigen = 0.7
     outtree = fortree
     short;
var &cluster_var;
run;
```

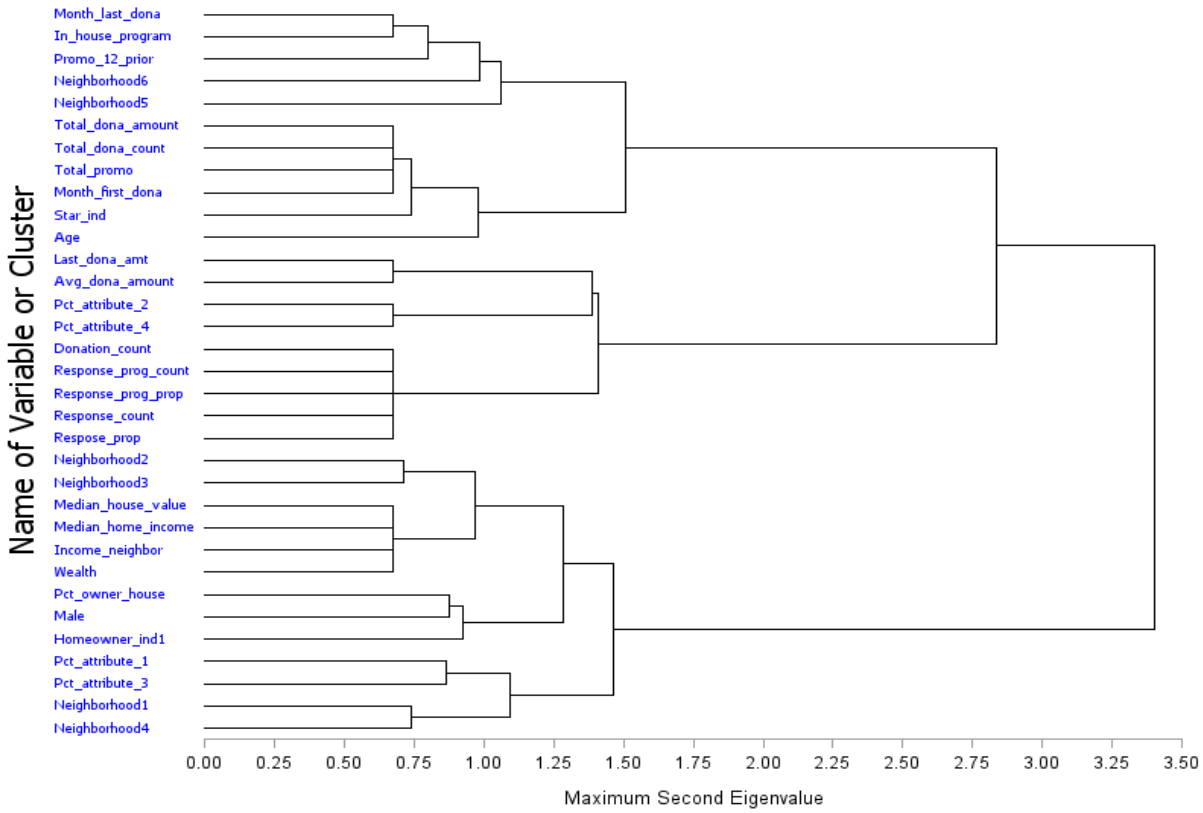


Figure1. Dendrogram from variable clustering process

Figure 1. is the output dendrogram from PROC VARCLUS. All inputs variables have been grouped by 20 clusters. In order to reduce the variables, representative variable needs to picked from each cluster based on smallest R-Square Ratio(1-R^2 ratio) value. For example,

Cluster	Variable	OwnCluster	NextClosest	RSquareRatio
Cluster 1	Donation_count	0.6795	0.106	0.3585
	Response_prog_count	0.8751	0.1133	0.1408
	Response_prog_prop	0.8505	0.0989	0.1659
	Response_count	0.9067	0.1211	0.1061
	Respose_prop	0.9146	0.1333	0.0985

Cluster1 is composed of 5 independent variables: Donation_count, Response_prog_count, Response_prog_prop, Response_count and Respose_prop. The variable Respose_prop has the smallest RSquareRatio value (0.0985) among the 5 variables, thus, it will be picked as the representative variable of cluster1. This is because, RSquareRatio= 1-R^2 ratio= (1-$R^2_{\text{own cluster}}$)/ (1-$R^2_{\text{next closet}}$). An ideal representative variable should have strong correlation within its own cluster