

Automated and Interactive Model Screening to Identify the Champion Model

Chong Ho Yu, Ph.D., D. Phil.

Professor and Director of Data Analytics

Azusa Pacific University

2021 IM Data Conference

A plethora of DSML methods

- Single methods (non-ensemble)
 - Support vector machine (SVM): linear, polynomial kernel, radial basis function (RBF), sigmoid.
 - Naïve Bayes
 - Decision tree: C5, Chi-square automatic interaction detection (CHAID), Quick, unbiased, efficient statistical tree (QUEST), Classification and regression tree (CRT)

A plethora of DSML methods

- Ensemble method
 - Bagging
 - Random forest
 - Boosting: Gradient boosting, XGBoost, Adaboost, LightGBM, CARboost
- Neural network

Which one should I use? Any consensus?

- Neural network is a black box; it is hard to interpret.
- In some situations, bagging outperforms boosting whereas in others the outcomes are reversed (Chandrabhasan et al. 2011, Dietterich 2000, Khoshgoftaar et al. 2011, Kotsiantis 2013, Wang et al. 2015, Zaman and Hirose 2011).
- The difference is minimal. In a study comparing between random forest and XGBoost in breast cancer risk prediction, random forest achieved 74.73% accuracy while XGBoost obtained 73.63% (Kabiraj et al. 2020).
- XGBoost is more widely used than gradient boost and Adaboost because of its higher accuracy, faster speed, and less sensitivity to noisy data (Deng et al. 2020, Niu 2020).

Model screening/Model comparison

- Run multiple models and select the champion model.
- Automatic or interactive (more human intervention)
- Two demos/illustrations
 - Classification problem (the DV is binary)
 - Regression problem (the DV is continuous)

Classification problem

- JMP Pro
- Predict diabetes
- It is always a good practice to include traditional statistical procedures as the baseline (e.g. logistic regression). You may be surprised!

Model Screening - JMP Pro

Fits many different predictive models and provides summaries of measures of fit.

Select Columns

▼ 14 Columns

- Y
- Y Binary
- Y Ordinal
- Age
- Gender
- BMI
- BP
- Total Cholesterol
- LDL
- HDL
- TCH
- LTG
- Glucose
- Validation

Cast Selected Columns into Roles

Y, Response Y Binary
optional

X, Factor Age
 Gender
 BMI
 BP

Weight *optional numeric*

Freq *optional numeric*

Validation Validation

By *optional*

Action

OK

Cancel

Remove

Recall

Help

Method

- Decision Tree
- Bootstrap Forest
- Boosted Tree
- K Nearest Neighbors
- Naive Bayes
- Neural
- Support Vector Machines
- Discriminant
- Fit Least Squares
- Fit Stepwise
- Logistic Regression
- Generalized Regression
- Partial Least Squares
- XGBoost

Options

- Remove Live Reports
- Log Methods
- Time Limit Each
- Set Random Seed

Folded Crossvalidation

Fit repeatedly with sequenced folds.

- K Fold Crossvalidation K
- Nested Crossvalidation K L
- Repeated K Fold

Modeling Options

- Add Two Way Interactions
- Add Quadratics
- Informative Missing
- Additional Methods

Classification problem

- Based on multiple criteria, the best two models are **logistic regression** and **SVM**.
- The bottom one is **Naïve Bayes**.
- But don't take it as final!

Diabetes - Model Screening of Y Binary - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Six Sigma Tools Tools Add-Ins View Window Help

Model Screening for Y Binary

Table: Diabetes Response: Y Binary Validation: Validation

Details

- Partition for Y Binary
- Bootstrap Forest for Y Binary
- Boosted Tree for Y Binary
- Naive Bayes
- Support Vector Machine
- XGBoost
- Nominal Logistic Fit for Y Binary
- Generalized Regression for Y Binary = High

Training

Validation

Method	N	Entropy RSquare	Misclassification Rate	AUC	RASE	Generalized RSquare
Nominal Logistic	133	0.3886	0.1504	0.8950	0.33448	0.5329
Generalized Regression Lasso	133	0.3886	0.1504	0.8947	0.33456	0.5329
Bootstrap Forest	133	0.3373	0.1880	0.8748	0.35876	0.4759
Support Vector Machines	133	0.3280	0.1429	0.8695	0.35345	0.4652
Boosted Tree	133	0.3062	0.2180	0.8609	0.36878	0.4397
Decision Tree	133	0.1967	0.2331	0.7974	0.39828	0.3006
XGBoost	133	-0.002	0.2556	0.8227	0.41615	-0.003
Naive Bayes	133	-0.183	0.2406	0.8620	0.42568	-0.352

