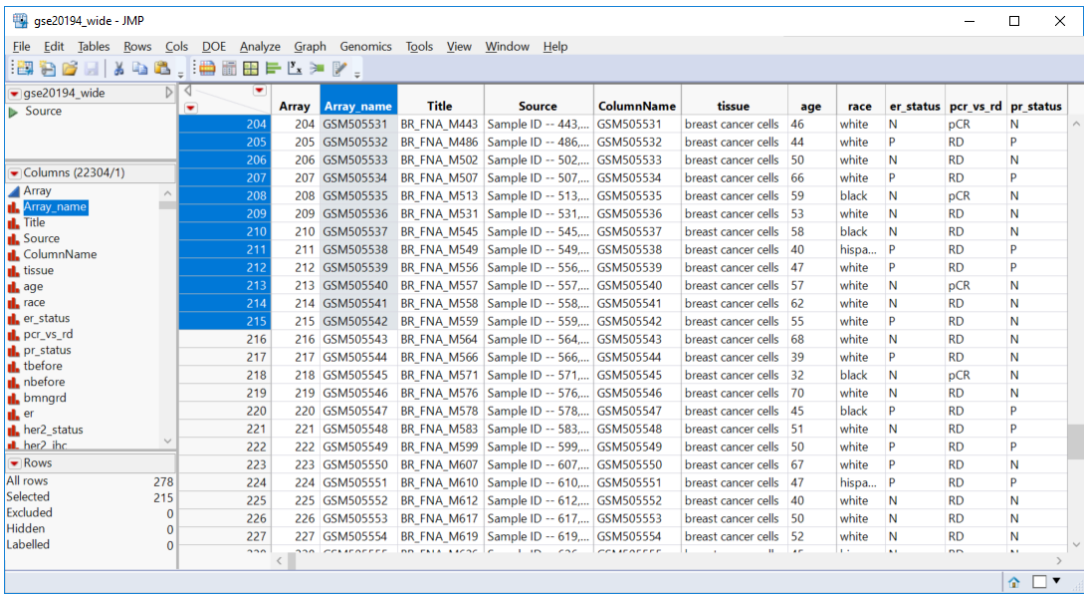
Subsetting a Test Data Set for Test Set Model Comparison

JMP Genomics has many tools for automating the cross validation and model comparison processes. However, if you wish to create your own **Test Set** from your data, JMP Genomics has a tool for that too. In this guide, we will create both a **Test Data Set** and a **Training Data Set** from the wide data set we created in the *Transposing a Data Set: Tall to Wide Step-Guide* and use the test set to compare the **Radial Basis** and **Partition Trees** models we created in the *Cross Validation Model Comparison Step-Guide*.

1. Open the ***gse20194_wide.sas7bdat*** data set created in the *Transposing a Data Set: Tall to Wide Step-Guide* by dragging it over the **Genomics Starter** window.
2. The data set will open in a new window. To open the subset options, click **Tables > Subset** from the toolbar at the top of the window.
3. A dialog box of subset options will appear. In our case, we want to subset by **Rows** (i.e. arrays/samples).
    - Our data has 278 rows corresponding to each sample. From our previous analysis in the *Learning Curves Step-Guide* we found that there were a sufficient number of samples to produce an effective model at around 77.5% of the samples contained in the entire data set. That is 215 of the 278 total rows in the training set, and 63 rows in the test set.
4. The **Subset** dialog has 4 options for subsetting by row.
    - One option is to highlight 215 rows in the data set window and then select **Selected Rows** to create the training set. Then to highlight the remaining 63 rows and follow the same process to create the test set.

