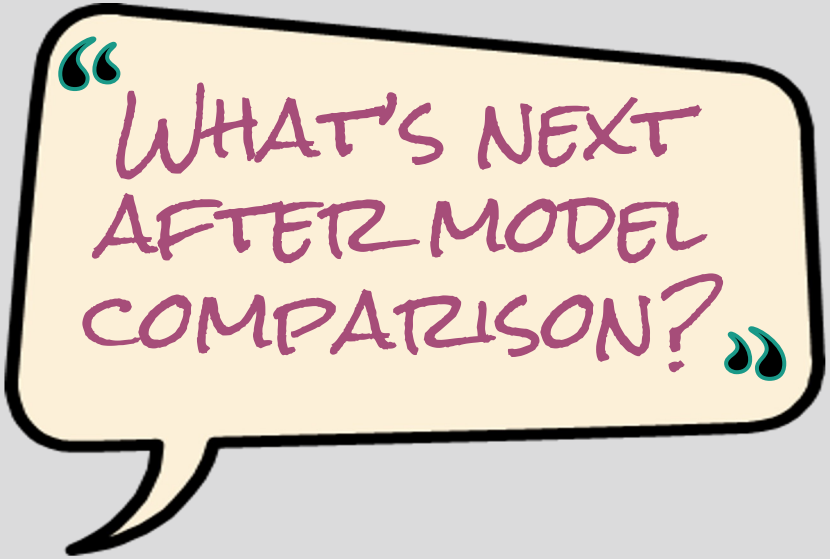
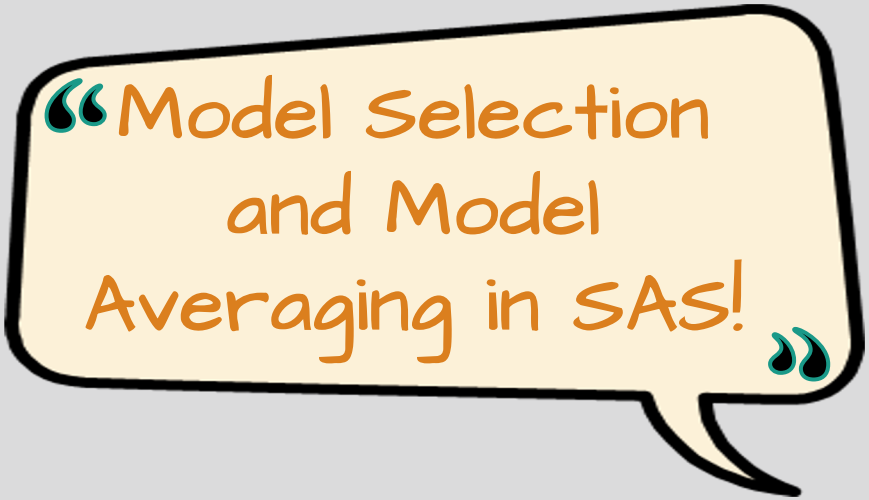


