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?

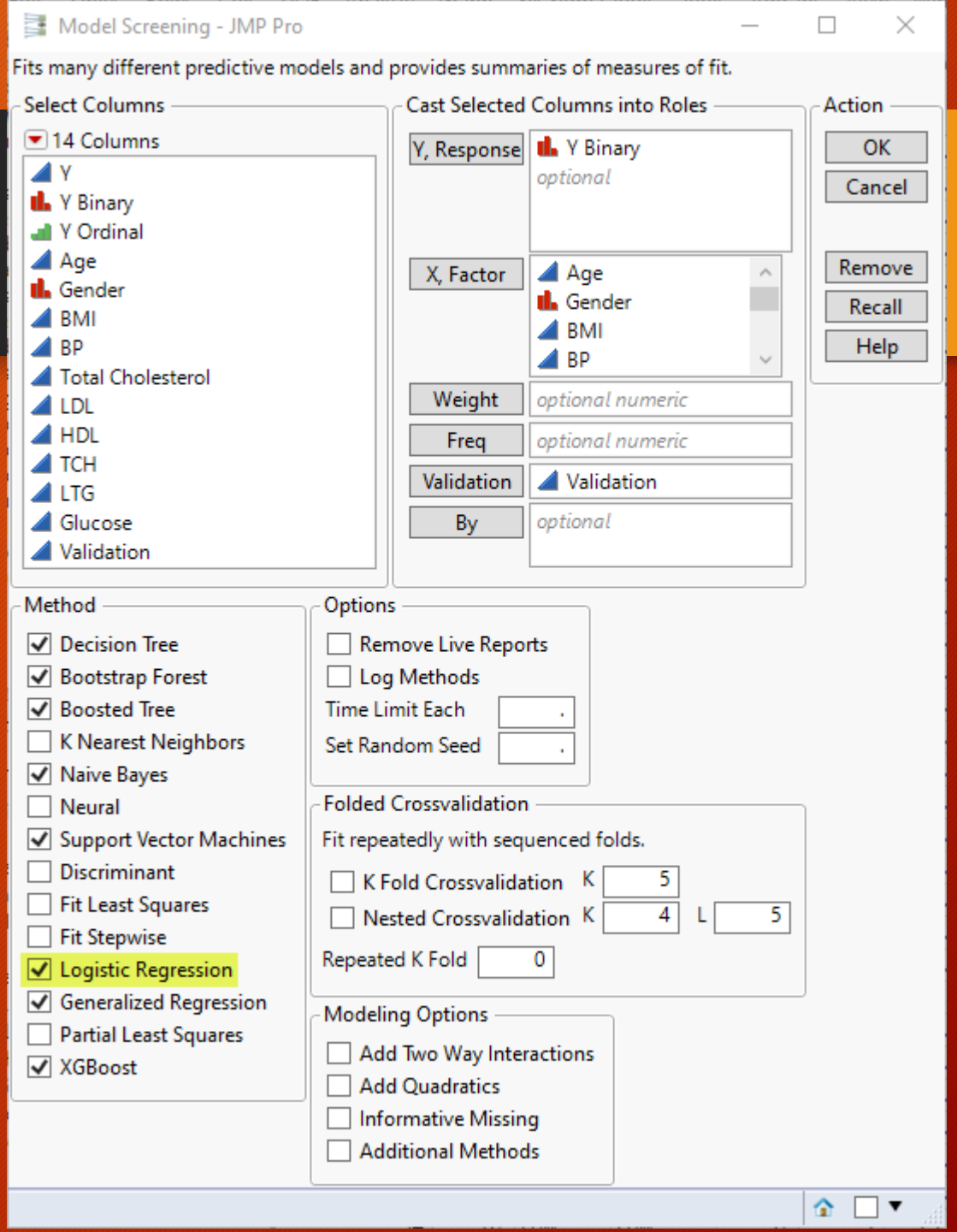
- Neural network is a black box; it is hard to interpret.
- In some situations, bagging outperforms boosting whereas in others the outcomes are reversed (Chandrahasan et al.2011, Dietterich 2000, Khoshgoftaar et al. 2011, Kotsiantis 2013, Wang et al. 2015, Zaman and Hirose 2011).
- The difference is 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 Adabost 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)

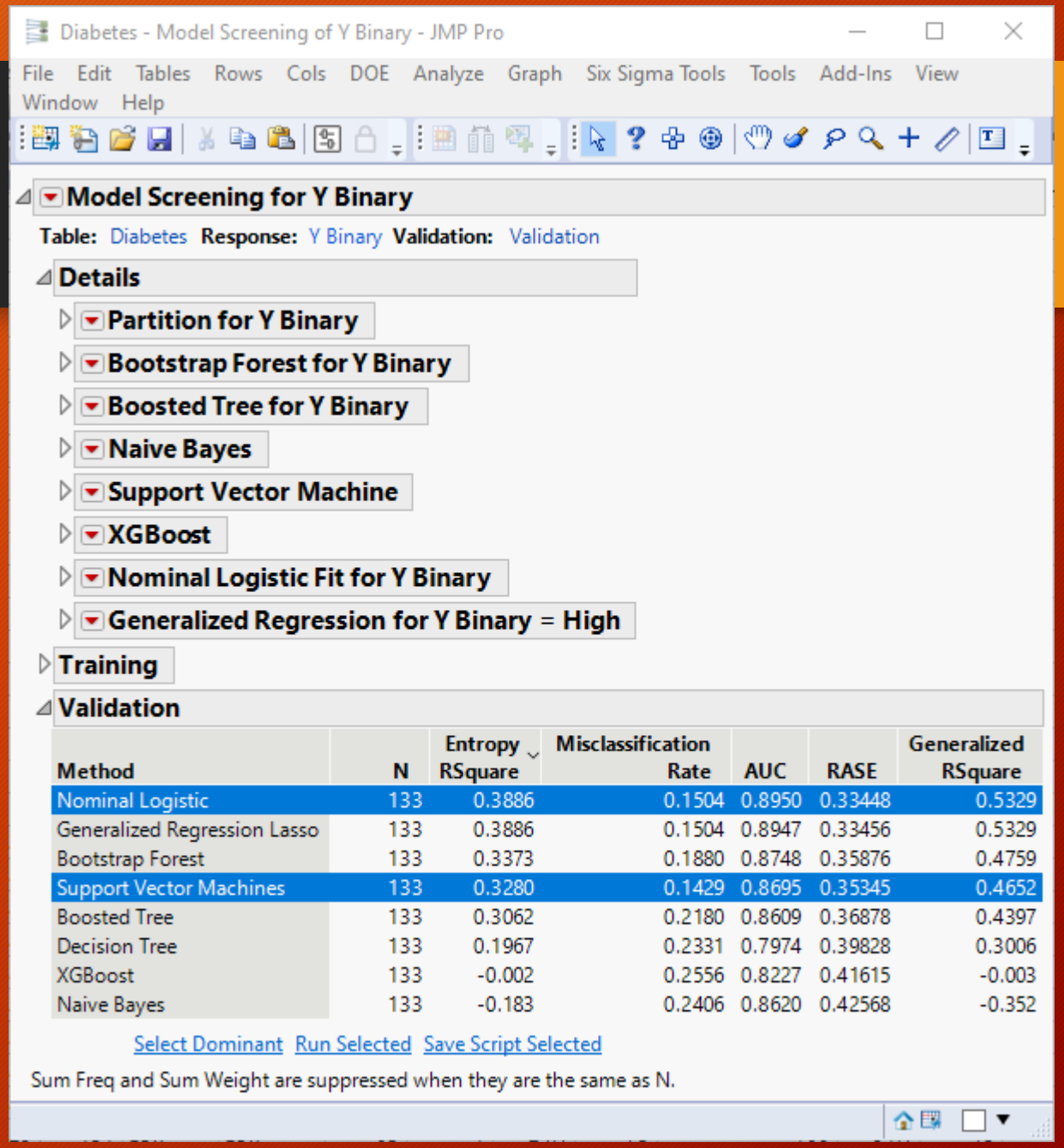
Classification problem

- 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!