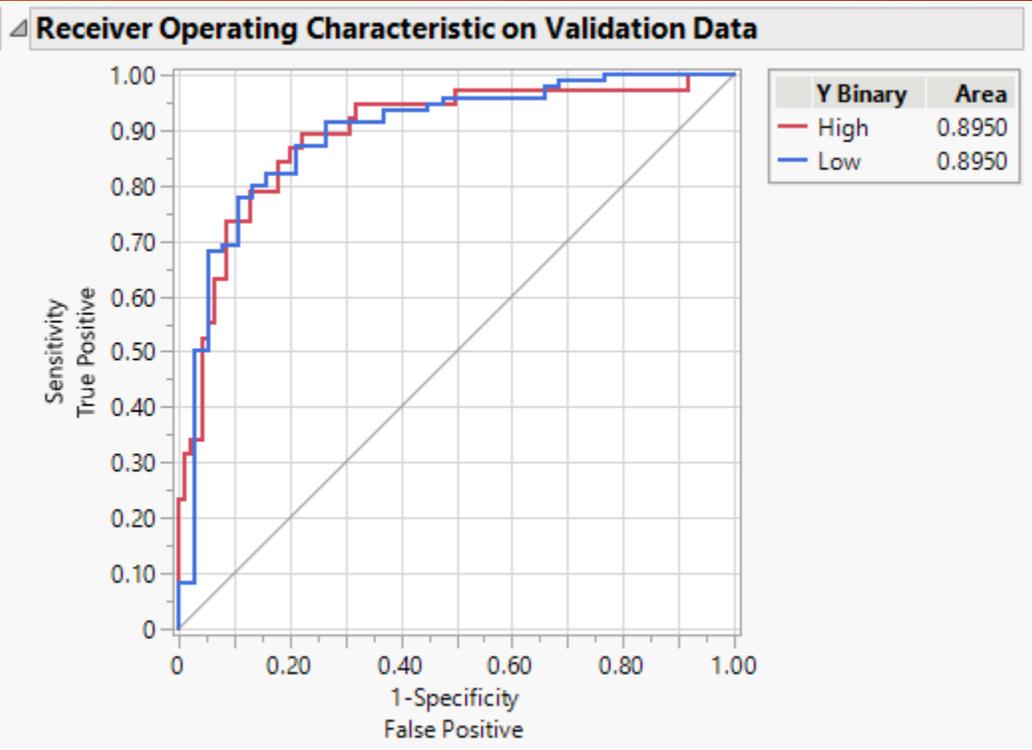
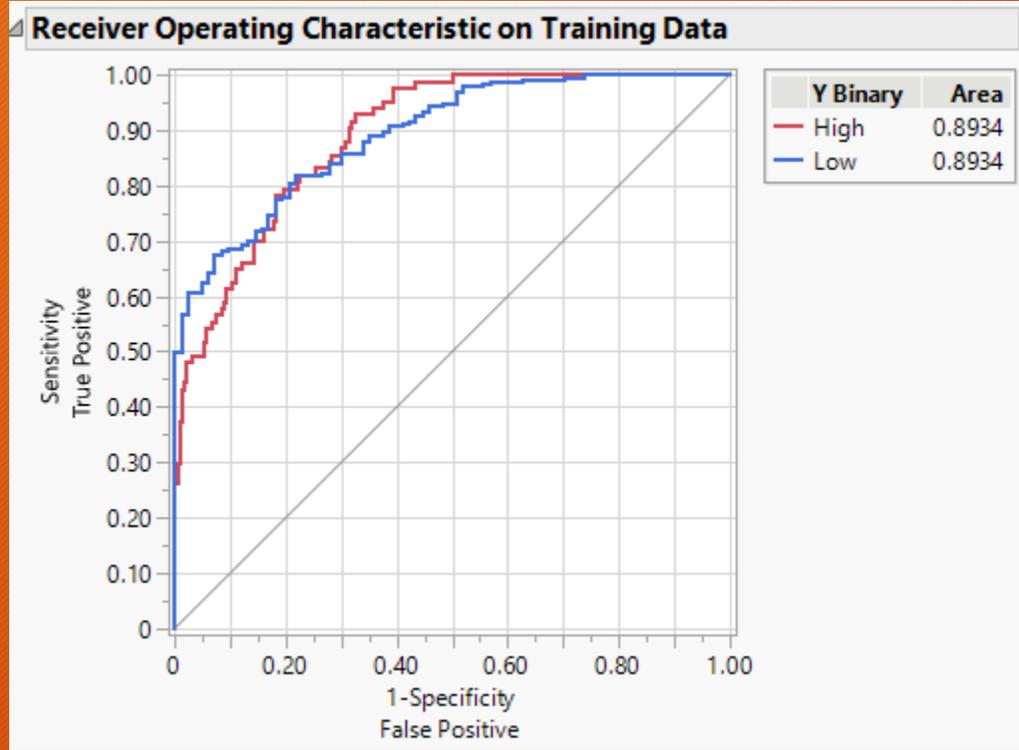
[Select Dominant](#) [Run Selected](#) [Save Script Selected](#)

Sum Freq and Sum Weight are suppressed when they are the same as N.