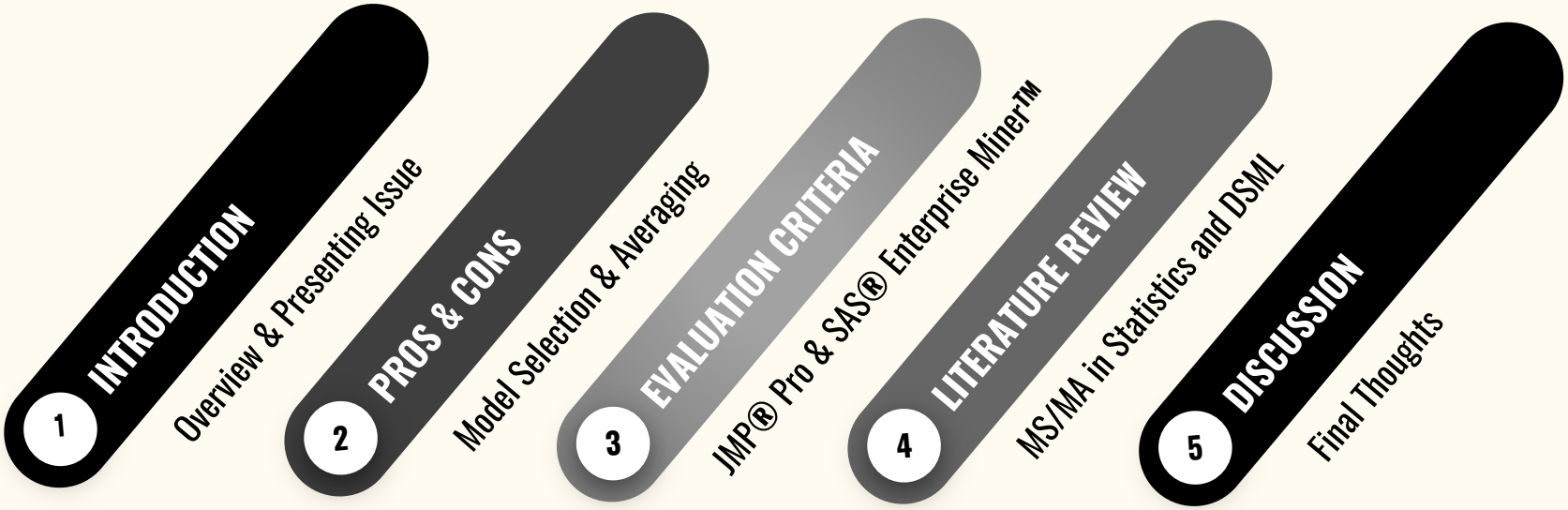


WUSS 2023

Chong Ho Alex Yu, PhD, DPhil

Charlene J Yang, MA

Table of Contents



The Presenting Issue

- A shared goal in research is to be able to replicate studies and enhance the **generalizability of findings**.
- Previous studies have utilized both model selection (MS) and model averaging (MA) techniques.
- There are mixed findings and preferences, with the **majority leaning towards MA**.
- No general consensus on **best practices**.
- **No documentation** on best practices for MS/MA in SAS.
- There is a need for greater understanding and research in this area.

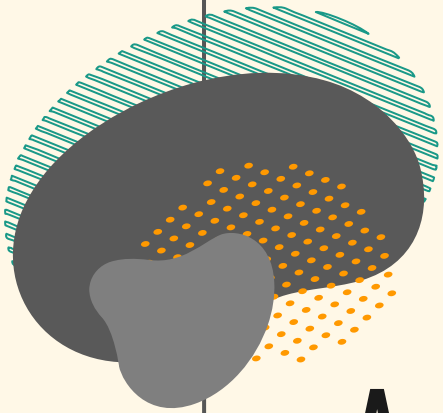
Introduction

Importance of Model Comparison

- Previous process of data analysis was a **one-shot process**.
- **Overfitting**, resulting in replication crisis.
- **Limitations** in traditional modeling methods.
- To address the aforementioned, **multiple models** need comparison.
  - Using neural networks, boosting, bagging, SVMs, etc.
  - Multiple analyses with various data subsets yield **more accurate models**.
- It serves as the initial phase of the **solution**.

After Model Comparison: Model Selection and Averaging

- Leveraging diverse outcomes. **Diversity** is key.
- **Choices are good.** The analyst can choose between two courses of action:
  - Model Selection → "best model"
  - Model Averaging → synthesis of multiple models to create final model
- MS and MA are not unique to data science.
- In contrast to DSML, MS/MA typically confined to single modeling technique.
- **Analyst's choice** depends on the problem, data, and objectives.



Advantages and Disadvantages:

Model Selection & Averaging

Model Selection

ADVANTAGES

Simplicity and Efficiency

- Choose the best model
- Straightforward

Interpretability

- Easier than interpreting an average of models.
- Helps to understand relationship between predictors and target variable.

