Comparing Predictive Model Performance with Confidence Curves

Bryan Fricke

JMP, bryan.fricke@jmp.com

Russ Wolfinger

JMP, russ.wolfinger@jmp.com

ABSTRACT

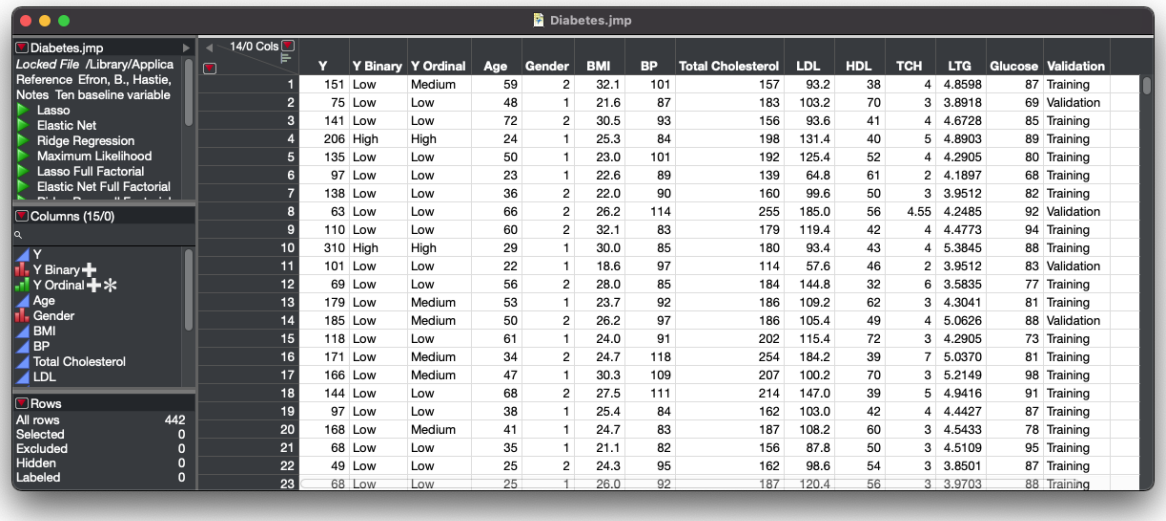
Repeated k-fold cross-validation is commonly used to evaluate the performance of predictive models. The problem is, how do you know when a difference in performance is sufficiently large to declare one model better than another? Typically, null hypothesis significance testing (NHST) is used to determine if the differences between predictive models are "significant", although the usefulness of NHST has been debated extensively in the statistics literature in recent years. In this paper, we discuss problems associated with NHST and present an alternative known as confidence curves, which has been developed as a new JMP Add-In that operates directly on the results generated from JMP Pro's Model Screening platform.

Keywords: confidence curve, significance test, p value, multiple comparisons, data visualization

THE MODEL SCREENING PLATFORM

The Model Screening platform introduced in JMP Pro 16 allows you to evaluate the performance of multiple predictive models using cross-validation in one setting. Prior to JMP Pro 16.0, you would need to launch the platforms for each predictive modeling method one at a time prior to using the Model Comparison platform to evaluate performance differences.

To see how the Model Screening platform works, start JMP Pro 16 and load the Diabetes data table available in the JMP Sample Data library. The first 3 columns in the data table, which is shown in Figure 1, represent disease progression in continuous, binary, and ordinal forms. In this paper, we will use the continuous column named **Y** as the response variable. We also use all the columns from **Age** to **Glucose** as predictors, or factors. We won't be using the **Validation** column since we leverage the cross-validation option built into the Model Screening platform.



	Y	Y Binary	Y Ordinal	Age	Gender	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose	Validation
1	151	Low	Medium	59	2	32.1	101	157	93.2	38	4	4.8598	87	Training
2	75	Low	Low	48	1	21.6	87	183	103.2	70	3	3.8918	69	Validation
3	141	Low	Low	72	2	30.5	93	156	93.6	41	4	4.6728	85	Training
4	206	High	High	24	1	25.3	84	198	131.4	40	5	4.8903	89	Training
5	135	Low	Low	50	1	23.0	101	192	125.4	52	4	4.2905	80	Training
6	97	Low	Low	23	1	22.6	89	139	64.8	61	2	4.1897	68	Training
7	138	Low	Low	36	2	22.0	90	160	99.6	50	3	3.9512	82	Training
8	63	Low	Low	66	2	26.2	114	255	185.0	56	4.55	4.2485	92	Validation
9	110	Low	Low	60	2	32.1	83	179	119.4	42	4	4.4773	94	Training
10	310	High	High	29	1	30.0	85	180	93.4	43	4	5.3845	88	Training
11	101	Low	Low	22	1	18.6	97	114	57.6	46	2	3.9512	83	Validation
12	69	Low	Low	56	2	28.0	85	184	144.8	32	6	3.5835	77	Training
13	179	Low	Medium	53	1	23.7	92	186	109.2	62	3	4.3041	81	Training
14	185	Low	Medium	50	2	26.2	97	186	105.4	49	4	5.0626	88	Validation
15	118	Low	Low	61	1	24.0	91	202	115.4	72	3	4.2905	73	Training
16	171	Low	Medium	34	2	24.7	118	254	184.2	39	7	5.0370	81	Training
17	166	Low	Medium	47	1	30.3	109	207	100.2	70	3	5.2149	98	Training
18	144	Low	Low	68	2	27.5	111	214	147.0	39	5	4.9416	91	Training
19	97	Low	Low	38	1	25.4	84	162	103.0	42	4	4.4427	87	Training
20	168	Low	Medium	41	1	24.7	83	187	108.2	60	3	4.5433	78	Training
21	68	Low	Low	35	1	21.1	82	156	87.8	50	3	4.5109	95	Training
22	49	Low	Low	25	2	24.3	95	162	98.6	54	3	3.8501	87	Training
23	68	Low	Low	25	1	26.0	92	187	120.4	56	3	3.9703	88	Training