Classification problem

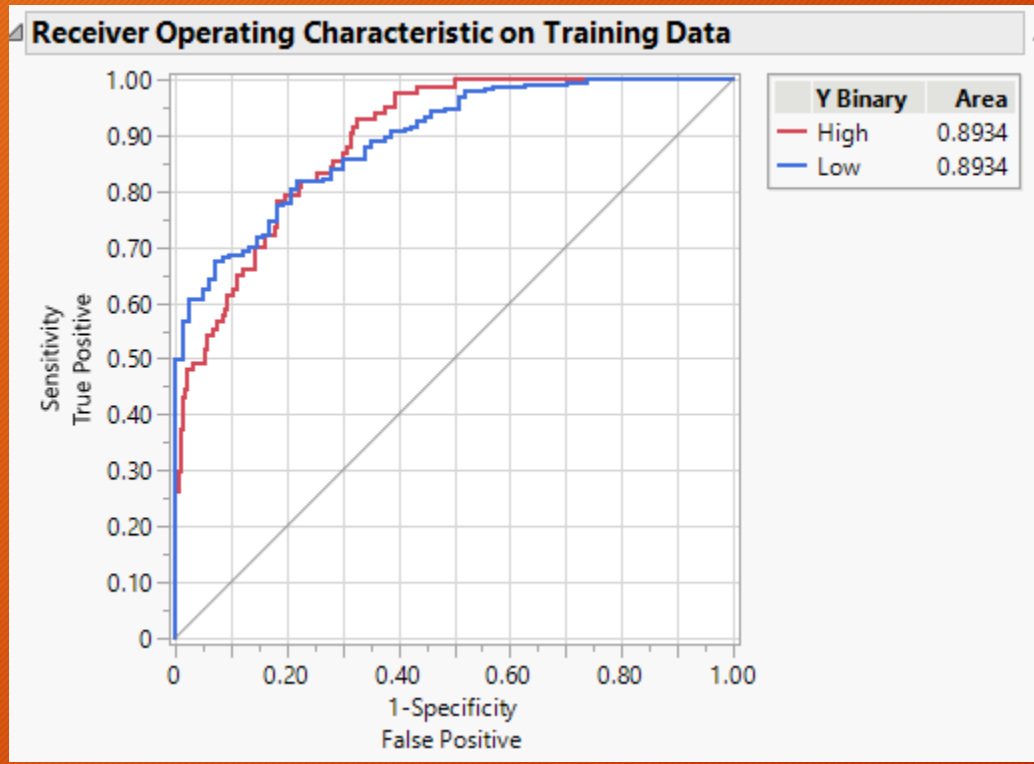
- Based on multiple criteria, the best two models are logistic regression and SVM.
- The bottom one is Naïve Bayes.
- But don't take it as final!

