SELECTPRED

Data mining problems often involve hundreds, or even thousands, of variables. As a result, the majority of time and effort spent in the model-building process involves examining which variables to include in the model. Fitting a computationally intensive model to a set of variables this large may require more time than is practical.

Predictor selection allows the variable set to be reduced in size, creating a more manageable set of attributes for modeling. Adding predictor selection to the analytical process has several benefits:

- Simplifies and narrows the scope of the variables essential to building a predictive model.
- Minimizes the computational time and memory requirements for building a predictive model because focus can be directed to a subset of predictors.
- Leads to more accurate and/or more parsimonious models.
- Reduces the time for generating scores because the predictive model is based upon only a subset of predictors.

Screening

This step removes variables and cases that do not provide useful information for prediction and issues warnings about variables that may not be useful.

The following variables are removed:

- Variables that have all missing values.
- Variables that have all constant values.
- Variables that represent case ID.

The following cases are removed:

- Cases that have missing target value.
- Cases that have missing values in all its predictors.

The following variables are removed based on user settings:

- Variables that have more than m_1% missing values.
- Categorical variables that have a single category counting for more than m_2% cases.
- Continuous variables that have standard deviation < m_3%.
- Continuous variables that have a coefficient of variation |CV| < m_4%. CV=standard deviation / mean.
- Categorical variables that have a number of categories greater than m_5% of the cases.

Values m_1, m_2, m_3, m_4, and m_5 are user-controlled parameters.

Ranking Predictors

This step considers one predictor at a time to see how well each predictor alone predicts the target variable. The predictors are ranked according to a user-specified criterion. Available criteria depend on the measurement levels of the target and predictor.

Categorical Target

All Categorical Predictors

The following notation applies:

X	The predictor under consideration with I categories.
Y	Target variable with J categories.
N	Total number of cases.
N_{ij}	The number of cases with X = i and Y = j.
$N_{i\cdot}$	The number of cases with X = i. $N_{i\cdot} = \sum_{j=1}^{J} N_{ij}$
$N_{\cdot j}$	The number of cases with Y = j. $N_{\cdot j} = \sum_{i=1}^{I} N_{ij}$

The above notations are based on non-missing pairs of (X, Y). Hence J, N and $N \cdot j$ may be different for different predictors.

P-value based on Pearson's Chi-square

Pearson's Chi-square is a test of independence between X and Y that involves the difference between the observed and expected frequencies. The expected cell frequencies under the null hypothesis of independence are estimated by $\hat{N}_{ij} = N_{i\cdot}N_{\cdot j}/N$. Under the null hypothesis, Pearson's Chi-square converges asymptotically to a Chi-square distribution χ^2_d with degrees of freedom d = (I−1)(J−1).

The p-value based on Pearson's Chi-square X^2 is calculated by p-value= Prob($\chi^2_d > X^2$), where:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(N_{ij} - \hat{N}_{ij}\right)^2 / \hat{N}_{ij}$$

Predictors are ranked by the following rules.

1. Sort the predictors by p-value in the ascending order
2. If ties occur, sort by Chi-square in descending order.

3. If ties still occur, sort by degree of freedom d in ascending order.
4. If ties still occur, sort by the data file order.

P-value based on Likelihood Ratio Chi-square

The likelihood ratio Chi-square is a test of independence between X and Y that involves the ratio between the observed and expected frequencies. The expected cell frequencies under the null hypothesis of independence are estimated by $\hat{N}_{ij} = N_{i.}N_{.j}/N$. Under the null hypothesis, the likelihood ratio Chi-square converges asymptotically to a Chi-square distribution χ^2_d with degrees of freedom d = (I−1)(J−1).

The p-value based on likelihood ratio Chi-square G^2 is calculated by p-value=Prob(χ^2_d> G^2), where:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} G_{ij}^2, \text{ with } G_{ij}^2 = \begin{cases} N_{ij}\ln\left(N_{ij}/\hat{N}_{ij}\right) & N_{ij} > 0 \\ 0 & \text{else} \end{cases}$$

Predictors are ranked according to the same rules as those for the p-value based on Pearson's Chi-square.

