Paper 111-2009

Predictive Models Based on Reduced Input Space That Uses Rejected Variables

Taiyeong Lee, David Duling, and Dominique Latour
SAS Institute Inc., Cary, NC

ABSTRACT

Because the number of variables is often tremendous in data mining applications, variable selection or dimension reduction is essential to produce models with acceptable accuracy and generalization. After applying variable selection methods, data miners often consider only selected variables for their tasks. However, the input space represented by the rejected variables might still contain potential for positive contribution to the model. This paper demonstrates the use of projection methods to combine both the selected variable space and the rejected variable space. The new input space, which is made by adding projections of the rejected variables to the set of selected variables, reduces the loss of input variable information while keeping interpretability of important individual variables. The proposed method provides an effective means for analyzing the additional information potential that is hidden in the rejected terms. We also introduce a SAS® Enterprise Miner™ extension node for random projection. This extension node enables easy creation of the new reduced input space.

INTRODUCTION

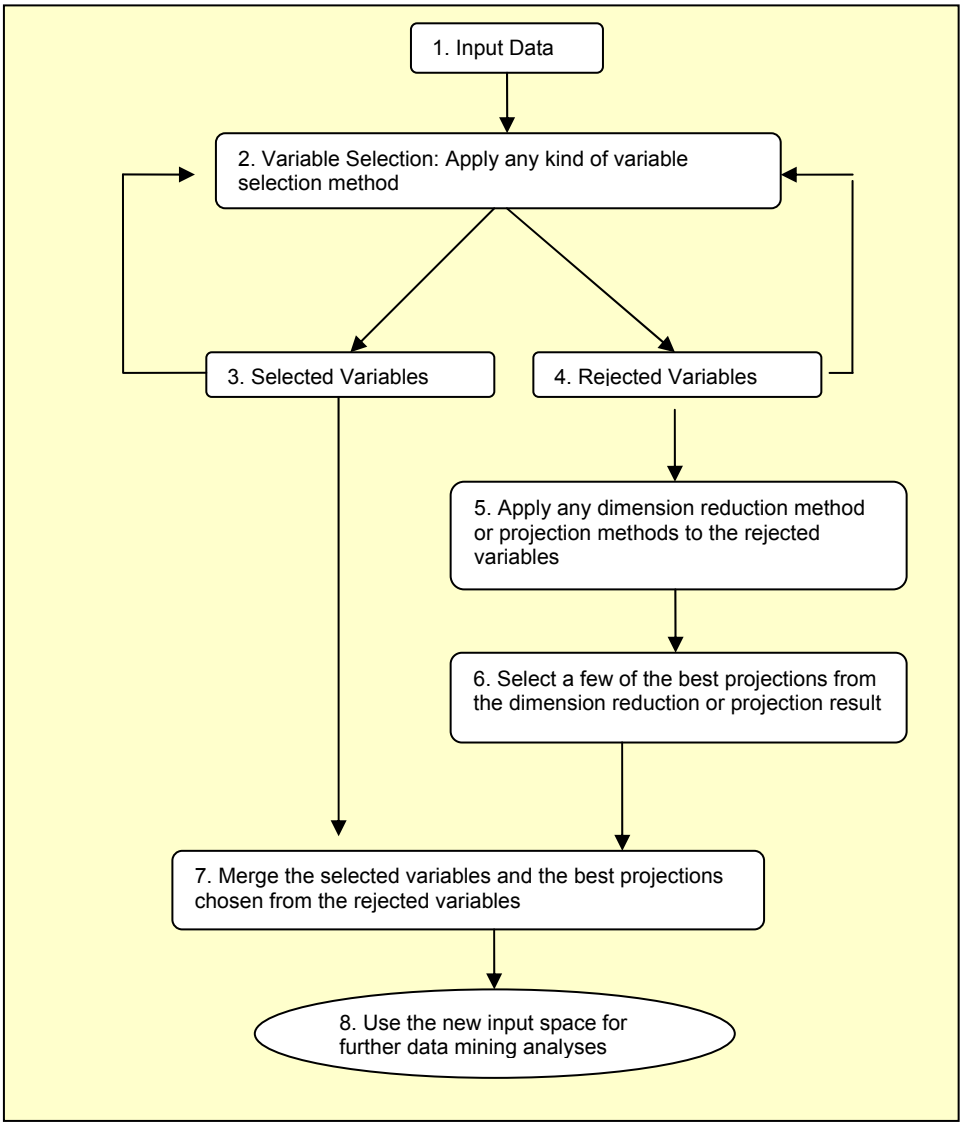
Dimension reduction of input space is essential to performing data mining tasks well. There are two categories of dimension reduction methods: direct selection of important variables (usually called variable selection) and dimension reduction by projection (usually called projection). In variable selection, data miners use stepwise regression, correlation, chi-square test, and so on; then they throw away the rejected variables. They consider only the selected variables for their tasks. Variable selection preserves good interpretability of individual variables. Another way to reduce the dimension of input space is to use a projection method such as principal components analysis, singular value decomposition, random projection, and so on. Since the projection methods usually use a linear combination of all input variables, no variable selection is needed. However, projection methods suffer from insufficient interpretation of individual variables.

Suppose we have 1,000 variables as inputs to construct a prediction model. A variable selection method might choose 100 variables among 1,000 variables and reject the other 900 variables. There is no guarantee that we can tell whether the rejected 900 variables have less information than each of the selected 100 variables does. We can only say the selected variables are more important than others based on the selection criterion. A projection method, such as principal component analysis, includes all 1,000 variables in forms of linear combinations, which make a reduced input space. However, the linear combinations of all input variables might ignore some meanings of individual variables and reduce the model interpretability.

In this paper, we introduce a new concept of reducing the dimension of input space, which incorporates advantages of both variable selection and projection methods into data mining tasks. We use the rejected variables to improve data mining tasks through projection techniques. First, we use variable selection methods and obtain important variables; this step preserves the maximum interpretability of the important variables. Second, we apply projection techniques on only the rejected space to retrieve additional information from the rejected variables; this step minimizes information loss of input variables. To avoid the increasing dimension problem at the second step, we also suggest incorporating some selection mechanisms to find a few of the best projections of rejected variables. We merge the chosen projections of rejected variables into the set of selected variables to make a new reduced input space. The new input space is used to make an additional comparable predictive model, and finally we choose the best predictive model among candidate models including this comparable model.