Computational Efficiency

- No further action required.

DISADVANTAGES

Risk of Overfitting

- If selection criterion is used to choose the most complex model.

Vagueness of “The Best” Model

- Subjective to analyst.
- Model can be best by a certain criterion, but that can easily change.

Model Uncertainty

- Most pervasive disadvantage.
- Might not capture true underlying relationship.

Ignoring Valuable Information

- A limitation when multiple models have complementary strengths.

Model Averaging

ADVANTAGES

Reduces Overfitting

- Enhances ability to generalize unseen data.

Accounts for Model Uncertainty

- This is *the most* cited reason for its use.
- Acknowledges multiple models may have similar predictive performances but different parameter estimates.

Improved Robustness

- Can lead to more robust predictions as they smooth out.

DISADVANTAGES

Complexity

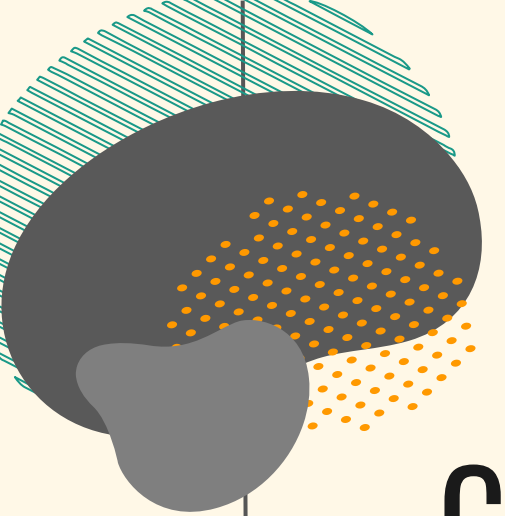
- More complex to implement and manage.

Loss of Interpretability

- Challenging to explain particular predictions.

Increased Computational Cost

- Requires more time and resources.
- However, it may not be an issue with high-performance computing.

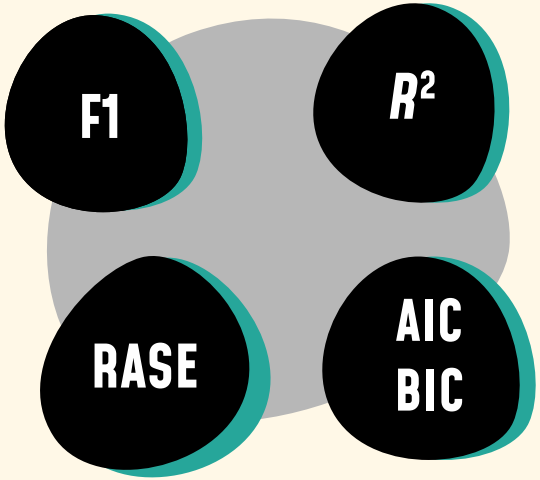


Criteria for Evaluating Model Goodness

Evaluation Criteria

F1 SCORE

Measures accuracy by balancing precision and recall; considers both false positives/negatives.



GENERALIZED R-SQUARED

Provides measure of the proportion of variance explained by the model.

ROOT AVERAGE SQUARED ERROR

Measures the difference between predictions and actual values which represents the average error magnitude.

AIC, AICc, and BIC

Estimate model quality based on balancing goodness of fit and complexity while favoring parsimony.

JMP® Pro

01

MODEL SCREENING

Test out model selection analyses.

02

MODEL AVERAGING

Model averaging performed in Model Comparison.

03

MODEL COMPARISON

Offers option to choose between MS or MA.