Cramer's V

Cramer's V is a measure of association, between 0 and 1, based upon Pearson's Chi-square. It is defined as:

$$V = \left(\frac{X^2}{N(\min\{I,J\}-1)}\right)^{1/2}$$

Predictors are ranked by the following rules:

1. Sort predictors by Cramer's V in descending order.
2. If ties occur, sort by Chi-square in descending order.
3. If ties still occur, sort by data file order.

Lambda

Lambda is a measure of association which reflects the proportional reduction in error when values of the independent variable are used to predict values of the dependent variable. A value of 1 means that the independent variable perfectly predicts the dependent variable. A value of 0 means that the independent variable is no help in predicting the dependent variable. It is computed as:

$$\lambda(Y|X) = \frac{\sum_i \max_j(N_{ij}) - \max_j(N_{.j})}{N - \max_j(N_{.j})}$$

Predictors are ranked by the following rules:

1. Sort predictors by Lambda in descending order.

2. If ties occur, sort by I in ascending order.
3. If ties still occur, sort by data file order.

All Continuous Predictors

If all predictors are continuous, p-values based on F-statistics are used. The idea is to perform a one-way ANOVA F-test for each continuous predictor; this tests if all the different classes of Y have the same mean of X.

The following notation applies:

N_j	The number of cases with Y = j.
$\overline{x}_j$	The sample mean of predictor X for target class Y = j.
s_j^2	The sample variance of predictor X for target class Y = j. $s_j^2 = \sum_{i=1}^{N_j} (x_{ij} - \overline{x}_j)^2/(N_j - 1)$
$\overline{\overline{x}}$	The grand mean of predictor X. $\overline{\overline{x}} = \sum_{j=1}^{J} N_j \overline{x}_j / N$

The above notations are based on non-missing pairs of (X, Y).

P-value based on F-statistics

The p-value based on F-statistics is calculated by p-value= Prob{F(J−1, N−J) > F}, where

$$F = \frac{\sum_{j=1}^{J} N_j \left(\overline{x}_j - \overline{\overline{x}}\right)^2/(J-1)}{\sum_{j=1}^{J} (N_j - 1)\, s_j^2/(N-J)}$$

and F(J-1, N-J) is a random variable follows a F-distribution with degrees of freedom J−1 and N−J. If the denominator for a predictor is zero, set the p-value=0 for the predictor.

Predictors are ranked by the following rules:

1. Sort predictors by p-value in ascending order.
2. If ties occur, sort by F in descending order.
3. If ties still occur, sort by N in descending order.
4. If ties still occur, sort by the data file order.

Mixed Type Predictors

If some predictors are continuous and some are categorical, the criterion for continuous predictors is still the p-value based on F-statistics, while the available criteria for categorical predictors are restricted to the p-value based on Pearson's Chi-square or the p-value based on the likelihood ratio Chi-square. These p-values are comparable and therefore can be used to rank the predictors.

Predictors are ranked by the following rules:

1. Sort predictors by p-value in ascending order.
2. If ties occur, follow the rules for breaking ties among all categorical and all continuous predictors separately, then sort these two groups (categorical predictor group and continuous predictor group) by the data file order of their first predictors.

Continuous Target

All Categorical Predictors

If all predictors are categorical and the target is continuous, p-values based on F-statistics are used. The idea is to perform a one-way ANOVA F-test for the continuous target using each categorical predictor as a factor; this tests if all different classes of X have the same mean of Y.

The following notation applies:

X	The categorical predictor under consideration with I categories.
Y	The continuous target variable. y_{ij} represents the value of the continuous target for the j^{th} case with X=i.
N_i	The number of cases with X=i.
$\overline{y}_i$	The sample mean of target Y in predictor category X = i.
$s(y)_i^2$	The sample variance of target Y for predictor category X=i. $s(y)i2 = \sum_{j=1}^{N_i} (y_{ij} - \overline{y}_i)^2/(N_i - 1)$
$\overline{\overline{y}}$	The grand mean of target Y. $\overline{\overline{y}} = \Sigma_{i=1}^{I} N_i \overline{y}_i / N$

The above notations are based on non-missing pairs of (X, Y).

