

K Nearest Neighbors

JMP PRO Use a proximity-based algorithm to predict a categorical outcome (classify) or prediction the value of a continuous outcome for new observations based upon the outcomes of similar observations (i.e., their nearest neighbors).

K Nearest Neighbors

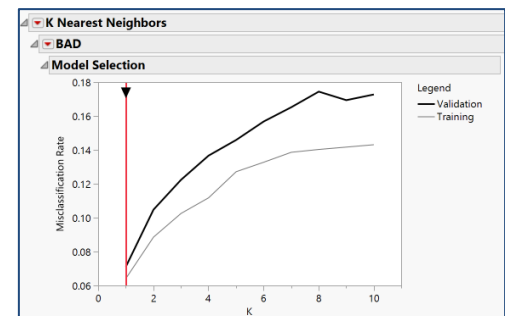
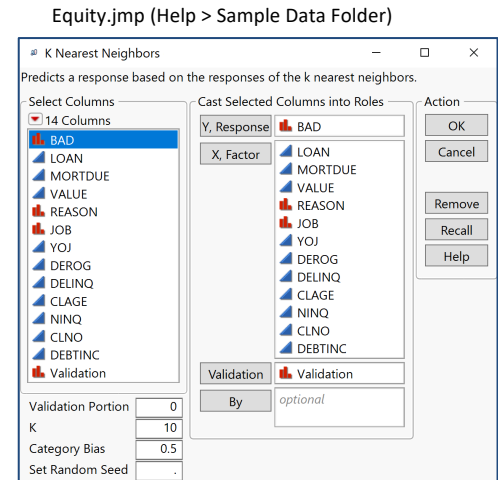
1. From an open JMP® table, select **Analyze > Predictive Modeling > K Nearest Neighbors**.
2. Select a categorical or continuous response variable from **Select Columns** and click **Y, Response**. Here, we illustrate using a categorical response variable.
3. Select candidate predictor variables and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

JMP displays:

- 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.

Results of the K Nearest Neighbors to predict the risk level (Bad/Good)

- There are 1,192 observations in the Validation Data. The misclassification rate is the lowest when the prediction is based on only 1 nearest neighbor: 85/1,192 (7.1%) were misclassified. Note that the misclassification rate increases as the number of nearest neighbors increase. Of these total misclassifications in the Validation Data, $3/(3+194) = 0.3\%$ of the Good Risk observations were misclassified as Bad Risk. $80/(80+195) = 29\%$ of the Bad Risk observations were misclassified as Good Risk.
- There were 1,192 observation set aside for the Test Data (Results not shown). The total misclassification rate for these observations using 1 nearest neighbor is 6.9%. 0.4% of the Good Risk observations were misclassified as Bad Risk and 3.2% of the Bad Risk observations were misclassified as Good Risk. These results are often considered to produce the most accurate estimate of what the misclassification rate will be in future data as these observations were not part of the model training nor selection process.



Training					Validation				
K	Count	Misclassification Rate	Misclassifications		K	Count	Misclassification Rate	Misclassifications	
1	3576	0.06432	230		1	1192	0.07131	85	
2	3576	0.08865	317		2	1192	0.10487	125	
3	3576	0.10263	367		3	1192	0.12248	146	
4	3576	0.11186	400		4	1192	0.13674	163	
5	3576	0.12724	455		5	1192	0.14597	174	
6	3576	0.13263	475		6	1192	0.15688	187	
7	3576	0.13870	496		7	1192	0.16527	197	
8	3576	0.14038	502		8	1192	0.17450	208	
9	3576	0.14178	507		9	1192	0.16946	202	
10	3576	0.14318	512		10	1192	0.17282	206	

Confusion Matrix for Best K=1									
Training					Validation				
Actual	Predicted Count				Actual	Predicted Count			
BAD	Good Risk	Bad Risk			BAD	Good Risk	Bad Risk		
Good Risk	2883	22			Good Risk	914	3		
Bad Risk	234	437			Bad Risk	80	195		
Actual	Predicted Rate				Actual	Predicted Rate			
BAD	Good Risk	Bad Risk			BAD	Good Risk	Bad Risk		
Good Risk	0.992	0.008			Good Risk	0.997	0.003		
Bad Risk	0.349	0.651			Bad Risk	0.291	0.709		

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.

Visit **Predictive and Specialized Models > K Nearest Neighbors** in **JMP Help** to learn more.