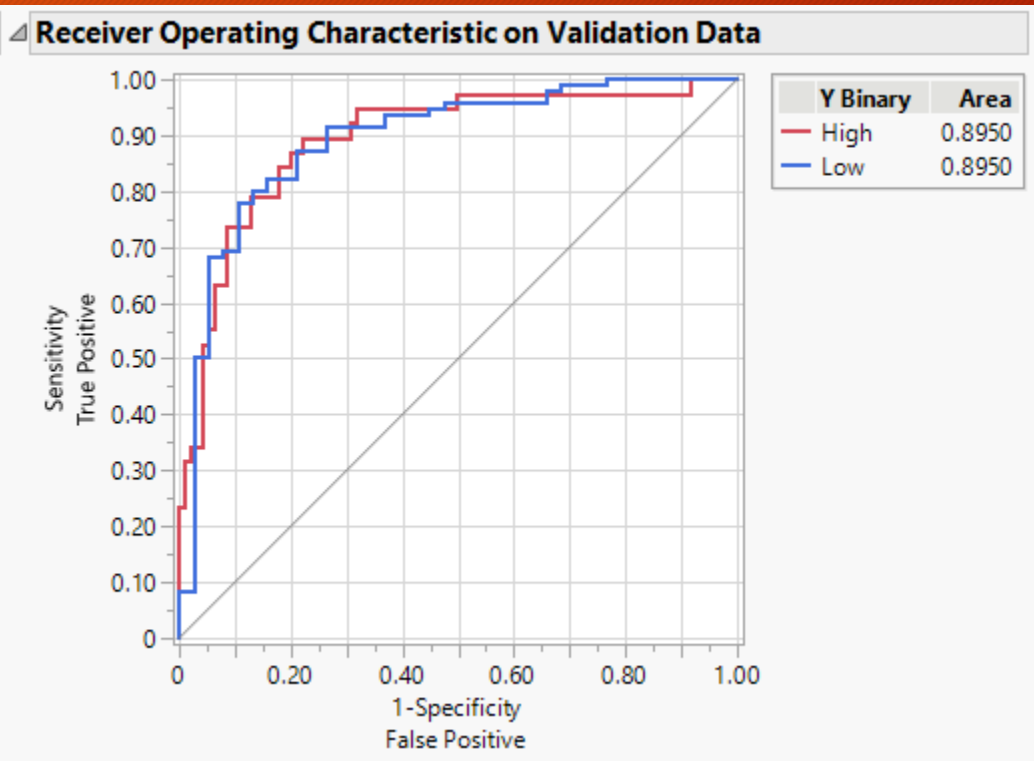


Run logistic regression

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



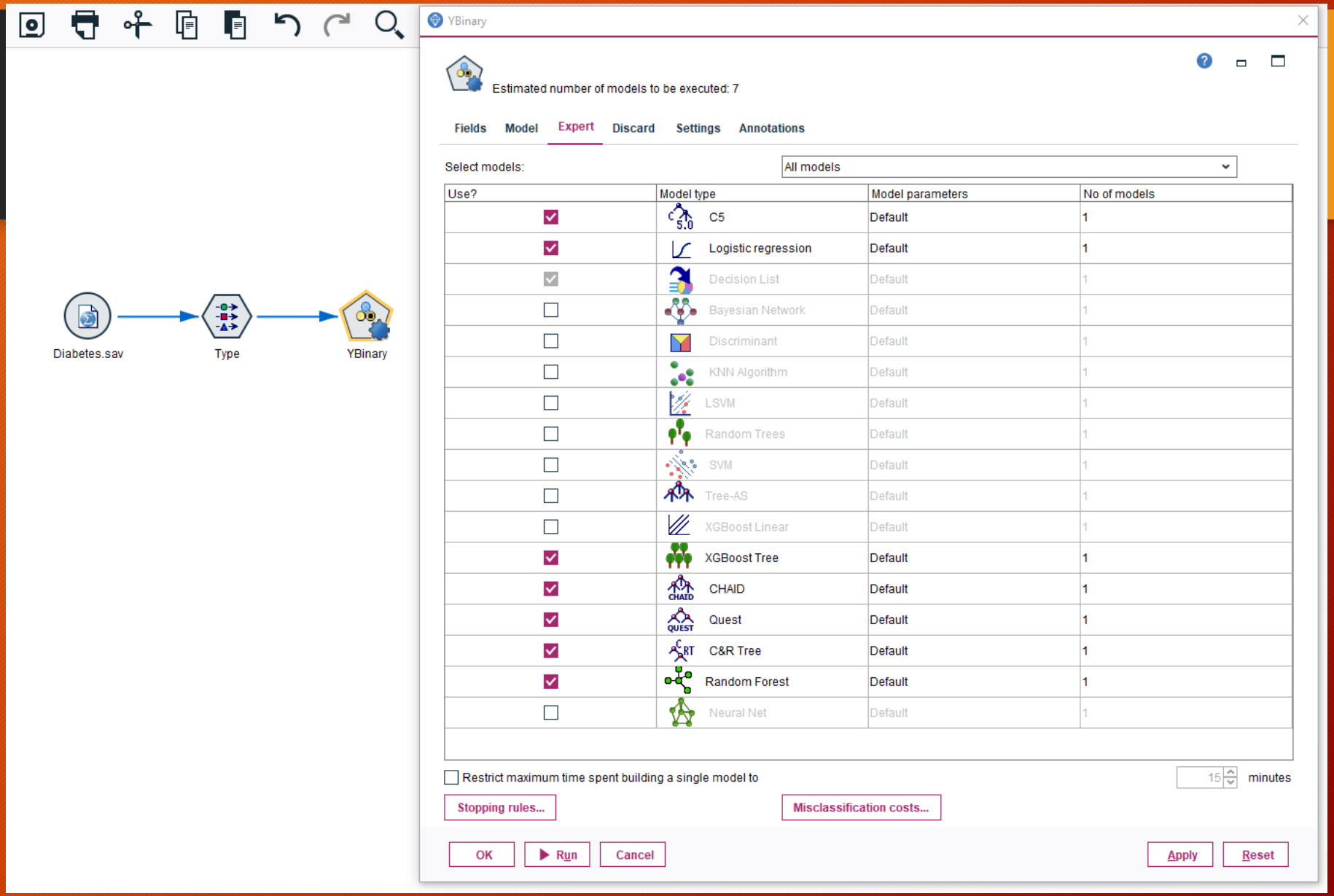


Run SVM

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

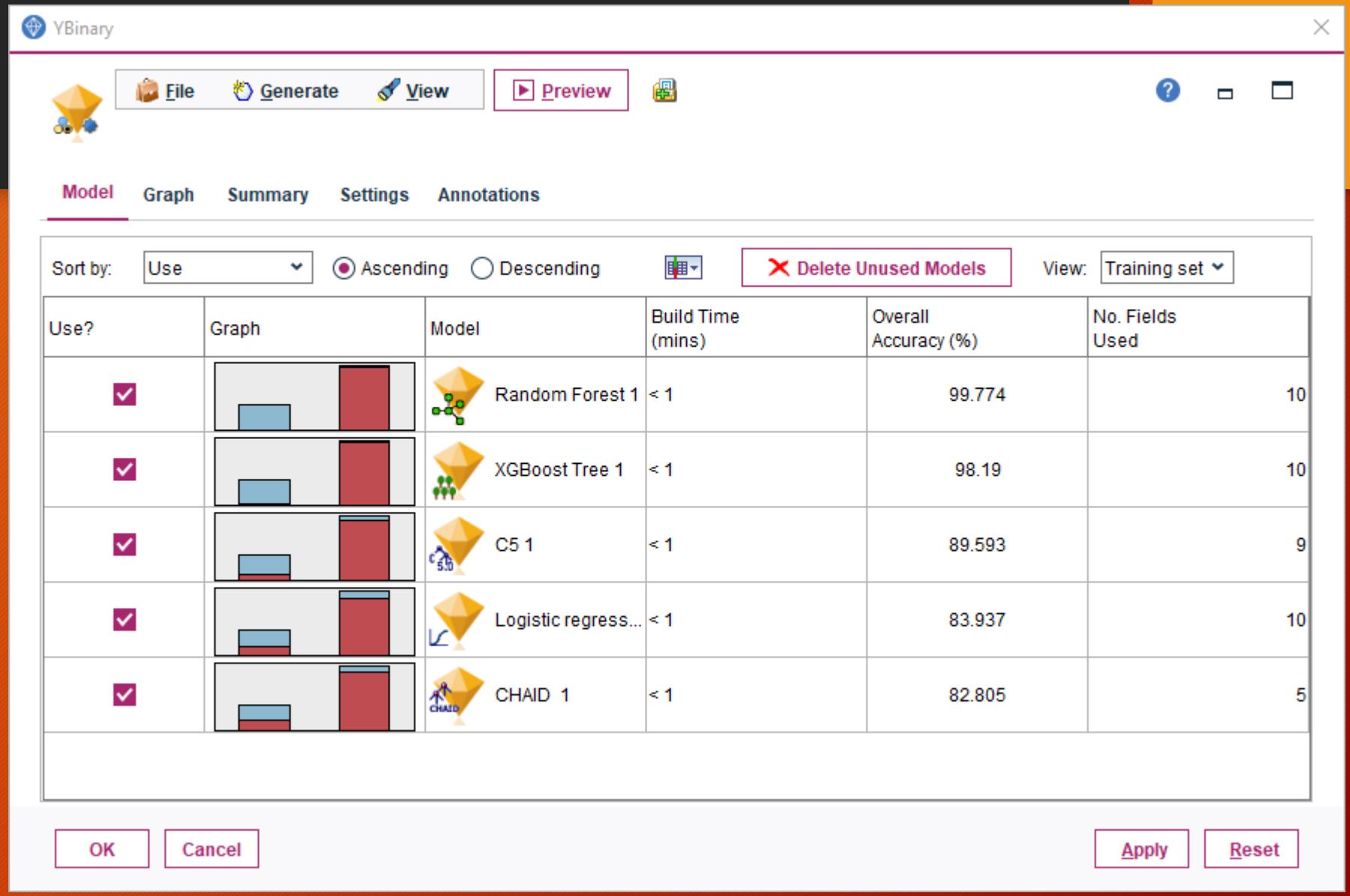
Classification problem