The p-value based on F-statistic is p-value= Prob{F(I−1, N−I) > F}, where

$$F = \frac{\sum_{i=1}^{I} N_i \left(\overline{y}_i - \overline{\overline{y}}\right)^2 / (I-1)}{\sum_{i=1}^{I} (N_i - 1)\, s(y)_i^2 / (N-I)}$$

in which F(I−1, N−I) is a random variable that follows a F-distribution with degrees of freedom I−1 and N−I. When the denominator of the above formula is zero for a given categorical predictor X, set the p-value = 0 for that predictor.

Predictors are ranked by the following rules:

1. Sort predictors by p-value in ascending order.
2. If ties occur, sort by F in descending order.
3. If ties still occur, sort by N in descending order.
4. If ties still occur, sort by the data file order.

All Continuous Predictors

If all predictors are continuous and target is continuous, p-values are based on the asymptotic t-distribution of a transformation t on the Pearson correlation coefficient r.

The following notation applies:

X	The continuous predictor under consideration.
Y	The continuous target variable.
$\overline{x} = \Sigma_{i=1}^{N} x_i / N$	The sample mean of predictor variable X.
$\overline{y} = \Sigma_{i=1}^{N} y_i / N$	The sample mean of target Y.
$s(x)^2$	The sample variance of predictor variable X.
$s(y)^2$	The sample variance of target variable Y.

The above notations are based on non-missing pairs of (X, Y).

The Pearson correlation coefficient r is:

$$r = \frac{\Sigma_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) / (N-1)}{\sqrt{s(x)^2 s(y)^2}}$$

The transformation t on r is given by

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

Under the null hypothesis that the population Pearson correlation coefficient $\rho = 0$, the p-value is calculated as

$$p - \text{value} = \begin{cases} 0 & \text{if } r^2 = 1 \\ 2\ \text{Prob}\{T > |t|\} & \text{else} \end{cases}$$

in which T is a random variable that follows a t-distribution with N−2 degrees of freedom. The p-value based on the Pearson correlation coefficient is a test of a linear relationship between X and Y. If there is some non-linear relationship between X and Y, the test may fail to catch it.

Predictors are ranked by the following rules:

1. Sort predictors by p-value in ascending order.
2. If ties occur in, sort by r^2 in descending order.
3. If ties still occur, sort by N in descending order.
4. If ties still occur, sort by the data file order.

Mixed Type Predictors

If some predictors are continuous and some are categorical in the data set, the criterion for continuous predictors is still based on the p-value value from a transformation and that for categorical predictors from the F-statistics.

Predictors are ranked by the following rules:

1. Sort predictors by p-value in ascending order.
2. If ties occur, follow the rules for breaking ties among all categorical and all continuous predictors separately, then sort these two groups (categorical predictor group and continuous predictor group) by the data file order of their first predictors.

Selecting Predictors

If the length of the predictor list has not been prespecified, the following formula provides an automatic approach to determine the length of the list:

Let L_0 be the total number of predictors under study. The length of the list L may be determined by:

$$L = \left[\min\left(\max\left(30, 2\sqrt{L_0}\right), L_0\right)\right]$$

where [x] is the closest integer of x. The following table illustrates the length L of the list for different value of the total number of predictors L_0.

L_0	L	L/L_0(%)
10	10	100.00%
15	15	100.00%

L_0	L	L/L_0(%)
20	20	100.00%
25	25	100.00%
30	30	100.00%
40	30	75.00%
50	30	60.00%
60	30	50.00%
100	30	30.00%
500	45	9.00%
1000	63	6.30%
1500	77	5.13%
2000	89	4.45%
5000	141	2.82%
10,000	200	2.00%
20,000	283	1.42%
50,000	447	0.89%