Model Comparison

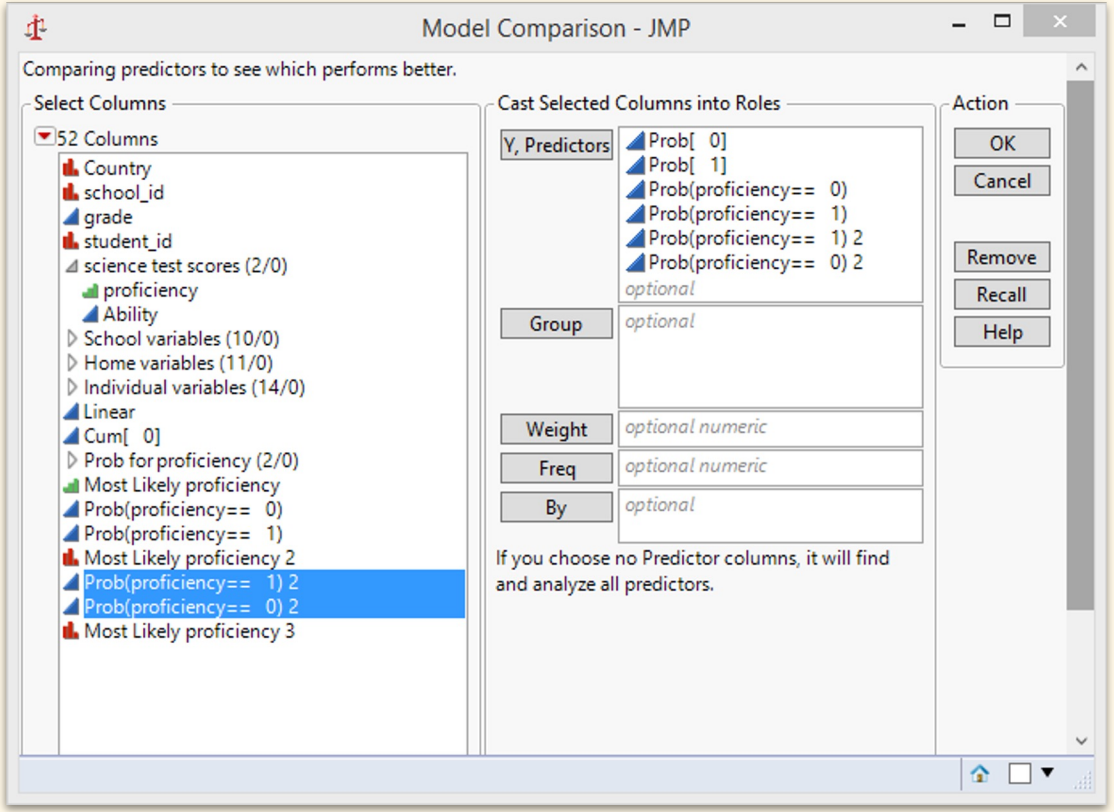
Model Comparison Functions

Model Selection performs in *Model Screening* whereas *Model Comparison* offers the choice between MS or MA.

After predicted outcomes for all models are generated, they can be inputted into *Model Comparison*.

Different criteria can be examined to determine which model is best.

