Large-scale international assessment by data visualization and rapid data mining

Chong Ho Yu, Ph.D., D. Phil. Zizhong (David) Xiao, M.S. Presentation at 2020 IDEAS Global AI Conference

AI vs. data exploration

- Many critics charge that while using traditional statistics many researchers blindly follow mechanical procedures without thinking (e.g. adopted p < .05 as the absolute criterion without checking the data pattern)
- Data mining, which is an extension of exploratory data analysis, aims to amend this problem by exploring the data.
- With the advance of AI some people hand over human judgment to AI.
- For example, neural network is said to be a black box. The analyst has no idea of what is happening inside the "hidden layers" (Deep learning has multiple hidden layers!)
- Are we going into a big circle?

Objective

- A demo of Rapid Predictive Modeling (RPM): Everything is automatic
- Data visualization: understand the data structure first so that appropriate methods can be selected.
- Manually run ensemble models and perform model comparison.

Data source

- 2018 Programme for International Student Assessment (PISA) by Organization for Economic and Collaboration Development (OECD)
- The test is administered to 80+ countries/regions every three years.
- Reading, math, and science.
- Also collected background information of the students (e.g. demographic, teacher info, school info, well-being).
- Focus: which wellbeing variables can predict math and science test performance?

Well-being

- Individual dimension (self-health, skills, and psychological functioning),
- School environment (social connections and school work),
- Out-of-school environment (social connection, material conditions, and leisure time)

Rapid Predictive Modeler

- Statistical Analysis System (SAS): one of the leading DS software applications
- Two ways to access RPM in SAS
- SAS on Demand (Online sever-based computing)
- SAS Enterprise Guide (Local computing)



Just one step!

Task	s Favorites Pro	gram Tools Help 🖺 🖬 🧀 🖬 🔛 🖴 🛍 🛍 🗙 🕪 🕬
	Browse	ss Flow 👻
	Data	🕨 n 👻 🗉 Stop Export 👻 Schedule 👻 📸 Project Log 📰 Properties 👻
	Describe	▶ <u></u>
	Graph	
	ANOVA	▶ 2018_RD
	Regression	▶ M
	Multivariate	•
	Survival Analysis	•
	Capability	•
	Control Charts	•
lín.	Pareto Chart	
	Time Series	•
	Data Mining	Model Scoring
	OLAP	Rapid Predictive Modeler
	Task Templates	Recency, Frequency, and Monetary Analysis OT8_RDM

• Define the roles of variables

8	Rapid Predictive Model for: H:\PISA2018\PISA2018_RDM.sas7bdat
Data Model	Data
Report Options Registration	Data source: H:\PISA2018\PISA2018_RDM.sas7bdat Task filter: None
	Input variables: Modeling roles:
	Name Dependent variable (Limit: 1)
	How_is_your_health_
	I_like_my_look_just_the_way_it_i I_consider_myself_to_be_attracti
	Lam_not_concerned_about_my_w By Excluded
	A I_like_my_body A I_like_the_way_my_clothes_fit_me A Student Gender
	▲ In_the_past_six_months_how_ofte
	In_the_past_six_months_how_ofte(
	▲ In_the_past_six_months_how_ofte;
	In_the_past_six_months_how_oftes
	A How_satisfied_are_you_with_each. ✓
	Variables table
	Run Save Cancel Help

. :

Modeling methods

- If "basic" is chosen, then only traditional and simple methods would be used (e.g. regression)
- If "advanced" is chosen, both traditional statistics and modern data science methods (e.g. neural networks, ensemble methods...etc.) will be used.

	Kapiu Fieui	cuve model n		AZUTO_KDIVI.Sas	Juai	
Data Model	Model					
Report Options Registration	Dependent variable:	PV_Science	Decisions and p	riors		
	Modeling method:	 Basic Intermediate Advanced 				
			Run	Save	Cancel	Help

 Model comparison is selected so that the algorithm can choose the best model (Champion)

2	Rapid Predictive Model for: H:\PISA2018\PISA2018_RDM.sas7bdat
Data Model Report Options Registration	Rapid Predictive Model for: H:\PISA2018\PISA2018_RDM.sas7bdat Report Report options Image: Model summarization Image: Variable ranking Image: Cross tabulations Image: Classification matrix Image: The statistices
	 Fit statistics Lift plot Model comparison Standard reports
	Run Save Cancel Help

Output



• The output is available in both HTML and PDF formats.