and have weak correlation between other clusters. Sometimes subject-matter consideration also plays important role to make the decision. After variable clustering, some further variable screening may be needed as well to minimize the variable size. Plot of rank of Pearson/Spearman vs. rank of Hoeffding's D is a good tool to screen variables. For example,

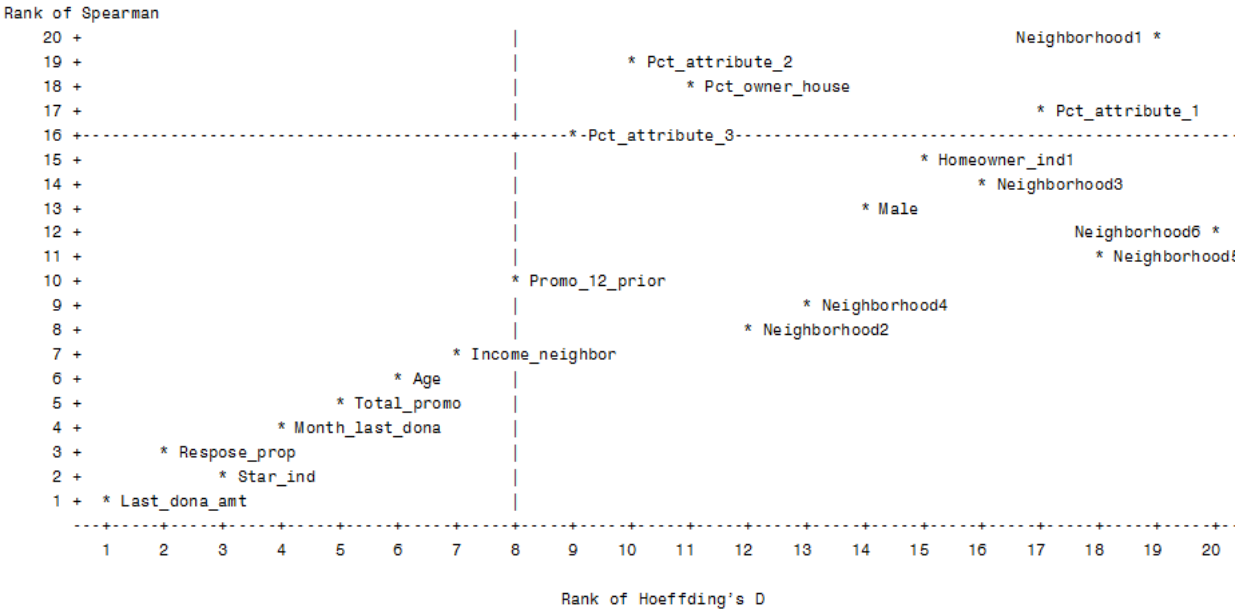


Figure2. Plot of Rank of Spearman correlation vs. Rank of Hoeffding's D statistics

Ranking of spearman is the rank of Spearman correlation of each variable against target variable. Compared to pearson correlation, spearman correlation does not require linearity assumption and is less sensitive to outliers. Hoeffding's D statistics is a nonparametric measure that detects more general departures from independence. The statistic approximates a weighted sum over observations of chi-square statistics for two-by-two classification tables. From Figure2, Variable-Pct_owner_house looks suspicious, as it has a rank of 18 on spearman correlation and a rank of 11 on Hoeffding's D statistics. Empirical logit plot is one of the diagnosis tools that can be used for further screening.

