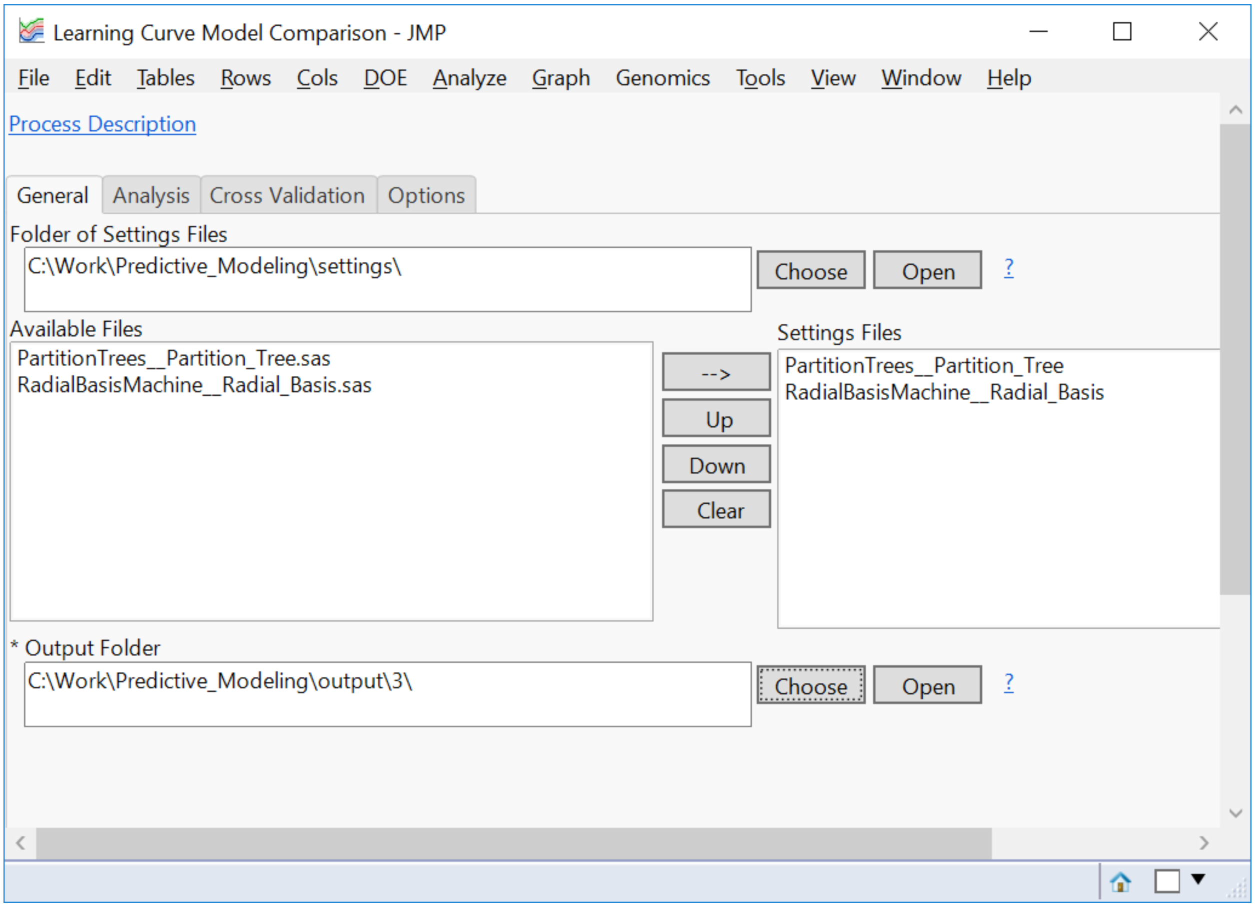
Learning Curves

A common issue in predictive modeling is having too few samples to generate a meaningful model. In JMP Genomics, we can use **Learning Curves** to determine if our data has an adequate number of samples to produce a model. Unlike the classical sample size/power calculations, the results from this analysis will indicate if more samples will improve model performance. For this example, we will work with the **Partition Tree** & **Radial Basis Machine** models, which we compared using cross-validation in the *Cross Validation Model Comparison Step-Guide.*

1. From the **Genomics Starter**, select **Predictive Modeling > Model Comparison > Learning Curve Model Comparison**.
2. Navigate to and select the "settings" folder that contains the models we created in the *Cross Validation Model Comparison Step-Guide.*
3. Select both the **Partition Tree** & **Radial Basis** models from the list of **Available Files**.
4. Create and select a new **Output Folder**.