Run logistic regression

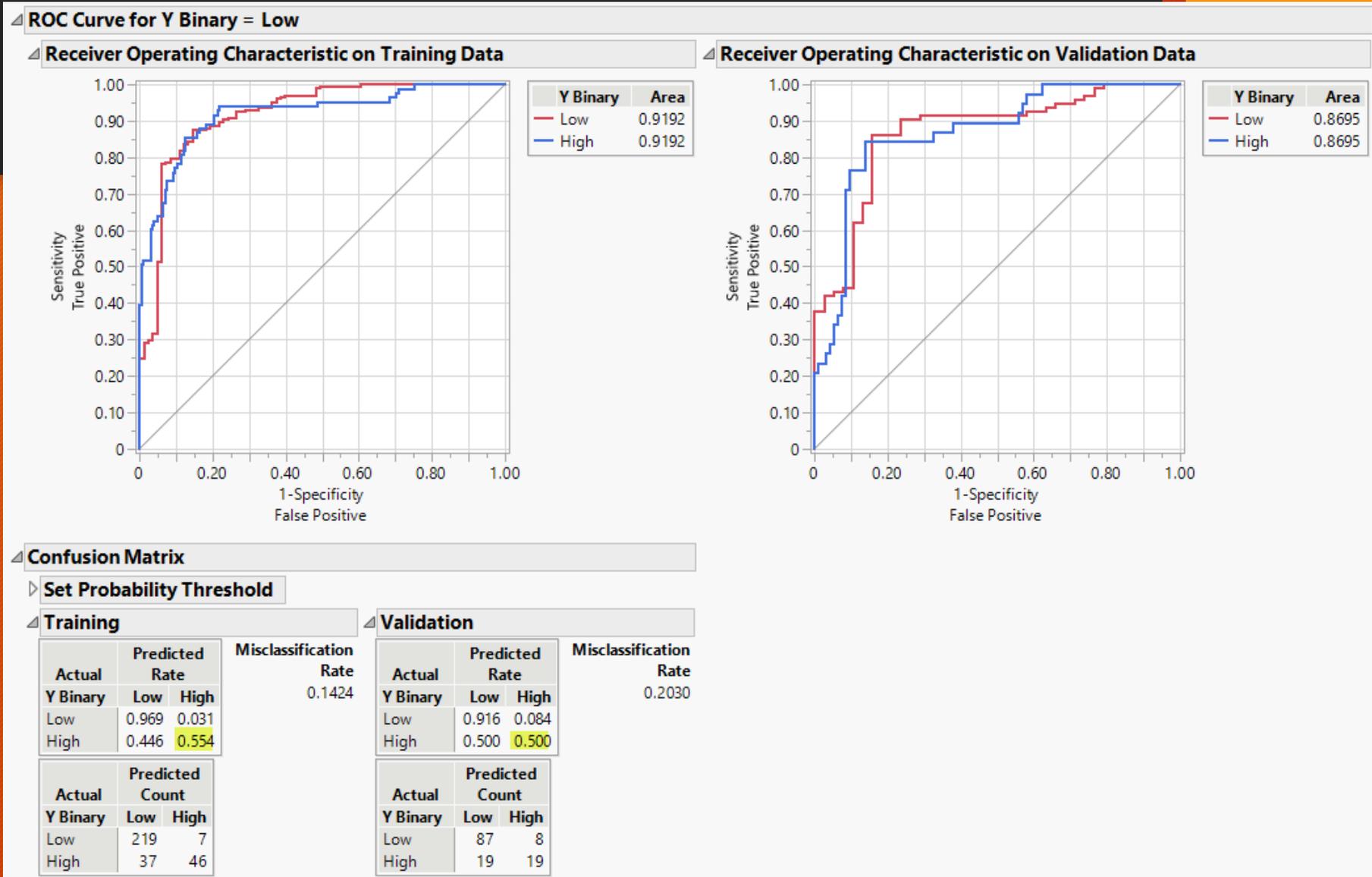
Source	LogWorth	PValue
BMI	3.892	0.00013
BP	3.182	0.00066
LTG	2.385	0.00413
Gender	1.102	0.07903
Total Cholesterol	0.971	0.10680
LDL	0.852	0.14066
HDL	0.353	0.44366
TCH	0.220	0.60189
Age	0.210	0.61634
Glucose	0.089	0.81415

[Remove](#) [Add](#) [Edit](#) FDR



Run SVM

- Predicted rate for low risk group: 50%
- You can flip a coin!



Classification problem

- IBM SPSS Modeler: Auto classifier.
- Again, include logistic regression as a baseline.



Diabetes.sav → Type → YBinary

YBinary

Estimated number of models to be executed: 7

Fields Model **Expert** Discard Settings Annotations

Select models: All models

Use?	Model type	Model parameters	No of models
<input checked="" type="checkbox"/>	C5	Default	1
<input checked="" type="checkbox"/>	Logistic regression	Default	1
<input checked="" type="checkbox"/>	Decision List	Default	1
<input type="checkbox"/>	Bayesian Network	Default	1
<input type="checkbox"/>	Discriminant	Default	1
<input type="checkbox"/>	KNN Algorithm	Default	1
<input type="checkbox"/>	LSVM	Default	1
<input type="checkbox"/>	Random Trees	Default	1
<input type="checkbox"/>	SVM	Default	1
<input type="checkbox"/>	Tree-AS	Default	1
<input type="checkbox"/>	XGBoost Linear	Default	1
<input checked="" type="checkbox"/>	XGBoost Tree	Default	1
<input checked="" type="checkbox"/>	CHAID	Default	1
<input checked="" type="checkbox"/>	Quest	Default	1
<input checked="" type="checkbox"/>	C&R Tree	Default	1
<input checked="" type="checkbox"/>	Random Forest	Default	1
<input type="checkbox"/>	Neural Net	Default	1

Restrict maximum time spent building a single model to 15 minutes

Stopping rules... Misclassification costs...

OK Run Cancel Apply Reset

Classification problem

- The best model is **random forest**.
- **Logistic regression** is near the bottom!
- It is different from the result of SAS/JMP!

YBinary

File Generate View Preview

Model Graph Summary Settings Annotations

Sort by: Use Ascending Descending Delete Unused Models View: Training set

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
<input checked="" type="checkbox"/>		Random Forest 1	< 1	99.774	10
<input checked="" type="checkbox"/>		XGBoost Tree 1	< 1	98.19	10
<input checked="" type="checkbox"/>		C5 1	< 1	89.593	9
<input checked="" type="checkbox"/>		Logistic regress...	< 1	83.937	10
<input checked="" type="checkbox"/>		CHAID 1	< 1	82.805	5

OK Cancel Apply Reset

Regression problem

- PISA 2018
- You can select multiple modeling methods, including traditional approaches (e.g. OLS regression & stepwise regression) and modern data science methods (e.g. decision tree, random forest, boosted tree, neural networks, XGBoost...etc.)