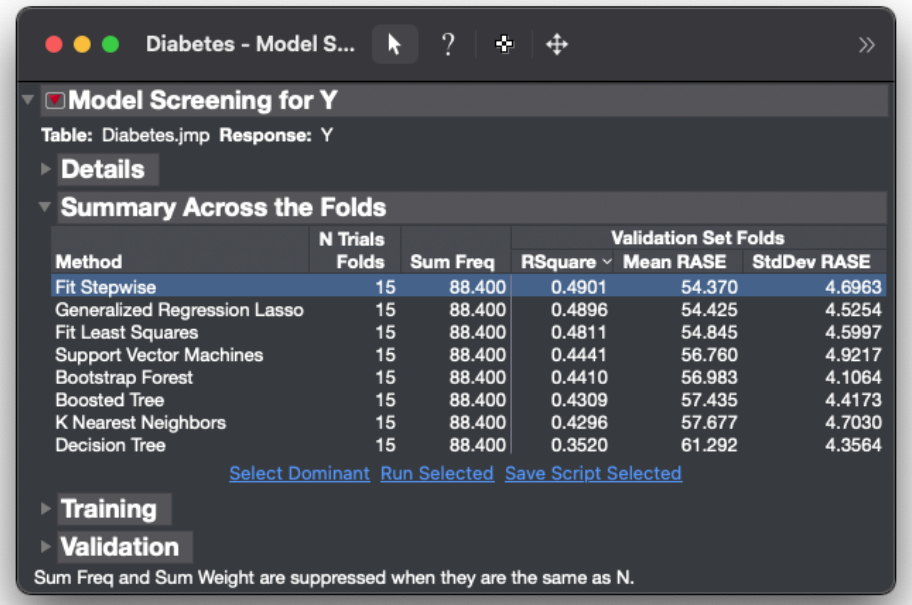
Figure 1. Diabetes data table

Now that the table is loaded, choose **Model Screening** from the **Analyze → Predictive Modeling** menu. JMP responds by displaying the **Model Screening** dialog. In the dialog, perform the following actions (see Figure 2):

1. Put the Y column in the **Y, Response** role.
2. Put all the columns from Age to Glucose in the **X, Factor** role.
3. Type **1234** in the **Set Random Seed** input box.
4. Select the checkbox next to **K Fold Crossvalidation**.
5. Type 5 into input box **K** next to **K Fold Crossvalidation**.
6. Type 3 into the input box next to **Repeated K Fold**.
7. In the **Method** list, unselect **Neural.**
8. Click **OK**.

Figure 2. Model Screening options dialog

JMP responds by training and testing models for each of the selected methods using their default parameter settings and cross-validation. After completing the training and testing process, JMP displays the results in a new window as shown in Figure 3.



Method	N Trials Folds	Sum Freq	Validation Set Folds RSquare	Mean RASE	StdDev RASE
Fit Stepwise	15	88.400	0.4901	54.370	4.6963
Generalized Regression Lasso	15	88.400	0.4896	54.425	4.5254
Fit Least Squares	15	88.400	0.4811	54.845	4.5997
Support Vector Machines	15	88.400	0.4441	56.760	4.9217
Bootstrap Forest	15	88.400	0.4410	56.983	4.1064
Boosted Tree	15	88.400	0.4309	57.435	4.4173
K Nearest Neighbors	15	88.400	0.4296	57.677	4.7030
Decision Tree	15	88.400	0.3520	61.292	4.3564

Select Dominant Run Selected Save Script Selected

Figure 3. Model Screening results window

For each modeling method, the Model Screening platform provides performance measures in the form of point estimates for the coefficient of determination, R-squared, the root average squared error, and the standard deviation for the root average squared error. Now click **Select Dominant**. JMP responds by highlighting the method that performs best across the performance measures. What is missing is a graphical depiction of the size of the performance difference between the dominant method and the other methods along with a visualization of the uncertainty associated with the measure of performance. As we shall see, confidence curves provide this missing graphic.

WHY NOT USE NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)?

But why not show a p-value and use NHST? Shouldn't a decision about whether one model is superior to another be based on significance? While a p-value and associated NHST can help determine if an observed difference is larger than we would expect under the null of no difference between models, relying completely on a p-value for decision making can be somewhat arbitrary. A p-value provides a probability based on a standardized difference, losing information on the raw difference itself. For example, suppose the mean difference in R-squared between method B and method A is 0.2 and the standard deviation is 0.11. Further assume that a standard normal statistic is applicable. In that case, $z = 0.20/0.11 = 1.8$ which is associated with a two-tailed p-value of $0.07 > 0.05$. So, we cannot reject the null hypothesis of equal performance. Now suppose the difference between method C and method A is 0.02 and the standard deviation

is 0.01. In that case, $z = 0.02/0.01 = 2$ which yields a p-value of $0.046 < 0.05$. Now we can reject the null hypothesis. But on average, method B accounts for 18% more of the variation than method C. Should we then decide that the difference between A and C is meaningful and the difference between A and B is not? If so, the size of the difference is conflated with statistical significance. If you are still not convinced, you can make the difference between method A and C arbitrarily small by multiplying the above difference and standard deviation by some arbitrarily small fraction while maintaining the same statistical significance. So, by itself, statistical significance tells us nothing about the size of the mean difference between methods. Unless two methods are equivalent, a large amount of data will often show a statistically significant difference even if the difference in models is negligible (Berrar 2017).

Even so, you may argue that it is pointless to show the difference between two methods before using NHST to detect whether the difference is real. Note that power determines our ability to correctly reject a null hypothesis. All other things being equal, we can increase power by increasing the alpha value used in NHST which simultaneously makes it more likely we will incorrectly reject the null hypothesis. We can also increase power by increasing the sample size. This is fine if we can collect more data and the difference in methods is substantial. Again, with sufficient data, it is possible to detect statistically significant differences that are not meaningful. If you are limited to the data and methods you have, your ability to affect power is limited. Schmidt and Hunter (1997) point out that the power of typical studies is typically between 0.40 to 0.60. If we assume the power to detect a difference between methods is 0.50 and the difference between methods is both real and meaningful, there is a 50% chance that such a difference will be labeled as insignificant. In such a case, NHST correctly identifies real differences no better than flipping an unbiased coin.

