WHITEPAPER

Variable Reduction Techniques

VALIANCE

Copyright @Valiance Analytics Pvt Ltd

Variable Reduction Techniques

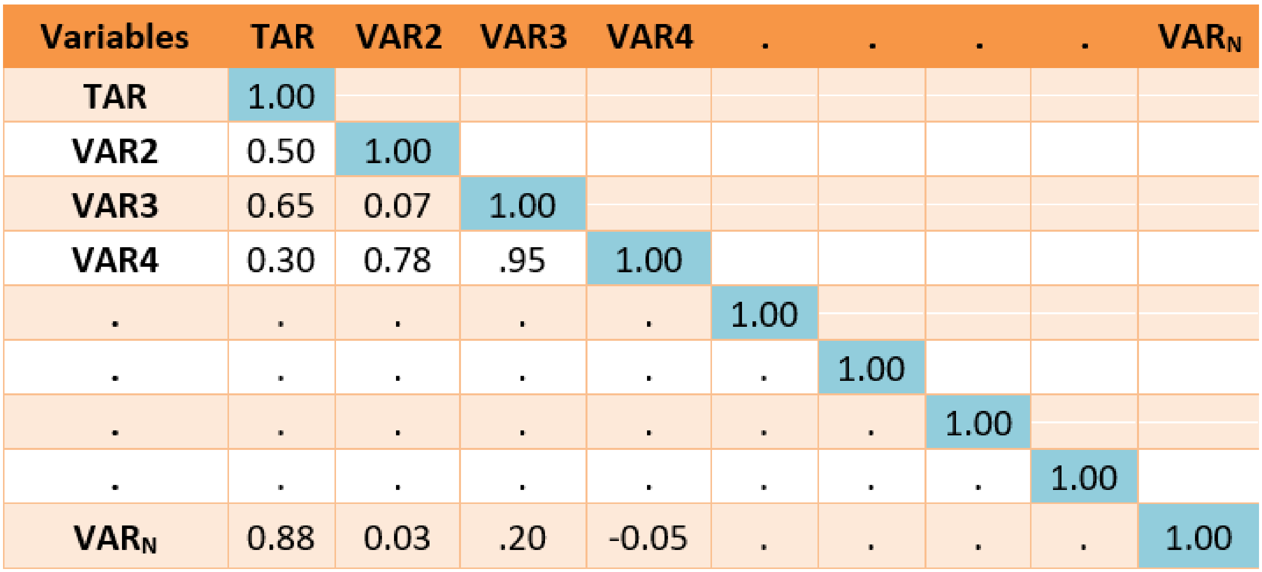
Abstract

Variable reduction is a crucial step for accelerating model building without losing potential predictive power of the data. With the advent of Big Data and sophisticated data mining techniques, the number of variables encountered is often tremendous making variable selection or dimension reduction techniques imperative to produce models with acceptable accuracy and generalization. The temptation to build an ecological model using all available information (i.e., all variables) is hard to resist. Ample time and money are exhausted gathering data and supporting information. Analytical limitations require us to think carefully about the variables we choose to model, rather than adopting a naive approach where we blindly use all information to understand complexity. The purpose of this paper is to illustrate the use of some techniques to effectively manage the selection of explanatory variables consequently leading to a parsimonious model with highest possible prediction accuracy. It may be noted that the following techniques may or may not be followed in the given order contingent on the data. The very basic step before applying following techniques is to execute univariate analysis for all the variables to get observations frequency count as well as missing value count. Variables with large proportion of missing values can be dropped upfront from the further analysis.

Correlation Analysis

We begin with developing a Correlation matrix between the dependent and independent variables, and between all the possible 2-pair combinations of independent variables. Correlation describes the strength of the linear association between two variables. It is measured by the correlation coefficient (r). The sign of the correlation coefficient indicates the direction of association and it always lies between -1 (perfect negative linear association) and 1 (perfect positive linear association). A zero value of r indicates no linear relationship. Let’s say we want to calculate a correlation matrix for variables all VAR1 VAR2 VAR3....VARN, including TAR - the target variable.

