
Article from

# Predictive Analytics and Futurism

June 2017
Issue 15

# Variable Selection in Predictive Modeling: Does it Really Matter?

**By Kailan Shang**

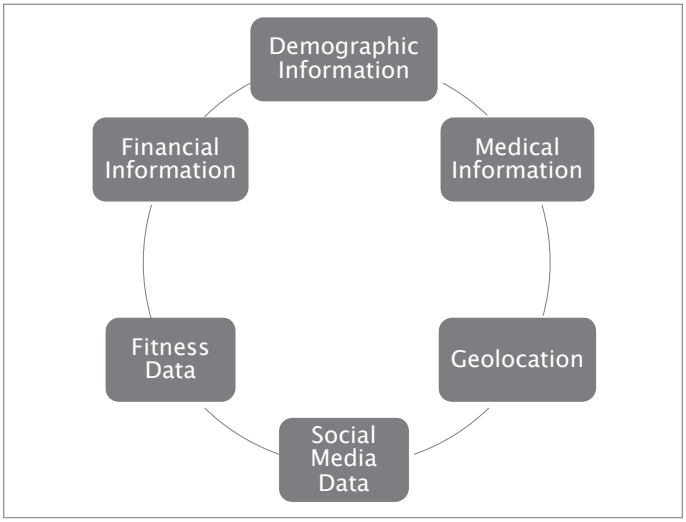
Many actuarial works have been expanded in the era of big data. Risk analyses are moving from the aggregate level to the individual level enabled by better data availability. For example, mortality risk can be assessed not only by traditional data such as age, gender, smoker/nonsmoker, occupation, face amount and basic medical information, but also new data, including location, detailed medical information, financial status, fitness data and even social media data. These new data sources can help us learn more about individuals or events that affect the mortality trend. In addition, some new data are categorical and cannot be used directly by predictive models like numerical data. For example, cancer patients have different tumor sites and medical treatments. An insurance client may participate in different types of sports. One categorical variable could become dozens of numerical variables, with each indicating the presence of a specific variable. The number of explanatory variables could easily exceed a few hundred.

## DO WE NEED VARIABLE SELECTION?

With so many variables, is it necessary to select a subset of variables with the best performance of prediction? For traditional predictive models used by actuaries, the answer is obviously positive. The robustness of linear regression models and generalized linear models (GLMs) can be low with the presence of collinearity caused by too many variables. The prediction results will be very sensitive to the input data. However, some machine learning models such as random forests and artificial neural networks (ANNs) were designed to handle large data input without prior assumption of the data relationship. Dimension reduction techniques such as principal component analysis (PCA) and autoencoder could also systematically reduce the number of explanatory variables. The needs for variable selection are less obvious for these models.

However, the benefits of variable selection go beyond model training and model selection. By selecting the best predictors, people can understand the most important relationships implied from the data. It is easier for people to assess these relationships at a small scale rather than being overwhelmed with hundreds of variables at the same time. Reducing the number of explanatory variables also decreases the chance of overfitting. Overfitting happens when too many variables are unintentionally used to explain the random noises instead of the relationships. The variance of prediction is large even though the accuracy of prediction may be high for the training data. Figure 2 illustrates an example of overfitting. A linear model with one explanatory variable X1 could explain the main relationship even though its accuracy is lower than a perfect matching nonlinear model with much more explanatory variables.

Figure 1
Data Sources for Individual Mortality Prediction




Overfitting can be overcome by analyzing the contribution of each variable to the prediction and removing variables with trivial contributions. Variable selection may not improve the model accuracy measured by the training data, but it can certainly improve the robustness of found relationships. Maintaining only the important variables in the predictive models also helps explain the model. The application of the model to new data will be more efficient. Less data collection, storage and calculation can be achieved by variable selection.

On the other hand, variable selection is challenging for big data. Will predictive models be able to identify important variables automatically? The answer is both yes and no. Predictive models are instrumental for identifying useful variables, but they are not always working in a desired way.