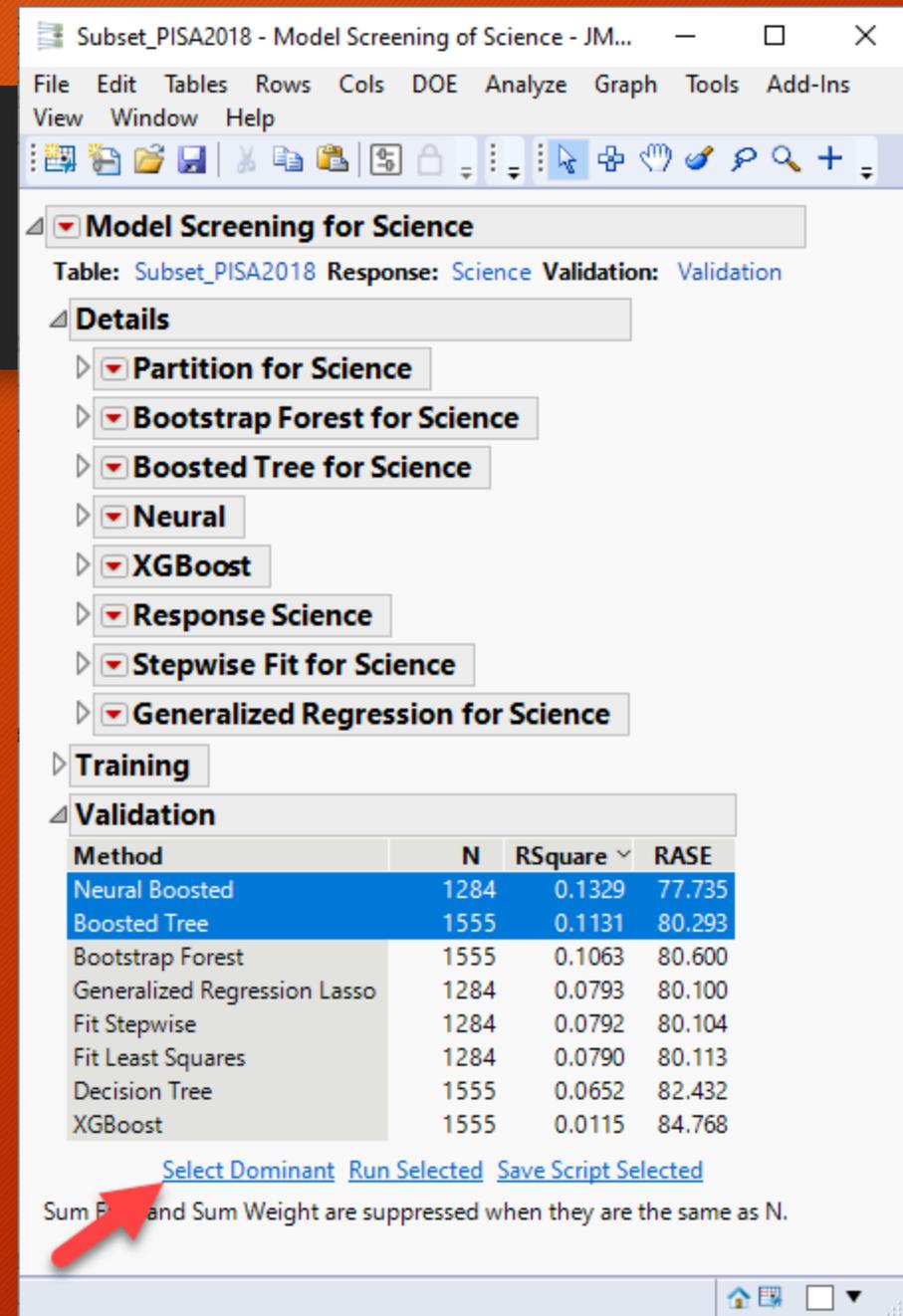
The screenshot shows the 'Model Screening - JMP Pro' dialog box. The title bar indicates the software version. Below the title, a subtitle reads 'Fits many different predictive models and provides summaries of measures of fit.' The dialog is divided into several sections:

- Select Columns:** A list of 14 columns is shown, with 'Validation' selected at the bottom. The columns include Country/region, Intl. School ID, Home possessions, Cultural possessions at home, Home educational resources, Family wealth, Teacher-directed instruction, Eudaemonia: meaning in life, Subjective well-being: Positive affect, Subjective well-being: S... of belonging to school, Social Connections: Parents, Math, Science, and Validation.
- Cast Selected Columns into Roles:** This section allows assigning roles to columns. 'Y, Response' is assigned to 'Science' (optional). 'X, Factor' is assigned to a list of columns: 'Home ...ssessions', 'Cultural...t home', 'Home e...sources', and 'Family wealth'. Other roles include 'Weight' (optional numeric), 'Freq' (optional numeric), 'Validation' (Validation), and 'By' (optional).
- Method:** A list of modeling methods with checkboxes. Checked methods include Decision Tree, Bootstrap Forest, Boosted Tree, Neural, Fit Least Squares, Fit Stepwise, Generalized Regression, and XGBoost. Unchecked methods include K Nearest Neighbors, Naive Bayes, Support Vector Machines, Discriminant, Logistic Regression, and Partial Least Squares.
- Options:** Includes checkboxes for 'Remove Live Reports' and 'Log Methods', and input fields for 'Time Limit Each' and 'Set Random Seed'.
- Folded Crossvalidation:** Includes a checkbox for 'Fit repeatedly with sequenced folds.' and options for 'K Fold Crossvalidation' (K=5), 'Nested Crossvalidation' (K=4, L=5), and 'Repeated K Fold' (0).
- Modeling Options:** Includes checkboxes for 'Add Two Way Interactions', 'Add Quadratics', 'Informative Missing', and 'Additional Methods'.

On the right side, there is an 'Action' panel with buttons for 'OK', 'Cancel', 'Remove', 'Recall', and 'Help'.