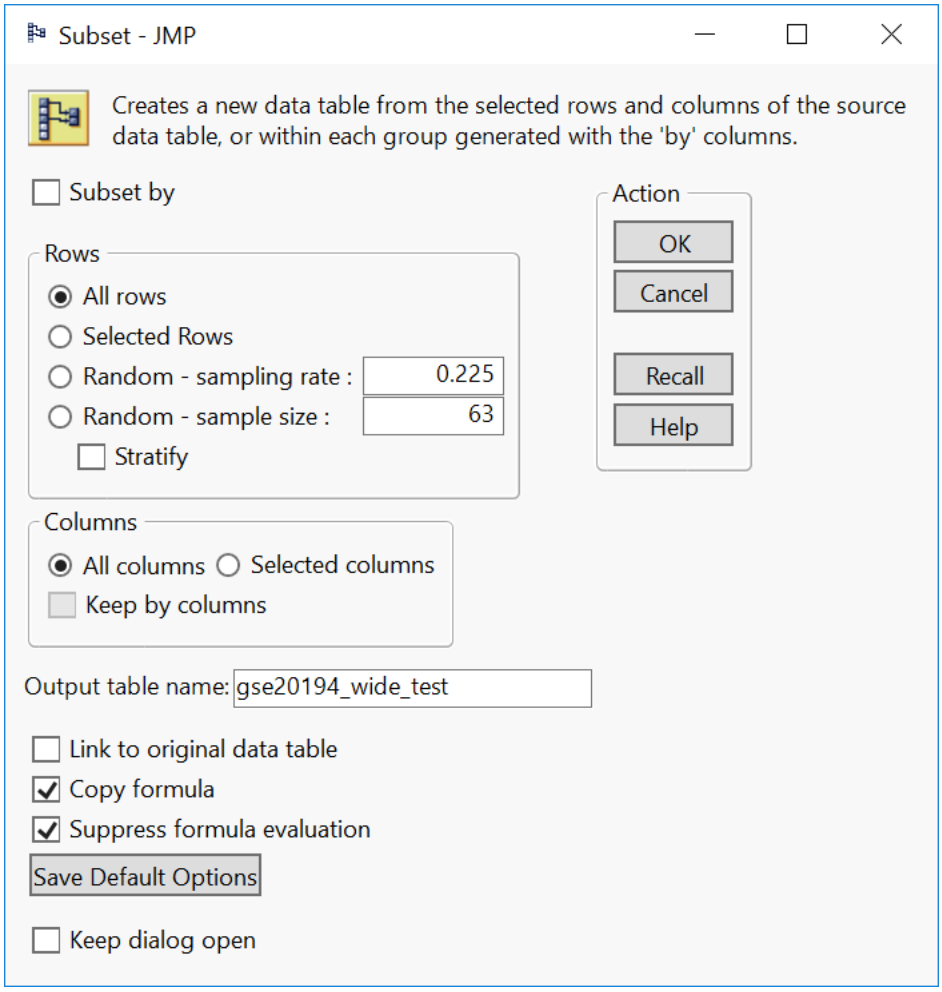


Array	Array_name	Title	Source	ColumnName	tissue	age	race	er_status	pcr_vs_rd	pr_status
204	GSM505531	BR_FNA_M443	Sample ID -- 443,...	GSM505531	breast cancer cells	46	white	N	pCR	N
205	GSM505532	BR_FNA_M486	Sample ID -- 486,...	GSM505532	breast cancer cells	44	white	P	RD	P
206	GSM505533	BR_FNA_M502	Sample ID -- 502,...	GSM505533	breast cancer cells	50	white	N	RD	N
207	GSM505534	BR_FNA_M507	Sample ID -- 507,...	GSM505534	breast cancer cells	66	white	P	RD	P
208	GSM505535	BR_FNA_M513	Sample ID -- 513,...	GSM505535	breast cancer cells	59	black	N	pCR	N
209	GSM505536	BR_FNA_M531	Sample ID -- 531,...	GSM505536	breast cancer cells	53	white	N	RD	N
210	GSM505537	BR_FNA_M545	Sample ID -- 545,...	GSM505537	breast cancer cells	58	black	N	RD	N
211	GSM505538	BR_FNA_M549	Sample ID -- 549,...	GSM505538	breast cancer cells	40	hispa...	P	RD	P
212	GSM505539	BR_FNA_M556	Sample ID -- 556,...	GSM505539	breast cancer cells	47	white	P	RD	P
213	GSM505540	BR_FNA_M557	Sample ID -- 557,...	GSM505540	breast cancer cells	57	white	N	pCR	N
214	GSM505541	BR_FNA_M558	Sample ID -- 558,...	GSM505541	breast cancer cells	62	white	N	RD	N
215	GSM505542	BR_FNA_M559	Sample ID -- 559,...	GSM505542	breast cancer cells	55	white	P	RD	N
216	GSM505543	BR_FNA_M564	Sample ID -- 564,...	GSM505543	breast cancer cells	68	white	N	RD	N