BASIC CONCEPT OF NEW REDUCED INPUT SPACE

To create a new reduced input space that uses rejected variables, we can apply any kind of existing variable selection method to all the input variables to choose the most predictable variables. Then we collect rejected variables for further analysis. After the variable selection, we apply any kind of dimension reduction methods or projection methods only on all the rejected variables. We select a few of the best projections from the dimension reduction or projection result based on a certain criterion such as large eigenvalues, R-square, or chi-square values associated with the target variable. The selection mechanism can be an interactive tool based on dimension reduction techniques. For example the best projection can be selected based on graphs or plots. Next, we merge the best projections chosen from the rejected variables into the set of the selected variables to form a new reduced input space. Figure 1 shows the steps for creating new combined input space.



.

Figure 1. Steps for Creating a Combined Input Space

PREDICTIVE MODELING PROCESS THAT USES NEW REDUCED INPUT SPACE

Figure 2 shows how we can use the new reduced input space to improve the predictive modeling. Step1 through step 5 in Figure 2 show a typical process of predictive model building, in which we choose variables that are most significant from various sources of variable selection and apply any modeling process to obtain the best predictive model based on the selected variables. Step 6 through step 13 use the new reduced input space to create a calibrated and comparable model which can be used for model comparison. We collect the rejected variables from the variable selection process and possibly some variables that were rejected during the modeling process. We use all the collected rejected variables to retrieve some additional information. We make a few projections of the rejected variable, add them to the input space, and rebuild a predictive model. The rebuilt predictive model is compared with the best model from the selected variable space. Finally we choose the best model between them.

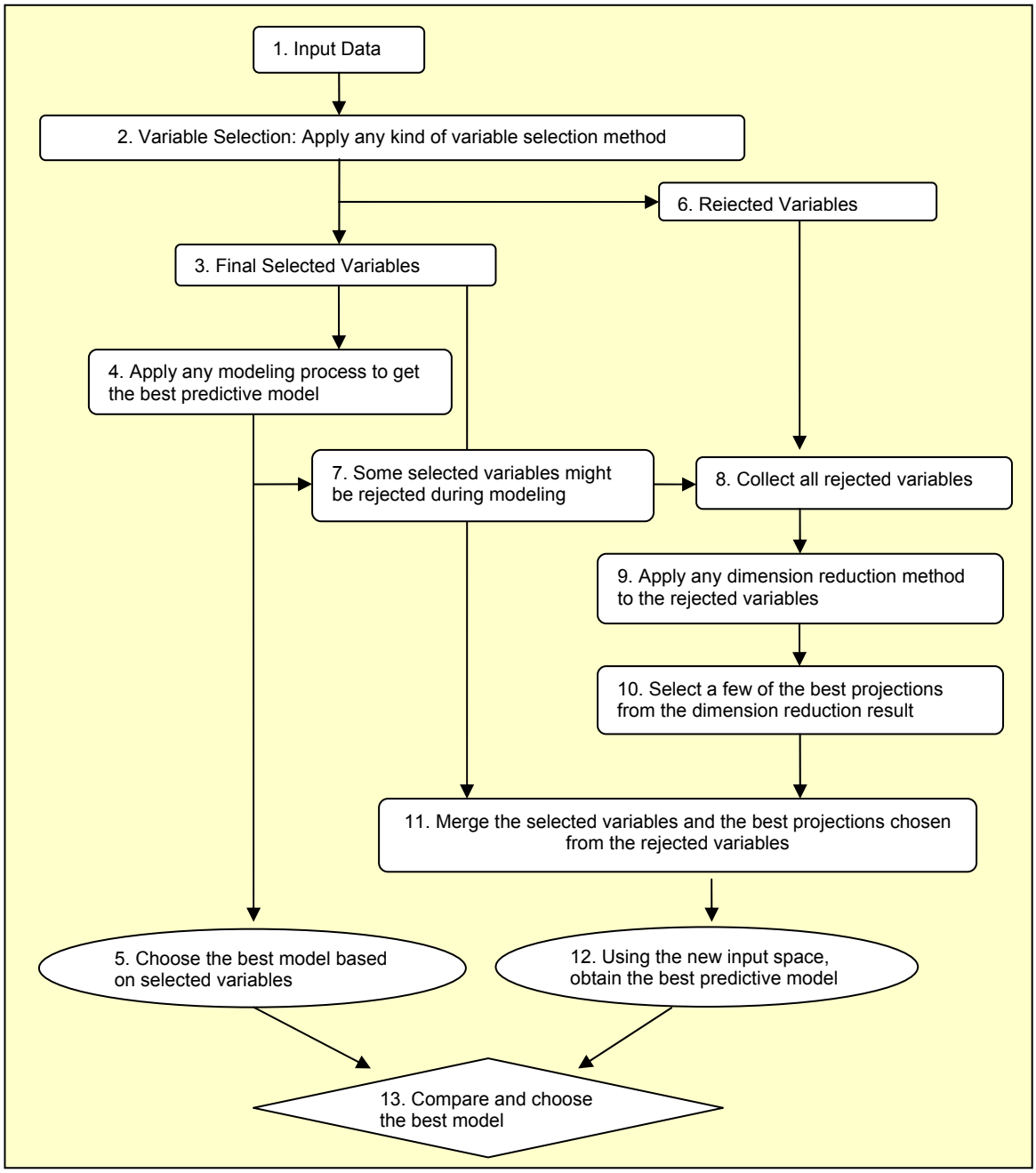


Figure 2. Steps for Using Reduced Input Space in Predictive Modeling

RANDOM PROJECTION NODE

SAS Enterprise Miner software introduces a new extension node for random projection. This tool node makes the proposed process of model building easy by using the rejected variables. Random projection is a technique that projects a set of points from a high-dimensional space to a randomly chosen low-dimensional subspace. Random projection approximately preserves pairwise distances with high probability. This idea of random projection has been motivated by the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984) as follows:

Johnson-Lindenstrauss Lemma

For any $0 < \varepsilon < 1$ *and an integer* n, *let* n *be a positive integer such that*

$$k \geq 4\ ln(n)\ (\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3})^{-1}$$

Then for any set V *of* n *points in* $\mathfrak{R}^d$, *there is a map* $f: \mathfrak{R}^d \rightarrow \mathfrak{R}^k$ *such that for all* $u, v \in \boldsymbol{V}$,

$$(1-\varepsilon)||u-v||^2 \leq ||f(u) - f(v)||^2 \leq (1+\varepsilon)||u-v||^2.$$

Using a random $d \times k$ $(k \ll d)$ matrix $\boldsymbol{R}$, the d-dimensional data, $\boldsymbol{X}$ is projected to a k-dimensional subspace,

$$\boldsymbol{Z}_{kxN} = \boldsymbol{R}_{kxd}\boldsymbol{X}_{dxN}.$$

A common random matrix can be obtained from a standard normal distribution. Let $\boldsymbol{R} = (r_{ij})$ be a random matrix such that each entry (r_{ij}) is chosen independently from $\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{1})$. Then a *d*-dimensional vector $u \in \mathfrak{R}^d$ is projected to $u' = \frac{1}{\sqrt{k}}\boldsymbol{R}^t u \in \mathfrak{R}^k$, where a constant $\frac{1}{\sqrt{k}}$ is a normalization factor.