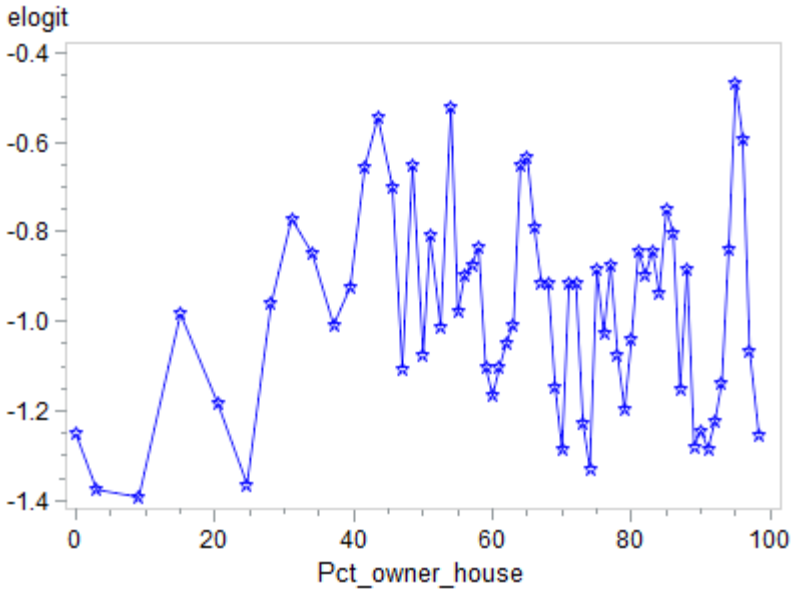


Figure3. Plot of Empirical logit vs. Pct_owner_house variable

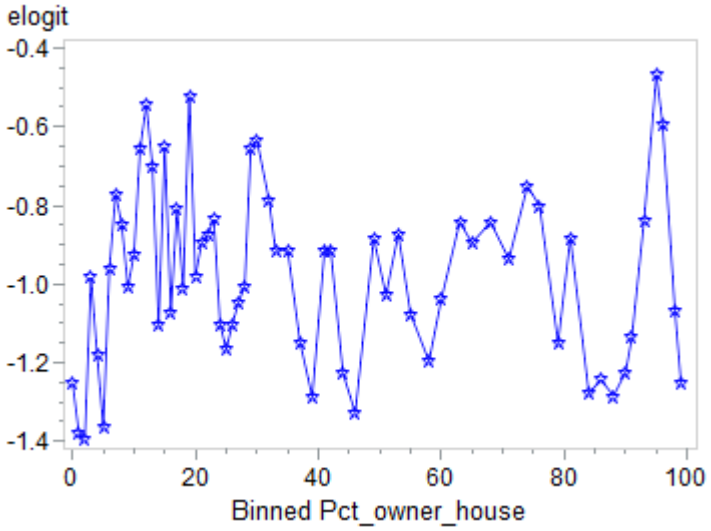


Figure4. Plot of Empirical logit vs. binned Pct_owner_house variable

Figure3 is the plot of Empirical logit against Pct_owner_house variable and Figure4 is the plot of Empirical logit against binned Pct_owner_house variable. Neither plot shows strong linear relationship against empirical logit, thus, Pct_owner_house variable should be removed from the model.

SAS ENTERPRISE MINER®

SAS Enterprise Miner® (EM) is a powerful application developed by SAS® company to conduct a variety of statistical and mathematical analysis to critical business or research issues, such as fraud detection, customer retention and attrition, database marketing, market segmentation, risk analysis, affinity analysis, customer satisfaction, bankruptcy prediction, and portfolio analysis. Enterprise miner is a visual programming tool with user friendly interface. People who use enterprise miner don't require knowledge of SAS programming and maybe have very little statistical expertise as it is as simple as selection icons or dragging a node from EM tool palette or menu bar. Yet, an expert statistician can still adjust the default settings and run their own specifications later. The tools for statistical analysis in Enterprise Miner are called Nodes. In this paper, only variable selection node, variable clustering node, and decision tree node will be introduced briefly. All examples demonstrated here are based on Enterprise Miner®13.1.

VARIABLE SELECTION NODE 

Property	Value
General	
Node ID	Varsel2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Max Class Level	100
Max Missing Percentage	50
Target Model	Chi-Square
Manual Selector	
Rejects Unused Input	Yes
Bypass Options	
Variable	None
Role	Input
Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.005
Stop R-Square	5.0E-4
Use AOV16 Variables	No
Use Group Variables	Yes
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default

The screenshot above is the property panel of variable selection node. Through property panel, setting can be adjusted, which is similar as options in the procedures.