Learning Curve Model Comparison - JMP
File Edit Tables Rows Cols DOE Analyze Graph Genomics Tools View Window Help
Process Description
General
Analysis
Cross Validation
Options
Choose
Open
?
-->
Up
Down
Clear
Choose
Open
?

Folder of Settings Files: C:\Work\Predictive_Modeling\settings\

Available Files:
PartitionTrees__Partition_Tree.sas
RadialBasisMachine__Radial_Basis.sas

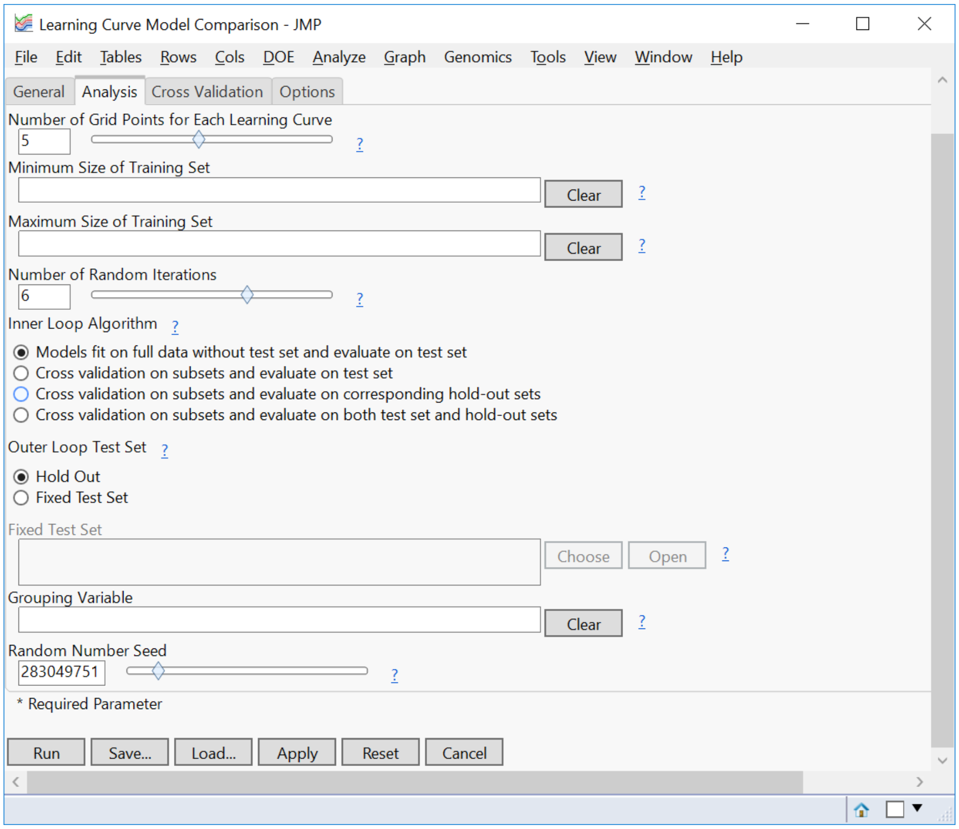
Settings Files:
PartitionTrees__Partition_Tree
RadialBasisMachine__Radial_Basis

* Output Folder: C:\Work\Predictive_Modeling\output\3\

5. On the **Analysis** tab, set the **Number of Grid Points for Each Learning Curve** to 5.

    - Comparing models via Learning Curves works by performing a series of CVMC runs along a fixed grid of subsets of the full data set. Here we are designating 5 points to be generated for each Learning Curve.