Achlioptas (2003) proposed a random matrix $\boldsymbol{R}$ (called *sparse random projection*) with i.i.d. entries in

$$r_{ij} = \begin{cases} 1 & with\ prob.\ \frac{1}{2s} \\ 0 & with\ prob.\ 1 - \frac{1}{s} \\ -1 & with\ prob.\ \frac{1}{2s} \end{cases}$$

where Achlioptas used $s = 1$ or $s = 3$.

Figure 3 shows the SAS Enterprise Miner Random Projection node and its properties. We can select Normal or Sparse from the Method property. The Sparse property corresponds to the s in the sparse random projection. The Dimension property is the dimension (k) to be reduced.

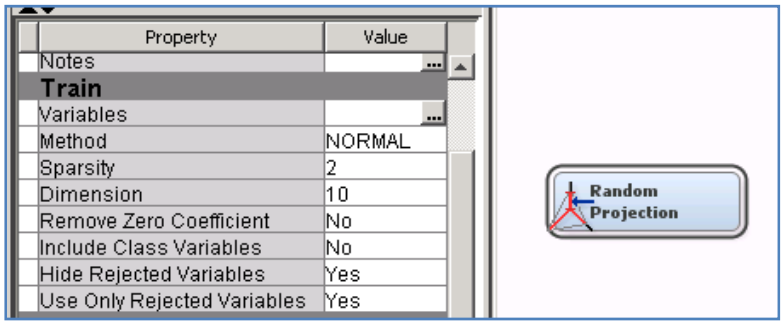


Figure 3. SAS Enterprise Miner Random Projection Node and Its Properties

The examples in the remaining sections of this paper illustrate how to use the Random Projection node. The Random Projection node is not a part of the production version of SAS Enterprise Miner 6.1, but the node is available upon reques.

EXAMPLES

We show six examples in this paper. The first three examples use a SAS Enterprise Miner example data set to show how to use the new reduced input space. The remaining three examples show how much the rejected variable space contributes to the predicted model after some variable selection techniques.

EXAMPLE 1: USING THE RANDOM PROJECTION NODE

This example demonstrates the key concept of the new reduced input space. It also illustrates the differences between a typical predictive modeling and the predictive modeling method with the new reduced input space.

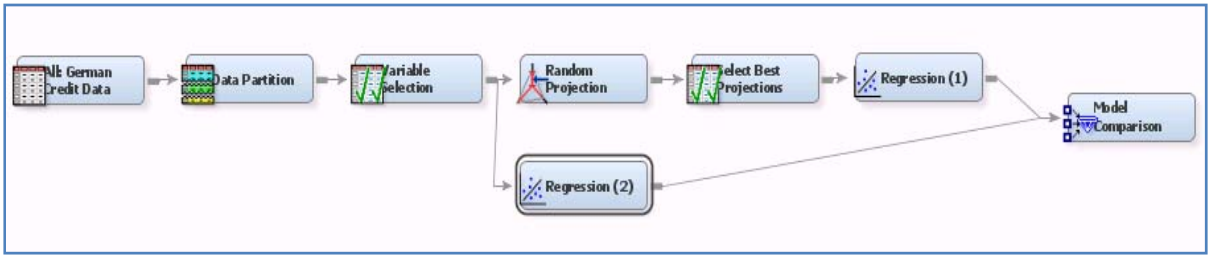


Figure 4. SAS Enterprise Miner Work Flow Diagram for Example 1

The input data set is called DMAGECR and is found in the SAS Enterprise Miner example data library. DMAGECR has 20 input variables and 1 target variable, and it contains 1,000 observations. The data has been partitioned for the purpose of model validation with 60% of data used for training and 40% of data used for validation. In the variable selection node, an R-square criterion has been applied through a forward stepwise logistic regression. Nine input variables have been selected from the 20 variables, and the remaining 11 variables have been rejected as shown in Figure 5.

Variable Selection

Name	Role	Level	Comment ▲
amount	Input	Interval	
checking	Input	Interval	
duration	Input	Interval	
foreign	Input	Interval	
history	Input	Interval	
installp	Input	Interval	
other	Input	Interval	
purpose	Input	Nominal	
savings	Input	Interval	
age	Rejected	Interval	Varsel:Small R-square value
coapp	Rejected	Interval	Varsel:Small R-square value
depends	Rejected	Interval	Varsel:Small R-square value
employed	Rejected	Interval	Varsel:Small R-square value
existcr	Rejected	Interval	Varsel:Small R-square value
housing	Rejected	Interval	Varsel:Small R-square value
job	Rejected	Interval	Varsel:Small R-square value
marital	Rejected	Interval	Varsel:Small R-square value
property	Rejected	Interval	Varsel:Small R-square value
resident	Rejected	Interval	Varsel:Small R-square value
telephon	Rejected	Interval	Varsel:Small R-square value

Figure 5: Variable Selection Showing Nine Selected Input Variables

We apply a random projection method on the rejected variables. The random projection matrix for the 11 rejected variables is shown in Figure 6, and the 10 random projection vectors at the matrix have been used to create 10 new variables (projections) from the 11 rejected variables.

Random Projection Matrix

RPV1	RPV2	RPV3	RPV4	RPV5	RPV6	RPV7	RPV8	RPV9	RPV10
0.767525	2.260774	-1.15644	-0.9506	0.160656	1.034888	0.033369	-0.6049	1.743889	-0.0428
0.990042	1.553993	0.689938	0.013239	-0.24471	0.968651	-2.0181	0.131461	-1.72067	0.80352
-0.67497	-0.90165	-1.28474	1.1013	0.756555	0.62684	-0.95448	0.73901	-0.17677	-0.98214
-0.16728	-0.49903	0.40206	-0.37098	0.218778	0.927773	-1.13131	-0.52473	-1.78842	-0.35035
0.502873	0.270932	-0.67262	0.82874	-0.10456	-1.10004	-0.40925	-0.25785	1.453048	0.504369
0.344871	0.673854	-0.91845	0.292544	-1.0031	-0.32774	0.028488	1.04108	0.46437	-1.01833
3.421082	1.471187	-0.31609	0.10717	0.231557	0.474024	0.461838	0.511607	0.589859	0.534024
-0.70568	0.787496	1.005162	1.579922	0.841892	2.385951	1.515943	0.076723	-2.79266	1.379481
-0.15074	-0.95802	-0.43457	0.765283	-0.24951	0.105303	-0.487	-1.02776	0.746796	0.539848
-0.08141	1.099105	0.466938	0.212444	-0.55704	0.189592	1.886047	-0.43662	-0.49222	-1.21716
1.55748	-0.8334	0.666113	-0.70426	1.17424	-1.04949	-1.08363	0.153772	0.515281	0.656777

Figure 6. Random Projection Matrix for Rejected Variables

For example, the scored projection _RP6 is calculated from the linear combinations of all the rejected variables using RPV6 as follows:

_RP6 = 1.034888067 × age +0.9686508181 × coapp +0.6268403247 × depends

+0.9277725607 × employed -1.100042561 × existcr -0.32773644 × housing

+0.4740244195 × job + 2.3859506457 × marital +0.1053032746 × property

+0.1895920912 × resident -1.049489085 × telephon.