217	GSM505544	BR_FNA_M566	Sample ID -- 566,...	GSM505544	breast cancer cells	39	white	P	RD	N
218	GSM505545	BR_FNA_M571	Sample ID -- 571,...	GSM505545	breast cancer cells	32	black	N	pCR	N
219	GSM505546	BR_FNA_M576	Sample ID -- 576,...	GSM505546	breast cancer cells	70	white	N	RD	N
220	GSM505547	BR_FNA_M578	Sample ID -- 578,...	GSM505547	breast cancer cells	45	black	P	RD	P
221	GSM505548	BR_FNA_M583	Sample ID -- 583,...	GSM505548	breast cancer cells	51	white	N	RD	P
222	GSM505549	BR_FNA_M599	Sample ID -- 599,...	GSM505549	breast cancer cells	50	white	P	RD	P
223	GSM505550	BR_FNA_M607	Sample ID -- 607,...	GSM505550	breast cancer cells	67	white	P	RD	N
224	GSM505551	BR_FNA_M610	Sample ID -- 610,...	GSM505551	breast cancer cells	47	hispa...	P	RD	P
225	GSM505552	BR_FNA_M612	Sample ID -- 612,...	GSM505552	breast cancer cells	40	white	N	RD	N
226	GSM505553	BR_FNA_M617	Sample ID -- 617,...	GSM505553	breast cancer cells	50	white	N	RD	N
227	GSM505554	BR_FNA_M619	Sample ID -- 619,...	GSM505554	breast cancer cells	52	white	N	RD	N

    - Another option is to take a random sample of the set by **Sampling Rate** or **Sampling Size**. For **Sampling Rate**, enter 0.225 into the box to make a

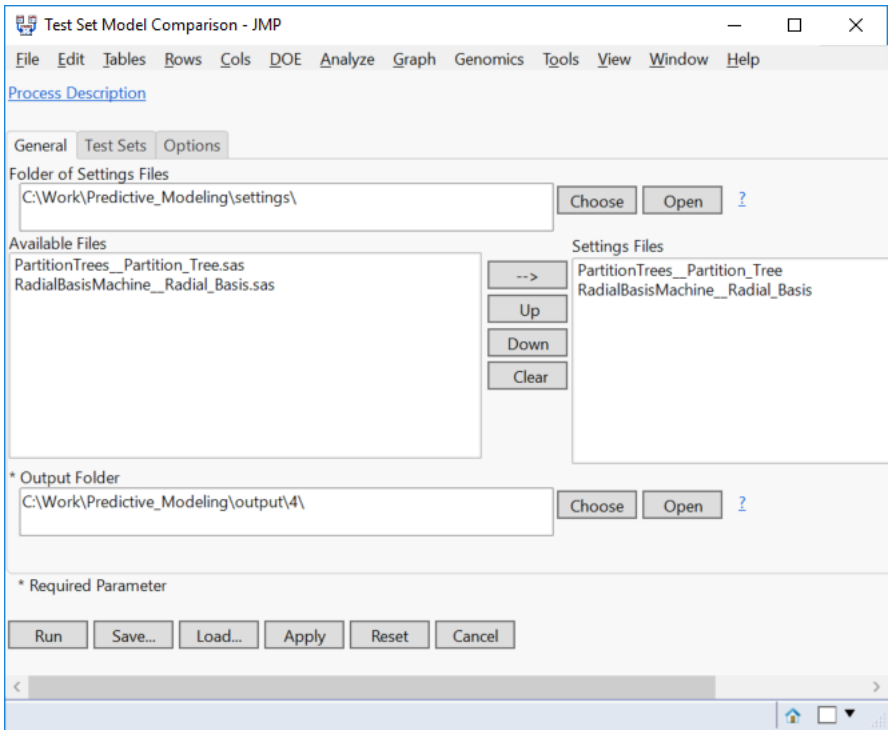
random test set subset. For **Sample Size**, enter 31 into the box to make a test set subset of the same size.