The following is an example of a correlation matrix:



Variables	TAR	VAR2	VAR3	VAR4	.	.	.	.	VAR_N
TAR	1.00								
VAR2	0.50	1.00							
VAR3	0.65	0.07	1.00						
VAR4	0.30	0.78	.95	1.00					
.	.	.	.	.	1.00				
.	.	.	.	.	.	1.00			
.	.	.	.	.	.	.	1.00		
.	.	.	.	.	.	.	.	1.00	
VAR_N	0.88	0.03	.20	-0.05	.	.	.	.	1.00

As a first screening check, we will analyse the correlation coefficient between the dependent variable (TAR) and the independent variables (VAR1, VAR2, VAR3,....,VARN). A correlation coefficient (abs) value of >=0.65 can be taken as a benchmark value for a significant linear association between the variables. The variables with a significant linear association with the target variable, indicated by a large correlation coefficient (r), should be included in the model at a prelim level. In the above example, VAR3 and VARN are significantly associated with TAR and should be included in the model.

We further move ahead by analysing the correlation coefficient of the 2-pair combinations of independent variables. A higher correlation coefficient (r) between two independent variables implies redundancy, indicating a possibility that they are measuring the same construct. In such a scenario, it would be prudent to select either of the two variables in consideration, or adopt an alternative approach to selection which involves two most widely used techniques viz. Principal Component Analysis (PCA) and Exploratory Factor Analysis.

Principal Component Analysis (PCA)

Principal Component Analysis is a variable reduction procedure and helps in obtaining a smaller number of variables called Principal Components, which account for most of the variance in the observed variables from a group of large number of redundant (correlated) variables.

In the above table, we can see that the variables VAR3 and VAR4 are highly correlated with r =0.95. Similarly, VAR2 and VAR4 with r =0.78 are significantly correlated. Principal Component Analysis can be performed on a set of correlated variables to obtain a new variable (Principal Component) which will have the properties of all the variables in question. A Principal Component is computed as a linear combination of optimally-weighted variables under consideration and can be used for subsequent analysis. One can compute as many principal components as the number of independent variables which can be further analysed and retained on the basis of the variability explained by them.

In SAS, a procedure called PRINCOMP is used for computing Principal Components where each component is a linear combination of the original variables (in our example VAR2, VAR3 and VAR4), with coefficients equal to the Eigenvectors of the correlation or covariance matrix.

Principal Component Analysis can also be used for exploring polynomial relationships and for multivariate outlier detection.

Exploratory Factor Analysis

Exploratory Factor Analysis is also a variable reduction procedure, similar to Principal Component Analysis in many respects but the underlying procedure for both the techniques remains the same. However, conceptually there are significant differences between the two techniques which are explained later in this section.

Factor analysis is a statistical techniques concerned with the reduction of a set of observable variables in terms of a small number of latent factors. The underlying assumption of factor analysis is that there exists a number of unobserved latent variables (or "factors") that account for the correlations among observed variables, such that if the latent variables are partialled out or held constant, the partial correlations among observed variables all become zero. In other words, the latent factors determine the values of the observed variables.

The term "common" in common factor analysis describes the variance that is analyzed. It is assumed that the variance of a single variable can be decomposed into common variance that is shared by other variables included in the model, and unique variance that is unique to a particular variable and includes the error component. Common factor analysis (CFA) analyzes only the common variance of the observed variables; principal component analysis considers the total variance and makes no distinction between common and unique variance.

The selection of one technique over the other is based upon several criteria. First of all, what is the objective of the analysis? Common factor analysis and principal component analysis are similar in the sense that the purpose of both is to reduce the original variables into fewer composite variables, called factors or principal components. However, they are distinct in the sense that the obtained composite variables serve different purposes. In common factor analysis, a small number of factors are extracted to account for the intercorrelations among the observed variables--to identify the latent dimensions that explain why the variables are correlated with each other. In principal component analysis, the objective is to account for the maximum portion of the variance present in the original set of variables with a minimum number of composite variables called principal components.

Secondly, what are the assumptions about the variance in the original variables? If the observed variables are measured relatively error free, (for example, age, years of education, or number of family members), or if it is assumed that the error and specific variance represent a small portion of the total variance in the original set of the variables, then principal component analysis is appropriate. But if the observed variables are only indicators of the latent constructs to be measured (such as test scores or responses to attitude scales), or if the error (unique) variance represents a significant portion of the total variance, then the appropriate technique to select is common factor analysis. Since the two methods often yield similar results, only CFA will be illustrated here.