Now we select a few of best projections from the candidate projections. We apply a binary split model using a chi-square test to select the best projections from the candidates. Two projections (_RP6 and _RP10) are selected out of the 10 projections as shown in Figure 7.

Variable Selection

Name	Label ▼	Role	Comment
_RP9	Random Projection Vector 9	Rejected	Varsel2:Small Chi-square value
_RP8	Random Projection Vector 8	Rejected	Varsel2:Small Chi-square value
_RP7	Random Projection Vector 7	Rejected	Varsel2:Small Chi-square value
_RP6	Random Projection Vector 6	Input	
_RP5	Random Projection Vector 5	Rejected	Varsel2:Small Chi-square value
_RP4	Random Projection Vector 4	Rejected	Varsel2:Small Chi-square value
_RP3	Random Projection Vector 3	Rejected	Varsel2:Small Chi-square value
_RP2	Random Projection Vector 2	Rejected	Varsel2:Small Chi-square value
_RP10	Random Projection Vector 10	Input	
_RP1	Random Projection Vector 1	Rejected	Varsel2:Small Chi-square value

Figure 7. Variable Selection Applied to Random Projection Vectors

This projection vector can be selected interactively as follows:

1. Produce a random vector.
2. Use the random vector to score the data and get a new variable.
3. Evaluate the new variable.
4. If the new variable is significant, keep it; if not, throw it away.
5. Generate a new random vector, and do the previous step again until you reach your goal.
6. The goal can be a fixed number of variables or a cutoff statistic for model improvement.

We combine the previously selected variables with the new chosen variables (projections). The merged variable list contains eight variables selected from original variable space and two variables created and selected from the projected space of the rejected variables. These 10 variables are used as the final input variables for data mining tasks. In this example they become independent variables for a regression modeling.

Variables - Reg

Name	Label	Role	Use	Report	Level
foreign		Input	Default	No	Interval
_RP10	Random Projection Vector 10	Input	Default	No	Interval
savings		Input	Default	No	Interval
history		Input	Default	No	Interval
purpose		Input	Default	No	Nominal
amount		Input	Default	Yes	Interval
installp		Input	Default	No	Interval
checking		Input	Default	No	Interval
other		Input	Default	No	Interval
_RP6	Random Projection Vector 6	Input	Default	No	Interval

Figure 8. Merged New Input Variables

When we compare the prediction performance between Regression (1) and Regression (2) in Figure 4, we see that Regression (1) uses the new input space, which has two more independent variables, and Regression (2) uses only the selected variables from the variable selection. Regression (1) and Regression (2) have exactly the same settings except for the new two variables. In this example, the predictive model with the new input space increases prediction accuracy about 1%, as shown in Figure 9. The improvement is not very impressive because we used simple logistic regression rather than stepwise, forward, and backward logistic regression. However, this example shows how the new input space works.

Fit Statistics
Model selection based on _VMISC_

Selected Model	Model Node	Valid: Misclassification Rate
Y	Reg1	0.215
	Reg2	0.225

Figure 9. Model Comparison for Example 1

EXAMPLE 2: USING PRINCIPAL COMPONENTS OF THE REJECTED VARIABLES

This example uses the same data set and settings of Example 1 up to variable selection, but it uses the principal components analysis (PCA) instead of random projections after the variable selection. Figure 10 shows the work flow diagram.

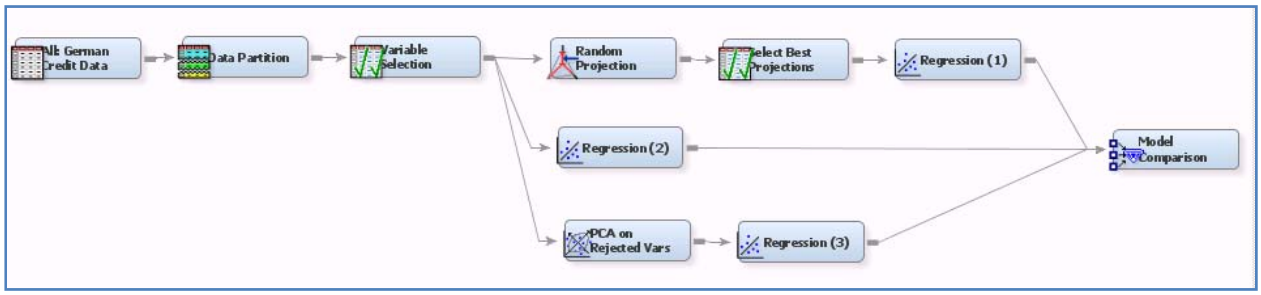


Figure 10. Work Flow Diagram with Principal Component Analysis

The rejected variables are used for principal components analysis. The first three principal components based on eigenvalues are chosen to explain the variation of rejected variables. The three principal components are merged with the already chosen input variables and are run through the regression node with the merged input data. The merged variable list at the Regression (3) node is shown in Figure 11.

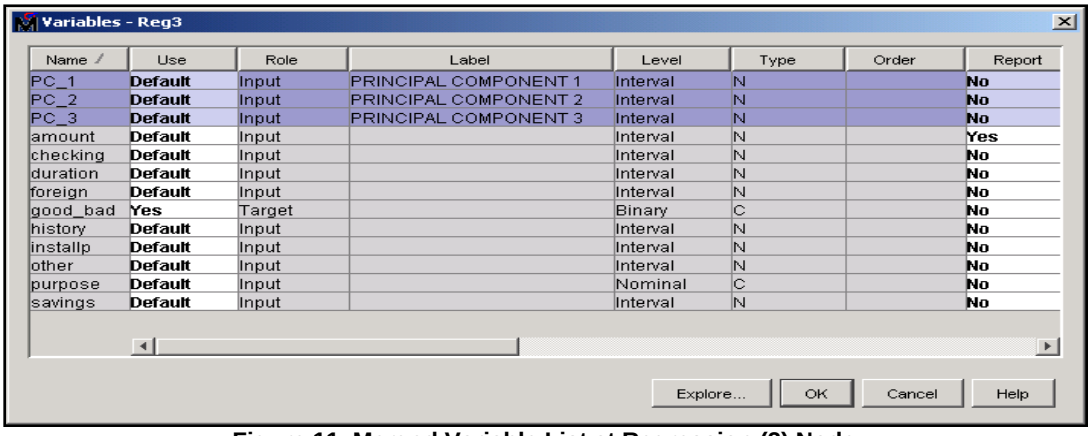


Figure 11. Merged Variable List at Regression (3) Node

We run the regression node and pass the result to the model comparison node. The model comparison produces the table shown in Figure 12. This table includes the model comparison from Example 1. Reg1 with random projection is still the winner, but Reg3 with the first three principal components also improves the accuracy of predictive model by 0.5%.

