CHAPTER 12
PREDICTIVE POWER AND VARIABLE REDUCTION I

12.1 INTRODUCTION

The reduction of the number of candidate independent variables of a predictive model is a good practice according to Occam's razor (Mitchell 1997). It states that the best solution is the one that is the simplest. In the case of predictive modeling, it would be the model with the fewest predictors, which is known as a *parsimonious* model. To achieve that, we need to do the following two things.

1. Remove all variables expected to have small or no contribution to the model.
2. For the remaining *good* predictors, *if possible,* find a set of transformations that will reduce their number but at the same time keep all, or most, of the information in them.

To achieve the first task, we need to define a *metric,* or set of metrics, to assess the *predictive power* of a variable. As for the second task, we will need to define another metric to measure the *information content* of a variable. These two concepts are somehow related because it is not acceptable to say, or worse to discover, that a variable with no information content has a high predictive power!

The definition of a metric of predictive power requires a dependent variable to define the predictive aspect of the question. Various measures of predictive power differ in how they weigh the different errors and how we plan to use the variable in the model. For example, the correlation coefficient between a continuous dependent variable and a candidate continuous predictor may be appropriate when we plan to

use linear regression. This is because correlation measures the degree of linear association between continuous variables. Similarly, the Gini measure is more suitable with a decision tree model.

Defining the information content of one or more variables is more tricky. In the case of a set of continuous variables, the concept of information content is realized using the *covariance matrix*. In this case, keeping the information of a set of variables translates to keeping the variance in the data that is expressed by the covariance matrix. This is the basis of factor analysis and principal component analysis.

In addition, many predictive modeling algorithms have mechanisms for selection of variables. For example, linear and logistic regression models can be used to iteratively select the variables by inserting and removing the different possible predictors and testing the contribution of each variable. This is the basic idea of stepwise variable selection algorithms in regression models. However, it is recommended that before importing a large number of variables into the modeling software, the number of actual variables be reduced. This is particularly important when using a large training partition.

However, before we use the different metrics to measure the predictive power of the candidate variables, we should do some simple checks to eliminate those that are guaranteed to show low value. This step involves considering the removal of the following variables:

- Constant fields—that is, variables with a cardinality of one. These variables will certainly not contribute to any model.
- Variables with high content of *missing values* (say, more than 99%). These variables are almost identical to constant fields, except that the constant value in this case is the missing value.
- Categorical variables with high cardinality. These variables cannot be easily used in regular models because they result in overfitted models. A typical example of these variables is the postal (zip) code portion of an address. In order to use this type of variable, it is usually necessary to transform it into another form, such as the distance between two points on the map or the expected driving time/distance between two locations. Another option is to group the categories of the zip code to a higher level with a smaller cardinality.

In the remainder of this chapter, we discuss the methods and metrics used for variable reduction. However, we defer the bulk of the SAS implementation to Chapter 17, after all the details of the needed metrics and reduction methods have been presented in Chapters 13 through 16.

12.2 Metrics of Predictive Power

The definition of a metric of the predictive power for a variable assumes that we have a well-defined dependent variable. As mentioned in Chapter 2, in classification

Table 12.1 Common predictive power metrics.

	Variable Y		
Variable X	*Nominal*	*Ordinal*	*Continuous DV*
Nominal	X^2 G, E	r, r_s G, E	r, r_s F-test G, E
Ordinal		r, r_s G, E	r, r_s F-test G, E
Continuous			r, r_s F-test G, E

problems the dependent variable will be either categorical or binary, and in regression or estimation problems it will be continuous. All metrics used to define the predictive power of a variable depend on measuring the level of association between the dependent variable and the candidate predictor in question. Table 12.1 summarizes the most common metrics.

In the table, X^2 is the Pearson Chi-squared statistic, r is the Pearson correlation coefficient, r_s is the Spearman correlation coefficient, G is the Gini variance, and E is the Entropy variance. (We discuss these measures in full detail in Chapters 13 through 16). Furthermore, the table is symmetric. We display only the diagonal and off-diagonal elements.

The use of the correlation coefficients to measure the predictive power of a variable is probably the easiest and most tempting method. However, it has been shown that variables exhibiting low correlation could still play a significant role in the final model when combined with other variables. This interesting finding can be clearly demonstrated in the case of linear regression (Wickens 1995). Therefore, it is not recommended to rely solely on the calculation of the correlation coefficients, nor any other single metric, in the selection of the variables.

The conservative and recommended approach is to use several metrics to assess all the variables. Then we select *all* the variables that show significant contributions to the model, using *any* metric. In this way, we minimize the chance of erroneous elimination of a possibly useful variable.

12.3 Methods of Variable Reduction

As mentioned in the introduction section, our plan is to reduce the variables over two stages. First, we remove the variables that do not show good prospects for contributing to the planned model. Second, we reduce the groups of good variables to smaller sets of new variables, without losing too much of the information content.

In the second stage, two methods are commonly used. They are (1) Principal Component Analysis and (2) Factor Analysis. Principal Component Analysis (PCA) aims at finding a set of linear transformations of a set of *continuous* variables such that the resulting set contains *most* of the *variance* in the original set within the first few terms. Factor Analysis (FA), on the other hand, attempts to find a smaller set of *hidden* variables, *factors,* such that performing a set of linear transformations on these factors would lead to the current set of variables.

PCA and FA are two sides of the same idea, finding a set of linear transformations. The difference is that the result of PCA, the principal components, is unique for any set of variables. FA, as detailed in Chapter 16, could result in different sets of factors depending on the criterion used to define these factors.

Furthermore, the factors resulting from FA could sometimes be given business interpretation. This was the original objective of developing factor analysis—to uncover hidden variables that govern the observed phenomena. For example, in banking applications, one could conduct factor analysis with the objective of uncovering a hidden factor that represents the *wealth* of each customer, their *willingness* to adopt new products, and so on. However, proper interpretation of factors is, in general, difficult. Therefore, we focus on using factor analysis for the task of data reduction.

12.4 Variable Reduction: Before or During Modeling

Many modeling techniques, such as decision trees, regression analysis, and some implementations of neural networks, offer systematic methods for the reduction of variables while building the model. The reduction algorithms adopt a simple optimization heuristic by attempting to optimize some model performance metric, such as R^2 (coefficient of multiple determination) in regression, or *RMSE* (Root of Mean Square Error) in neural networks.

These methods could offer an alternative to systematic variable reduction, using the methods described in this chapter and in Chapter 17. In most real business applications, however, the number of variables in the final mining view is too large to effectively use the modeling features of variable reduction. Therefore, it is recommended that systematic variable reduction methods be adopted as part of the data preparation procedure, so that the modeling technique would focus on the final tuning and selection from the best set of variables.