6. Leave the **Maximum/Minimum Size of Training Set** blank.
    - This field allows the grid points specified in step 5 to be logarithmically spaced between the minimum and maximum points entered here.
    - If these fields are left blank, the points will be spaced between 1/10 size and the full size of the data set.
7. The **Number of Random Iterations** is the number of iterations for each point on the grid. Set this to 6 for a total of 30 iterations. This number will be multiplied by the **Number of Random Holdout Iterations** in the **Cross Validation** tab later.
8. The **Inner Loop Algorithm** type specifies the type of cross validation that will be done on the data set. Options to add further rigor can be added by cross-validating on the subset training sets themselves, then specifying the set for the evaluation set. Also, the model can be fit to test data from the hold out, or to a separate test data set. Here, we choose the first option to fit the model to the full data set without the test set and evaluate its fit on the test set.
9. For the **Outer Loop Test Set**, choose Hold Out. This randomly holds out observations to be used as the test set, rather than specifying a fixed test set.



Number of Grid Points for Each Learning Curve: 5

Minimum Size of Training Set:

Maximum Size of Training Set:

Number of Random Iterations: 6

Inner Loop Algorithm

● Models fit on full data without test set and evaluate on test set
○ Cross validation on subsets and evaluate on test set
○ Cross validation on subsets and evaluate on corresponding hold-out sets
○ Cross validation on subsets and evaluate on both test set and hold-out sets

Outer Loop Test Set

● Hold Out
○ Fixed Test Set

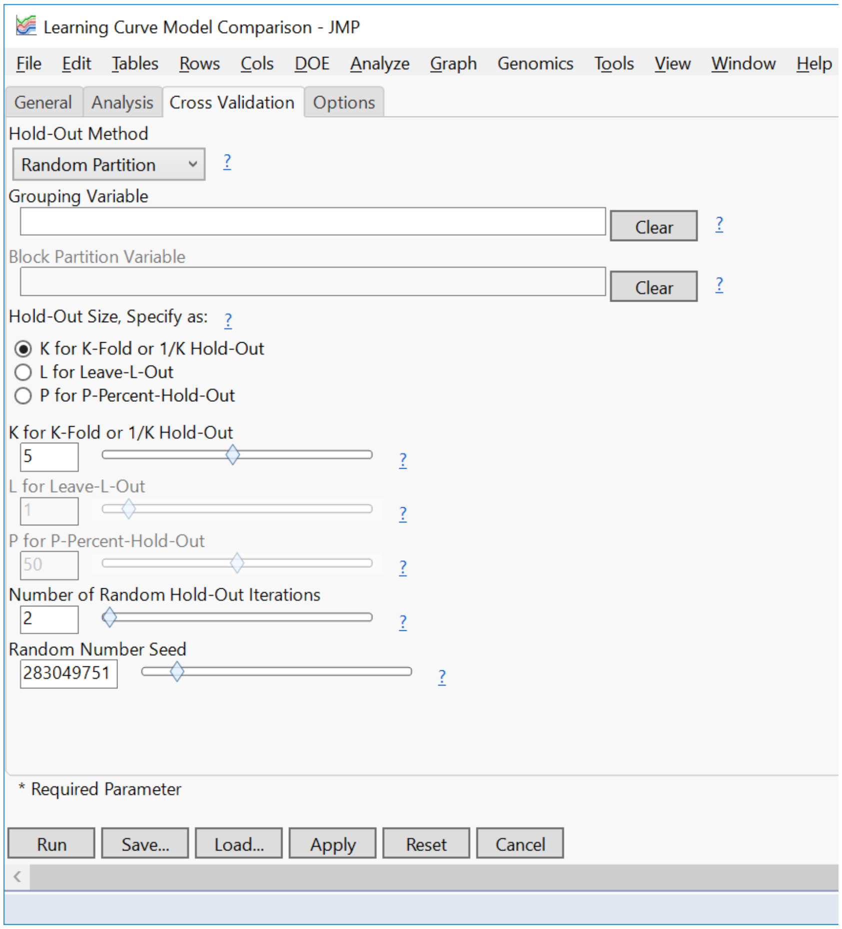
Fixed Test Set

Grouping Variable

Random Number Seed: 283049751

* Required Parameter

10. The **Cross Validation** tab will be filled out similarly to the same tab in the *Cross Validation Model Comparison Step-Guide*. Select Random Partition for the **Hold-Out Method**, and K for K-Fold or 1/K Hold-Out for the **Hold-Out Size**. Set the **K for K-Fold or 1/K Hold-Out** slider to 5.