JMP Pro: Model screening

- The best two are **neural boosted** and **gradient boosting**.
- Suppose XGBoost should outperform gradient boosting, but it is at the bottom!



Subset_PISA2018 - Model Screening of Science - JM...

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins
View Window Help

Model Screening for Science

Table: Subset_PISA2018 Response: Science Validation: Validation

Details

- Partition for Science
- Bootstrap Forest for Science
- Boosted Tree for Science
- Neural
- XGBoost
- Response Science
- Stepwise Fit for Science
- Generalized Regression for Science

Training

Validation

Method	N	RSquare	RASE
Neural Boosted	1284	0.1329	77.735
Boosted Tree	1555	0.1131	80.293
Bootstrap Forest	1555	0.1063	80.600
Generalized Regression Lasso	1284	0.0793	80.100
Fit Stepwise	1284	0.0792	80.104
Fit Least Squares	1284	0.0790	80.113
Decision Tree	1555	0.0652	82.432
XGBoost	1555	0.0115	84.768

[Select Dominant](#) [Run Selected](#) [Save Script Selected](#)

Sum P and Sum Weight are suppressed when they are the same as N.

SPSS Modeler

- Keep OLS regression as the baseline model.

The diagram shows a workflow in SPSS Modeler. It starts with a data source node labeled 'PISA2018.sav', followed by a 'Type' node, and finally a 'Science' node. The nodes are connected by blue arrows pointing from left to right. The 'Science' node is highlighted with a yellow border.

The 'Science' dialog box is open, showing the 'Expert' tab. The 'Estimated number of models to be executed: 6' is displayed at the top. The 'Fields' tab is selected, and the 'All models' dropdown is visible. The table below lists the model types and their parameters.

Use?	Model type	Model parameters	No of models
<input checked="" type="checkbox"/>	Regression	Default	1
<input checked="" type="checkbox"/>	Generalized Linear	Default	1
<input type="checkbox"/>	Generalized linear ...	Default	1
<input type="checkbox"/>	KNN Algorithm	Default	1
<input type="checkbox"/>	Linear-AS	Default	1
<input type="checkbox"/>	LSVM	Default	1
<input checked="" type="checkbox"/>	Random Trees	Default	1
<input type="checkbox"/>	SVM	Default	1
<input type="checkbox"/>	Tree-AS	Default	1
<input type="checkbox"/>	XGBoost Linear	Default	1
<input checked="" type="checkbox"/>	XGBoost Tree	Default	1
<input type="checkbox"/>	Linear	Default	1
<input type="checkbox"/>	CHAID	Default	1
<input checked="" type="checkbox"/>	C&R Tree	Default	1
<input type="checkbox"/>	XGBoost-AS	Default	1
<input type="checkbox"/>	Random Forest	Default	1
<input checked="" type="checkbox"/>	Neural Net	Default	1

Restrict maximum time spent building a single model to minutes

Stopping rules...

OK Run Cancel Apply Reset

SPSS Modeler

- XGBoost is the best!
- It is opposite to the result of SAS/JMP!

The screenshot shows the SPSS Modeler interface with a table of models. The table columns are: Use?, Graph, Model, Build Time (mins), Correlation, No. Fields Used, and Relative Error. The XGBoost model is the top performer.

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
<input checked="" type="checkbox"/>		XGBoost Tre...	< 1	0.664	9	0.648
<input checked="" type="checkbox"/>		Random Tre...	< 1	0.633	9	0.633
<input checked="" type="checkbox"/>		Neural Net 1	< 1	0.289	9	0.918
<input checked="" type="checkbox"/>		Regression 1	< 1	0.279	9	0.922
<input checked="" type="checkbox"/>		Generalized ...	< 1	0.279	9	0.922

Why?

- Different software packages have **different default tuning parameters** and also their algorithms are slightly different.
- Automatic model comparisons using different software packages with different default parameters might lead to **very different results**.
- Many software packages offer both automated and interactive model comparison e.g.
 - IBM SPSS Modeler
 - JMP Pro
 - SAS Enterprise Miner
 - SAS Viya: Model Studio

JMP Pro: Model comparison

PISA2006_USA_Canada - Model Comparison - JMP

Model Comparison

Predictors

Measures of Fit for proficiency

Creator	.2	.4	.6	.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Ordinal Logistic					0.0765	0.1322	0.6209	0.4642	0.4313	0.3407	13829
Bootstrap Forest					0.1345	0.2239	0.5832	0.4448	0.4259	0.2807	16236
Boosted Tree					0.0627	0.1096	0.6325	0.4699	0.4525	0.3468	16602