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



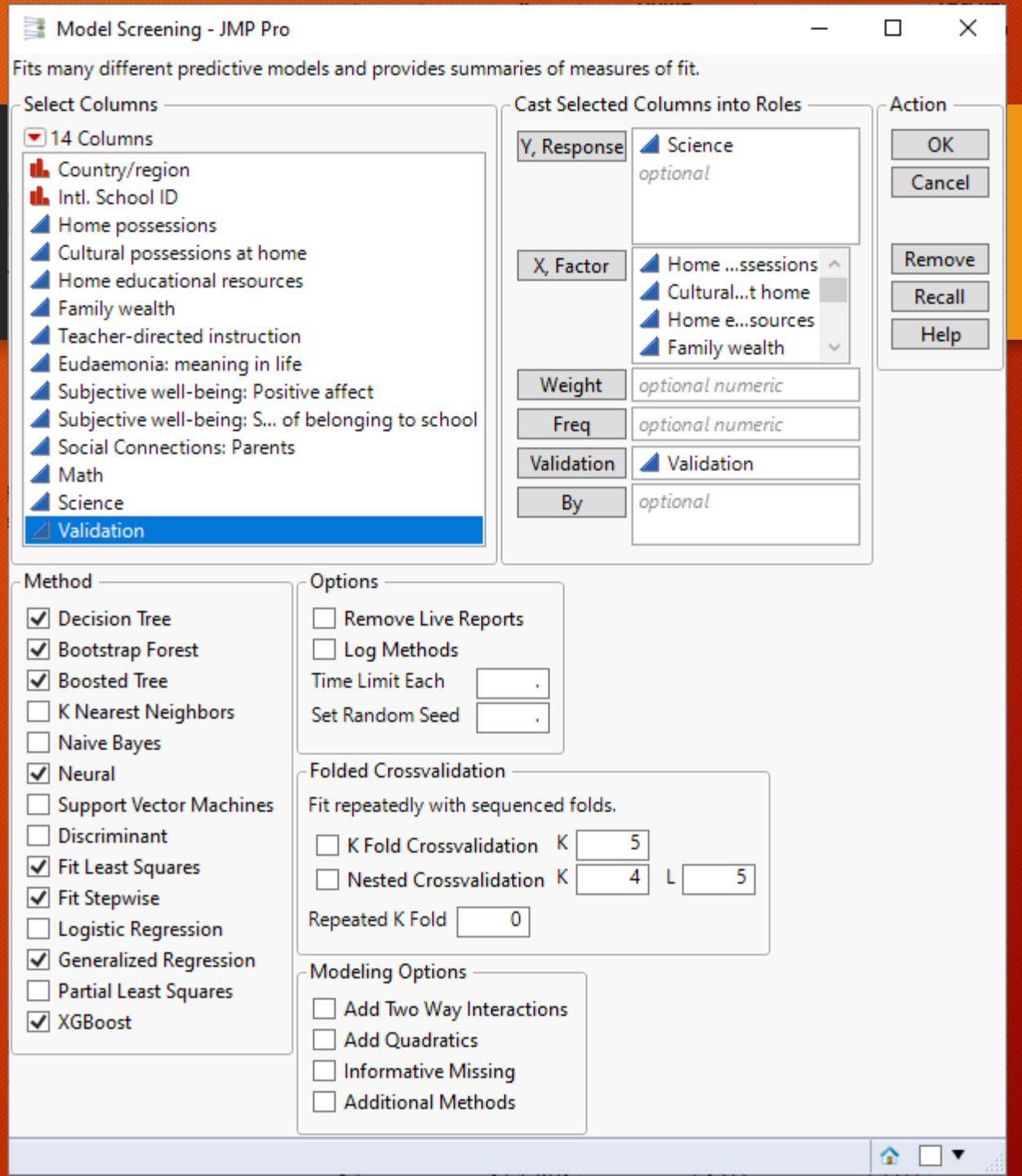
Classification problem

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



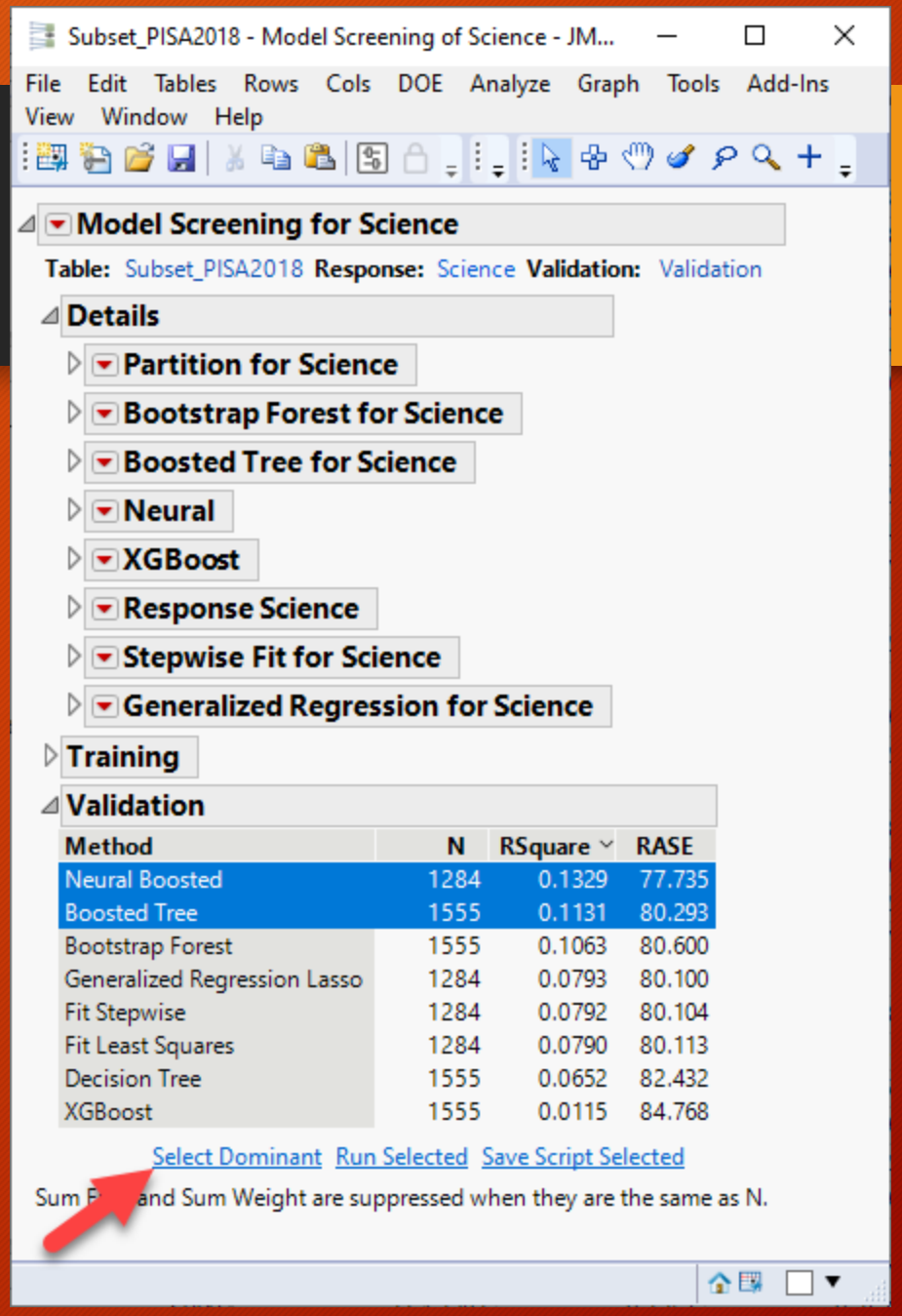
Regression problem

- 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.)