Multicollinearity Check - Variance Inflation Factor (VIF)

Next we look at Multicollinearity, which occurs when independent variables are highly correlated among themselves. For instance, we have 5 independent variables – VAR1, VAR2, VAR3, VAR4, and VAR5. If any one of these variables can be expressed as a linear/non-linear function of other variable(s), then we say that data suffers from multicollinearity. In such a scenario, the coefficient estimates may change erratically in response to small changes in the data. The presence of multicollinearity affects the validity of individual predictor's estimated coefficient. The Variance Inflation Test (VIF) is recommended for a more thorough solution to the problem.

Variance Inflation Factor (VIF) is defined as:

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

For each explanatory variable i, R-square is defined as the coefficient of determination in a regression model where independent variable i is considered as target variable and all other independent variables are explanatory variables. Higher R-square results in higher VIF and indicates high correlation between the target variable (i.e. independent variable i) and all other independent variables.

The VIF provides information on how large the standard error is compared with what it would be if the variables were uncorrelated with the other predictor variables in the model. It is calculated for each explanatory variable and those with high values are removed. The definition of 'high' is somewhat arbitrary but a common thumb-rule classifies a VIF value of >=5 significantly high implying high multicollinearity. A cut-off VIF valueof <=2 is used by most businesses since it offers a more stringent and clear rule.

Now, once we have decided on the cut-off value for VIF, the next step is to check and compare the VIF values of the observed explanatory variables. Variables with a VIF value greater than the cut-off value may be dropped from the model. If for instance, the VIF values for all the explanatory variables is greater than the cut-off value then one can choose to keep the variables with the lowest VIFs. However, this is not a thumb-rule to address the problem of collinearity in the data. Different practitioners use different ways of handling the problem of multicollinearity and the probable success of the different methods depend on the severity of the collinearity problem and the business problem at hand.

WALD CHI Square

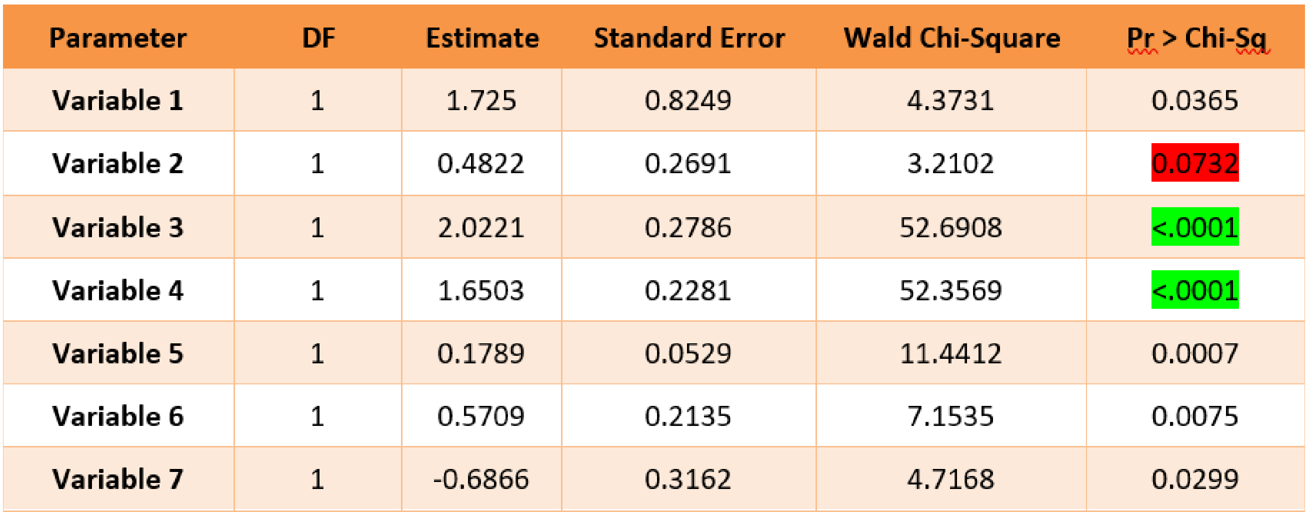
Wald Chi-Square is another popular technique which assists in variable selection. The Wald Chi-Square test statistic is the squared ratio of the Estimate to the Standard Error of the respective predictor. The probability that a particular Wald Chi-Square test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is given by Pr > Chi-Sq.

For instance an business analyst, who has data on sales and number of sales executive in a retail outlets, wonders whether sales is associated with number of sales executive in a retail outlets. Say is the average increase in sales for outlets having greater than 50 executives as compared to outlets having less than 50 executives: then the Wald test can be used to test whether is 0 (in which case sales has no association with number of sales executive in a retail outlets) or non-zero (sales varies with respect to number of executives presents in the outlet).