Fit Statistics
Model selection based on _VMISC_

Selected Model	Model Node	Valid: Misclassification Rate	Train: Average Squared Error
Y	Reg1	0.215	0.15853
	Reg2	0.225	0.15981
	Reg3	0.220	0.15922

Figure 12. Model Comparison for Example 2

EXAMPLE 3: USING PRINCIPAL COMPONENTS SELECTED BY THEIR TARGET ASSOCIATION

This example uses the same data set and settings as Example 2, but instead of choosing principal components with the top three largest eigenvalues as was done in Example 2, we use the R-square values between the target variable and principal components as a selection criterion, as shown in Figure 13.

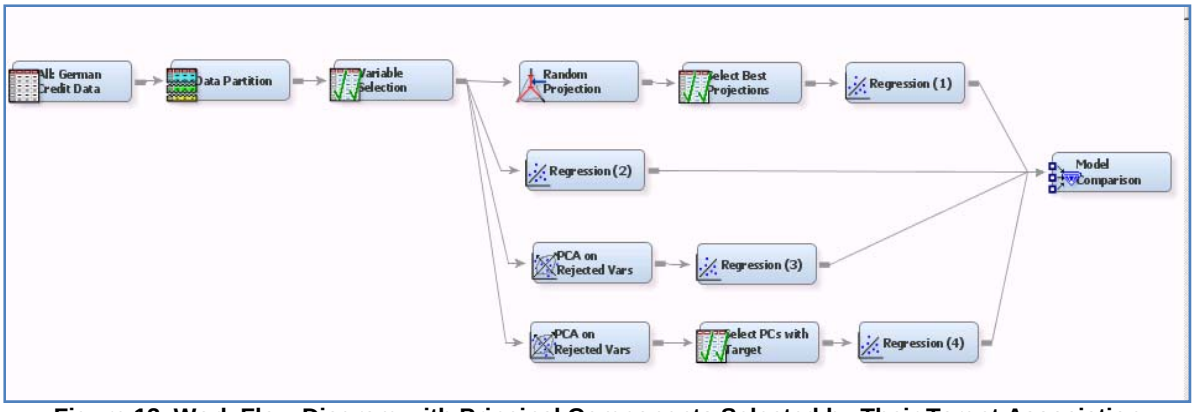


Figure 13. Work Flow Diagram with Principal Components Selected by Their Target Association

We create principal components, but no selection has been made. So 10 principal components are passed to the "Select PCA with target (Variable Selection)" node. We run the variable selection node with an R-square criterion. The two principal components, PC_7 and PC_2, are selected and added to the existing input variables. The variable

editor at Regression (4) node shows the combined new input variable list, as shown in Figure 14.

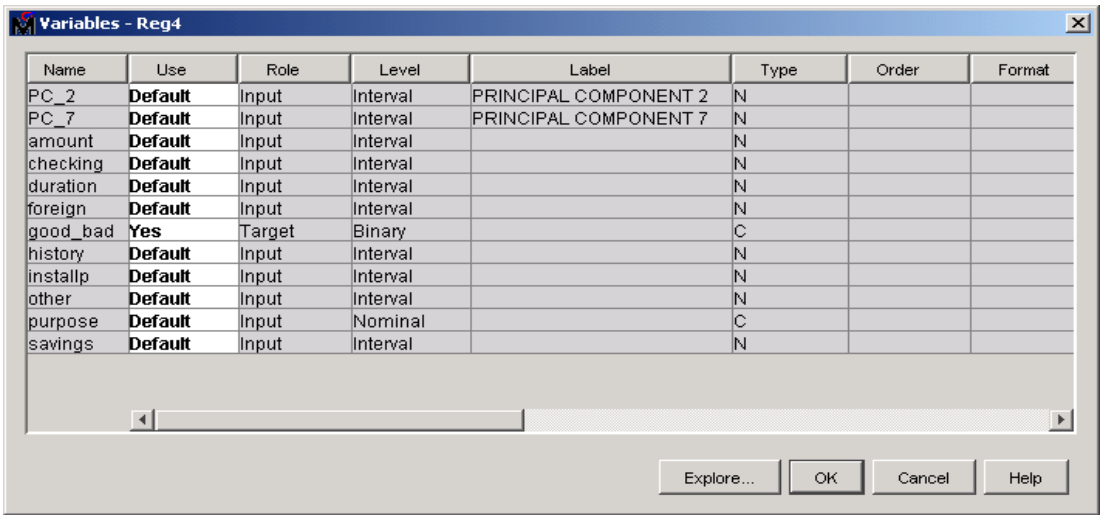


Figure 14. Merged Variable List at Regression (4)

We run the Regression (4) node using the new input variables and compare the result with the previous ones. Based on the misclassification of validation data shown in Figure 15, Reg1 is still the best, but Reg4 also shows a slight improvement in the model accuracy.

Fit Statistics
Model selection based on _VMISC_

Selected Model	Model Node	Valid: Misclassification Rate	Train: Average Squared Error	Valid: Average Squared Error
Y	Reg1	0.2150	0.15853	0.15345
	Reg2	0.2250	0.15981	0.15504
	Reg3	0.2200	0.15922	0.15300
	Reg4	0.2225	0.15778	0.15224

Figure 15. Model Comparison for Example 3

EXAMPLE 4: RANDOM PROJECTION WITH STEPWISE REGRESSION VARIABLE SELECTION

This example uses the Sonar data from the UCI Machine Learning Repository used by Gorman and Sejnowski (1998) in their study of the classification of sonar signals using a neural network. The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock. The data set contains 111 patterns obtained by bouncing sonar signals off a metal cylinder and 97 patterns obtained from rocks. Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time. The label associated with each record contains the letter "R" if the object is a rock and "M" if it is a mine (metal cylinder). The data set contains 208 observations, 60 input variables (VAR1…VAR60), and 1 binary target variable.