RPM used both traditional and modern methods

Model Selecti	ion based o	on Model Node			
Selected Model	Model Node	Model Description	Target Label	Train: Average Squared Error	Valid: Average Squared Error
Y	Neural	Neural Network	PV Science	5521.25	5610.22
	Ensmbl	Ensemble_Champion	PV Science	5926.17	6000.79
	Reg	Main Effects Regression	PV Science	5961.75	6037.51
	Reg2	Forwards	PV Science	8363.95	8417.82

- I didn't choose the data mining methods. SAS made the decision for me.
- Traditional: Main effects regression and forward-selection regression
- Modern data science: Neural networks and ensemble champion

The best and the worst

Model Fit Statistics

Statistic			Train		Validation	
Akaike's Inform	nation Criterion	1	414265.2713			
Schwarz's Bay	esian Criterion	1	417250.1164			
Average Squa	red Error		5521.2514		5610.2184	
Maximum Abs	olute Error		389.1706		460.1698	
Sum of Freque	encies		48000.0000		32000.0000	
Root Average	Square Error		74.3051		74.9014	
Sum of Square	e Érrors		265020067.79		179526989.97	
Mean Squared	d Error		5560.6393		5610.2184	
Root Mean Sq	uared Error		74.5697		74.9014	
Average Error	Function		5521.2514		5610.2184	
Model Selectio	on based on Mo	del Node			Pagel	Break
				Train:	Valid:	
Selected	Model		Target	Average	Average	
Model	Node	Model Description	Label	Squared	Squared	
				Error	Error	
Y	Neural	Neural Network	PV Science	5521.25	5610.22	
	Ensmbl	Ensemble_Champion	PV Science	5926.17	6000.79	
	Reg	Main Effects Regression	PV Science	5961.75	6037.51	
	Reg2	Forwards	PV Science	8363.95	8417.82	

- Averaged squared error
- Worst: Forwardselection regression
- "Best": Neural network
- But we paid a "high price."

Page Break

Many variables are selected! Too complicated!

Selected Variable Importance

ICT resourc Cultural possessions at hor

Eudaemonia: meaning in I Home educational resourc

How did you feel the last time you spent time outside your home with your friends? Nervous How did you feel the last time you did your homework/studied for school? Motivated o General fear of failu

How did you feel the last time you did your homework/studied for school? Nervous How did you feel the last time you attended a mathematics class at school? Motivated o Thinking about your parents or guardians, how often do they: Treat me li How did you feel the last time you did your homework/studied for schoo How did you feel the last time you attended a [test language lesson] at schoo How satisfied are you with each of the following? How you use Social Connections: Parer

How satisfied are you with each of the following? All the things In the past six months, how often have you had the following? Irritability or ba When was the last time you spent time outside your home with you

Outside of school, during the past 7 days, on how many days did you engage in: Vigorous physical ac When was the last time you attended a mathematics class a

How often do you talk to your friends on the phone, send them text messages or have contact through socia Outside of school, during the past 7 days, on how many days did you engage in: Moderate physical ac How often do you worry about how much money your fai

- Neural network is well-fitted to the data and achieve high predictive accuracy.
- But the end result is too complicated!
- We need a concise model for actionable items.

Weighted Likelihood Estimates

- I like my look just the way it is
- I consider myself to be attractive
- I am not concerned about my weight
- 🚽 l like my body
- I like the way my clothes fit me.

- 199 items in the well-being survey
- It is very tempting to collapse many variables into a few by principal component analysis (PCA). After all, they are all conceptually related (wellbeing).
- Example: Group all responses on the left into "Body image"
- WLE: Weighted Likelihood Estimates

Data Visualization: Overplotting!

- JMP Pro, a leading DV software developed by SAS Institute.
- Too many data! It obscures us from seeing the real relationship between science test performance and body image.
- If a regression line is fitted to the data, the *p* value is <.0001.
- But regression assumes a linear relationship.
- The residuals are too high.

PV Science	800 700 600 500 400 300 200					
		-2	-1 Body ima	0 ige (WLE)	1 2	2
	— Line	ar Fit				
⊿L	inear Fi	t				
P١	/ Scienc	e = 459.6	9184 - 5.613	0118*Body im	age (WLE)	
⊿	Summa	ary of Fit				
	RSquar RSquar Root M Mean o Observa	e e Adj ean Squa of Respon ations (or	are Error Ise r Sum Wgts)	0.003452 0.003439 98.7055 459.0691 75497		
\triangleright	Lack O	f Fit				
Δ	Analys	is of Vari	iance			
	Courses	DE	Sum of	Mann Course	C Datia	
	Model	1	2547600	2547600	261 4962	
	Error	75495	735530781	9743	Prob > F	
	C. Total	75496	738078479		<.0001*	

Data reduction: Binning



Data visualization: Median smoothing

- Look at the trend of the median
- At first test score improves as body image goes up.
- Later it goes downward!
- Those who feel very good about their body (rightmost) are worse than those who feel bad (leftmost) in science test.