5. Create 3 test sets. One using each subsetting method.
6. Name your new data sets in the **Output table name** box. Name the first test set ***gse20194_wide_ test_1*** and the others 2 and 3, respectively. Name the training set ***gse20194_wide_training***.
    - Note you will have to make these data sets separately.

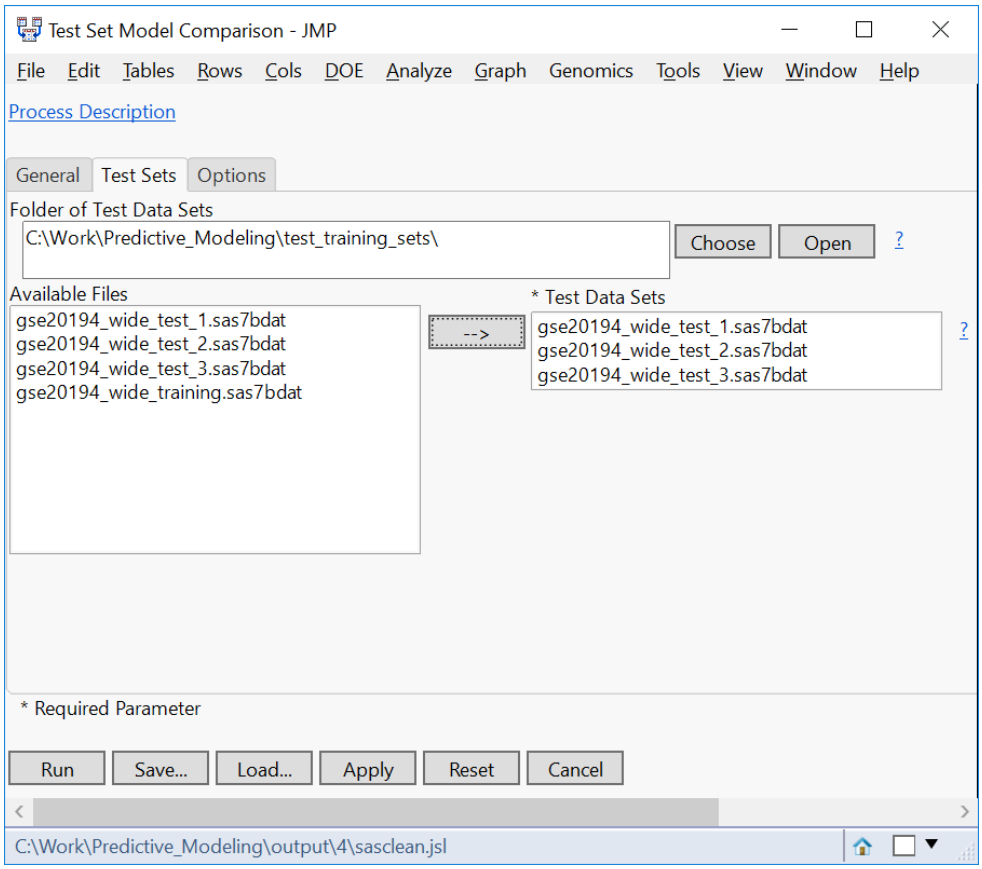


7. Click **OK** and the subset data will open in a new window. From the toolbar, click **File > Save As…** Under the file name field, select SAS Data Set (*.sas7bdat) as the file type from the **Save as type** drop down menu.
    - Note: When creating more than one test set, be sure to save all test sets to the same folder.
8. Once the **Test Sets** and a **Training Set** have been saved, select **Predictive Modeling > Model Comparisons > Test Set Model Comparison** from the **Genomics Starter** menu.
9. In the **General** tab, find and select the **Folder of Settings** where the two models created in the *Cross Validation Model Comparison Step-Guide* were saved.