Learning Curve Model Comparison - JMP
File Edit Tables Rows Cols DOE Analyze Graph Genomics Tools View Window Help
General Analysis Cross Validation Options
Clear
Clear
Run Save... Load... Apply Reset Cancel

Hold-Out Method: Random Partition

Grouping Variable

Block Partition Variable

Hold-Out Size, Specify as:

● K for K-Fold or 1/K Hold-Out
○ L for Leave-L-Out
○ P for P-Percent-Hold-Out

K for K-Fold or 1/K Hold-Out: 5

L for Leave-L-Out: 1

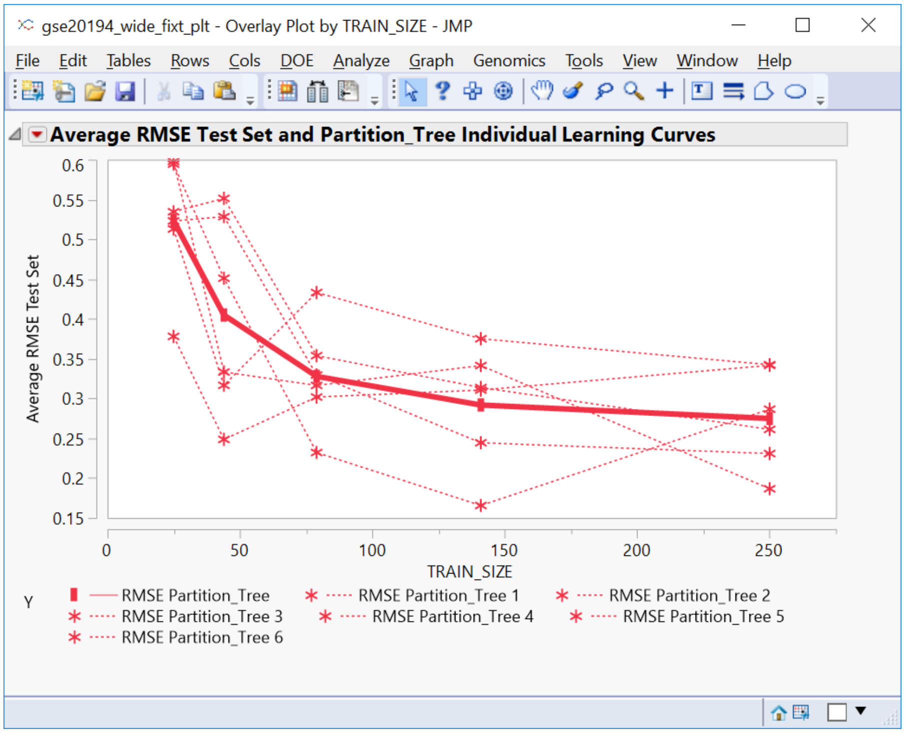
P for P-Percent-Hold-Out: 50

Number of Random Hold-Out Iterations: 2

Random Number Seed: 283049751

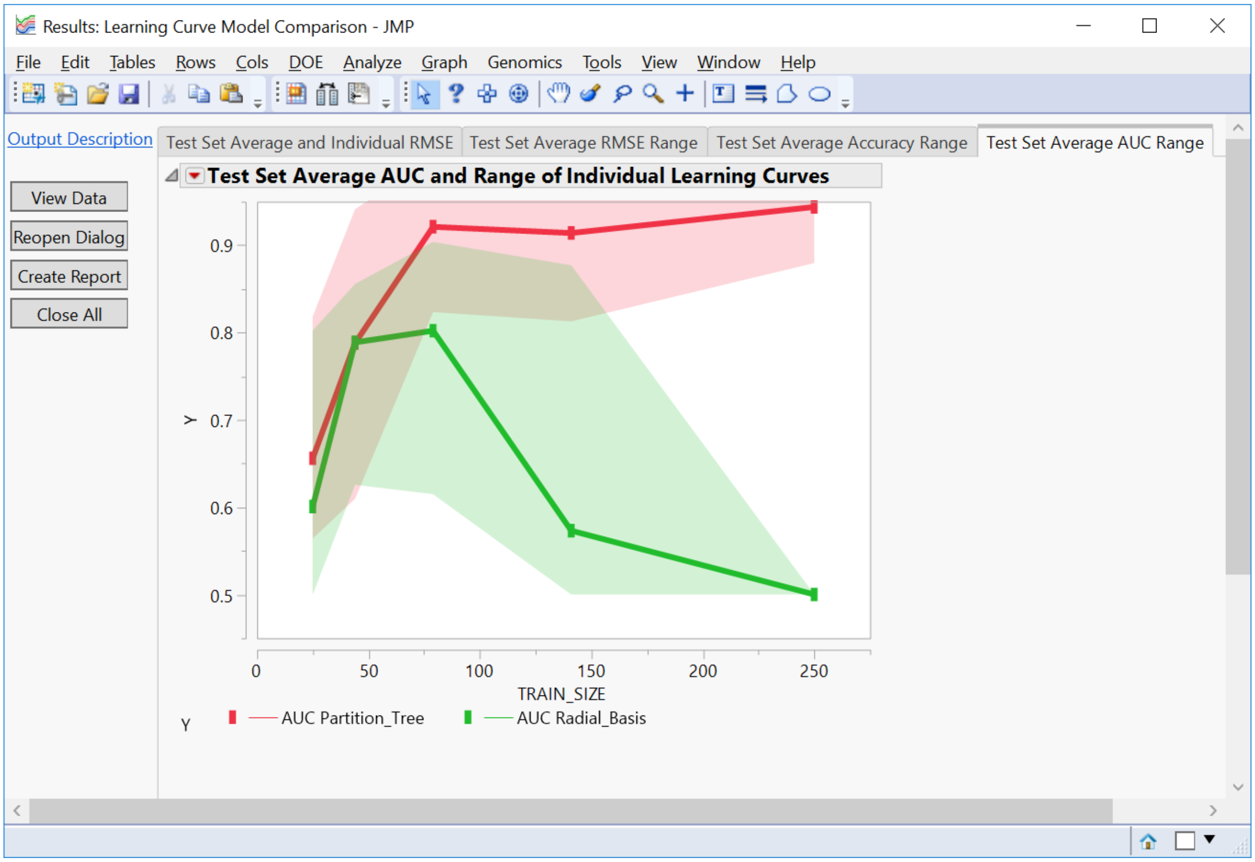
* Required Parameter

11. Once all tabs match the screenshots above, click **Run**.
12. When the analysis is complete the Random Mean Squared Error for each iteration of each model is shown. As seen in the CVMC results, the Partition Tree model performs best. Click the red triangles above the plot to select only the **Partition_Tree** model for the **RMSE** statistic. A plot will open in a new window.



- Each dotted line shows an iteration of Cross-Validation with the solid line being the mean of all iterations. Each point on these lines represents the sample size from the 5 grid points for each learning curve.
- The curve flattens out at the point where a larger sample size does not improve the model's performance.

13. Subsequent tabs (like the AUC plot below) show the mean of the iterations of cross-validation with the range of the iterations shaded around it. Once again, the plot shows the sufficient number of samples where the curve begins to flatten.