AUC Comparison

AUC Comparison for proficiency= 1

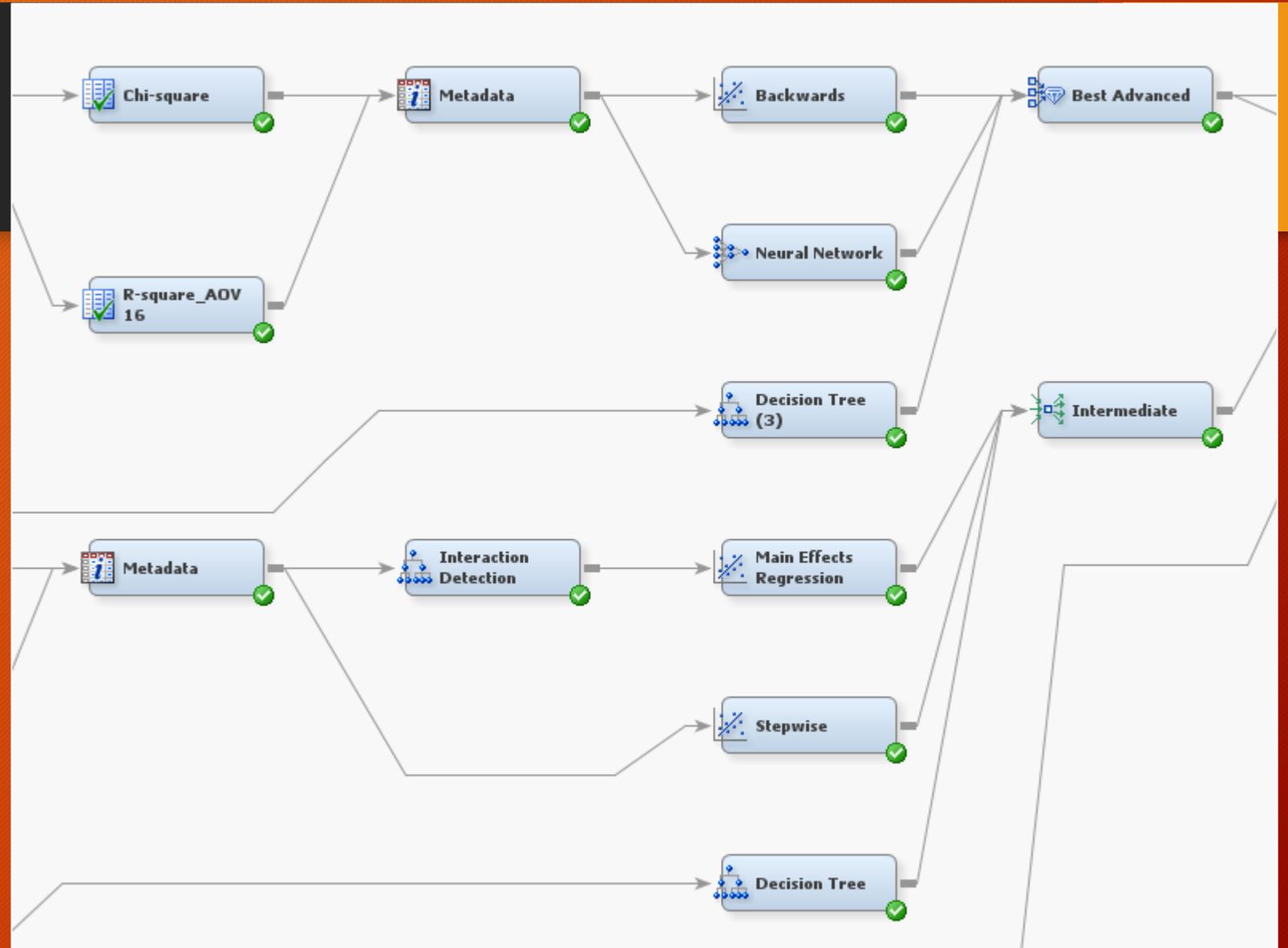
Predictor	AUC	Std Error	Lower 95%	Upper 95%
Prob[1]	0.6834	0.0046	0.6743	0.6924
Prob(proficiency== 1)	0.7801	0.0036	0.7729	0.7872
Prob(proficiency== 1) 2	0.6743	0.0042	0.6660	0.6826

Predictor	vs. Predictor	AUC Difference	Std Error	Lower 95%	Upper 95%	ChiSquare	Prob> ChiSq
Prob[1]	Prob(proficiency== 1)	-0.097	0.0019	-0.100	-0.093	2708.1	<.0001*
Prob[1]	Prob(proficiency== 1) 2	0.0091	0.0007	0.0076	0.0105	153.23	<.0001*
Prob(proficiency== 1)	Prob(proficiency== 1) 2	0.1058	0.0019	0.1020	0.1095	3061.7	<.0001*