Wald chi-square is calculated to check the association between the dependent variable and the independent variable. We use univariate Logistic regression to calculate the Wald Chi-square statistics for each independent variable. Wald chi-square value greater than 6 is considered to be better as higher the value higher is the association between the dependent and independent variable. Variables having chi-square value less than 6 can be dropped from the model as they do not have significant association with the dependent variable.

A Fashion brand in US uses direct sales agents to sell products directly to the customer. Agents host the collection of the company in their local areas .The Company reported a decline in the revenues for past couple of years, soto expand and grow the business company wants to understand the attributes which affect the performance of an agent so that they can bring efficient agents on board.

After running univariate logistic regression for all the independent variable on dependent variable, summary table having Wald chi-square statistic for each independent variable is made.



Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Sq
Variable 1	1	1.725	0.8249	4.3731	0.0365
Variable 2	1	0.4822	0.2691	3.2102	0.0732
Variable 3	1	2.0221	0.2786	52.6908	<.0001
Variable 4	1	1.6503	0.2281	52.3569	<.0001
Variable 5	1	0.1789	0.0529	11.4412	0.0007
Variable 6	1	0.5709	0.2135	7.1535	0.0075
Variable 7	1	-0.6866	0.3162	4.7168	0.0299

From the above table we can see that variables having Wald chi-square statistic greater than 6 are more significant as compared to variables having chi-square value less than 6 i.e. Variable 3 and variable 4 are highly significant as compared to variable 1 and variable 2.

Variables having Wald chi-square statistic less than 6 can be dropped from the model building exercise. It will enhance the model performance and there will be very less loss of information because of dropping those variables. This technique can also be used to check the impact of dropping variable(s) on the model's predictive accuracy, though to be implemented at the later stages of model development.

Variable Clustering Using Proc Varclus

The VARCLUS procedure can also be used as a variable-reduction method. A large set of variables can often be replaced by the set of cluster components with little loss of information. A given number of cluster components does not generally explain as much variance as the same number of principal components on the full set of variables, but the cluster components are usually easier to interpret than the principal components, even if the latter are rotated.

The VARCLUS procedure divides a set of numeric variables into either disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster, which may be either the first principal component or the centroid component. PROC VARCLUS tries to maximize the sum across clusters of the variance of the original variables that is explained by the cluster components.

By default, PROC VARCLUS begins with all variables in a single cluster. It then repeats the following steps:

Step 1: A cluster is chosen for splitting. Depending on the options specified, the selected cluster has either the smallest percentage of variation explained by its cluster component (using the PERCENT= option) or the largest eigenvalue associated with the second principal component (using the MAXEIGEN= option).

Step 2: The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation (raw quartimax rotation on the eigenvectors), and assigning each variable to the rotated component with which it has the higher squared correlation.

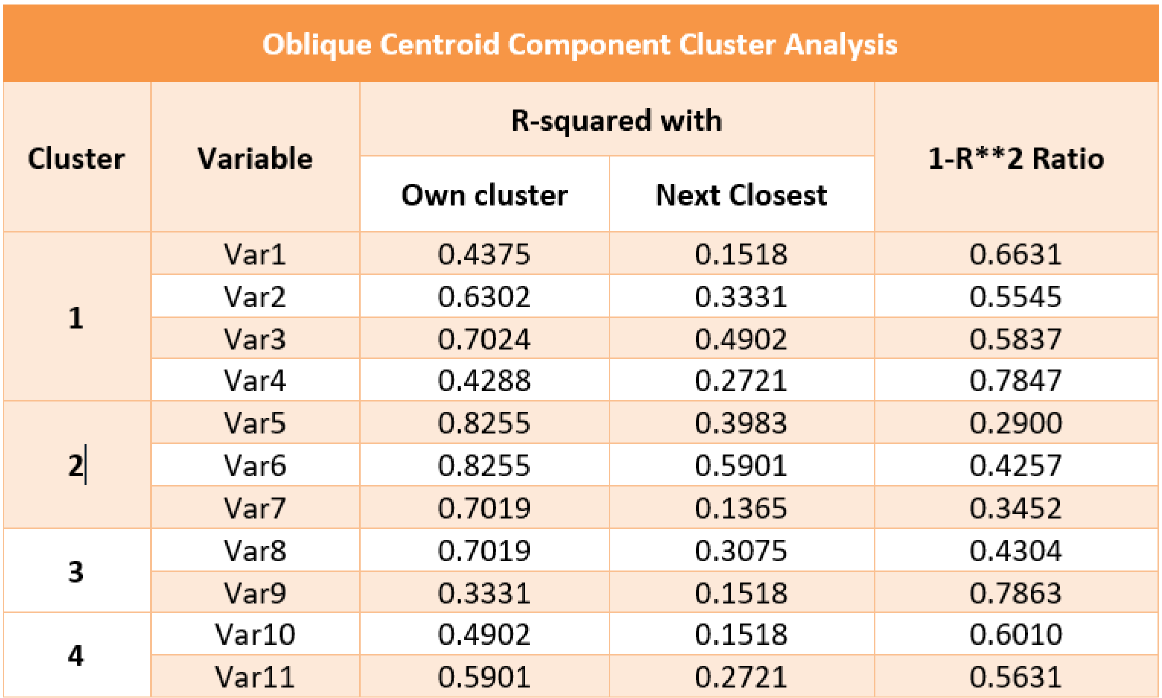
Step3: Variables are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components. The reassignment may be required to maintain a hierarchical structure.