As an alternative to NHST, Cohen (1994) and Schmidt (1996) have suggested replacing significance testing with point estimates and confidence intervals. One objection to doing so is that point estimates and confidence intervals can be seen as another form of NHST (Schmidt and Hunter 1997). Even if point estimates and confidence intervals are interpreted as NHST, point estimates and confidence intervals improve upon NHST by showing the size of the difference and reporting the extent of the uncertainty. So, both the magnitude of the difference and the range of uncertainty are put front and center whereas a lone p-value conceals them both. That said, point estimates and confidence intervals need not be interpreted as NHST. Fisher only began promoting NHST in the 1930s, while point estimates and confidence intervals have been used as “error bands” by the likes of Bernoulli and Poisson as early as the 1700s. Moreover, the physical sciences continue to use the equivalent of point estimates and confidence intervals in lieu of NHST. So, it can be argued that point estimates and confidence intervals should not be interpreted as NHST (Schmidt and Hunter 1997).

WHY USE CONFIDENCE CURVES?

As mentioned, authors such as Cohen (1994) and Schmidt (1996) have strongly recommended replacing NHST with point estimates and confidence intervals. As we shall see, confidence curves provide both. Even so, the recommendation to use confidence intervals begs the question, which ones do we show? Showing only the 95% confidence interval would likely encourage you to interpret it as another form of NHST. The solution provided by confidence curves is to literally show all confidence intervals up to an arbitrarily high value confidence level.

HOW DO I GET THE CONFIDENCE CURVES ADD-IN?

To conveniently create confidence curves in JMP, install the Confidence Curves Add-In. In a web browser, bring up the JMP Community home page. Type "Confidence Curves" into the search window embedded in the home page as shown in Figure 5 and select the first entry that appears in the results.

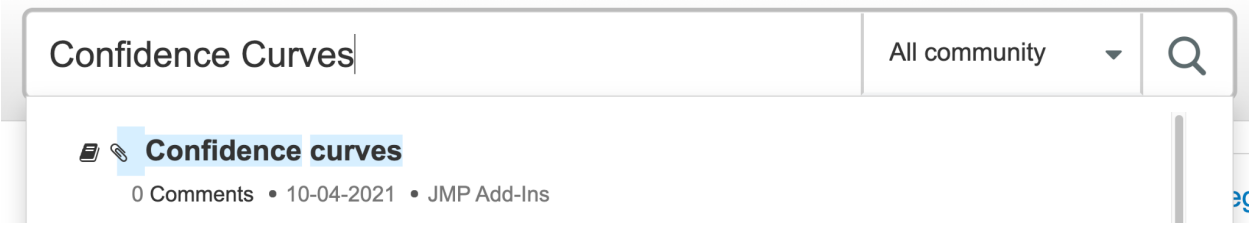


Figure 5. Confidence Curves article shown in JMP Community search control

Once the Confidence Curves page is loaded, click the download icon next to **Confidence Curves.jmpaddin** as shown in Figure 6.

Figure 6. The Confidence Curves Add-In download on JMP Community

Now install the Confidence Curves Add-In by double-clicking the downloaded file in either Windows File Explorer or Mac Finder. When JMP prompts you, choose the option to **Install**. Afterwards, you can access the confidence curves capability via the **Add-Ins** menu.

HOW ARE CONFIDENCE CURVES GENERATED?

To generate confidence curves for this report, select **Save Results Table** from the top red triangle menu located on the **Model Screening** report window as shown in Figure 7.

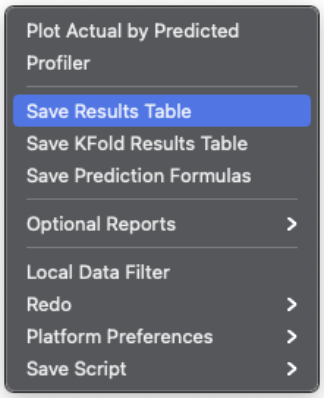


Figure 7. Model Screening red triangle menu

JMP responds by creating a new table containing, among others, the following columns:

- Trial
- Fold
- Method
- N

Note that the Trial column will be missing if the number of repeats is exactly 1, in which case, the Trial column is neither created nor needed. Save for the preceding exception, these columns are essential for the Confidence Curves Add-In to function properly.

In addition to these columns, you need one column that provides the metric to compare between methods. In this paper, we will use R-squared as the metric of interest.

After JMP generates the Model Screening results table, click **Add-Ins** from JMP's main menu bar. The first dialog that appears requests you to select the name of the table that was generated when you chose **Save Results Table** from the Model Screening report's red triangle menu. Select the appropriate table name and then select **OK**.

Next, a dialog is displayed that requests the name of the method that will serve as the baseline from which all the other method metrics are measured. We suggest starting with the method that was selected when you clicked the **Select Dominant** option in the Model Screening report window. Select the method of your choice and then select **OK**.

Finally, a dialog is displayed that requests you to select the metric to be compared between the various methods. As mentioned earlier, in this paper, we use R-squared as the metric for comparison. After selecting a metric, select **OK**.

JMP responds by creating a confidence curve table that contains p-values and corresponding confidence levels for the mean metric difference between the chosen baseline method and each of the other methods. More specifically the generated table has columns for the following:

- Model: Name of the modeling method whose performance is evaluated relative to the baseline method.
- P-Value: The probability associated with a performance difference at least as extreme as the value shown in the Difference in RSquare column.
- Confidence Interval: The confidence level we have that the true mean is contained in the associated interval.
- Difference in RSquare: The maximum or minimum of the expected difference in R-squared associated with the confidence level shown in the Confidence Interval column.

From this table, confidence curves are created and shown in a Graph Builder graph as shown in Figure 8.