First, we run a stepwise logistic regression to find which variables are important. We partition the data so that 50% of the data are used for training and 50% of the data are used for testing. Each data partition has 104 observations.

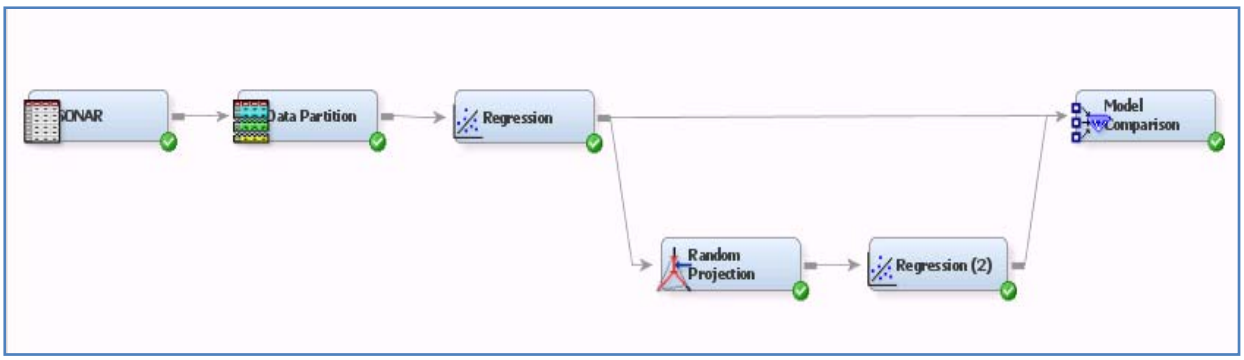


Figure 16. Work Flow Diagram for Example 4

We use a stepwise regression to produce the nine variables shown in the first column in Figure 17 as important variables.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	4.8489	1.6070	9.11	0.0025		127.600
VAR1	1	-70.6976	24.7408	8.17	0.0043	-0.8506	0.000
VAR12	1	-15.5630	5.4482	8.16	0.0043	-1.1399	0.000
VAR13	1	11.5780	5.2271	4.91	0.0268	0.8675	999.000
VAR17	1	4.7648	2.0725	5.29	0.0215	0.6895	117.303
VAR20	1	-7.1074	2.0622	11.88	0.0006	-0.9630	0.001
VAR36	1	3.9058	1.8424	4.49	0.0340	0.5610	49.690
VAR40	1	7.7383	2.6601	8.46	0.0036	0.7552	999.000
VAR43	1	-11.0496	3.9673	7.76	0.0053	-0.8081	0.000
VAR49	1	-39.0800	15.5421	6.32	0.0119	-0.6993	0.000

Figure 17. Variable Selection from Stepwise Regression

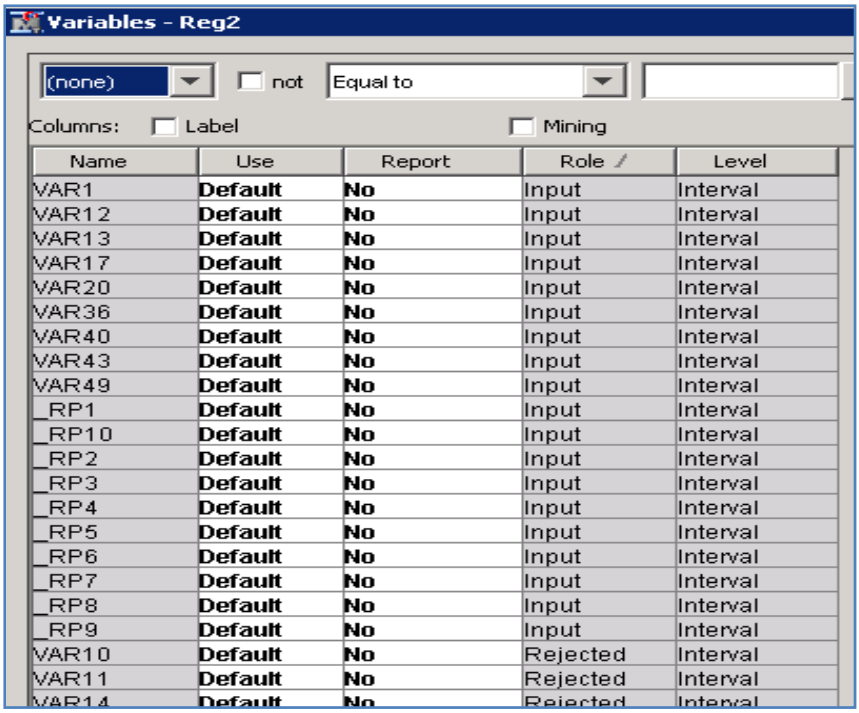
Using random projections of rejected variables and the previously selected variables, we build a predictive model Regression (2) as shown in Figure 16. The random projection node produces 10 scored projections from the 51 rejected variables. In other words, the rejected 51 variables are projected to a reduced space with dimension 10. The scored projections are added to the set of previously selected input variables. This can be done very easily using the Random Projection node with the property of Use Only Rejected Variables Yes. The random vectors are also shown in Figure 18.

Random Projection Matrix

Random Projection Vector 1	Random Projection Vector 2	Random Projection Vector 3	Random Projection Vector 4	Random Projection Vector 5	Random Projection Vector 6	Random Projection Vector 7	Random Projection Vector 8	Random Projection Vector 9	Random Projection Vector 10
-0.62724	-1.14743	-1.26532	-0.61026	0.747364	0.0313	-2.215	1.743698	0.969467	0.672834
-1.14018	-0.45766	0.104644	-0.36776	1.396928	0.125304	1.465335	0.744799	-0.54853	-0.80088
0.346269	-0.26714	0.674329	1.81211	-0.26528	0.669999	-0.04573	-0.04572	-0.4493	-0.44253
-0.56771	-0.89075	-1.02158	-0.40989	-0.50746	-0.9064	-0.78596	0.335356	-0.79274	0.063282
-3.33628	-0.60131	-1.21858	0.114353	-1.34331	2.366692	1.869308	-1.37029	0.817864	-0.39474
-1.90497	-0.79805	-0.54568	-0.94969	-0.05927	0.154652	1.970148	-1.14513	0.002958	0.901316
0.650718	0.233258	0.754659	-0.19534	1.255857	-0.22828	1.400377	1.39827	-2.13655	-0.98798
-0.3308	0.787826	0.057794	1.725283	0.36242	1.769296	0.376992	-0.06436	2.674475	0.34878
-2.3876	2.657913	-0.75556	0.306669	0.709478	0.404907	0.097509	0.335786	1.748537	0.167269
-0.11743	-0.33549	0.974007	-0.82794	-1.35206	1.076487	-1.62434	0.743458	-2.12126	0.345359
-0.17318	-0.74232	0.408963	-0.7843	-0.79636	0.844035	0.353806	-0.89119	0.64075	-0.04081
-0.74952	0.527002	0.172948	-0.16633	1.081312	0.738999	1.691635	1.04638	2.15941	-0.54376

Figure 18. Random Matrix from Random Projection Node

Next, we run the same stepwise logistic regression with this new combined set of input variables. The Regression (2) node in Figure 16 has a new set of input variables with 10 random projections (_RP1 through _RP10) shown in Figure 19.