The procedure stops when each cluster satisfies a user-specified criterion involving either the percentage of variation accounted for or the second eigenvalue of each cluster. By default, PROC VARCLUS stops when each cluster has only a single eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying factor dimension.

The following statements create the variable clusters:

```
PROC VARCLUS DATA=SAMPLE CENTROIDMAXCLUSTERS= N;
VAR PREDICTOR VARIABLES;
RUN;
```

The output will include the total number of clusters created, the number of variables used in the analysis, the number of observations, and the maxeigen threshold used to create the clusters. It will show the number of final clusters PROC VARCLUS has created. PROC VARCLUS will also show which variables have been assigned to the various clusters.



Oblique Centroid Component Cluster Analysis				
Cluster	Variable	R-squared with		1-R**2 Ratio
		Own cluster	Next Closest	
1	Var1	0.4375	0.1518	0.6631
	Var2	0.6302	0.3331	0.5545
	Var3	0.7024	0.4902	0.5837
	Var4	0.4288	0.2721	0.7847
2	Var5	0.8255	0.3983	0.2900
	Var6	0.8255	0.5901	0.4257
	Var7	0.7019	0.1365	0.3452
3	Var8	0.7019	0.3075	0.4304
	Var9	0.3331	0.1518	0.7863
4	Var10	0.4902	0.1518	0.6010
	Var11	0.5901	0.2721	0.5631

The above table shows the final output of PROC VARCLUS. We can select variables from each cluster - if the cluster contains variables which do not make any business sense, the cluster can be ignored. A variable selected from each cluster should have a high correlation with its own cluster and a low correlation with the other clusters. The 1-R**2 ratio can be used to select these types of variables. Small values of this ratio indicate good clustering. Variables having low 1-R**2 ratio can be selected. Two or more variables can also be selected from the cluster.