Test	ChiSquare	DF	Prob> ChiSq
All AUCs equal	3067.02	2	<.0001*

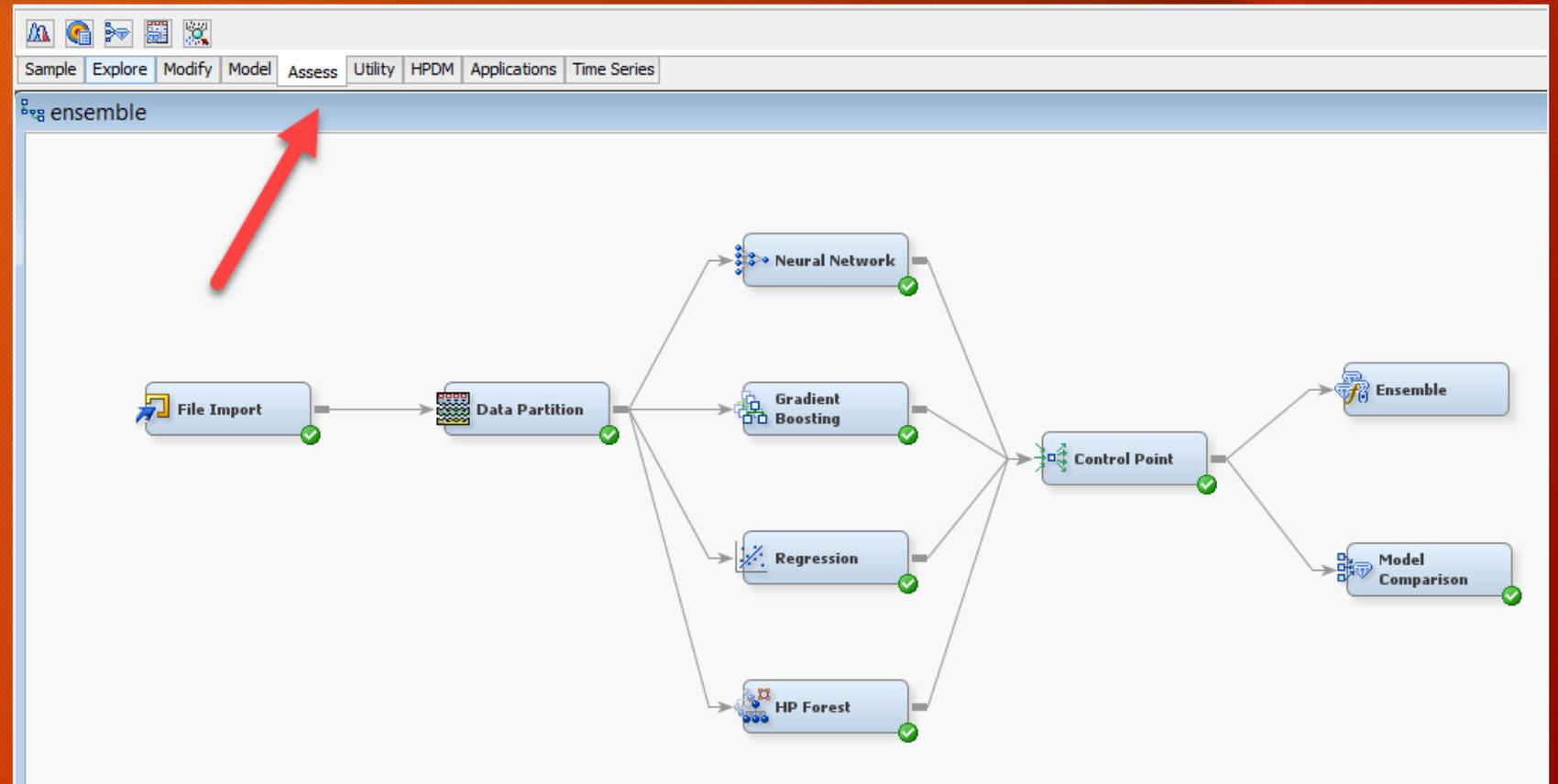
SAS Enterprise Guide

- Rapid Data Mining
- Totally automatic
- Just a few clicks

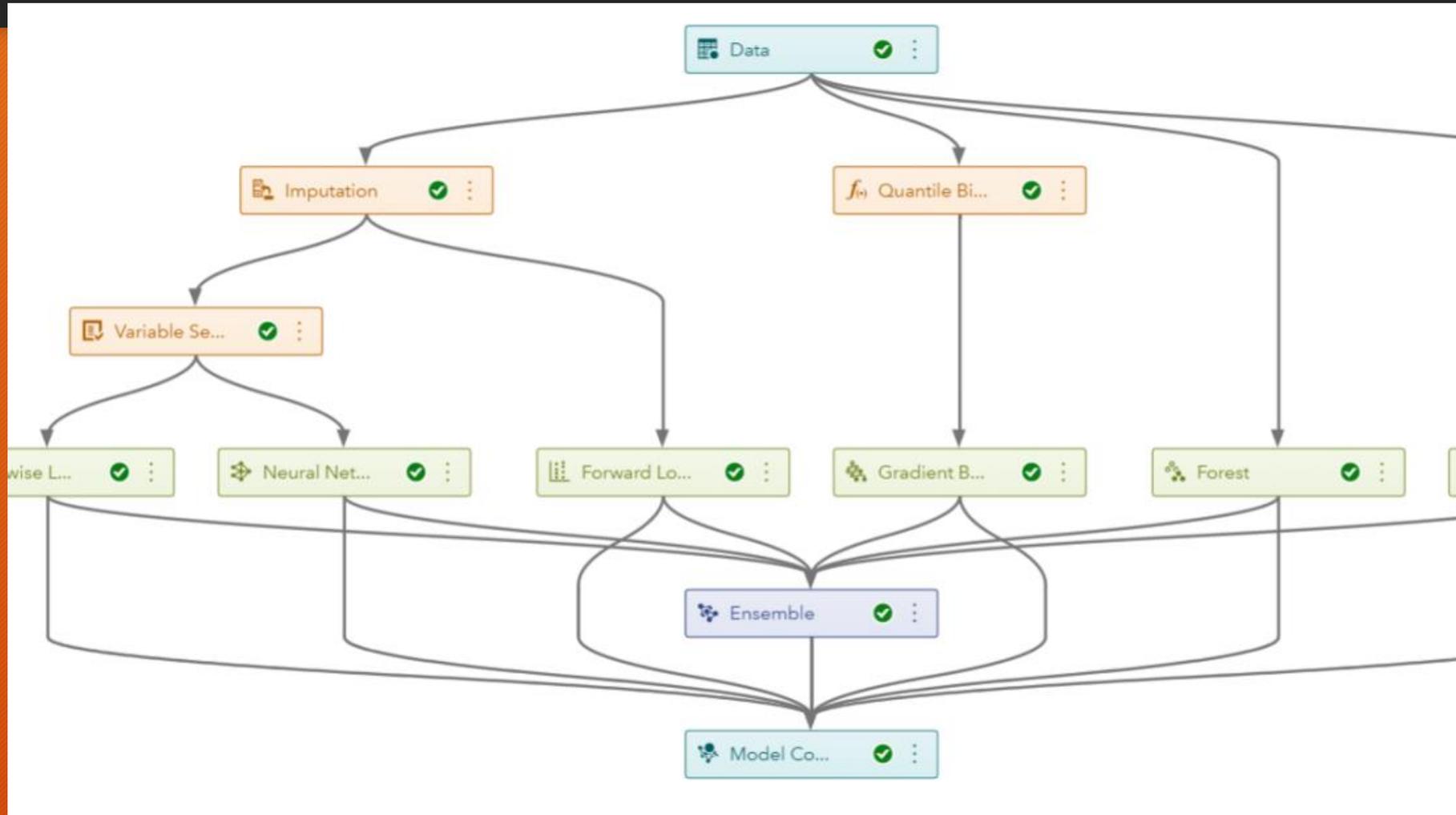


SAS Enterprise Miner

- Interactive model comparison
- Change parameters along the way.