There are three major variable selections methods available in “Target Model” filed: R-square, Chi-square and R & Chi-square combined. R-square is suitable for continuous target variable, while Chi-square is suitable for binary or discrete target variable. If both the R-square and Chi-square selection is chosen, the system will choose based on the type of target variable. If target variable is continuous, then only R-square is applied. If the target variable is categorical variable, both R-square and Chi-square will be applied, which is too restrict. Since target variable in the example is binary, Chi-square criterion was applied in the “Target Model” as shown above.

Figure 5 shows the list of final variables after screening with relative importance in descending order.

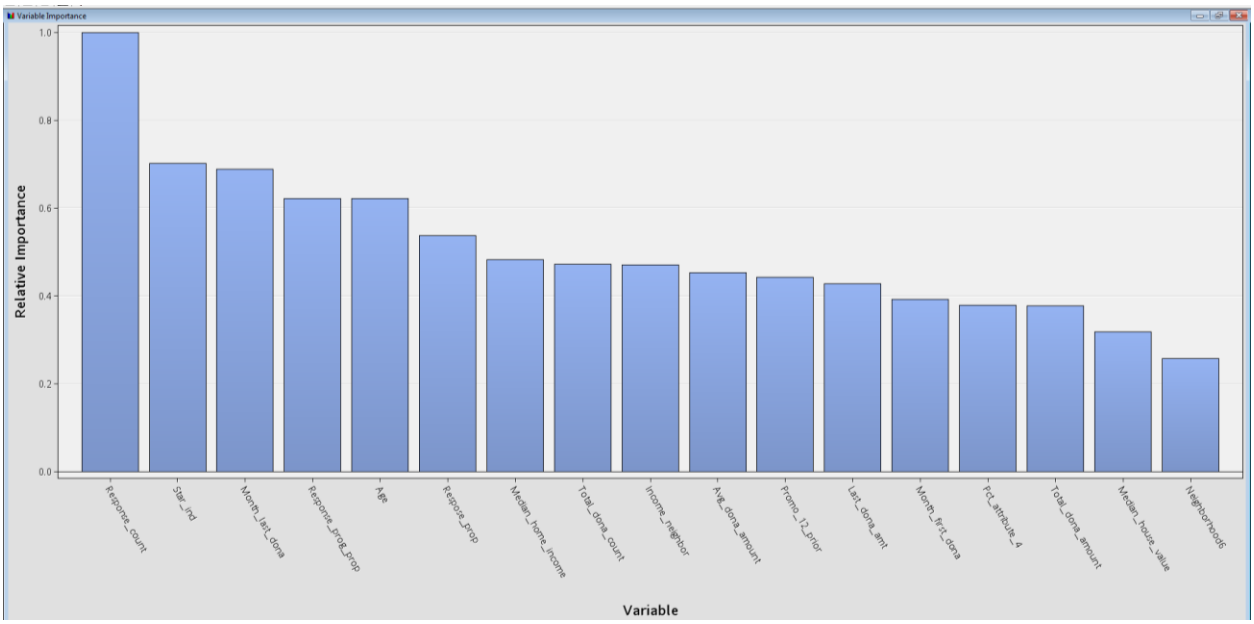
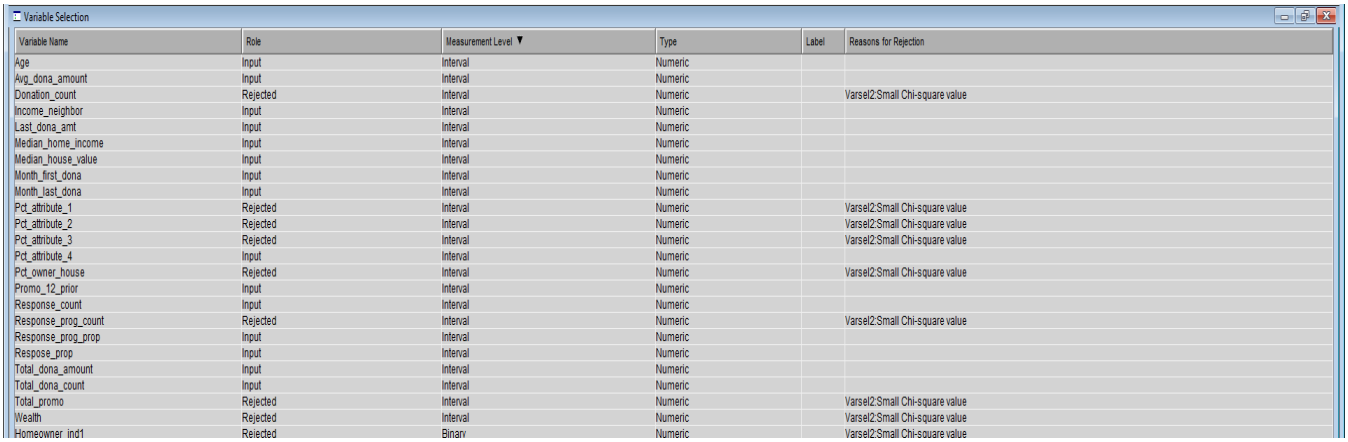


Figure5. Variable relative importance chart

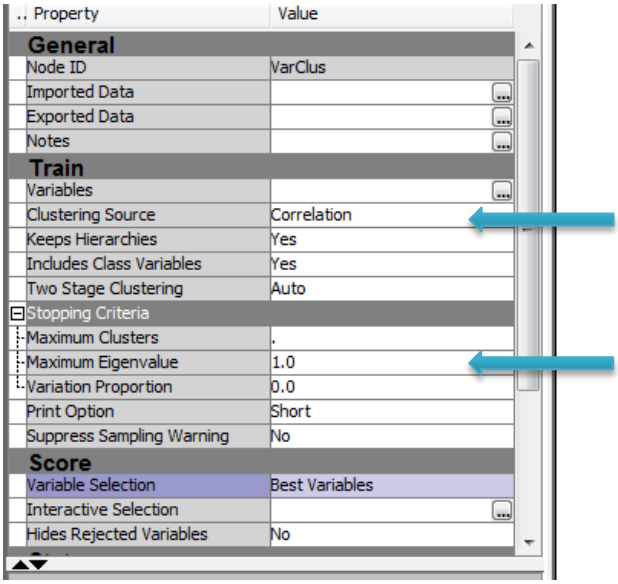
Figure 6 lists all input variables with final decisions (selected or rejected). It also includes the reasons for rejection.


Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
Age	Input	Interval	Numeric		
Avg_dona_amount	Input	Interval	Numeric		
Donation_count	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Income_neighbor	Input	Interval	Numeric		
Last_dona_amt	Input	Interval	Numeric		
Median_home_income	Input	Interval	Numeric		
Median_house_value	Input	Interval	Numeric		
Month_first_dona	Input	Interval	Numeric		
Month_last_dona	Input	Interval	Numeric		
Pct_attribute_1	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Pct_attribute_2	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Pct_attribute_3	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Pct_attribute_4	Input	Interval	Numeric		
Pct_owner_house	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Promo_12_prior	Input	Interval	Numeric		
Response_count	Input	Interval	Numeric		
Response_prog_count	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Response_prog_prop	Input	Interval	Numeric		
Respose_prop	Input	Interval	Numeric		
Total_dona_amount	Input	Interval	Numeric		
Total_dona_count	Input	Interval	Numeric		
Total_promo	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Wealth	Rejected	Interval	Numeric		Varsel2:Small Chi-square value
Homeowner_ind1	Rejected	Binary	Numeric		Varsel2:Small Chi-square value

Figure6. Summary result table of variable selection node

VARIABLE CLUSTERING NODE 



.. Property	Value
General	
Node ID	VarClus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Clustering Source	Correlation
Keeps Hierarchies	Yes
Includes Class Variables	Yes
Two Stage Clustering	Auto
Stopping Criteria	
Maximum Clusters	.