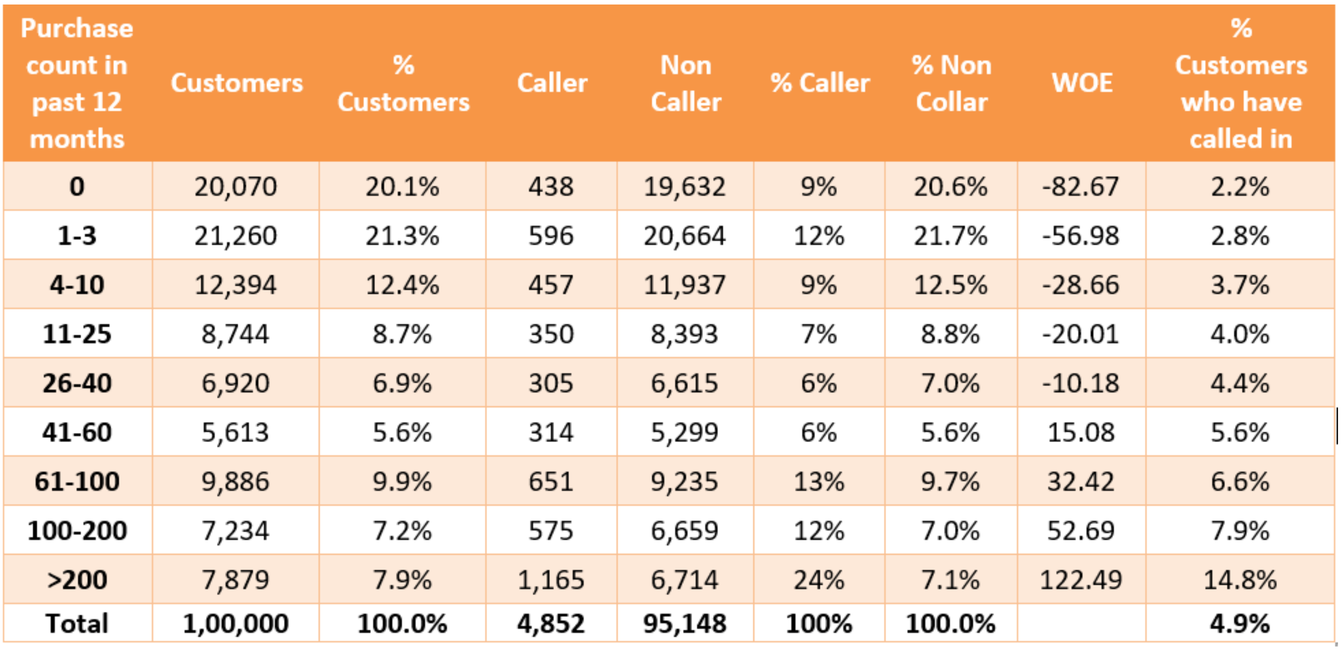
$$\text{1-R**2 ratio} = \frac{\text{1-R}^2\text{ own cluster}}{\text{1-R}^2\text{ next closest}} = \frac{1 - \uparrow}{1 - \downarrow} \Rightarrow \frac{\downarrow}{\uparrow} \Rightarrow \downarrow$$

Information Value (IV) & Weight Of Evidence (WOE)

Last but not the least, Information Value (IV)and Weight of Evidence (WOE) technique is very useful for variable selection in model building process. The main advantage of this technique is that it can assess both continuous and categorical variables.

Weight of Evidence analyzes the predictive power of a variable in relation to the targeted outcome, Information Value assesses the overall predictive power of the variable being considered, and therefore can be used for comparing the predictive power among competing variables.

The Following tables illustrate how the weight of evidence and information value is calculated.



Purchase count in past 12 months	Customers	% Customers	Caller	Non Caller	% Caller	% Non Collar	WOE	% Customers who have called in
0	20,070	20.1%	438	19,632	9%	20.6%	-82.67	2.2%
1-3	21,260	21.3%	596	20,664	12%	21.7%	-56.98	2.8%
4-10	12,394	12.4%	457	11,937	9%	12.5%	-28.66	3.7%
11-25	8,744	8.7%	350	8,393	7%	8.8%	-20.01	4.0%
26-40	6,920	6.9%	305	6,615	6%	7.0%	-10.18	4.4%
41-60	5,613	5.6%	314	5,299	6%	5.6%	15.08	5.6%
61-100	9,886	9.9%	651	9,235	13%	9.7%	32.42	6.6%
100-200	7,234	7.2%	575	6,659	12%	7.0%	52.69	7.9%
>200	7,879	7.9%	1,165	6,714	24%	7.1%	122.49	14.8%
Total	**1,00,000**	**100.0%**	**4,852**	**95,148**	**100%**	**100.0%**		**4.9%**