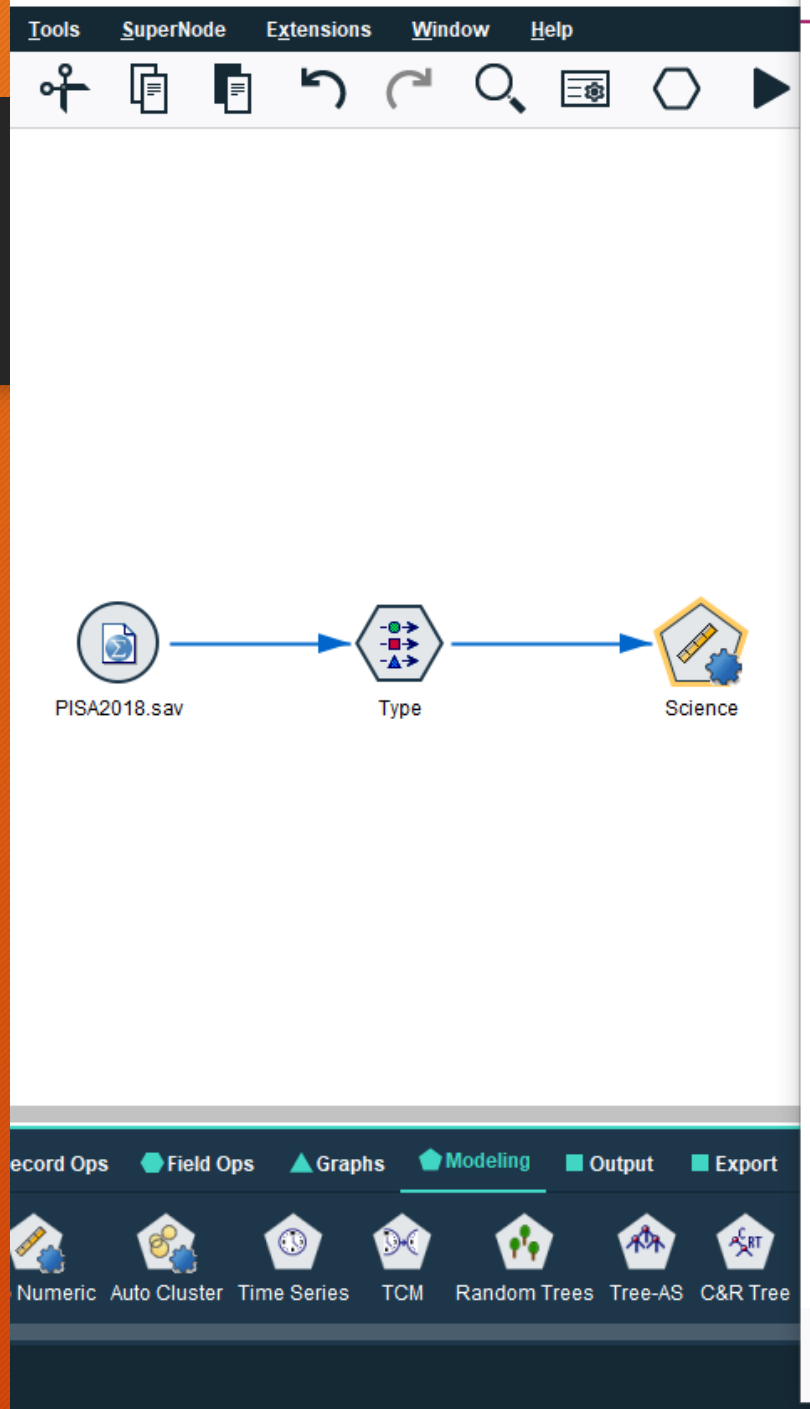
JMP Pro: Model screening

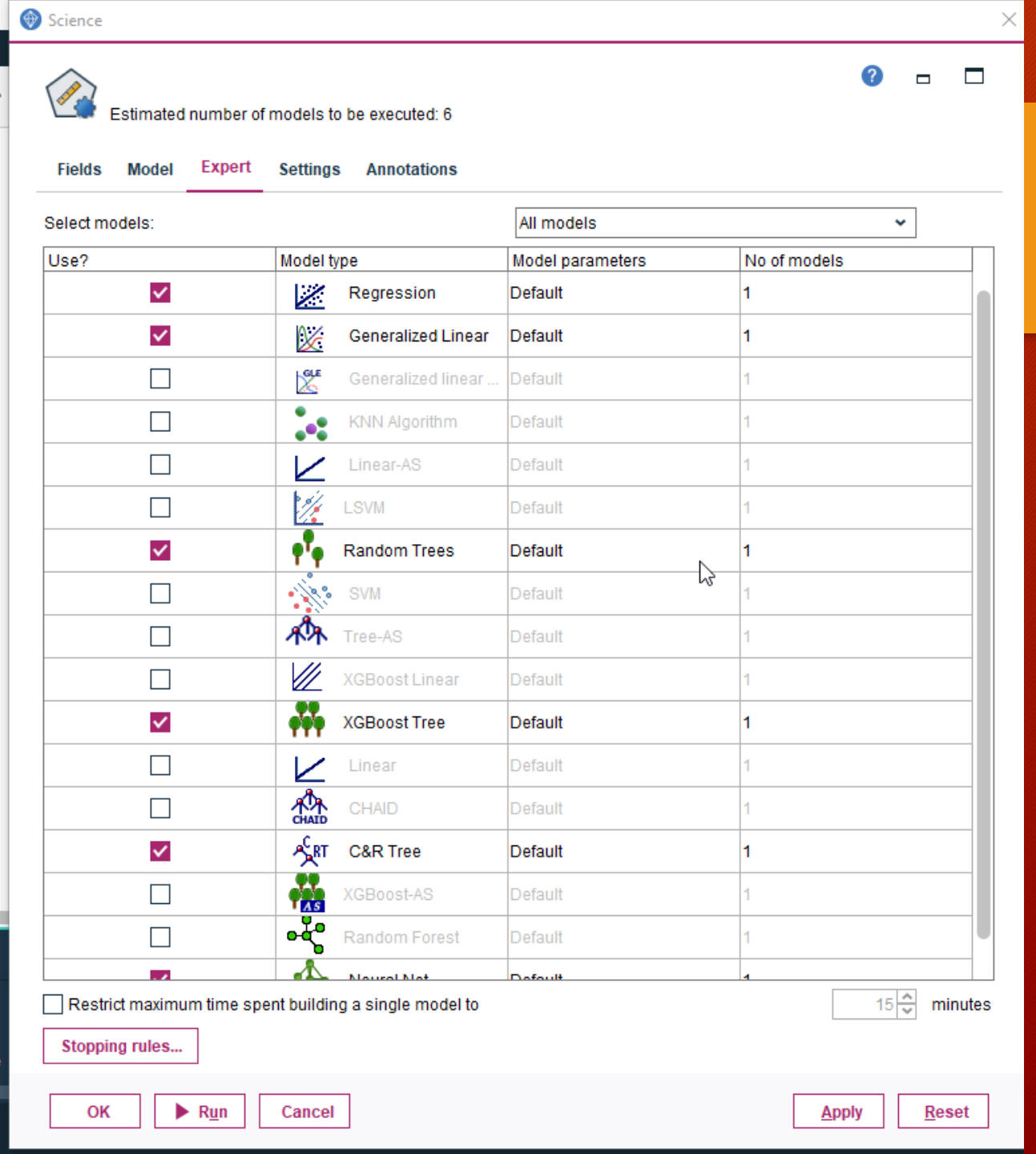
- The best two are neural boosted and gradient boosting.
- Suppose XGBoost should outperform gradient boosting, but it is at the bottom!