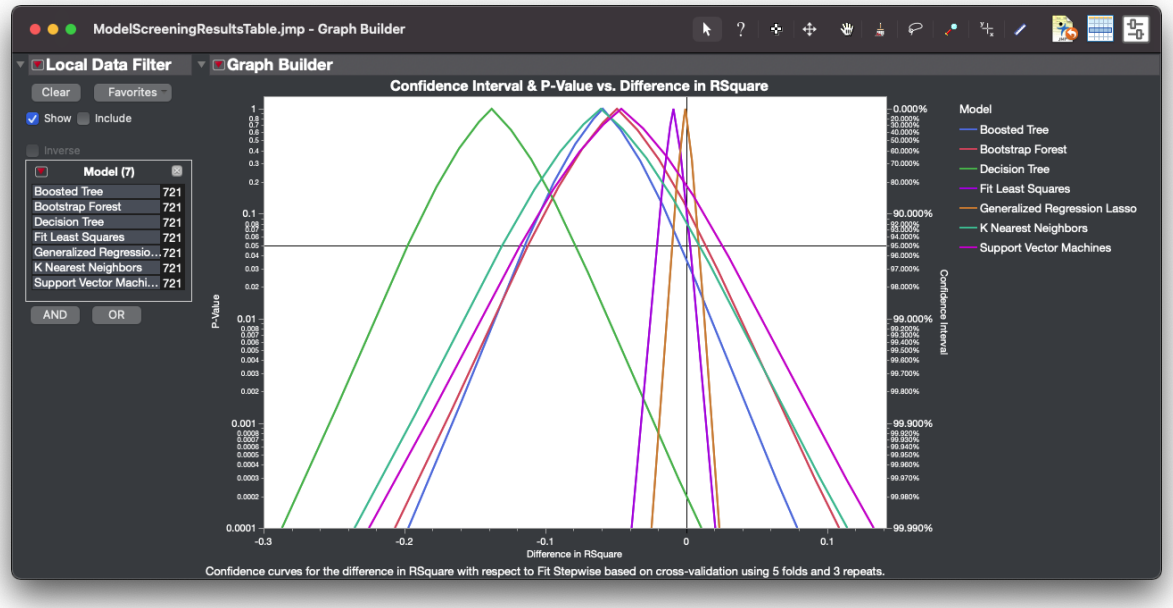


Figure 8. Confidence Curves Graph Builder report

Notice the generated report contains a **Local Data Filter** that allows you to visualize one or more of the confidence curves rather than all of them simultaneously. Also, there is text at the bottom of the report that indicates the performance metric being compared, the method used as the reference or baseline, and the number folds and repeats used to generate the performance differences.

WHAT ARE CONFIDENCE CURVES?

To clarify the key attributes of a confidence curve, hide all but one confidence curve using the **Local Data Filter**. More specifically, click on **Support Vector Machines** in the **Local Data Filter**.

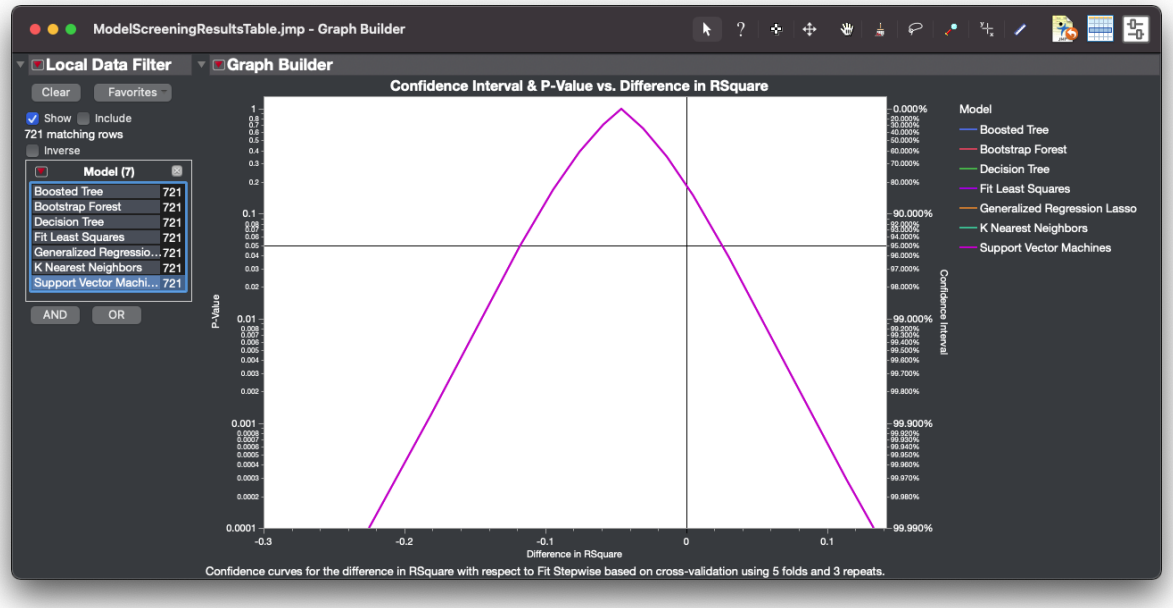


Figure 9. Use the local data filter to limit the number of displayed confidence curves

JMP responds by showing only the Support Vector Machines confidence curve.

A confidence curve plots the extent of each confidence interval from the generated table between the 0% and 99.99% confidence level. Along the left y-axis, p-values associated with the confidence intervals are shown. Along the right y-axis, the confidence level associated with each confidence interval is shown. The y-axis uses a log scale and, for that reason, the confidence curve table records a constant number of confidence intervals, specifically 90, at each order of magnitude between 0% and 99.99% confidence. By default, a confidence curve only shows the lines that connect the extremes of each confidence interval. To see the points, select **Show Control Panel** from the red triangle menu located next to the **Graph Builder** text in the title bar. Now shift-click the points icon as shown in Figure 10.

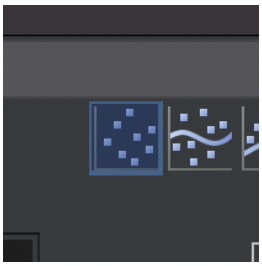


Figure 10. Graph Builder option for showing the points in the graph

JMP responds by displaying the end points of the confidence intervals that make up the confidence curves. If you hover the mouse pointer over any of these points, a hover label shows the p-value, confidence interval, difference in the size of the metric, and the method used to generate the model being compared to the reference model as shown in Figure 11.