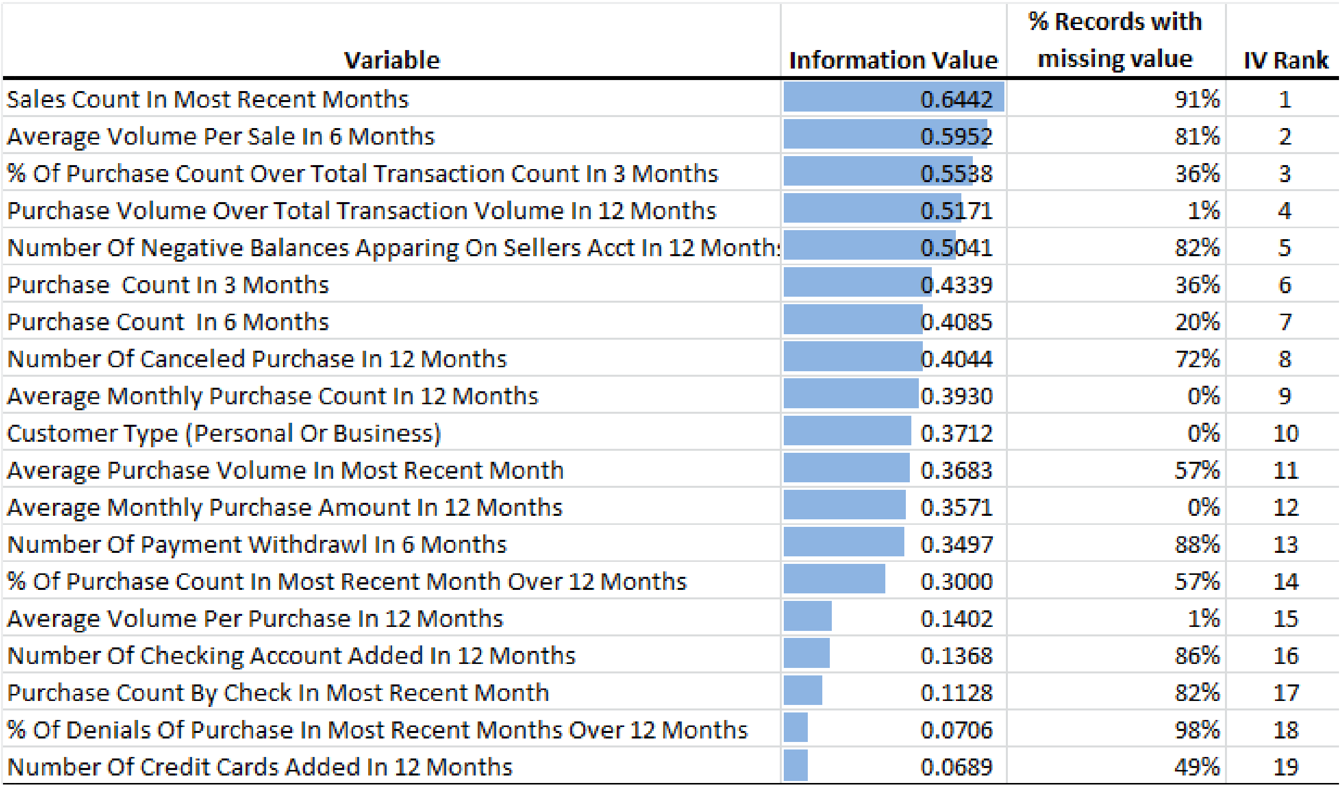
% Caller: number of callers of the tier/number of all callers
% nonCaller: number of noncallers of the tier/all noncallers

$$\mathrm{WOE} = \left[\ln\left(\frac{\%Caller_i}{\%nonCaller_i}\right)\right] \times 100$$

Information value (IV) tells the predictive power of a variable. IV is calculated using the following formula:

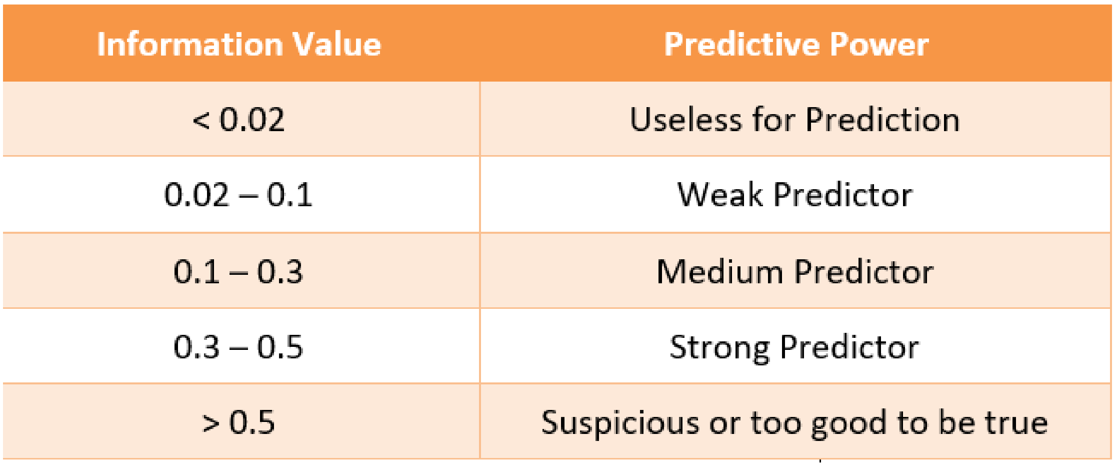
$$IV = \sum_{n=1}^{\infty} \left((\%Caller_i - \%nonCaller_i) \times \ln\left(\frac{\%Caller_i}{\%nonCaller_i}\right) \right)$$