SPSS Modeler

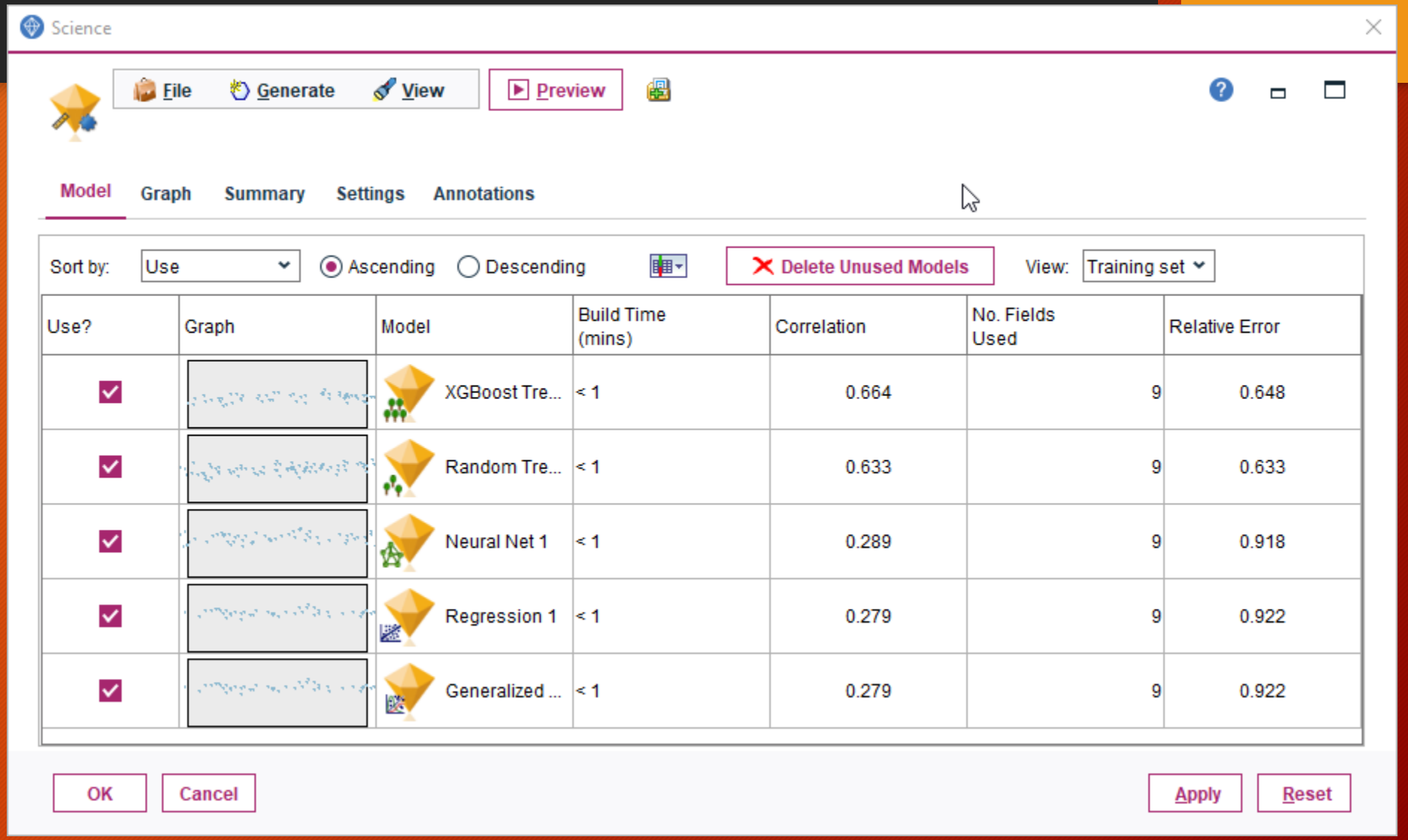
- Keep OLS regression as the baseline model.





SPSS Modeler

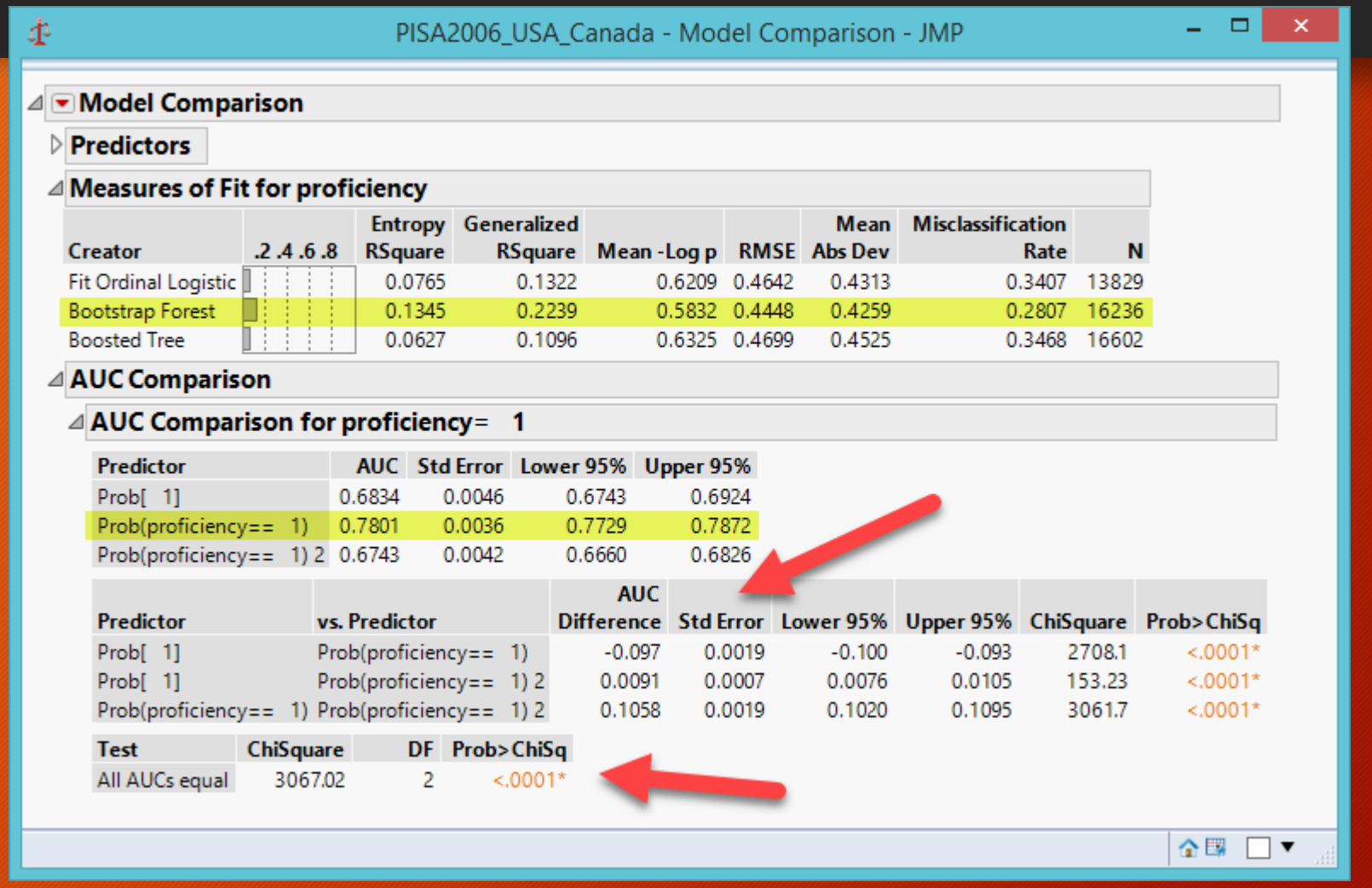
- XGBoost is the best!
- It is opposite to the result of SAS/JMP!