Variables - Reg2

(none) not Equal to

Columns: Label Mining

Name	Use	Report	Role	Level
VAR1	Default	No	Input	Interval
VAR12	Default	No	Input	Interval
VAR13	Default	No	Input	Interval
VAR17	Default	No	Input	Interval
VAR20	Default	No	Input	Interval
VAR36	Default	No	Input	Interval
VAR40	Default	No	Input	Interval
VAR43	Default	No	Input	Interval
VAR49	Default	No	Input	Interval
_RP1	Default	No	Input	Interval
_RP10	Default	No	Input	Interval
_RP2	Default	No	Input	Interval
_RP3	Default	No	Input	Interval
_RP4	Default	No	Input	Interval
_RP5	Default	No	Input	Interval
_RP6	Default	No	Input	Interval
_RP7	Default	No	Input	Interval
_RP8	Default	No	Input	Interval
_RP9	Default	No	Input	Interval
VAR10	Default	No	Rejected	Interval
VAR11	Default	No	Rejected	Interval
VAR14	Default	No	Rejected	Interval

Figure 19. Variable Set with Random Projections

As the result of stepwise regression, the new model drops five variables from the previously selected inputs and adds two projection variables as significant predictors to the model, as shown in Figure 20.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	6.5974	1.8197	13.14	0.0003		733.195
VAR1	1	-68.8541	21.7238	10.05	0.0015	-0.8284	0.000
VAR12	1	-5.1270	2.4887	4.24	0.0394	-0.3755	0.006
VAR36	1	2.9163	1.3617	4.59	0.0322	0.4189	18.474
VAR49	1	-20.6267	9.1746	5.05	0.0246	-0.3691	0.000
_RP2	1	-1.3975	0.4686	8.89	0.0029	-0.6462	0.247
_RP4	1	0.5487	0.2446	5.03	0.0249	0.3943	1.731

Figure 20. Finally Selected Variables for Example 4

The model comparison shown in Figure 21 demonstrates a slight improvement by random projection and indicates that the rejected variables from the first stepwise regression still contain significant information for prediction.

Fit Statistics

Model Selection based on Test: Misclassification Rate (_TMISC_)

Selected Model	Model Node	Model Description	Test: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate
Y	Reg2	Regression (2)	0.26667	0.13557	0.21359
	Reg	Regression	0.28571	0.10578	0.15534

Figure 21. Model Comparison for Example 4

EXAMPLE 5: RANDOM PROJECTION WITH VARIABLE SELECTION NODE

This example uses the same data set and settings as Example 4 and adds a variable selection node. Figure 22 shows the work flow diagram.

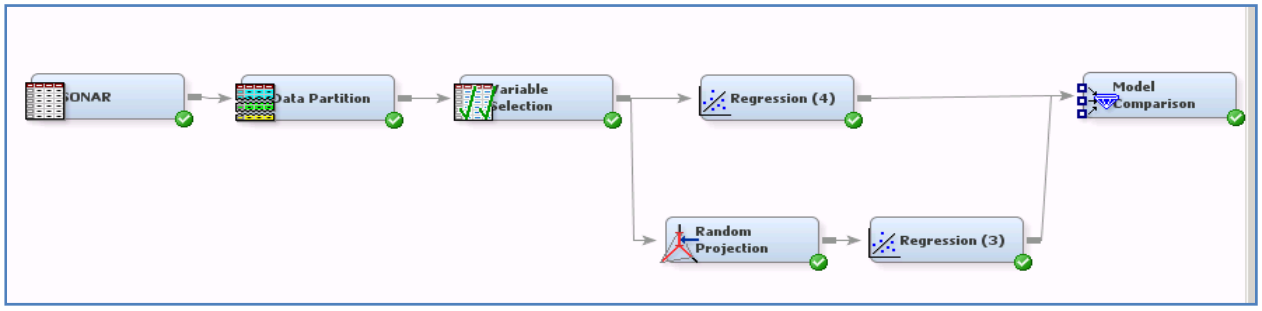


Figure 22. Work Flow Diagram for Example 5

In the variable selection node, R-square and chi-square criteria are used to select the following variables: VAR1, VAR12, VAR13, VAR17, VAR20, VAR36, VAR40, VAR43, and VAR49. These variables are different from the variables selected in the stepwise regression in Example 4: VAR5, VAR9, VAR10, VAR12, VAR18, VAR19, VAR31, VAR54, and VAR58. Only VAR12 is selected from both. Modeling with the same stepwise regression with these variables chosen from variable selection might not return a good result because the variable selection might miss significant predictors. Therefore, we need to calibrate the model by using random projections of rejected variables.

In Figure 22, Regression (4) runs a stepwise regression with the selected variables from variable selection and Regression (3) runs a stepwise regression with the selected variables and 10 random projections of rejected variables. The final model from Regression (4) shows VAR12, VAR19, and VAR18 as significant predictors as shown in Figure 23.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	VAR12	1	1	17.7276		<.0001
2	VAR19	1	2	4.5854		0.0322
3	VAR18	1	3	11.4299		0.0007

Figure 23. Summary of Stepwise Selection for Regression (4)

Instead of VAR19 and VAR18, the final model from Regression (3) selects the two random projections of rejected variables as shown in Figure 24.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	VAR12	1	1	17.7276		<.0001
2	_RP10	1	2	9.3042		0.0023
3	_RP8	1	3	7.6522		0.0057

Figure 24. Summary of Stepwise Selection for Regression (3)

The result shown in Figure 24 demonstrates that some of rejected variables have much more significant information than some of selected variables, so the new calibrated model has been improved in predictability. The model accuracy has been increased by 13% in the test data set. The model comparison in Figure 25 shows the model with the combined new input space (Regression (3)) performs much better than the other model (Regression (4)).

Fit Statistics
Model Selection based on Test: Misclassification Rate (_TMISC_)

Selected Model	Model Node	Model Description	Test: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate
Y	Reg3	Regression (3)	0.22857	0.16968	0.25243
	Reg4	Regression (4)	0.36190	0.17326	0.27184

Figure 25. Model Comparison for Example 5

EXAMPLE 6: RANDOM PROJECTION WITH VARIABLE CLUSTERING NODE

This example uses the same data set that was used in Example 5, but it uses a variable clustering node to select important variables and do a similar analysis. Figure 26 shows the work flow diagram.