Median smoothing raw variables: Same pattern



Using the variables as is

- PISA does not have WLEs for all well-being constructs
- Rather than running PCA and then do binning on all of them, we used the raw variables as is.
- Bagging
- Boosting

Bagging

- Bootstrap Aggregation
- An ensemble of many decision-trees obtained from repeated sampling with replacement from one data set.
- Afterward, the multitude of results are then combined to form a converged conclusion.

Bagging Results: PISA Well-being variables and Science Scores

- Complicated!
- Retain many more variables than its boosting counterpart.

-					,			
	Specific	ations						
-	Target Colun	nn:		PV Sc	ience	Training Rows:	68397	
						Validation Rows:	29481	
I	Number of T	rees in the	Forest:		15	Test Rows:	0	
	Number of T	erms Sam	pled per Spli	t:	20	Number of Terms:	82	
				Bootstrap Samples:	68397			
			Minimum Splits per Tree:	10				
						Minimum Size Split:	97	
▼	Overall	Statisti	cs					
	Individual Trees	RMS	E					
	In Bag Out of Bag	74.8086 97.0344	52 13					
		RSquare	RMSE	Ν				
	Training Validation	0.089 0.083	95.421399 95.415635	68397 29481				
	Cumula	tive Val	idation					
	Per-Tree	e Sumn	naries					
▼	Column	Contri	butions					
	Term							Nun of S
		e						

Now think of the last time you had a break between classes at school. How did you feel? Nervous or tense How did you feel the last time you spent time outside your home with your friends? Nervous or tense (R) How easy is it for you to talk to the following people about things that really bother you? Your teachers How did you feel the last time you did your homework/studied for school? Motivated or inspired

Thinking about your parents or guardians, how often do they: Treat me like a baby When was the last time you spent time outside your home with your friends? How many days a week do you usually spend time with your friends right after school? Thinking about your parents or guardians, how often do they: Try to control everything I do How did you feel the last time you attended a [test language lesson] at school? Nervous or tense

Number of Splits	SS	Portion
203	3975469.33	0.0802
196	3798937.52	0.0767
206	3245432.48	0.0655
211	3205043.13	0.0647
266	2910990.85	0.0587
269	2644726.74	0.0534
157	2600567.38	0.0525
164	2493957.22	0.0503
139	2476344.41	0.0500
180	2141007.93	0.0432
269	2053140.34	0.0414

Bagging Results: PISA Well-being variables and Math Scores

- Complicated!
- Retain many more variables than its boosting counterpart.

 Bootst 	rap For	est for P	V Mat	th		
Specific	cations					
Target Colur	mn:		PV M	ath	Training Rows: Validation Rows:	68397 29481
Number of 1	Trees in the	Forest:		23	Test Rows:	0
Number of 1	Terms Sam	pled per Spl	it:	20	Number of Terms:	82
					Bootstrap Samples:	68397
					Minimum Splits per Tree:	10
					Minimum Size Split:	97
Overall	Statisti	cs				
Individual Trees	RMS	E				
In Bag Out of Bag	77.615 100.386	57 52				
	RSquare	RMSE	Ν			
Training Validation	0.075 0.069	98.917861 98.795643	68397 29481			
	tive Val	idation				
Per-Tre	e Summ	naries				
Column	Contri	butions				

	Number		
Term	of Splits	SS	Portion
Now think of the last time you had a break between classes at school. How did you feel? Nervous or tense	301	3580419.9	0.0785
This school year, on average, on how many days do you attend physical education classes each week?	83	3195879.98	0.0701
How did you feel the last time you spent time outside your home with your friends? Bored	293	3193773.85	0.0700
Now think of the last time you had a break between classes at school. How did you feel? Lonely	298	3138136.4	0.0688
How did you feel the last time you did your homework/studied for school? Motivated or inspired	412	2916652.72	0.0640
How did you feel the last time you spent time outside your home with your friends? Nervous or tense	325	2778975	0.0609
(R) How easy is it for you to talk to the following people about things that really bother you? Your teachers	340	2697437.5	0.0592
When was the last time you spent time outside your home with your friends?	249	2025966.4	0.0444

Boosting

- Gradient boost tree
- A sequential method.
- Initially, the algorithm assigns all observations the equal weight before producing a statistical model. If the model fails to classify some of the observations correctly, then these observations will be assigned a heavier weight so they are more likely to be selected in the subsequent model.
- Each model is revised and updated constantly to successfully classify all the observations

Boosting Results: PISA Well-being and Science scores

Boostad Tree for DV Salance

Compact!