Maximum Eigenvalue	1.0
Variation Proportion	0.0
Print Option	Short
Suppress Sampling Warning	No
Score	
Variable Selection	Best Variables
Interactive Selection	
Hides Rejected Variables	No

Variable clustering node is another tool that can be used in variable selection. It is analogous to PROC VARCLUS procedure. In its property panel, users can choose either correlation or covariance in the setting of “Cluster Source”. “Maximum Eigenvalue” specifies the largest permissible threshold for the second eigenvalue of each cluster. The variable cluster node stops splitting when the second eigenvalue of a cluster exceeds that threshold. In the example, correlation is set in “Cluster Source” and value 1 is set in “Maximum Eigenvalue”. Please note, “Maximum Eigenvalue” must be greater than or equal to 1 in EM.

Figure7 and 8 are the selected outputs after running variable clustering node. The result from variable clustering node had different result from PROC VARCLUS. Instead of having 20 clusters from PROC

VARCLUS, variable clustering node generated 17 clusters. This is because, in PROC VARCLUS, 0.7 was used in “maxeigen”. If maxeigen is equal to 1, both methods will produce identical results.

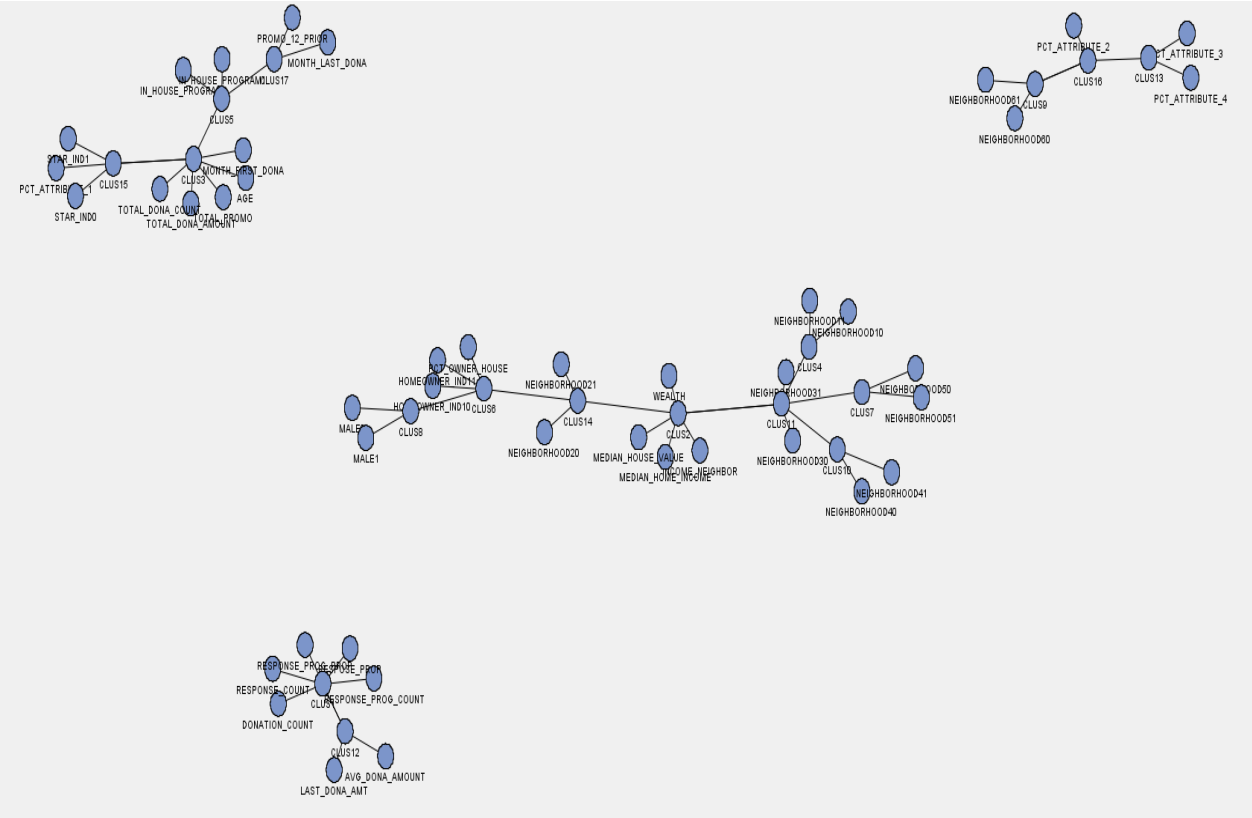


Figure7. Cluster Plot

Cluster	Variable	Label	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	1-R2 Ratio	Variable Selected
CLUS1	RESPOSE_PROP		0.920362	CLUS12	0.106316	Variable	0.089112	YES
CLUS10	NEIGHBORHOOD40	Neighborhood4=0	1	CLUS11	0.078209	Variable	0	YES
CLUS11	NEIGHBORHOOD30	Neighborhood3=0	1	CLUS2	0.129019	Variable	0	YES
CLUS12	LAST_DONA_AMT		0.904033	CLUS1	0.092228	Variable	0.105717	YES
CLUS13	PCT_ATTRIBUTE_3		0.763311	CLUS9	0.058837	Variable	0.251486	YES
CLUS14	NEIGHBORHOOD20	Neighborhood2=0	1	CLUS2	0.106882	Variable	0	YES
CLUS15	STAR_IND0	Star_ind=0	0.997637	CLUS3	0.25737	Variable	0.003182	YES