SAS Viya: Model Studio



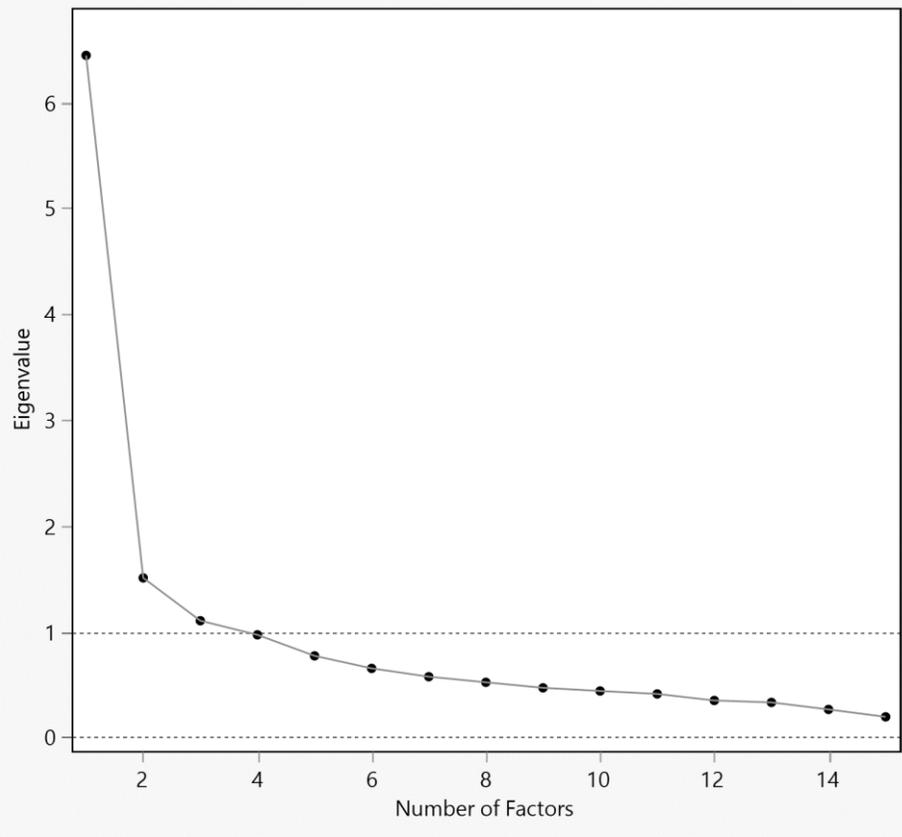
Challenges

- When the data set are massive or/and the analytical tasks are complicated, running multiple models in one job (model screening or model comparison) can take a long time.
- Solution: **High performance computing (HPC)**
- designed to utilize **multi-threading**.
- Complex analytical tasks are divided across processing nodes in a **distributed system**, and at the end the results are assembled into a single, final presentation.
- Drawback: if HP procedures are run on an environment that **do not have HPC resources**, it will take longer or cannot run at all!

Challenges

- If HPC resources are NOT available, do variable pre-screening!
- Is it necessary to collect so many data (e.g. 400-500 fields)?
- Is it necessary to include all 400-500 features (variables)?
 - **Variable selection**: drop the variables that are less important or unimportant e.g. stepwise regression (traditional, not recommended), generalized regression, and predictor screening (better)
 - **Dimension reduction**: Collapse variables into a few dimensions e.g. principal component analysis (PCA), partial least square.
- Use the remaining for model comparison.

After select the champion model...



Column Contributions

Term	Number of Splits	SS	Portion
Home possessions	1228	34269551	0.3928
Family wealth	1319	12172007.2	0.1395
Teacher-directed instruction	629	8623031.84	0.0988
Internet and computer technology resources	548	6649776.28	0.0762
Eudaemonia: meaning in life	466	4135997.67	0.0474
Subjective well-being: Sense of belonging to school	721	3943683.12	0.0452
Interest in Internet and computer technology	365	3867270.18	0.0443
Cultural possessions at home	688	2513755.89	0.0288
Mastery goal orientation	503	1848368.94	0.0212
Home educational resources	424	1397594.56	0.0160
Information about careers	374	1209676.69	0.0139
Parents' emotional support perceived by student	266	1068124.79	0.0122
Resilience	428	1040483.96	0.0119
Body image	378	1030046.39	0.0118
Perceived feedback	436	901777.22	0.0103
General fear of failure	353	801114.583	0.0092
Subjective well-being: Positive affect	297	699484.756	0.0080
Social Connections: Parents	225	695923.813	0.0080
Perception of cooperation at school	177	328041.924	0.0038
Parents' emotional support	28	42727.8656	0.0005

Conclusion

- Do **pre-screening** to cut down the number of predictors.
- Using **automated model comparison** is OK, but should be used with caution.
- Include **traditional modeling methods** as the baseline (e.g. logistic regression, OLS regression, stepwise regression...etc.)
- Use **more than one software packages**. If they don't agree, turn to interactive model comparison.
- Use **HP procedures** if resources are available.
- After selecting the best model, retain predictors by looking for the **inflection point**.