Figure 1. Model Comparison in JMP® Pro



Model Comparison

Results

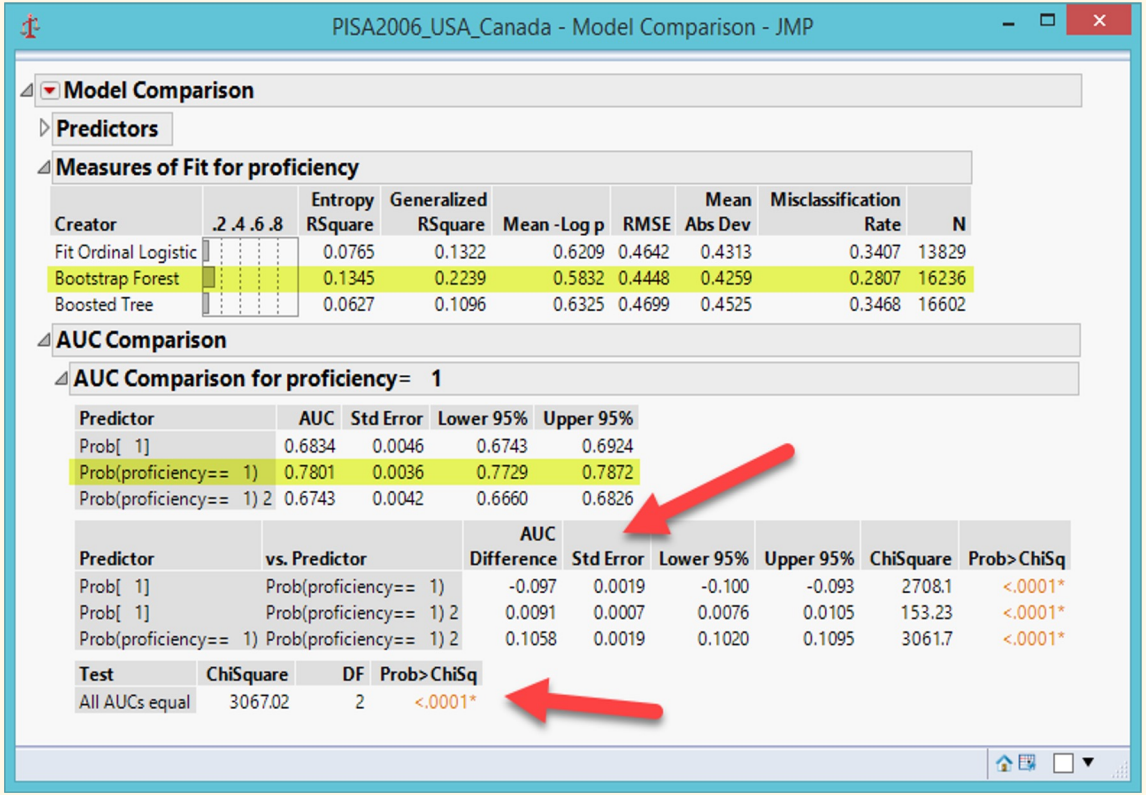
Table here shows,

- Results of null hypothesis test.
- Hypothesis: All AUCs are not significantly different, but they are.
- See multiple comparison results.
- All pairs are significantly different from each other.

In this example,

- Bootstrap forest model has highest Entropy R-square.
- Lowest RMSE
- Lowest misclassification rate
- Highest AUC
- Lowest SE

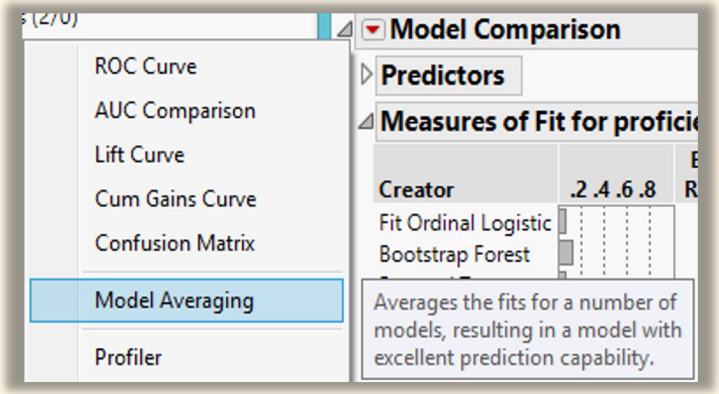
Figure 2. Model Comparison Results in JMP® Pro



Averaging

Model Averaging can create a new field of the arithmetic mean of the predicted values across models.