Why?

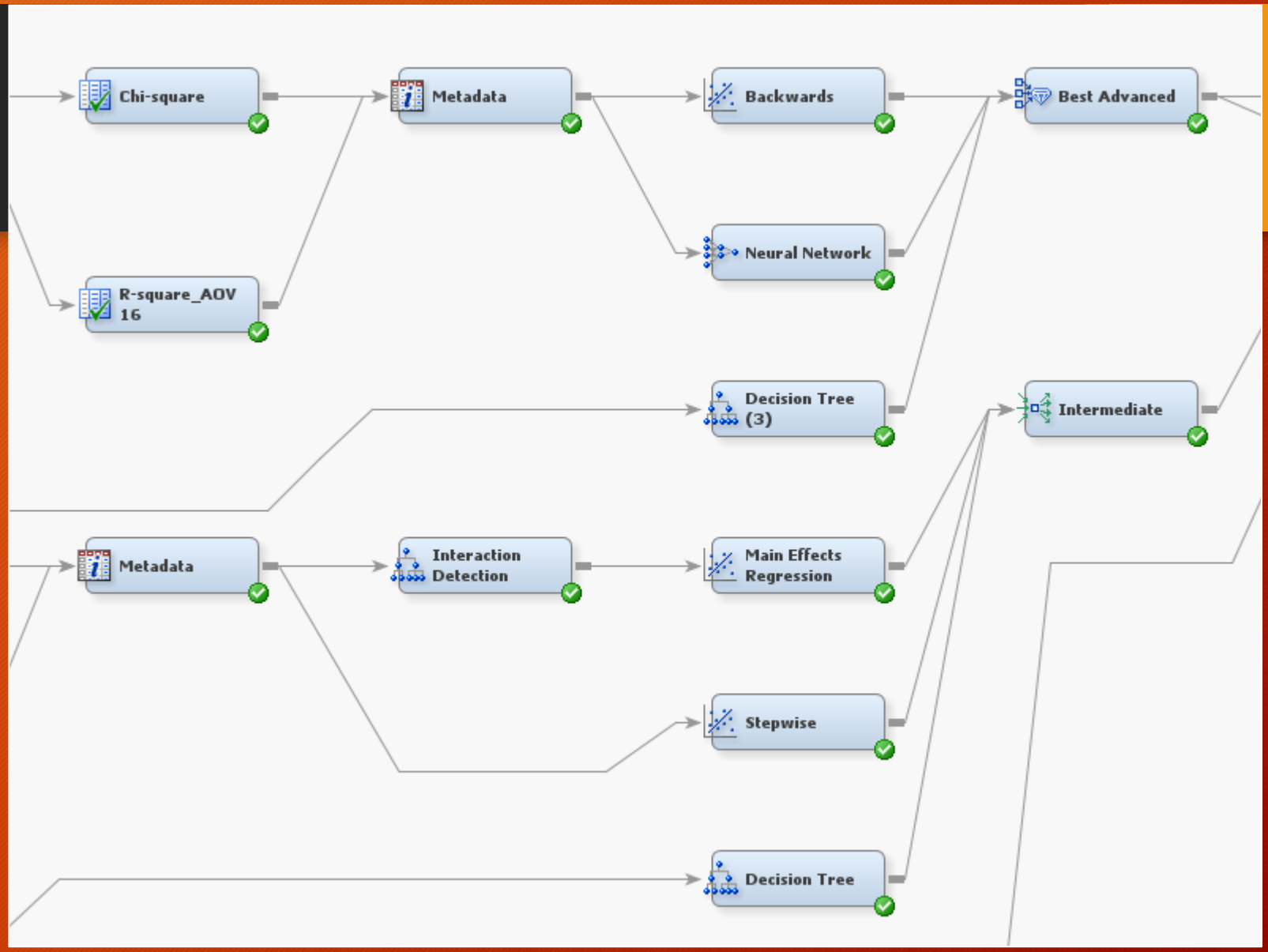
- Different software packages have different default tuning parameters and also their algorithms are slightly different.
- Automatic model comparisons using different software packages with different default parameters might lead to very different results.
- Many software packages offer both automated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