CLUS16	PCT_ATTRIBUTE_2		1	CLUS9	0.126927	Variable	0	YES
CLUS17	PROMO_12_PRIOR		0.680423	CLUS5	0.055014	Variable	0.338182	YES
CLUS2	INCOME_NEIGHBOR		0.83876	CLUS11	0.080793	Variable	0.175412	YES
CLUS3	TOTAL_PROMO		0.812622	CLUS5	0.242826	Variable	0.247747	YES
CLUS4	NEIGHBORHOOD10	Neighborhood1=0	1	CLUS11	0.084188	Variable	0	YES
CLUS5	IN_HOUSE_PROGRAM0	In_house_program=0	1	CLUS3	0.137253	Variable	0	YES
CLUS6	HOMEOWNER_IND10	Homeowner_ind1=0	0.989095	CLUS14	0.018987	Variable	0.011116	YES
CLUS7	NEIGHBORHOOD50	Neighborhood5=0	1	CLUS11	0.046926	Variable	0	YES
CLUS8	MALE0	Male=0	1	CLUS6	0.00247	Variable	0	YES
CLUS9	NEIGHBORHOOD60	Neighborhood6=0	1	CLUS16	0.126927	Variable	0	YES

Figure8. Summary table from variable clustering node

DECISION TREE NODE 

Decision tree node is the last node that will be covered in this paper. Decision tree node is a simple, but powerful form of multiple variable analyses, which provides a unique capability to supplement, complement and substitute for traditional statistical model. Decision tree is a flexible tool that can handle numerical and categorical variable, missing value and non-missing value, correlated and uncorrelated data. It doesn't require the assumption of data distribution or checking the multi-collinearity. Decision tree splits the data into subgroups that are as homogeneous as possible with respect to the target variable through a recursive process.

At first, it takes the original data set as one segment, then partitions the whole segment into two or more subgroups by applying a series of criteria, such as Chi-squares, variance or Gini index etc. The whole process will last until no more partitioning is available. Eventually a tree dendrogram will be produced based on the criterion used in the splitting. A tree developing using Chi-Square criterion is called CHAID (Chi-squared Automatic Interaction Detection).

