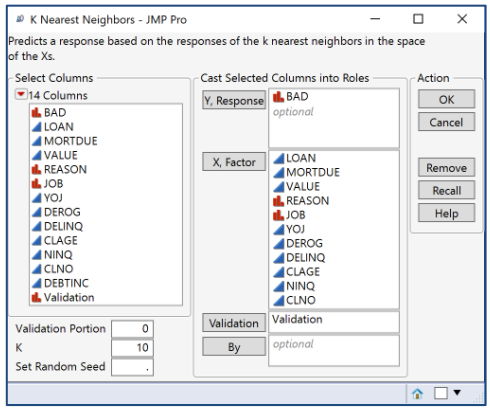
K Nearest Neighbors

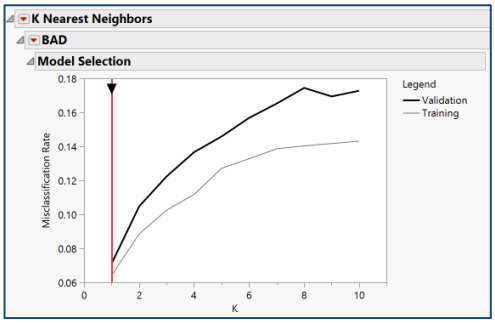
JMP PRO Use this predictive modeling technique to predict (classify) a categorical (nominal or ordinal) response variable or predict the value of a continuous response variable as a function of candidate categorial and/or continuous predictor variables. K Nearest Neighbors make predictions for an observation by utilizing the outcomes of other observations that are similar to it.

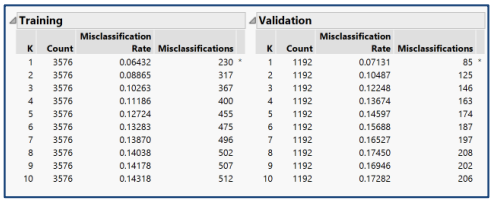
K Nearest Neighbors

Equity.jmp (Help > Sample Data Library)

1. From an open JMP® table, select **Analyze > Predictive Modeling > K Nearest Neighbors**.
2. Select a categorical or continuous response variable from **Select Columns** and click **Y, Response**. Here, we illustrate using a categorial response variable.
3. Select candidate predictor variables and click **X, Factor**.
4. If desired, enter the **Validation Portion** or select a validation column and click **Validation**.
5. Click **OK**. JMP displays:







Training

K	Count	Misclassification Rate	Misclassifications
1	3576	0.06432	230 *
2	3576	0.08865	317
3	3576	0.10263	367
4	3576	0.11186	400
5	3576	0.12724	455
6	3576	0.13283	475
7	3576	0.13870	496
8	3576	0.14038	502
9	3576	0.14178	507
10	3576	0.14318	512

Validation

K	Count	Misclassification Rate	Misclassifications
1	1192	0.07131	85 *
2	1192	0.10487	125
3	1192	0.12248	146
4	1192	0.13674	163
5	1192	0.14597	174
6	1192	0.15688	187
7	1192	0.16527	197
8	1192	0.17450	208
9	1192	0.16946	202
10	1192	0.17282	206

Confusion Matrix for Best K=1

Training

Actual BAD	Predicted Count Good Risk	Bad Risk
Good Risk	2894	11
Bad Risk	219	452

Validation

Actual BAD	Predicted Count Good Risk	Bad Risk
Good Risk	917	0
Bad Risk	85	190

- Graph and table showing the misclassification rates and counts across a range of values for K.
- Confusion Matrix detailing the classification performance for the value of K with the smallest misclassification rate.
- Mosaic plots (not shown here) which graphically shows the values in the confusion matrix.

Resuts of the K Nearest Neighbors to predict the risk level (Bad/Good) from the 5,960 customers:

- There are 3,576 observations in the Training Data. The misclassifcation rate is the lowest when the prediction is based on only 1 nearest neighbor: 230/3576 = 6% were misclassified. Note that the misclassification rate increases as the number of nearest neighbors increase. Of these total misclassifications, 11/(2894+11) = 0.4% of the Good Risk observations were misclassified as Bad Risk. 219/(219+452) = 33% of the Bad Risk observations were misclassifed as Good Risk.
- There are 1,192 observations in the Validation Data. The misclassifcation rate is the lowest when the prediction is based on only 1 nearest neighbor: 85/1192 = 7% were misclassified. Of these total misclassifications, 0/(917+0) = 0% of the Good Risk observations were misclassified as Bad Risk. 85/(85+190) = 31% of the Bad Risk observations were misclassifed as Good Risk.

Notes:

Additional options, such as **Lift Curves**, **Saving Predicteds**, **Save Prediction Formula**, and **Publish Prediction Formula** are accessible from the **red triangle** near the top next to the response variable name.

For more information on using K Nearest Neighbors, see the book ***Predictive and Specialized Models*** (under **Help > Books**) or search for “k nearest neighbors” in the JMP Help.