Figure 3. Model Averaging in JMP® Pro



proficiency	0 Avg Predictor	proficiency	1 Avg Predictor
	0.6089894538		0.3910105462
	0.5698144172		0.4301855828
	0.3208607923		0.6791392077
	0.5072586939		0.4927413061
	0.4536519821		0.5463480179
	0.3260864315		0.6739135685
	0.5267876581		0.4732123419
	0.6061649692		0.3938350308

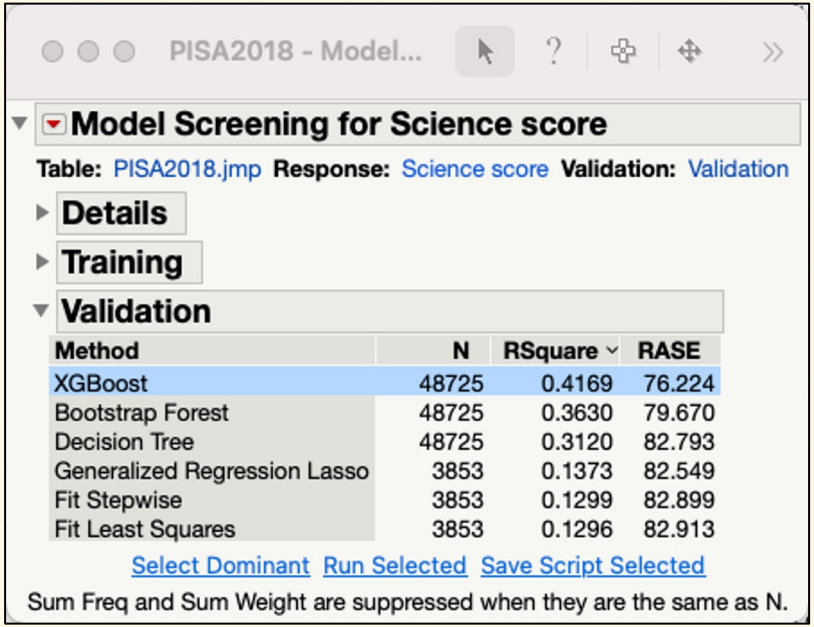
Model Screening

MS in Model Screening

Quick Tool for *Model Selection*,

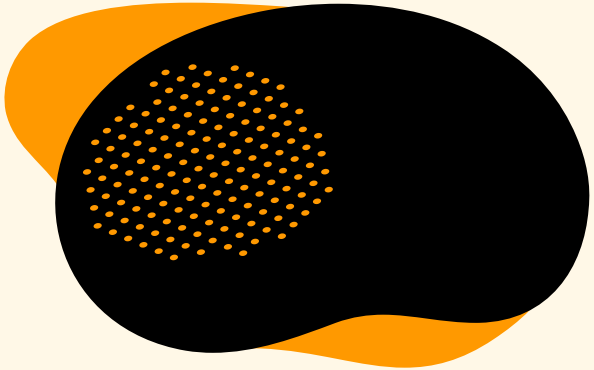
- Multiple methods employed to analyze same data set in tandem.
- Summary table displayed for selecting the dominant model.
- Information on model adequacy presented in *Model Screening* is less than that of *Model Comparison*.
- XGBoost is the dominant model.
- LSR is the weakest model.
- No model averaging allowed in *Model Screening*.

Figure 4. Model Screening in JMP® Pro



SAS® Enterprise Miner™

Offers the capability of utilizing both MS and MA techniques.



01

ENSEMBLE NODE

- In Ensemble node, all modeling results are merged.
- Harmonizes component models to create ultimate model solution.
- Newly created model is employed for scoring new data.

02

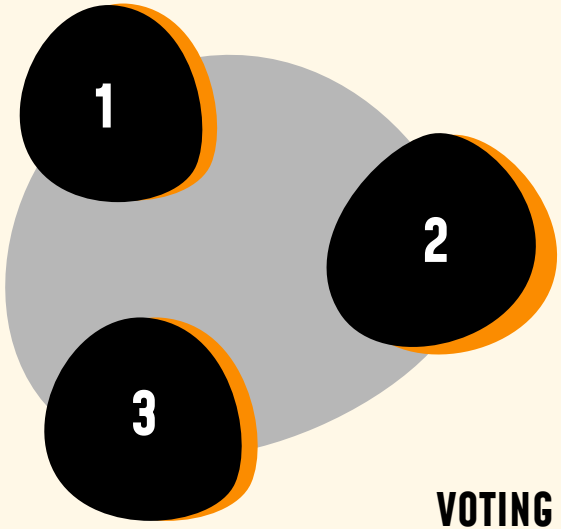
CENTRAL POINT

Where the results are stored for model comparison.