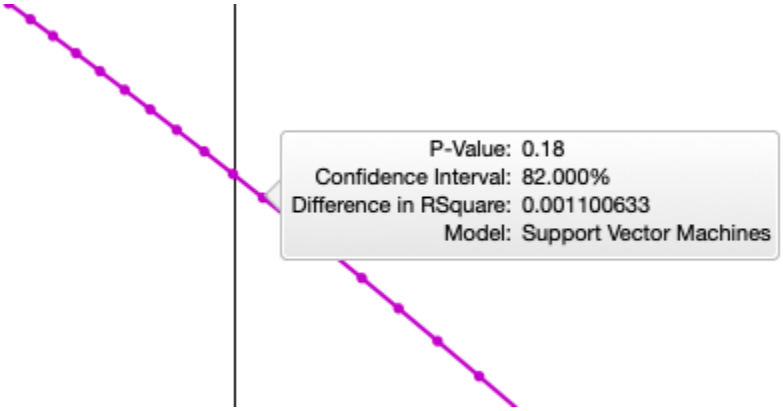


Figure 11. Hover label for a confidence interval extreme

Each of these points is connected to the next larger and smaller confidence interval extreme by a line segment. Thereby, we represent the full continuum of confidence intervals from a 0% up to a 99.99% confidence level. After examining a few of these points, hide the points by again shift-clicking the points icon. Even though the individual points are not shown, you can still view the associated hover label by placing the mouse pointer over the confidence curve.

The point estimate for the difference in performance is shown at the 0% confidence level which is the mean value of the computed differences. We denote the mean value as *d* as shown in Figure 12.

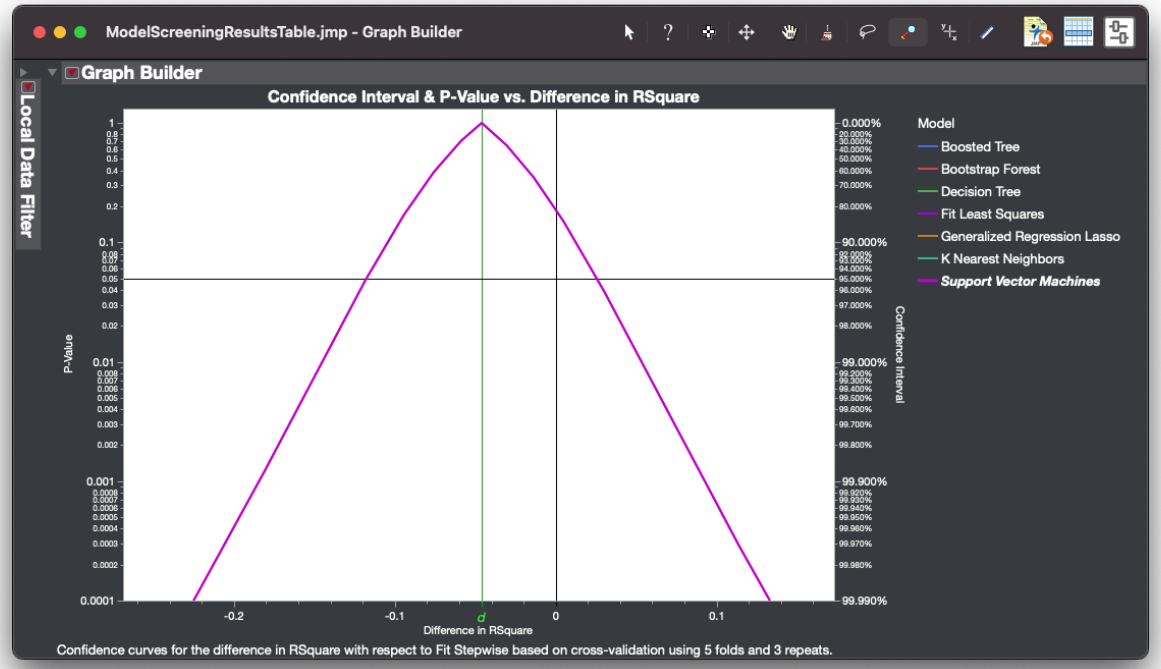


Figure 12. Mean difference in R-squared is denoted as *d*

By default, two reference lines are plotted alongside a confidence curve. The vertical line represents the traditional null hypothesis of no difference in effect. Note you can change the vertical line position and, thereby, the implicit null hypothesis in the axis settings. The axis settings are accessed by selecting **Axis Settings…** on the popup menu that appears when you right-click the x-axis. The horizontal line passes through the conventional 95% confidence interval. As with the vertical reference line, you can change the horizontal line position and, thereby, the implicit level of significance, by selecting **Axis Settings…** on the popup menu that appears when you right-click the y-axis. If a confidence curve crosses the vertical line above the horizontal line, you cannot reject the null hypothesis using NHST. On the other hand, if the confidence curve crosses the vertical line below the horizontal line, you can reject the null hypothesis using NHST.

HOW ARE CONFIDENCE CURVES COMPUTED?

The current implementation of confidence curves assumes the differences are computed using r-times repeated k-fold cross-validation. The extent of each plotted (1 - α)100% confidence interval is computed using the variance-corrected resampled t test (Nadeau and Bengio 2003),

$$d \pm t_{v,1-\frac{1}{2}\alpha} \times s\sqrt{\frac{1}{kr}+\frac{n_2}{n_1}}$$

where

- t is the critical value of Student's distribution with ν = kr − 1 degrees of freedom and
- k is the number of cross-validation folds
- r is the number of repetitions
- $n2$ is the number of cases in one validation set
- $n1$ is the number of cases in the corresponding training set (where n1 ≈ 5n2)

The standard deviation s is calculated as

$$s = \sqrt{\frac{\sum_i^r \sum_j^k \left(d_{ij} - \bar{d}\right)^2 1}{kr - 1}}$$

Where,

- d_{ij} is the difference between the metric computed by the reference model and the metric computed by a given alternative model in the ith repetition of the jth cross-validation fold;
- $\overline{d}$ is the average of the differences

Note that a corrected resampled t-test is typically used in cases where training sets are 5 or 10 times larger than validation sets.

A confidence curve c(x, d) is the graphical depiction of an infinite number of confidence intervals. This curve can be defined using the cumulative distribution function (CDF) F in the following manner.

$$c(x,d) = \begin{cases} 2F(d-x) & if\ d \leq x \\ 2[1-F(d-x)] & if\ d > x \end{cases}$$

Where,

- $F(d-x) = \int_{-\infty}^{d-x} f(u)du$
- $f(u)$ is the probability density function for a Student's t-distribution

As the degrees for freedom increase (v > 30), the t-distribution approaches a normal distribution. The value of the CDF is multiplied by 2 since we are using a two-sided p-value as the default alternative hypothesis is assumed to be a non-zero difference between methods (Berrar 2017).

