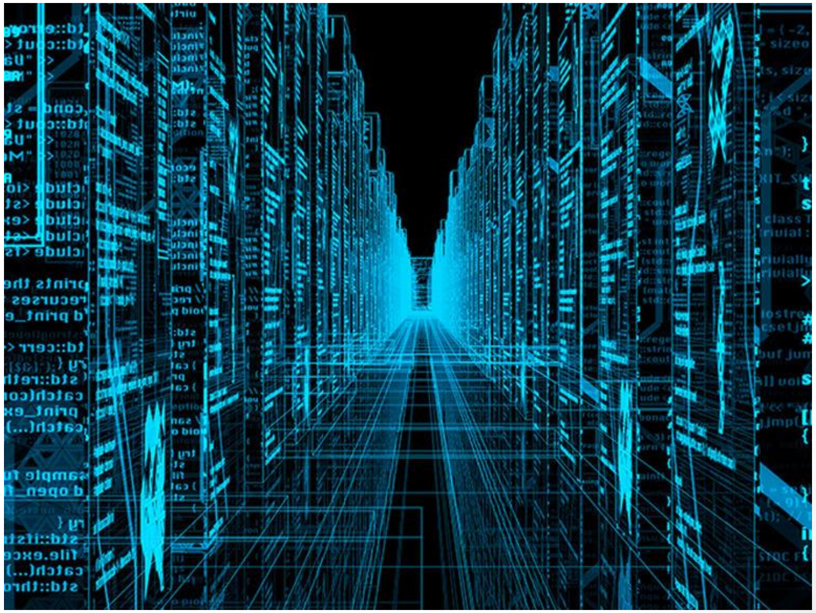
The Ensemble Method and Model Comparison for Predictive Modeling with Big Data

Siyan Gan, Pepperdine University
Hyun Seo Lee, Azusa Pacific University
Emily Brown, Azusa Pacific University
Chong Ho Yu, Ph.D., D. Phil., Azusa Pacific University

Presented at Southern California Big Data Discovery Summit, November 2017, Azusa, USA

Movement toward Big Data

- The size of digital data will double every two years.
- High volume
  - Thousands of rows or columns, can often result in problems with data storage, data management, and data analysis.
- High velocity
  - Non-stop data feed that has the potential to overwhelm a conventional database server.
- High variety
  - Different types of data (e.g. numbers, texts, images, audio files, video clips…etc.).



Types of Data

Unstructured Data

- Webpages and digital footprints on social media.
- Extracting data from Websites.
- Challenging to work with.

Structured Data

- **Could be used by social science researchers for nationwide or cross-cultural studies.**
- **Usually survey data, stored in a conventional row X column matrix.**

Data Mining

- Most social science researchers are trained in the traditional Fisherian statistics (***hypothesis testing***).
- Most of the time, it should not be used for big data analysis.
  - Shows inaccurate “significant” results.
  - Imposes strong assumptions on the data structure and the distribution.
- **Data mining** = ***Data-driven, not hypothesis-driven***

Ensemble Methods in Big Data Analytics/ Data Mining

- Big data set are separated into many subsets and multiple analyses are run.
- In each run the model is refined by previous "training.”
  - Results are from replicated studies.
- Machine learning (based on Artificial intelligence): Learning from previous analysis
- The Ensemble Method: Merging multiple analyses
  - Compares, complements, and combines multiple methods in the analysis.
  - Better predictive outcome than using just one analysis.

Boosting vs. Bagging

Boosting

- Increases the predictive accuracy.
- Creates a working model from the subsets of the original data set.
- Adjusts weak models so they are combined to be a strong model.

Bagging

- *(**B**ootstrap **Agg**regation)*
- Repeated multisets of additional training data from the original sample
- Increases the size of these generated data
- Minimizes the variance of prediction by decreasing the influence of extreme scores

Bagging as voting

- Imagine that there are 1,000 analysts. Each one randomly draws a sub-sample from the big sample and then run an analysis.
- The results must be diverse.
- Now these 1,000 analysts meet together and vote.
- “How many of us found Variable A as a crucial predictor? Please raise your hand.” And then move on to B, C…etc.
- At the end we have a list of top 10 or Top 20.

Boosting as gradual tuning

- Imagine that you are a cook. You put spices into the dish and taste it.
- If it is not salty enough, you put more salt and pepper next time.
- But next time if it is too spicy, then you put less hot sauce.
- You make gradual improvement every time until you have the final recipe.
- Similarly, boosting is a gradual tuning process.