The Following table has all the variables Information Value in Descending order



Variable	Information Value	% Records with missing value	IV Rank
Sales Count In Most Recent Months	0.6442	91%	1
Average Volume Per Sale In 6 Months	0.5952	81%	2
% Of Purchase Count Over Total Transaction Count In 3 Months	0.5538	36%	3
Purchase Volume Over Total Transaction Volume In 12 Months	0.5171	1%	4
Number Of Negative Balances Apparing On Sellers Acct In 12 Month:	0.5041	82%	5
Purchase Count In 3 Months	0.4339	36%	6
Purchase Count In 6 Months	0.4085	20%	7
Number Of Canceled Purchase In 12 Months	0.4044	72%	8
Average Monthly Purchase Count In 12 Months	0.3930	0%	9
Customer Type (Personal Or Business)	0.3712	0%	10
Average Purchase Volume In Most Recent Month	0.3683	57%	11
Average Monthly Purchase Amount In 12 Months	0.3571	0%	12
Number Of Payment Withdrawl In 6 Months	0.3497	88%	13
% Of Purchase Count In Most Recent Month Over 12 Months	0.3000	57%	14
Average Volume Per Purchase In 12 Months	0.1402	1%	15
Number Of Checking Account Added In 12 Months	0.1368	86%	16
Purchase Count By Check In Most Recent Month	0.1128	82%	17
% Of Denials Of Purchase In Most Recent Months Over 12 Months	0.0706	98%	18
Number Of Credit Cards Added In 12 Months	0.0689	49%	19

Now the question arises how to interpret the IV ? The Below table states the rule of thumb for IV interpretation.



Information Value	Predictive Power
< 0.02	Useless for Prediction
0.02 – 0.1	Weak Predictor
0.1 – 0.3	Medium Predictor
0.3 – 0.5	Strong Predictor
> 0.5	Suspicious or too good to be true

Mostly, Variables with medium and strong predictive power are selected for model building. This is one of the most efficient technique for variable reduction.

Business Scenario

In one of exercises, our team had 200 variables initially in the dataset. It's not viable to build a model putting all the 200 variables. So we used the above discussed techniques to reduce down to a manageable number for model building. After analyzing all the variable reduction techniques result and as per business understanding we came down to 50 variables for model building. 50 variables is also a high number but manageable for building a model.

All the 50 variables are put in to the model building process, various selection techniques i.e. forward/ backward/step-wise are deployed while building the model. After doing this exercise, let say we came down to 15 variables which are statistically significant. Ideally a good model should not have more than 10 predictor variables. So now we will reduce down from 15 to 10 on the basis of business understanding as to which variable are highly explainable both statistically and as per business implication or we can choose the top 10 variables on the basis of the wald chi-square statistic . Now we form different combination of variables to reduce down from 15 to 10 variables and build model for each set of variables which gives us 'N' number of models. Now these 'N' models will be relatively compared with eachother basis the model performance parameters i.e. lift chart, K-S Statistics, Hosmer and Lemeshow test, Gini coefficient. On the basis of Models performancecomparison, we select the best predictive model. This best model has the final set of predictor variables which are highly significant statistically and also as per business implications.

Conclusion

In this paper we have discussed many vital techniques for variable reduction. Each technique has their own relevance and importance. So, variable selection is an art as well as science. Use of variable selections techniques differ with respect to different business scenarios and also depends upon the intellectual decision making of an analyst.

Valiance is a global AI & Data analytics firm helping clients build cutting edge digital transformation solutions for operational excellence, process improvement and enabling new business models. We are present in India & US with a delivery team comprising 120+ professionals across Machine Learning, Data Engineering & Cloud expertise. We are also partners with hyperscalers like AWS & GCP and building new partnerships with data platforms. Our client portfolio comprises global indian conglomerates, fortune 1000 companies and digitally native companies. Some notable mentions are Rio-Tinto, Coke, Aditya Birla Group etc

Key Technology Partners aws Google Cloud Azure

WRITE TO US : info@valiancesolutions.com VISIT US : www.valiancesolutions.com FOLLOW US :

Artificial Intelligence Machine Learning | Data Engineering | Data lake | Data warehousing | IoT Analytics