## USING PREDICTIVE MODELS

A few approaches can be used to select important variables by running multiple models. The forward approach starts from an empty model and adds one variable at a time. At each step, the variable with the biggest accuracy improvement is chosen. The

Figure 2
Overfitting Illustration

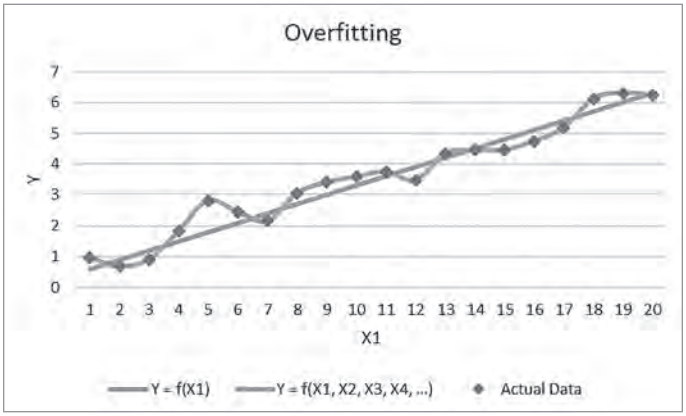
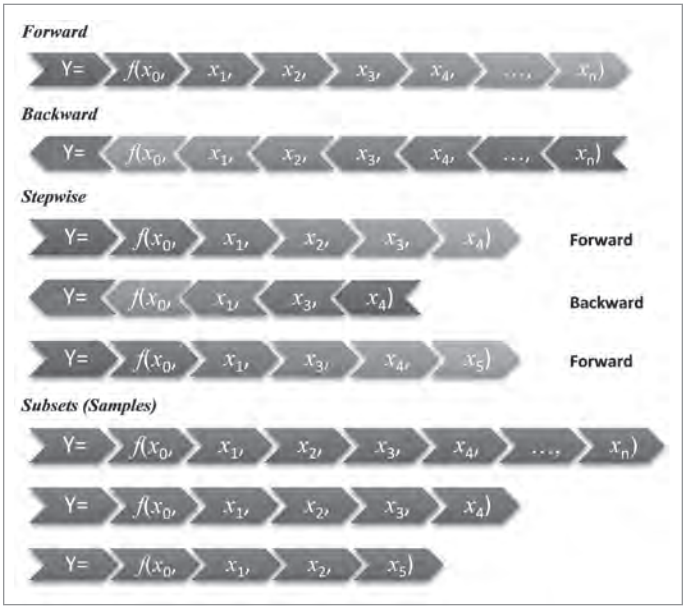



Figure 3
Variable Selection Methods




forward process ends when the model accuracy stops improving or the improvement is trivial. The backward approach starts from a full model with all variables and removes one variable at a time. At each step, the variable with the biggest negative impact or the smallest positive impact is removed, until the model accuracy stops improving or reaches the desired level. However, the problem with both the forward and backward approach is that the sequence of the explanatory variables matters. Adding a new variable to the model could change the importance of existing variables. The stepwise approach addresses this issue by combining the forward approach and the backward approach. At each step, an additional variable is added, and then the new model works backward to remove any existing variables that have a negative or trivial impact on model accuracy. Another more comprehensive yet costly approach is to iterate through all possible combination of explanatory variables and choose the subset with the smallest set of variables given that the model accuracy meets the target.

When applying these approaches, many measures can be used to represent model accuracy. The measures are used in two places: the target above which variable selection process will stop and the minimum positive improvement deciding whether a variable should be added or dropped. Table 1 lists a few measures for regression and/or classification models.

However, these four approaches are expensive given the number of models that need to be run. It could be very challenging for big data with many variables. Table 2 lists the maximum number of models that need to be trained to finish the variable selection process for each approach assuming $n$ explanatory variables. The actual number of models could be smaller than the maximum number because the process could stop once the target accuracy is achieved.