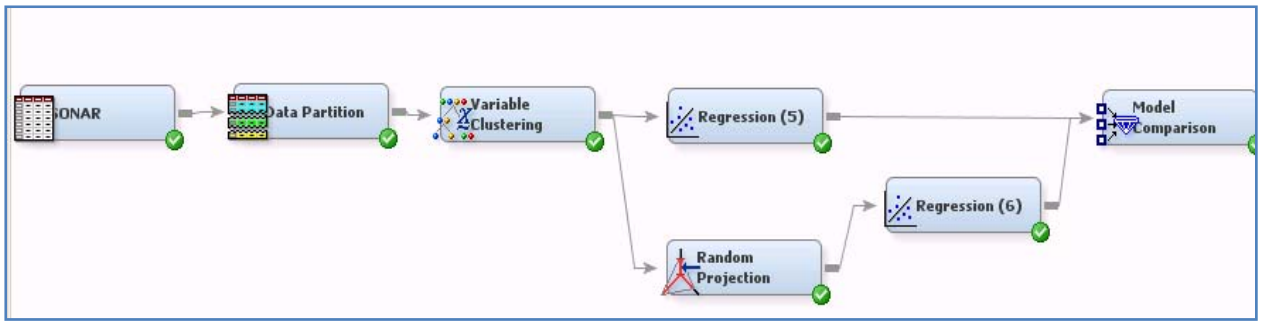


Figure 26. Work Flow Diagram for Example 6

The Variable Clustering node produces 14 clusters and selects the most significant variable from each cluster. The 14 selected variables are shown in Figure 27.

Variable Selection Table

Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	VAR4	0.80244	CLUS9	0.176159	Variable		0.239803	YES
CLUS10	VAR19	0.902291	CLUS3	0.30112	Variable		0.139809	YES
CLUS11	VAR48	0.835968	CLUS2	0.343886	Variable		0.250006	YES
CLUS12	VAR26	0.921928	CLUS6	0.296848	Variable		0.111031	YES
CLUS13	VAR12	0.905168	CLUS5	0.315497	Variable		0.138541	YES
CLUS14	VAR6	0.883783	CLUS3	0.169913	Variable		0.140006	YES
CLUS2	VAR43	0.769616	CLUS11	0.264938	Variable		0.313421	YES
CLUS3	VAR16	0.918484	CLUS10	0.308846	Variable		0.117941	YES
CLUS4	VAR37	0.824764	CLUS2	0.143628	Variable		0.204626	YES
CLUS5	VAR9	0.888739	CLUS14	0.233167	Variable		0.145092	YES
CLUS6	VAR28	0.881572	CLUS12	0.195877	Variable		0.147276	YES
CLUS7	VAR23	0.959469	CLUS12	0.193255	Variable		0.05024	YES
CLUS8	VAR33	0.854309	CLUS4	0.251237	Variable		0.194575	YES
CLUS9	VAR59	0.636505	CLUS1	0.183653	Variable		0.44527	YES
CLUS1	CLUS1	1	CLUS9	0.265702	ClusterComp	Cluster 1	0	NO
CLUS1	VAR3	0.792237	CLUS9	0.186175	Variable		0.255292	NO
CLUS1	VAR2	0.783524	CLUS9	0.2082	Variable		0.273398	NO

Figure 27. Variable Selection Table from Variable Clustering

Using the two stepwise regressions (Regression (5) and Regression (6) nodes), we can evaluate the selected

variables from variable clustering. The stepwise regression from the Regression (5) node chooses three variables out of 14 selected variables as final predictors.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	VAR12		1	1	17.7276		<.0001
2	VAR19		1	2	4.5854		0.0322
3	VAR16		1	3	4.3749		0.0365
4	VAR9		1	4	4.2927		0.0383
5		VAR9	1	3		3.7072	0.0542

The selected model is the model trained in the last step (Step 5). It consists of the following effects:

Intercept VAR12 VAR16 VAR19

Figure 28. Stepwise Variable Selection after Variable Clustering

When we use random projections of 46 rejected variables from variable clustering node, the stepwise regression chooses two variables and one random projection as its best predictors and it improve the model by 6% based on the test data set.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	VAR12	1	1	17.7276		<.0001
2	_RP6	1	2	9.2573		0.0023
3	VAR19	1	3	4.4595		0.0347

The selected model is the model trained in the last step (Step 3). It consists of the following effects:

Intercept VAR12 VAR19 _RP6

Figure 29. Stepwise Variable Selection Including Random Projections after Variable Clustering

Fit Statistics

Model Selection based on Test: Misclassification Rate (_TMISC_)

Selected Model	Model Node	Model Description	Test: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate
Y	Reg6	Regression (6)	0.25714	0.17242	0.22330
	Reg5	Regression (5)	0.31429	0.18482	0.24272

Figure 30. Model Comparison for Example 6

CONCLUSION

We showed how to use rejected variables to create a new reduced input space for predictive models. The new input space is formed by adding projections of rejected variables to the set of already selected variables. The suggested new input space uses the advantages of both variable selection and projection of variables: interpretability of individual important variables and minimization of information loss from rejecting variables. We can always build one more comparable predictive model after a conventional predictive modeling process is done. The new calibrated and comparable predictive model increases the model accuracy by well chosen projections of rejected variables, and it also provides a good tool for evaluating the selected variables through the projections. Even though random projection is a fast and efficient projection of rejected variables, more research is still required on how fast we can generate and select a few of the best predictable projections of rejected variables.

REFERENCES

Achlioptas, D. (2003), “Database-Friendly Random Projections: Johnson-Lindenstrauss with Binary Coins,” *Journal of Computer and System Sciences*, 66(4) pp.671−687.

Dasgupta, S. and Gupta A. (1999), “An Elementary Proof of the Johnson–Lindenstrauss Lemma,” Technical report 99−006, U. C. Berkeley, March.

Gorman, R. P., and Sejnowski, T. J. (1988), “Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets", *Neural Networks*, Vol. 1, pp. 75−89.

Johnson W. and Lindenstrauss J. (1984), “Extensions of Lipschitz Maps into a Hilbert Space,” *Contemporary Mathematics*, 26, pp.189−206.

CONTACT INFORMATION

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

Taiyeong Lee

SAS Institute Inc.

SAS Campus Drive

Cary, NC 27513

919-531-2186

E-mail:Taiyeong.Lee@sas.com

David Duling

919-531-5267

E-mail:David.Duling@sas.com

Dominique Latour

919-531-6312

E-mail:Dominique.Latour@sas.com