Results: Learning Curve Model Comparison - JMP
File Edit Tables Rows Cols DOE Analyze Graph Genomics Tools View Window Help
Output Description
Test Set Average and Individual RMSE
Test Set Average RMSE Range
Test Set Average Accuracy Range
Test Set Average AUC Range
View Data
Reopen Dialog
Create Report
Close All
Test Set Average AUC and Range of Individual Learning Curves
Y
0.9
0.8
0.7
0.6
0.5
0
50
100
150
200
250
TRAIN_SIZE
Y
AUC Partition_Tree
AUC Radial_Basis

Summary

The purpose of the **Learning Curves** tool is to determine if our data has an adequate number of samples to produce an effective model. Also, it serves to show when adding more data points no longer substantially improves the model. From this analysis, we see that the **Partition Trees** model is robust, and that it has an adequate number of data points at the 4th grid point of the data. Recall that because we did not specify a minimum and maximum at which to limit the grid points, they began at 1/10 the size of the entire data set and were set incrementally (5 increments in this example) up to the entire size of the data set. Note that the type of cross validation and therefore the power of this tool can be adjusted by changing the **Inner Loop Algorithm** as well as the **Number of Random Hold-Out Iterations** for CVMC. For the next example of comparing predictive models, see the *Test Set Model Comparison Step-Guide*.