JMP Pro: Model comparison



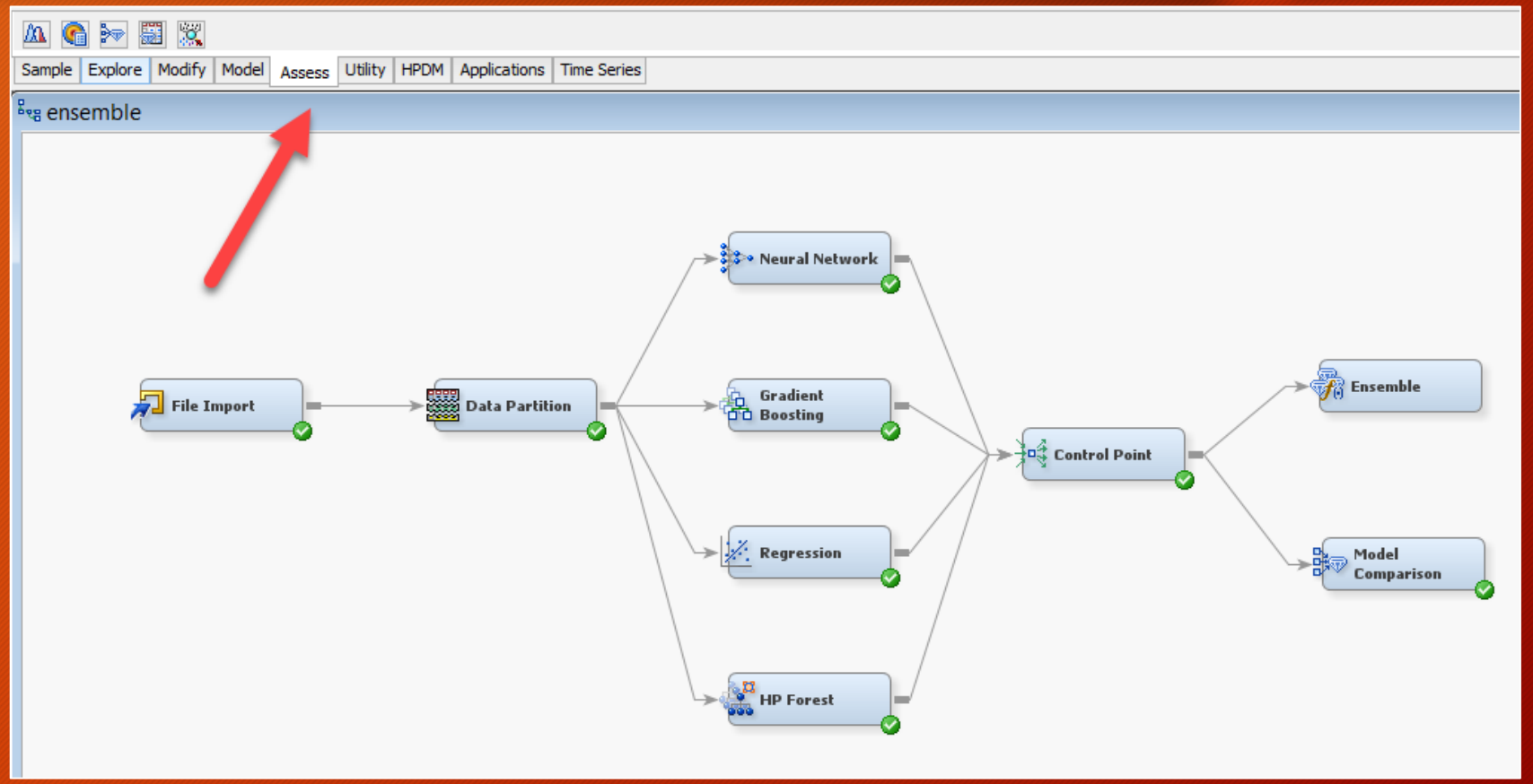
SAS Enterprise Guide

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

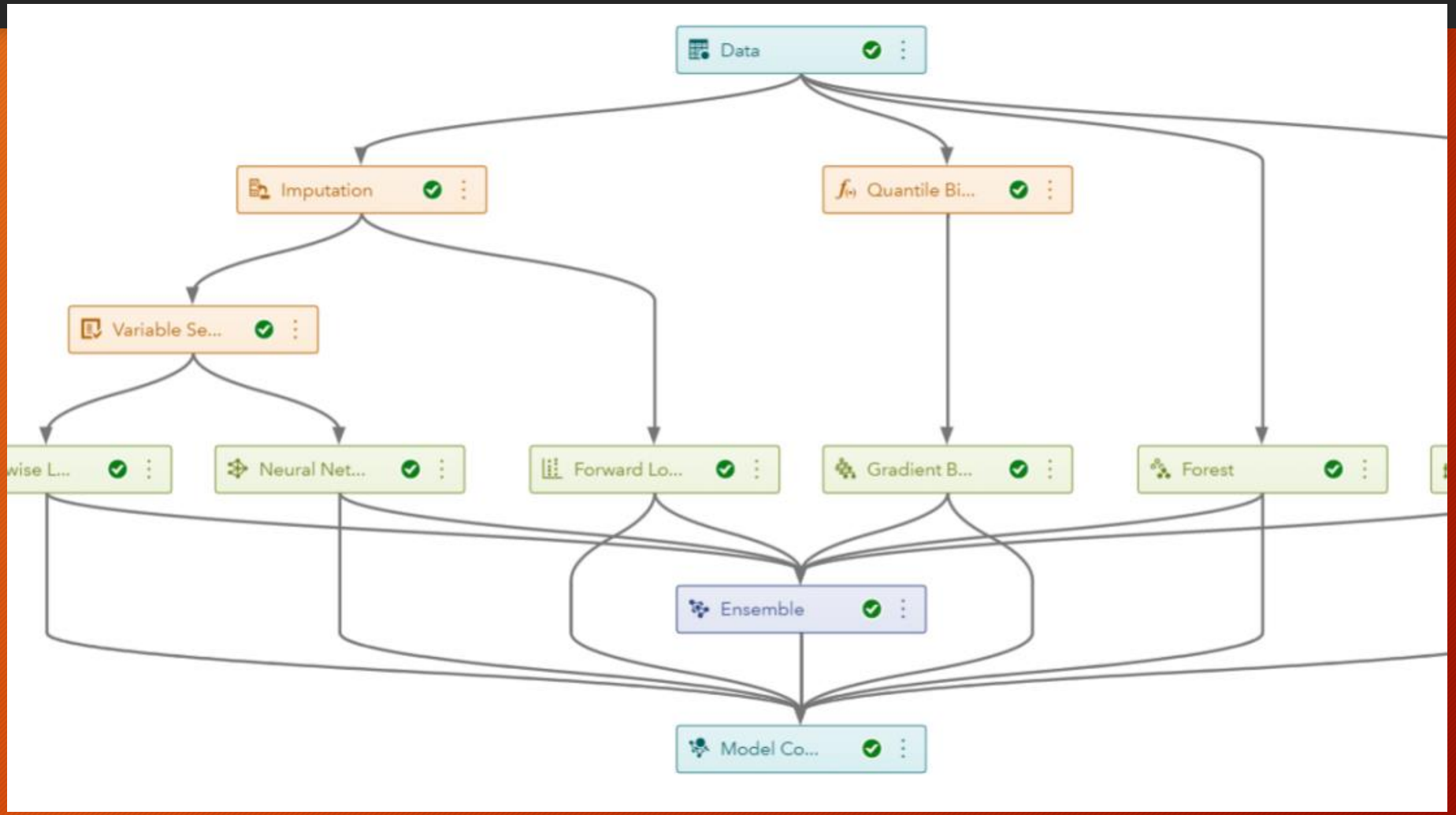


SAS Enterprise Miner

- Interactive model comparison
- Change parameters along the way.



SAS Viya: Model Studio



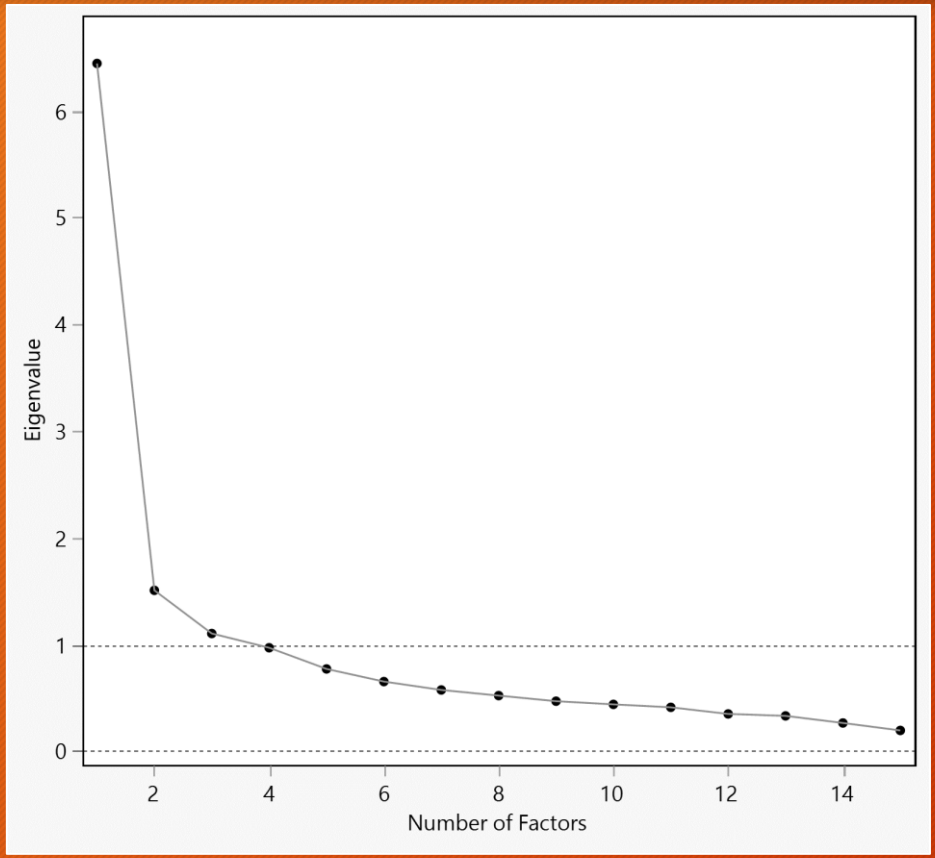
Challenges

- 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.