Data Visualization

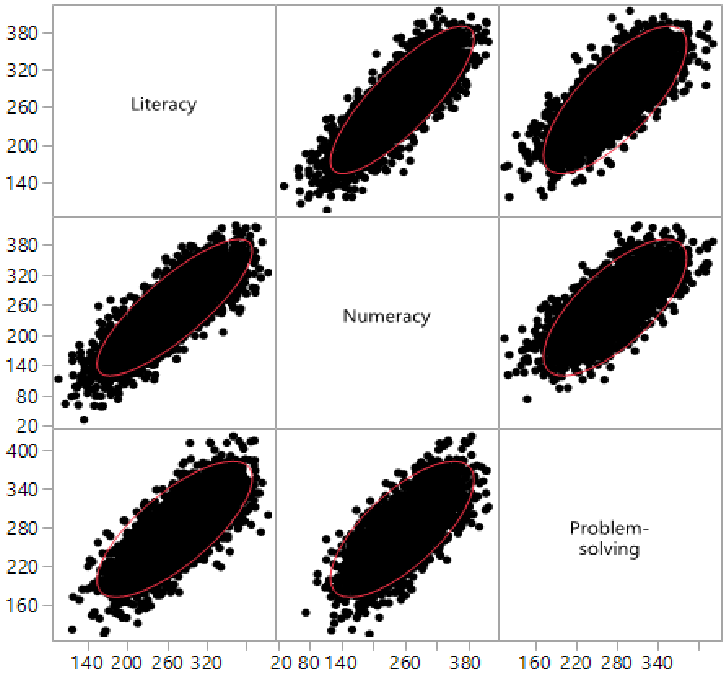
- Presentation
- Unveil undetected patterns

PIAAC Study

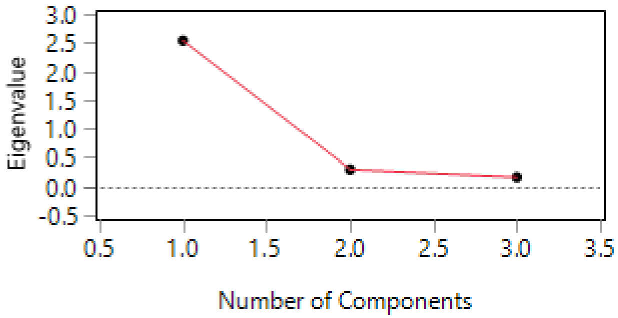
- *Programme for the International Assessment of Adult Competencies* (PIAAC).
  - Developed by *Organization for Economic and Cooperation and Development* (OECD).
- In 2014, PIAAC collected data from 33 participating nations (OECD, 2016).
- U.S. adults were behind in all three test categories:
  - Literacy, numeracy, and problem solving in technology-rich environments.
- Survey items included factors related to learning:
  - Readiness to learn, cultural engagement, political efficacy, and social trust.

Variables

- The scores of literacy, numeracy, and technology-based problem-solving strongly correlated.
- All three skills were combined into one component.
- Composite score of literacy, numeracy, and problem-solving was treated as the dependent variable.



Correlation matrix of literacy, numeracy, and problem-solving.



Screen plot of PCA of literacy, numeracy, and problem solving.

Bagging vs. Boosting

	Bagging	Boosting
Sequent	Two-step	Sequential
Partitioning data into subsets	Random	Give misclassified cases a heavier weight
Sampling method	Sampling with replacement	Sampling without replacement
Relations between models	Parallel ensemble: Each model is independent	Previous models inform subsequent models
Goal to achieve	Minimize variance	Minimize bias, improve predictive power
Method to combine models	Weighted average	Majority vote
Requirement of computing resources	Highly computing intensive	Less computing intensive

Model Comparison

Subset type	Method	R^2	RASE	AAE
No subset	OLS regression	0.1647	43.692	34.603
Training	Boosting	0.2058	42.708	34.031
Training	Bagging	0.4813	34.515	26.979
Validation	Boosting	**0.1791**	**43.488**	**34.597**
Validation	Bagging	0.1685	43.768	34.689

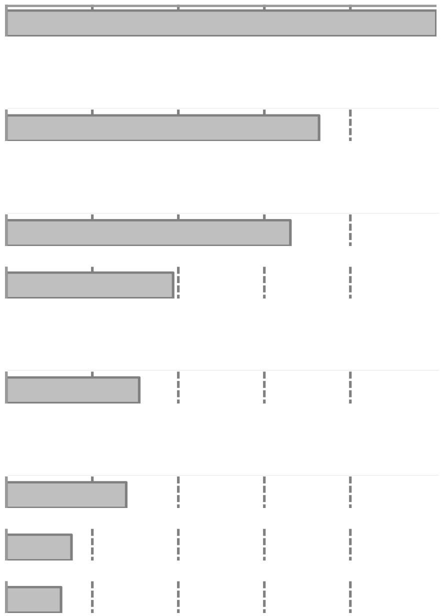
- Bagging and boosting outperformed than OLS regression in variance explained and error rate.
- In training the bootstrap method yielded overfitted models because the R^2 is unreasonably high.
- The boosted tree model outperformed the bagging approach (higher variance explained and lower error).
    - *Using R-square, RASE, and AAE*