3 Techniques for Amalgamating Data

AVERAGE

Calculates the mean of the posterior probabilities or predicted values, offering it as a prediction in the Ensemble node.



MAXIMUM

Selects the highest posterior probabilities or the maximum predicted values, presenting it as the prediction from the Ensemble node.

VOTING

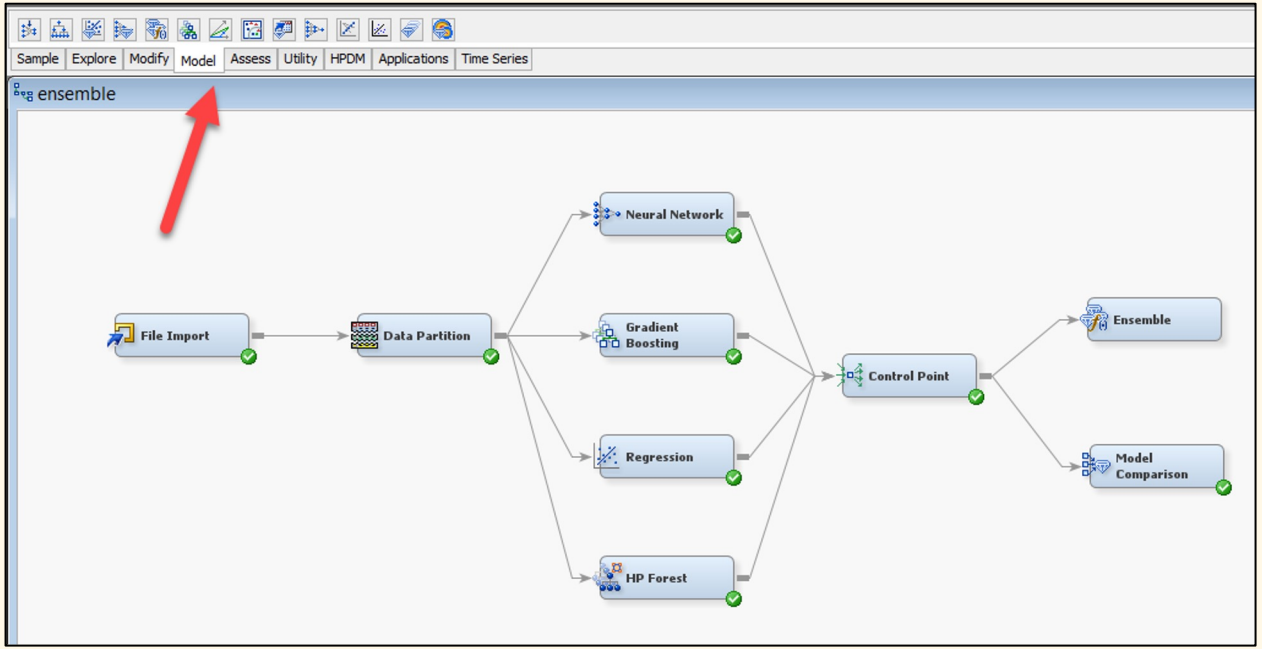
Facilitates the computation of posterior probabilities. Two methods available: *Average* and *Proportion*.