• Only three variables!

Specific	ations				
Farget Colur	nn: PV S	Science	Number o	of training rows:	68397
Number of L	ayers:	50	Number o	f validation rows:	29481
Splits per Tr	ee:	3			
_earning Ra	te:	0.1			
Overall	Statistic	S			
		DUGE			
	RSquare	RMSE	N		
Training	RSquare 0.037	98.101691	N 68397		

Column Contributions

	Number		
Term	of Splits	SS	Portion
How did you feel the last time you spent time outside your home with your friends? Bored	73	87229455.2	0.6532
How did you feel the last time you spent time outside your home with your friends? Nervous or tense	50	26138195.5	0.1957
Now think of the last time you had a break between classes at school. How did you feel? Nervous or tense	21	17121605.4	0.1282
Now think of the last time you had a break between classes at school. How did you feel? Lonely	6	3049721.29	0.0228
Law is using baskbo	^	0	0.0000

Boosting Results: PISA Well-being and Math Scores

Boosted Tree for PV Math

Specificatio	ns		
Target Column:	PV Math	Number of training rows:	6839
Number of Layers:	50	Number of validation rows:	2948
Splits per Tree:	3		
Learning Rate:	0.1		
Overall Stat	istics		

	RSquare	RMSE	N
Training	0.031	101.25089	68397
Validation	0.024	101.14068	29481

Cumulative Validation

Column Contributions

V

	Number				
Term	of Splits	SS			Portion
How did you feel the last time you spent time outside your home with your friends? Bored	80	83376586.6			0.7001
Now think of the last time you had a break between classes at school. How did you feel? Nervous or tense	19	17675884.5			0.1484
How did you feel the last time you spent time outside your home with your friends? Nervous or tense	50	17351229.4			0.1457
Now think of the last time you had a break between classes at school. How did you feel? Lonely	1	683530.232			0.0057
How is your health?	0	0			0.0000
I like my look just the way it is	0	0			0.0000
I consider myself to be attractive	0	0			0.0000
Law wat as a second about we containt	0	0		 1	0.0000

Model Comparison: Between Bagging and Boosting

7 (Model Comp	arison						
	Predictors							
▼	Measures of I	Fit for PV Ma	th					
Predictor Creator .2.4.6.8 RSquare RASE AAE								
	PV Math Predictor	Bootstrap Forest		0.0715	98.990	80.207	97878	
	PV Math Predictor 2	Boosted Tree		0.0247	101.45	82.332	97878	

	Model Compar	ison					
Þ	Predictors						
V	Measures of Fit	for PV Scier	nce				
	Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
	PV Science Predictor	Bootstrap Forest		0.0766	95.978	78.115	97878
	PV Science Predictor 2	Boosted Tree		0.0299	98.375	80.181	97878

Data Visualization - Medium Smoothing





- Automated rapid data mining is quick, but it tends to produce a complicated model, which is not practical.
- Automated rapid data mining includes traditional OLS regression modeling, which is not necessary because modern data science methods always outperform OLS regression analysis.

Model Node	Model Description	Target Label	Train: Average Squared Error	Valid: Average Squared Error
Neural Ensmbl	Neural Network Ensemble_Champion	PV Science PV Science	5521.25 5926.17	5610.22 6000.79
Reg	Main Effects Regression	PV Science	5961.75	6037.51
Reg2	Forwards	PV Science	8363.95	8417.82

- Some traditionalists would like to look at the results of regression. Regression modeling may be necessary for the purpose of comparison. If so, we should use generalized regression.
- Cannot effectively deal with collinearity
- Tend to overfit
- Generalized regression amends these problems by imposing a penalty on a complicated model.
- The coefficient of unimportant variable will be zeroed out (in Elastic net).

- According to rapid predictive modeling, neural network produces the "best" model.
- In our manual data mining the bootstrap forest outperforms gradient boost tree.
- But should we choose the so-called "best" model?
- If I tell you to do 15-25 things to improve the current situation, can you do it?
- The gradient boosted tree suggest only three most important predictors of science/math test performance.
- It is helpful to inform policy-making and educational practice.

- We should not blindly follow the numeric findings; pattern-seeking data visualization is indispensable.
- Before running predictive modeling, we visualize the data to find solutions for overplotting.
- After running predictive modeling, again we visualize the data to find out whether the relationships are linear or nonlinear.