OLS Regression Result

Predictor	Estimate	Std. Error	t Ratio	p
Relate new ideas into real-life	13.07	0.85	15.32	<.0001*
Like learning new things	1.93	1.02	1.89	0.0595
Attribute something new	1.54	0.98	1.56	0.1180
Get to the bottom of difficult things	1.80	0.91	1.96	0.0497*
Figure out how different ideas fit together	-3.46	0.96	-3.61	0.0003*
Looking for additional info	0.56	0.95	0.59	0.5576
Voluntary work for non-profit organizations	4.50	0.56	7.97	<.0001*
No influence on the government	-3.08	0.53	-5.85	<.0001*
Trust only few people	-3.57	0.61	-5.84	<.0001*
Other people take advantage of you	-3.28	0.73	-4.50	<.0001*

Final Boosted Tree Model for the USA sample

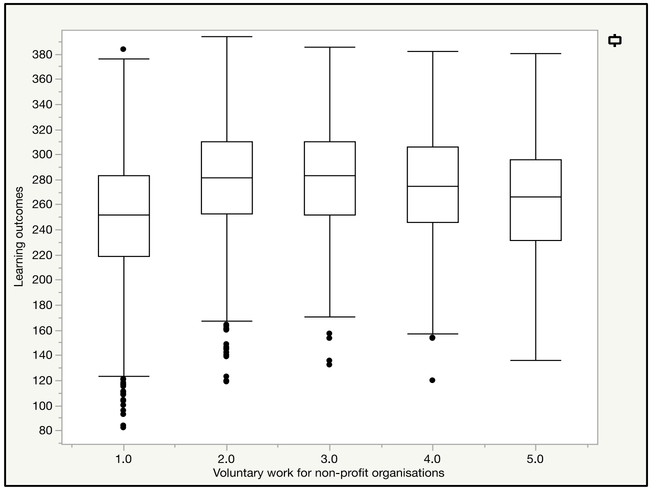
Variable	Number of Splits	Sum of squares	Variable
Voluntary work for non-profit organizations	17	1.1594e+11	
Other people take advantage of you	29	8.5015e+10	
Like learning new things	23	7.687e+10	
Figure out how different ideas fit together	20	4.5563e+10	
Get to the bottom of difficult things	16	3.6352e+10	
No influence on the government	17	3.2498e+10	
Looking for additional info	16	1.7984e+10	
Trust only few people	12	1.5299e+10	

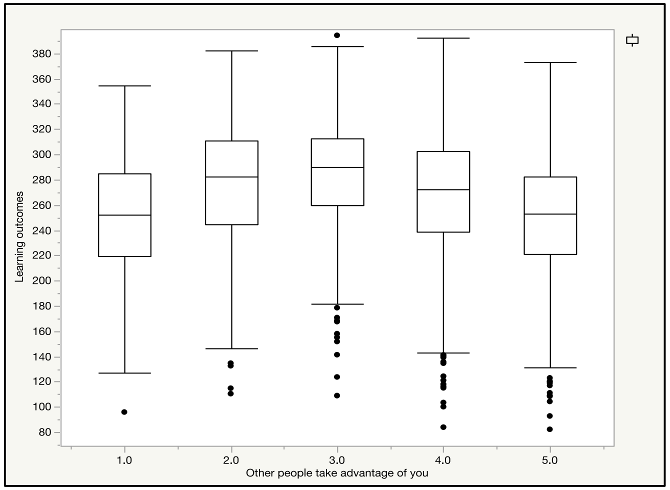


Top predictors: cultural engagement (voluntary work for non-profit organizations), social trust (other people take advantage of you), and readiness to learn (like learning new things).

Median smoothing plots

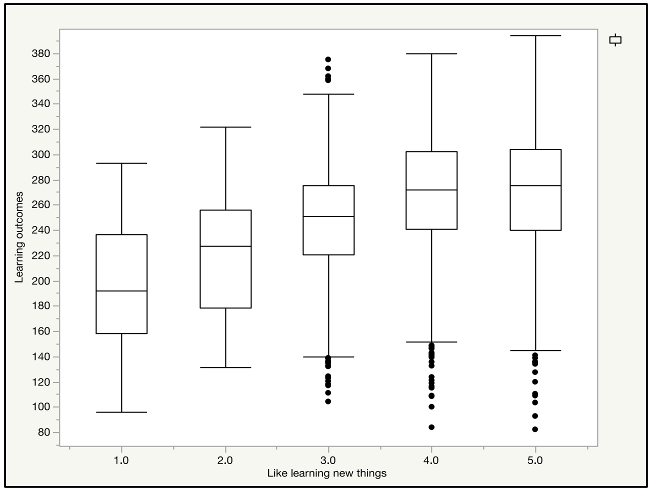
- Learning outcomes and social trust in the US sample.





- Learning outcomes and cultural engagement in the US sample.

- Learning outcomes and readiness to learn in the US sample.



Discussion

- Method choice and model goodness should be assessed on a case-by-case basis.
  - Run both bagging and boosting, then choose the best result according to the model comparison.
- Big data analytics fixes the problem of hypothesis testing by using model building and data visualization
- *When the ensemble method, model comparison, and data visualization are employed side by side, interesting patterns and meaningful conclusions can be found from a big data set.*