Average, Maximum & Voting

Figure 5 shows utilization of four distinct modeling techniques:

- Neural Networks
- Gradient Boosting
- Regression
- High-Performance Forest

Figure 5. Ensemble and Model Comparison in SAS® Enterprise Miner™



Literature Review

Literature Review

- **9 out of 20** papers reviewed endorsed model averaging
- **6** were inconclusive or said either was fine
- **2** favored contextual usage
- **2** recommended combining the two
- **1** favored model selection

- This tally was counted by mixing MS and MA in **TR statistics**, **Bayesian statistics**, and **DSML**.
- When considering only articles related to **DSML**, it became evident that all of them **preferred MA to MS**.
- Most studies overwhelmingly relied on **AIC, BIC, or both.**

Lit Review Summaries

Author(s) and Paper	Traditional (TR), Bayesian, or DSML Methods	Criteria for Comparison	Favors Model Selection	Favors Model Averaging	Either Way or Inconclusive	Depends on the Conditions	Favors Combining MS and MA	Notes
Aoki et al. (2013)	DSML	AIC & BIC	X					Four methods are introduced that combine multiple candidate dose-response models: model selection, bootstrap model selection, model averaging, and bootstrap model averaging. Bootstrap model selection performed best overall. It had good accuracy for dose finding, decision-making, and estimating the probability of achieving the target response. Model averaging reduced bias compared to just using a single-selected model, but was still outperformed by bootstrap approaches.
Berge (2017)	Bayesian & DSML	Quadratic Probability Score & ROC/AUC		X				Four methods are compared: equally weighted forecast, Bayesian Model Averaging (BMA), Linear Boosted Model, and Nonlinear Boosted Model. BMA and boosted models performed much better than equally weighted forecasts, as they can discriminate between useful predictors.
Buatois et al. (2018)	TR	AIC		X				In an informative design, MA and MS provided similar predictive performances and led to accurate prediction of target dose. However, with less informative designs, by estimating weights for a predefined set of NLMEMs, MA showed better overall predictive performances than MS, increasing the likelihood of accurately characterizing the dose-response relationship.
Gao et al. (2015)	TR	AIC, BIC & MSFE		X				This investigation applied six commonly used MS criteria, including AIC, BIC, HQC, Mallows' C_p, LooCV, and LsoCV, and six FMA methods, namely S-AIC, S-BIC, S-HQC, JMA, LsoMA, and AFTER. The MSFE (mean-squared forecast error) was calculated for each selected model and MA method. The mean of MSFE was used to rank performance of each method, and from the ranking, the LsoMA method was found to be the best.
Grainger et al. (2017)	TR, Bayesian & DSML	AIC & AICc					X	To address the problem of food waste, both MS and MA were employed to retain the most promising models out of 16,384 potential candidates.
Gu et al. (2018)	TR	AIC, BIC & APRESS					X	The result of combining MS and MA is more robust; thus, both should be used.
Haggag (2014)	TR	AIC, BIC, TIC, HQC & Mallows' C_p		X				Results showed smaller values of bias, variance, and PMSE for regression coefficient estimates of MA than that of MS.
Okoli et al. (2018)	Bayesian	AIC & RMSE			X			These authors compared MS and two types of MA, arithmetic (unweighted) MA and weighted MA. When the sample size was small, both MS and MA outperformed a single model. When the sample size was large, MS and MA (weighted or unweighted) had similar variances.

Discussion

Choosing between MS and MA depends on the goal and availability of resources.

Generally speaking, **MODEL SELECTION** should be considered:

- ➔ When the goal is to identify a single model that can be used for both prediction and inference.
- ➔ When the number of candidate models is small.
- ➔ When the computational resources are limited.

On the other hand, **MODEL AVERAGING** should be taken into account:

- ➔ When the goal is to improve the predictive performance of the model.
- ➔ When the number of candidate models is large.
- ➔ When the computational resources are available.

Final Thoughts

Both AIC and BIC favor simplicity…but is the simplest model always the best?

That's a great question. Simple is good, but a complex model also has its place.
However…
…if it shouldn't be complex--punishment!

Questions?

Thank you!

CONTACT INFORMATION

Chong Ho Alex Yu, PhD, DPhil

Hawai‘i Pacific University

cayu@hpu.edu

chonghoyu@gmail.com

https://creative-wisdom.com/index.html

https://scholar.google.com/citations?user=mdGny3EAAAAJ&hl=en

Charlene J Yang, MA

Azusa Pacific University

cyang20@apu.edu

charlenejyang@gmail.com