HOW ARE CONFIDENCE CURVES INTERPRETED?

First, a confidence curve graphically depicts the mean difference in the metric of interest between a given method and a reference method at the value *d* as shown in Figure 12. So, we can evaluate whether the mean difference between methods is meaningful. If the mean difference isn't meaningful, there is little point in further analysis of the given method versus the reference method with respect to the chosen metric. What constitutes a meaningful difference depends on the metric of interest as well as the intended scientific or engineering application. In Figure 13, you can see the model developed with the decision tree method is, on average, more than 13% worse than Fit Stepwise, which, arguably, is a meaningful difference.

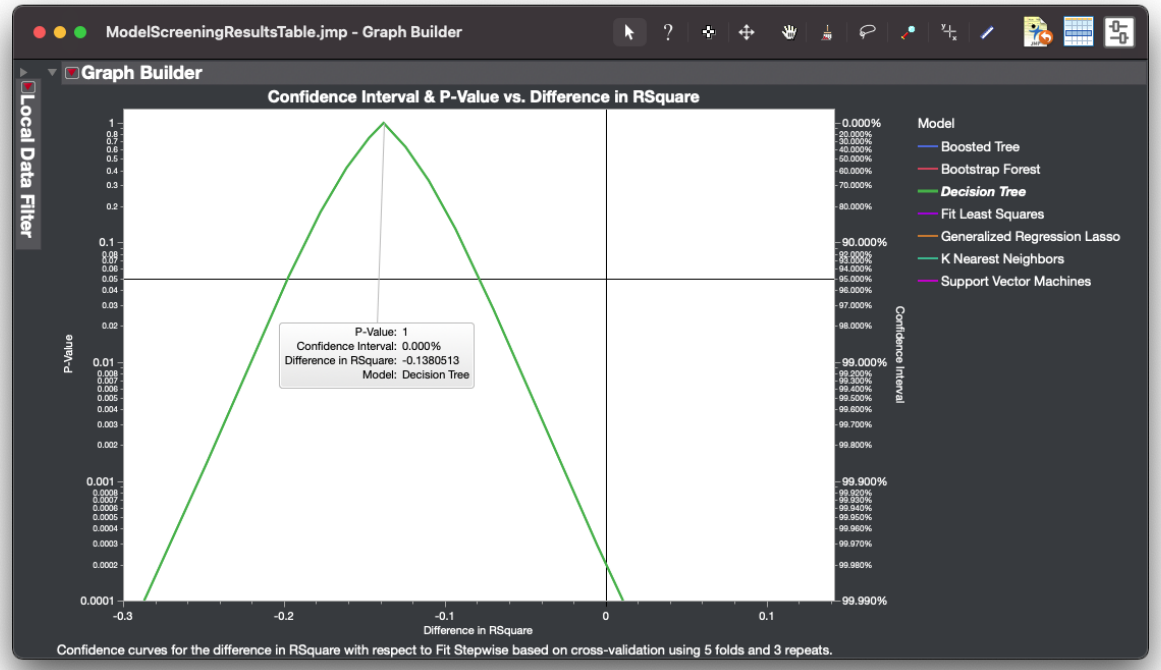


Figure 13. Mean difference in R-squared between decision tree and fit stepwise models

If the difference is meaningful, we can evaluate how precisely the difference has been measured by evaluating the width of the associated confidence interval. For any confidence interval not crossing the default vertical axis, we have at least that level of confidence that the mean difference is non-zero. So, at what level of confidence should a non-zero difference be expected? At a minimum, we suggest some confidence level greater than or equal to 50%. If there is no

reason to prefer one method versus another apart from their performance, type I and type II errors are of equal value. So, if you can estimate the power of the experiment at different p-values, we suggest taking that into consideration. Also, if you have knowledge outside of the current analysis, for example, you know that in your domain one method generally outperforms the other, we encourage you to consider that as well. In Figure 8, you can see we are at least 80% confident the fit least squares model is at least as good as every other method other than Generalized Regression Lasso.

Now let's consider multiple confidence curves. If two or more confidence curves significantly overlap one another and the value of *d* of each is not meaningfully different from the other, the data suggest each method performs about the same as the other with respect to the reference model. For example, in Figure 14, we see that on average the support vector machines model performs less than 0.5% better than bootstrap forest and the confidence intervals do not overlap until about the 4% confidence level, which suggests these values would be expected if both methods really do have about the same difference in performance with respect to the reference.