To reduce the burden of additional model training, variable selection can be done based on the result of the complete model with a couple of adjustments. After the model is trained, the importance of each variable can be measured to determine its contribution to the prediction. Variables are then selected based on their importance. Several adjustments to the modeling process can be made to address the issue of overfitting in one model training:

1. Collinearity/multicollinearity checking. Variables with high correlation, either positive or negative, can be reduced. If the absolute value of correlation coefficient exceeds a threshold such as 95 percent, one variable of the pair can be removed. For multicollinearity where one explanatory variable can be explained very well by other explanatory variables, the explanatory variable can be removed as well because its information can be provided by the remaining variables. Multicollinearity can be assessed using the variance inflation index (VIF). For an explanatory variable $x_i$, a linear regression can be run against other explanatory variables:

$x_i = \alpha + \beta_1 x_1 + \quad + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \quad + \beta_n x_n$. Its VIF is calculated as

$$\frac{\sum_{j=1}^{m} \left( x_i^j - \overline{x_i} \right)^2}{\sum_{j=1}^{m} \left( \hat{x}_i^j - x_i^j \right)^2}$$

Table 1
Variable Selection Measures

| | | |
|---|---|---|
| **Akaike Information Criterion (AIC)** | $2p - 2\log(likelihood)$ | All |
| **Bayesian Information Criterion (BIC)** | $plog(m) - 2\log(likelihood)$ | All |
| **Adjusted $R^2$** | $1 - \frac{\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2 / (n-p-1)}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2 / (n-1)}$ | Regression |
| **Mean Square Error (MSE)** | $\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2$ | Regression |
| **Mean Absolute Error (MAE)** | $\frac{1}{m}\sum_{i=1}^{m}\lvert Y_i - Y_i \rvert$ | Regression |
| **Precision** | $\frac{True\ Positive}{True\ Positive + False\ Positive}$ | Classification |
| **Recall** | $\frac{True\ Positive}{True\ Positive + False\ Negative}$ | Classification |
| **F-Measure** | $\frac{2 \bullet Precision \bullet Recall}{Precision + Recall}$ | Classification |

**Notes:**
$p$: number of variables
$m$: number of data records
$Y_i$: actual value of explained variable for the $i$th data record
$\hat{Y}_i$: predicted value of explained variable for the $i$th data record
$\bar{Y}_i$: average value of the explained variable

Table 2
Models under Four Variable Selection Approaches

| | | | | |
|---|---|---|---|---|
| **Maximum No. of Models** | $\frac{n(n+1)}{2}$ | $\frac{n(n+1)}{2}$ | $\mathbb{O}(n^3)$ | $2^n - 1$ |

where

*m:* number of data records

$x_i^j$ : the value of $x_i$ for the *j*th data record

$\overline{x_i}$ : the average value of $x_i$ for the *m* data records

$\hat{x}_i^j$ : the predicted value of $x_i$ based on other $(n-1)$ explanatory variables for the *j*th data record.

Kutner et al. (2004: 408–409) suggest that a VIF greater than 10 or the mean value of VIF for all explanatory variables greater than 1 indicates the existence of multicollinearity.

2. Data normalization. To facilitate variable importance measurement, explanatory data can be normalized into the same value range. By doing this, variable importance can be determined by the magnitude of the model parameter for that variable. For example, in an linear equation such as $Y = 0.5 + x_1 + 4x_2$. If both $x_1$ and $x_2$ are within the same value range, we may simply conclude that $x_2$ is four times more important than $x_1$ in the prediction. For nonlinear relationship, it is more complicated but normalization is still useful for a consistent comparison. Normalization can be done in different forms such as feature scaling and standard score:

   Feature scaling: $\frac{X - X_{min}}{X_{max} - X_{min}}$

   Standard score: $\frac{X -}{\sigma}$

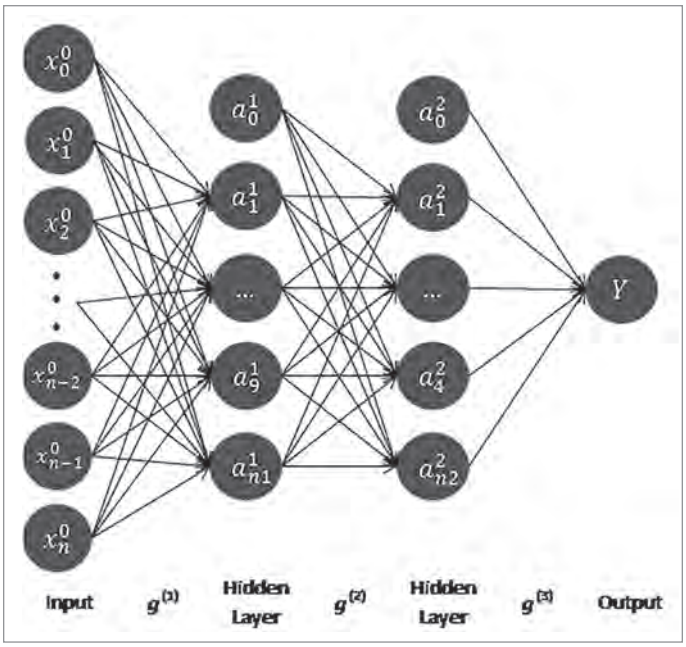
3. Regularization is often used in models that can handle many variables to address the issue of overfitting. By introducing the penalty for model complexity, it does not explicitly select variables in the model but limits the value of model parameters. For example, ridge regression intends to minimize the sum of squared errors and squared parameters. Parameter λ controls the weight of the penalty:

$$\min_{\beta} \sum_{j=1}^{m} \left( Y_j - \sum_{i=1}^{n} x_i^j \beta_i \right)^2 + \lambda \sum_{i=1}^{n} \beta_i^2$$

Normal regularization includes L1 regularization, which uses the sum of the absolute value of parameters, and L2 regularization, which uses the sum of the squared value of parameters, as in the ridge regression. Models such as random forest do not have model parameters for each variable. Other approaches are used for regularization such as controlling the maximum depth of the trees to avoid overfitting.

After all these adjustments, variable importance can be measured and used for variable selection. For model with normalized data, the absolute value of coefficients can be used for models like linear regression and GLMs to determine the relative importance of variables. For more complicated models, the calculation of relative importance is more complicated. For example, for an ANN model with two hidden layers, the impact of the explanatory variables is determined through three sets of parameters: $g^{(1)}$, $g^{(2)}$ and $g^{(3)}$, as illustrated in Figure 4.

Figure 4
ANN Model Structure




A possible measure is to consider the impact of the explanatory variable through three layers, including the two hidden layers and the output layer:

$$Imp(x_i) = \sum_{j=1}^{n1} \sum_{k=1}^{n2} \frac{\left|\theta_{ij}^0\right|}{\sum_{r=1}^{n} \left|\theta_{rj}^0\right|} \cdot \frac{\left|\theta_{jk}^1\right|}{\sum_{s=1}^{n1} \left|\theta_{sk}^1\right|} \cdot \frac{\left|\theta_{kY}^2\right|}{\sum_{t=1}^{n2} \left|\theta_{tY}^2\right|} \quad (19)$$

where

$x_i$ : The *i*th input variable

$n1$ : The number of neurons in the first hidden layer

$n2$ : The number of neurons in the second hidden layer

$n$ : The number of explanatory variables

$\theta_{ij}^0$ : The parameter that determines the weight of the *i*th input variable applied to the *j*th neuron in the first hidden layer

$\theta_{jk}^1$ : The parameter that determines the weight of the *j*th neuron in the first hidden layer applied to the *k*th neuron in the second hidden layer

$\theta_{kY}^2$ : The parameter that determines the weight of the *k*th neuron in the second hidden layer applied to the output variable *Y*

This measure also has its disadvantages because it cannot tell whether the relationship is positive or negative. It also does not consider the specific function used to link the layers.