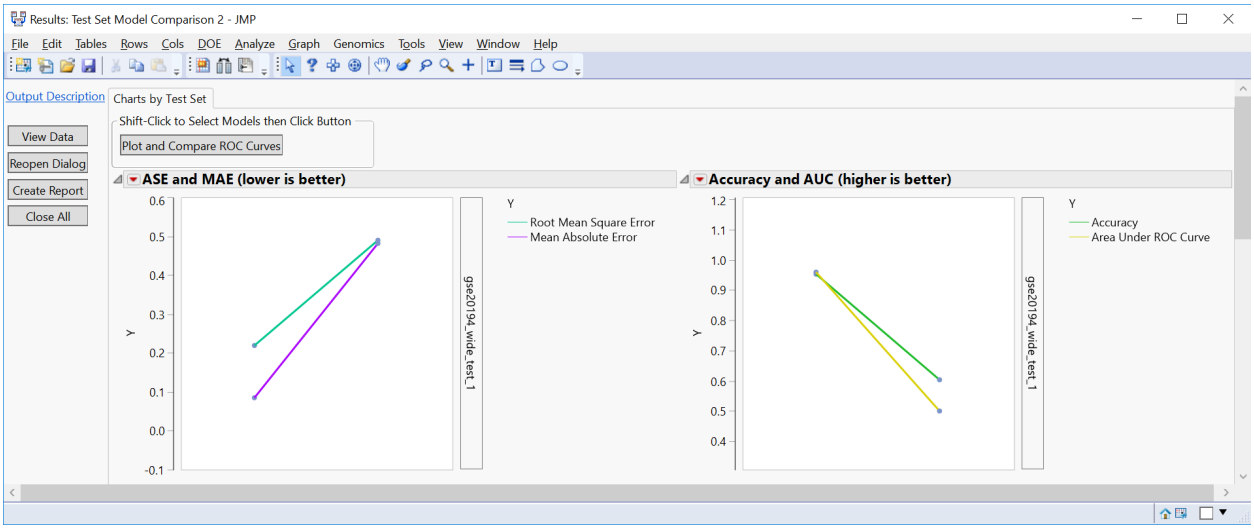
10. Select both the **Partition Trees** and **Radial Basis Machine** models from the **Available Files** box.
11. Specify an **Output Folder**.



12. In the **Test Sets** tab, choose the folder containing the **Test Set** files saved in step 6 as the **Folder of Test Data Sets**.
13. Select the test sets, ***gse20194_wide_test_1.sas7bdat, gse20194_wide_test_2.sas7bdat,*** and ***gse20194_wide_test_3.sas7bdat*** from the **Available Files** box as the **Test Data Sets** to use.



14. In the **Options** tab, there are options for the output to be separated by model or by test set. Select **Separate Charts for Each Test Set**.
15. Click **Run**.



16. Four analyses are performed for each test set and model: Accuracy, Area Under Curve, Average Square Error and Mean Absolute Error. Each Test Set is shown in a different plot. These results show that the **Partition Trees** model performs significantly better on all three Test Sets.

Summary

Using the **Test Set Model Comparison** tool is a quick way to evaluate the effectiveness of multiple models by fitting them to multiple sets of hold-out data called **Test Sets**. This guide showed a few different methods for subsetting data to create these **Test Sets**, as well as how to use the **Test Set Model Comparison** tool and interpret the results. Note that other tools in JMP Genomics such as **Cross Validation Model Comparison** (See: *Cross Validation Model Comparison Step-Guide*) compare models through similar methods and automate the test set creation process. However, these methods are much more computationally and time intensive. For the last step in the predictive modeling pipeline, we will fit the **Partition Trees** model to the training set in the *Whole Model Fit Step-Guide.*