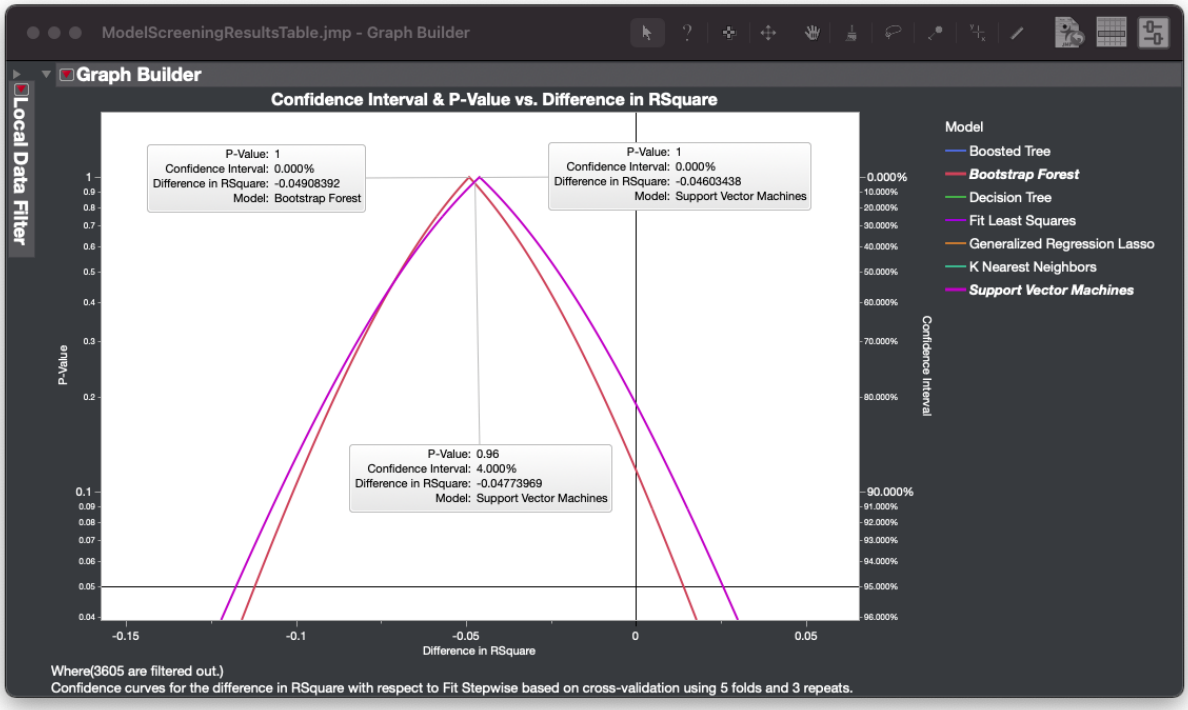


Figure 14. Confidence curves showing no meaningful difference

If the value *d* is about the same for two confidence curves, but the confidence intervals don't overlap much, the data suggest the methods perform about the same as each other with respect to the reference model; however, the differences are precisely measured. On the other hand, if the value of *d* for each of the two confidence curves is meaningfully different from the other, and the

confidence curves have little overlap, the data suggest the methods perform differently from one another with respect to the reference method. An example of the latter can be seen in Figure 15 where, on average, the Generalized Regression Lasso method predicts about 13.8% more of the variation in the response than does the Decision Tree. Moreover, the intervals for the true mean difference between each method and Neural Boosted don't overlap until about the 99.9% confidence level, which suggests the results are quite unusual if the methods actually perform about the same with respect to the reference.

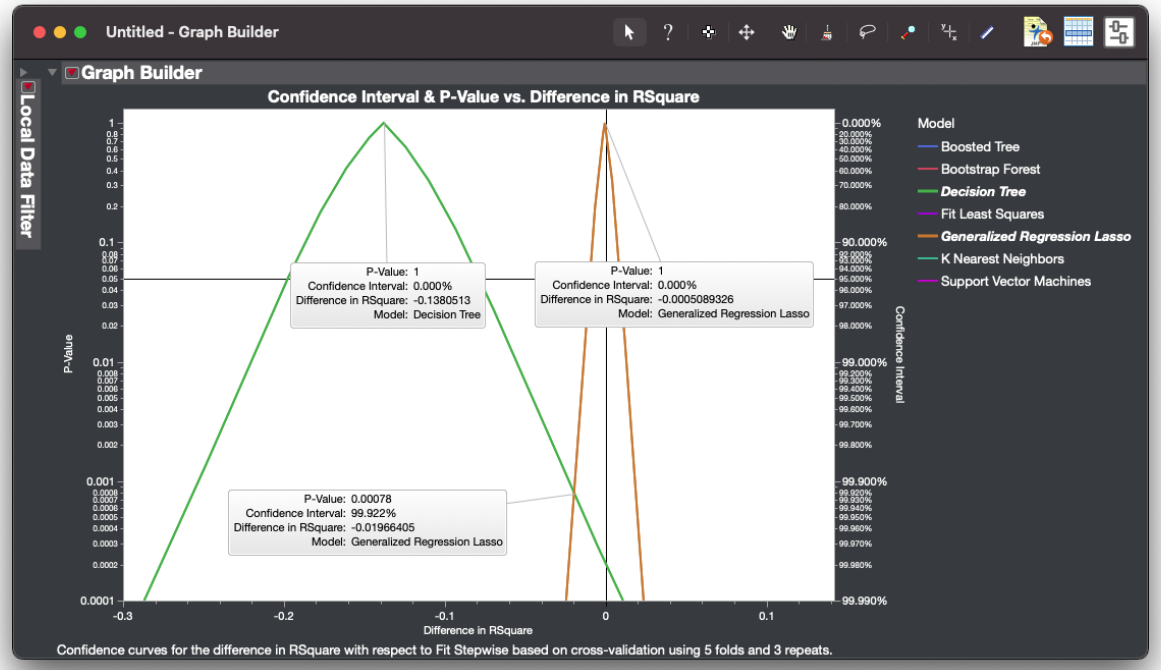


Figure 15. Confidence curves showing a meaningful difference

Finally, if the *d* values of two confidence curves are meaningfully different from one another, and have considerable overlap, the data suggest that, while the methods performed differently from one another with respect to the reference in the analysis, it would not be surprising if that difference is spurious. For example, in Figure 16, we can see that, on average, Support Vector Machines predicted about 1.4% more of the variance in the response than did K Nearest Neighbors; however, the confidence intervals begin to overlap at about the 17% confidence level, which suggests it would not be surprising if the difference in performance between each method and the reference is actually smaller than measured. Simultaneously, it would not be surprising if the actual difference is larger than measured or if the direction of the difference is actually reversed. In other words, the difference in performance is imprecisely measured.