For tree-type models like random forests, the measurement of variable importance is different and even more complicated. A possible measure is the Gini importance measured by the improvement of the Gini impurity index. The Gini index is defined as

$$G(T) = \sum_{i=1}^{n} p_i (1 - p_i)$$

where

$p_i$ is the probability that the data belongs to category *i*

*n* is the number of categories in the data

*T* is the data set based on which Gini index is calculated

For each split based on the variable, the Gini importance is measured as the reduction in the Gini index:

$$Imp(x_i) = n(T)G(T) - n(T_L)G(T_L) - n(T_R)G(T_R)$$

where

$x_i$ is the variable for the split

$T_L$ is the data subgroup of the split's left branch

$T_R$ is the data subgroup of the split's right branch

*n* is the number of data points in the data set

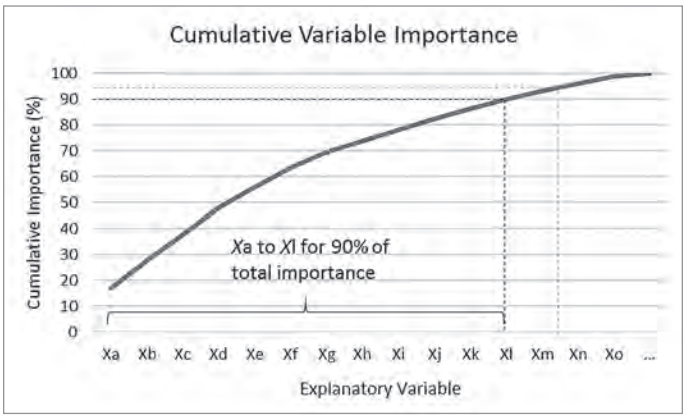
*p* is the portion of the data subgroup in the data set before splitting

If the variable is used in multiple splits, the Gini importance is aggregated for the variable. For random forests with multiple trees, the mean Gini importance across all trees can be used to measure variable importance.

Permutation importance can also be used to measure variable importance in tree models. The prediction is revised by permutating the value of the variable, and the loss of prediction accuracy is used as the importance measure for that variable. When variables are highly correlated, conditional permutation can be used to maintain the correlation. However, this is less of a concern after collinearity/multicollinearity checking.

After the variable importance is calculated, the top variables can be selected for future prediction. The threshold can be set based on a specified portion of total importance that selected variables explain in aggregate. Figure 5 illustrates the variable selection based on the cumulative variable importance. Variable importance is scaled so that the total importance is 100 percent.

Figure 5
Variable Selection Based on Variable Importance




## THE ROLE OF EXPERT OPINION

Although using predictive models to automatically search for important variables is a convenient and consistent approach, human judgments are needed at various stages of the process. At the initial stage, explanatory variables need to be screened one by one to assess their relevance to the explained variable. Both a blacklist and a whitelist of the variables can be created. If strong evidence exists for the irrelevance of an explanatory variable, the variable can be added to the blacklist and removed from the entire process. On the contrary, for variables that are believed to have a strong relationship with the explained variable, they can

be added to the whitelist and kept in the model. In the collinearity analysis, when a pair of variables are found highly correlated, human judgment is also needed to decide which one is more likely the root cause and should be retained in the analysis.

After the variable selection finishes, the reasonableness of the relationships derived from the data needs to be assessed. Sometimes even if the model accuracy is satisfactory, the relationship could be inconsistent with past experience, scientific findings and common sense. Additional work needs to be done to before accepting or rejecting the relationships. More data collection, model adjustments and different variable selections could be triggered by human judgment.

## CONCLUSION

Although many models can address the overfitting issues caused by too many variables by regularization, variable selection is still meaningful. The model and data are more parsimonious, and it is easy for people to assess, understand and explain the relationships derived from the data. Variable selection can be done through either multiple models or measures based on the complete model with adjustments. Measures might be complicated and different depending on the model, but they are computationally cheaper than multiple model runs. Human judgment is also important in the process of variable selection to incorporate expert opinions based on existing knowledge and experience.

Kailan Shang, FSA, CFA, PRM, SCJP, is managing director of Swin Solutions Inc. in Kitchener, Ontario. He can be reached at *kailan.shang@swinsolutions.com.*

### REFERENCE

Kutner, Michael H., Christopher J. Nachtsheim, John Neter and William Li. 2004. *Applied Linear Regression Models*. New York: McGraw-Hill.