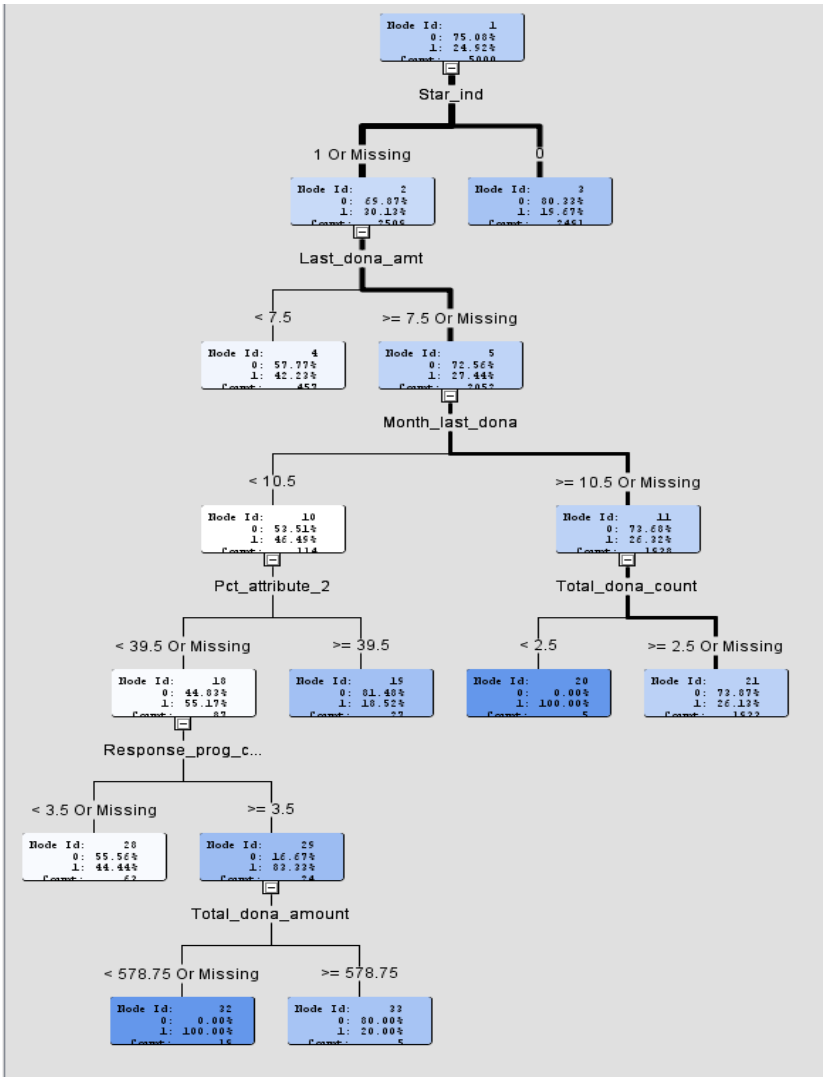


Figure9. Tree dendrogram from Decision Tree Node

CONCLUSION

Redundant variables reduction and screening is an important step before actually building a statistical model. In this paper, several statistical methods were introduced by either SAS/STAT® module or SAS/ENTERPRISE MINER®. In this paper, only three commonly used nodes/tools were selected from EM. However, besides these nodes, there are other nodes that can be used in variable selection. Some statisticians recommend applying data manipulations steps, such as imputation or transformation, before reducing redundant variables.

REFERENCES

Matignon, Randall(2007), Data Mining Using SAS Enterprise Miner, http://www.lexjansen.com/wuss/2007/analyticsstatistics/anl_matignon_datamining.pdf

Nelson, Bryan, D., Variable Reduction for Modeling using PROC VARCLUS, SUGI26, Paper 261-26

Predictive Modeling Using Logistic Regression, Course Note, 2008, Cary, NC: SAS Institute Inc.

Sarma, Kattamura, S.(2013), Predictive Modeling with SAS® Enterpriprise Miner™, Practical Solutions for Business Applications, Second Edition. Cary, NC: SAS Institute Inc.

SAS OnlineDoc 9.2, SAS/STAT(r) User's Guide, Second Edition.

ACKNOWLEDGMENTS

I would like to acknowledge Jim Jones, VP Healthcare Analytics, AmeriHealth Caritas, and Wanzhen Gao, Director, Health Care Analytics, AmeriHealth Caritas for their support and encouragement.

Amerihealth Caritas is the nation's leader in health care solutions, with more than 30 years of experience managing care for individuals and families in publicly funded programs.

CONTACT INFORMATION

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

Xinghe Lu
AmeriHealth Caritas Family of Companies
3rd Floor, 200 Stevens Drive, Philadelphia, PA 19113.
E-mail: xlu@amerihealthcaritas.com
Web: http://www.amerihealthcaritas.com/