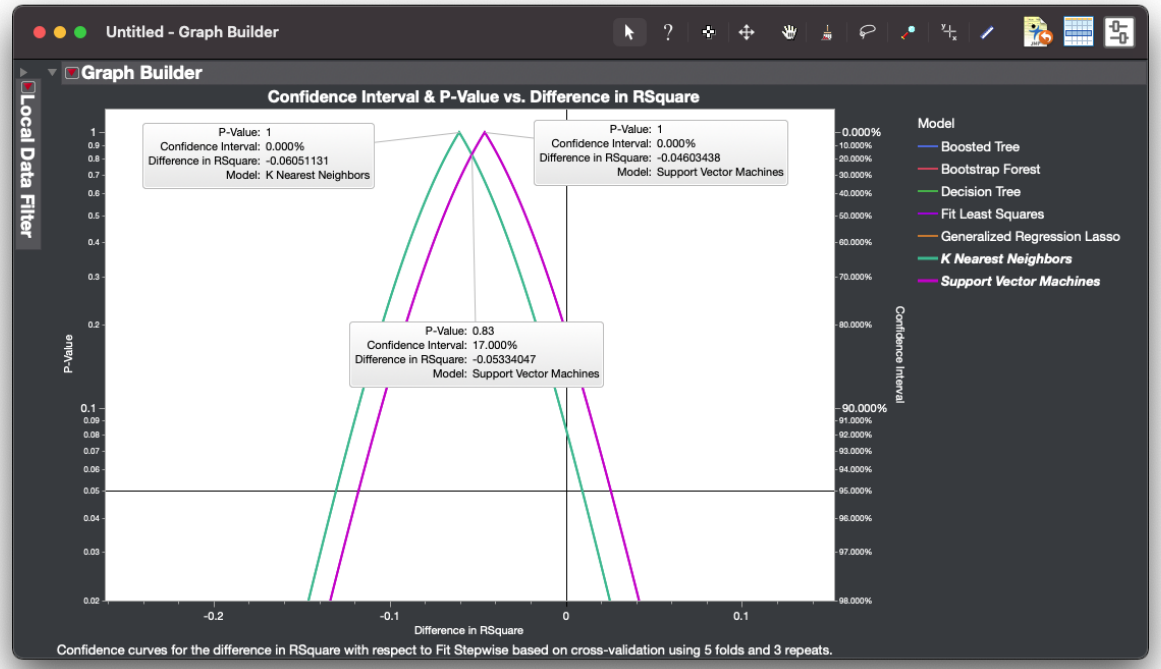


Figure 16. Confidence curves not clearly delineated

Note that it is not possible to assess the variability in performance between two methods relative to one another using confidence curves that are relative to a third method. To compare the variability in performance between two methods relative to one another, one of the two methods must be the reference method from which the differences are measured.

WHAT ABOUT MULTIPLE COMPARISONS?

Suppose you wish to perform NHST using confidence curves. It is widely recognized when performing multiple comparisons that the p-values need to be adjusted to control the family-wise type I error rate. If we adjust p-values to control for type I error rate, we simultaneously increase the type II error rate. In the case at hand, is it really more preferable to err on the side of saying there is no difference between methods when there really is? Given the suggestion to only consider differences that are meaningful, not adjusting p-values in order to increase power is somewhat justified when exploring your data, keeping in mind the dangers of cherry-picking and p-hacking.

Nevertheless, let's assume you believe that p-values should be adjusted when there are multiple comparisons. Let's further assume you have set the random seed in an experiment in order to be able to replicate your comparison results. Finally, assume you conduct an analysis considering only two methods, A and B, and find there is a difference that is both meaningful and statistically

significant. You share your findings with your associates and are asked to consider a third method, C, using the same data and random seed. You then rerun the analysis and modify the p values since you are now performing multiple comparisons and find the difference between method A and B is no longer significant. Would you then retract your earlier statement about the difference between A and B being statistically significant? Rothman (1990) suggests you shouldn't have to pay a "penalty for peeking" since the introduction of another comparison has no bearing on the difference observed between A and B. Berrar elaborates by suggesting that adjustments are needed in confirmatory studies where a goal is pre-specified, but not in exploratory studies. This suggests using unadjusted multiple confidence curves in an exploratory fashion and only a single confidence curve generated from different data to confirm your finding of a significant difference between two methods when using NHST (Berrar 2017).

SUMMARY

The Model Screening platform introduced in JMP Pro 16.0, provides a means to simultaneously compare the performance of predictive models created using different methodologies. JMP has a long standing goal to provide a graph with every statistic and confidence curves help to fill that gap for the Model Screening platform (Sall et al. 2017).

You might naturally expect to use NHST to differentiate between the performance of the various methods being compared; however, p-values have come under increased scrutiny in recent years for obscuring the size of performance differences. In addition, p-values are often misinterpreted as the probability the null hypothesis is true. Instead, a p-value is the probability of observing a difference as or more extreme assuming the null hypothesis is true. The probability of correctly rejecting the null hypothesis when it is false is determined by power or 1 - β. We have argued that it is not uncommon to only have a 50% chance of correctly rejecting the null hypothesis with an alpha value of 0.05. As an alternative, a confidence interval could be shown instead of a lone p-value; however, the question would be left open as to which confidence level to show.

Confidence curves address the above concerns by showing all confidence intervals up to an arbitrarily high level of confidence. The mean difference in performance is clearly visible at the 0% confidence level and that value is the one most consistent with the data. All other things being equal, type I and type II errors are equivalent, so confidence curves don't embed a bias toward trading type I errors for type II. Even so, by default, a vertical line is shown in the confidence curves graph for the standard null hypothesis of no difference along with a horizontal line that delineates the 95% confidence level, which readily affords a typical NHST analysis if desired. The defaults for said lines are easily modified if a different null hypothesis and confidence level is desired. Even so, we encourage you to use confidence curves to evaluate both the size of the mean performance difference as well as the uncertainty of the measurement (Berrar 2017). Thereafter, either use independent data to confirm your findings or make a preliminary judgment based on the analysis as well as posteriori knowledge.

REFERENCES

Berrar, Daniel. 2017. “Confidence Curves: An Alternative to Null Hypothesis Significance Testing for the Comparison of Classifiers.” *Machine Learning* 106 (6): 911–49.

Cohen, Jacob. 1994. “The Earth Is Round (P<. 05).” *American Psychologist* 49 (12): 997.

Nadeau, Claude, and Yoshua Bengio. 2003. “Inference for the Generalization Error.” *Machine Learning* 52 (3): 239–81. https://doi.org/10.1023/A:1024068626366.

Rothman, Kenneth J. 1990. “No Adjustments Are Needed for Multiple Comparisons.” *Epidemiology* 1 (1): 43–46.

Sall, John, Mia L. Stephens, Ann Lehman, and Sheila Loring. 2017. *JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP*. Sas Institute.

Schmidt, Frank L. 1996. “Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers.” *Psychological Methods* 1 (2): 115.

Schmidt, Frank L., and John E. Hunter. 1997. “Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data.” *What If There Were No Significance Tests*, 37–64.