Paper 51-2018

Analyzing international assessments:

An ensemble and model comparison approach

Chong Ho Yu, Azusa Pacific University; Hyun Seo Lee, Azusa Pacific University;
Siyan Gan, Pepperdine University; Emily Lara, Azusa Pacific University

ABSTRACT

Learners in Asian countries and religions are among the top performers in the Programme of International Assessment of Adult Competencies (PIAAC). In the past, numerous studies had been conducted to identify the predictors of their outstanding performances. However, this type of analysis is challenging due to the large sample size. To rectify the situation, this study utilized ensemble methods (bagging and boosting) in SAS and JMP to analyze these international assessment data. Bagging can minimize variance but may inflate bias whereas boosting can reduce bias and improve predictive power, but cannot control variance. In order to identify the best model, both methods were employed and different criteria were examined for the model selection.

INTRODUCTION

Lack of replicability is one of the major challenges in social science research. After replicating one hundred psychological studies, Open Science Collaboration (OSC) (2015) found that a large portion of the replicated results were not as strong as what were reported in the original studies, in terms of significance (i.e. p-values) and magnitude (i.e. effect sizes). Specifically, 97% of the original studies reported significant results ($p < .05$), but only 36% of the replicated studies yielded significant results. Further, the average effect size of the replicated studies was only half of that which was found in the initial studies. Nevertheless, this problem is not surprising; single analyses tend to overfit data to models, often generating non-replicable results. To rectify this situation, big data analytics involves partitioning of a big data set into many subsets, on which multiple analyses are run. In each run the model is refined by previous "training." As such, results of big data analyses are considered the product of replicated studies. The process of learning from previous analysis is called "machine learning," whereas the process of merging multiple analyses is known as, "the ensemble method." To be more specific, the ensemble approach compares, complements, and combines multiple methods in the analysis, enabling one to conclude a better predictive outcome than what the analyst could have obtained, using just one solitary analysis (Chen, Lin, & Chou, 2011, Polikar, 2006; Rokach, 2010; Skurichina & Duin, 2002). The ensemble approach can be implemented in JMP Pro, SAS programming environment, SAS Studio, SAS Enterprise Guide and SAS Enterprise Miner. Due to space constraints, this paper focuses on the JMP Pro SAS syntax, and Enterprise Miner only.

MACHINE LEARNING AS A REMEDY TO BIAS AND VARIANCE

Given the emergence and advance of machine learning algorithms in the field of predictive analytics, an ensemble approach of several different machine learning methods has received its due importance. In the field of statistical analysis, the trade-off of bias and variance is a well-known problem. The bias is quantified by the error which results from missing a target. For example, if an estimated mean is 3, but the actual population value is 3.5, then the bias value is 0.5. The variance is the error which results from noise or random fluctuation. When the variance of a model is high, this model is considered unstable. A complicated model tends to have low bias but high variance. Conversely, a simple model is more likely to have a higher bias and a lower variance.

Among many machine learning methods, bagging is popularly utilized to decrease the variance whereas boosting is widely used to weaken the bias in the process of building a predictive model. Bagging, which stands for Bootstrap Aggregation, creates repeated multisets of additional training data from the original sample (Breiman, 1996; Büchlmann & Yu, 2002). Hence, bagging increases the size of these generated data and effectively minimizes the variance of prediction by decreasing the influence of extreme scores (Miller, Lubke, McArtor, & Bergeman, 2016). In contrast, boosting serves a different purpose: increasing

the predictive accuracy. The boosting method first creates a working model from the subsets of the original data set, and then augments the performances of weak models so that they are eventually combined to be a strong model (Breiman, 1998; Schapire, Freund, Bartlett, & Lee, 1998). Depending on the characteristics of the data and the specific aim of the predictive model, these two methods show varying degrees of suitability. Thus, a visual representation of information is conducive to optimizing accessibility to and communication of quantitative message: Data visualization is a powerful method that aids in detection and meaningful interpretation of certain distribution, pattern, and/or relation found in the data (Aparicio & Costa, 2015). The ensemble approach of bagging, boosting, and data visualization in efforts to synthesize the results significantly enhances the overall accuracy and understanding of analyzed material (Skurichina & Duin 2002).

This study demonstrated how one can utilize a variety of data mining techniques, including bootstrap forest, boosted tree, and data visualization, to unveil patterns in the large-scaled and imbalanced data set of Programme for the International Assessment of Adult Competencies (PIAAC). Developed by Organization for Economic and Cooperation and Development (OECD), this international assessment measure and evaluate the basic skills and competencies of adults around the globe. The results of the latest PIAAC (OECD, 2016), collected from 33 participating nations in 2014, indicated that the U.S. adults were lagging behind their international counterparts in all three test categories, namely, literacy, numeracy, and problem solving in technology-rich environments. This alarming if not disturbing report gave an impetus to probe exclusively the U.S. sample. In addition to test items which measured literacy, numeracy, and problem-solving in technology-rich environments, PIAAC also comprised multiple survey items believed to be relevant to learning and therefore supposedly conducive to test outcomes. In the analysis, this study purposefully took account of several of those related items, which are readiness to learn, cultural engagement, political efficacy, and social trust.

METHOD

VARIABLES

The learning outcomes recorded in PIAAC were literacy, numeracy, and technology-based problem-solving scores. The scores of these three domains in the US sample are strongly correlated (Figure 1). Further, as shown by the scree plot (Figure 2), a principal component analysis indicated that all three skills can be combined into one component (eigenvalue = 2.54). Taking all of the above into consideration, the composite score of literacy, numeracy, and problem-solving (the overall learning outcomes) was treated as the dependent variable.

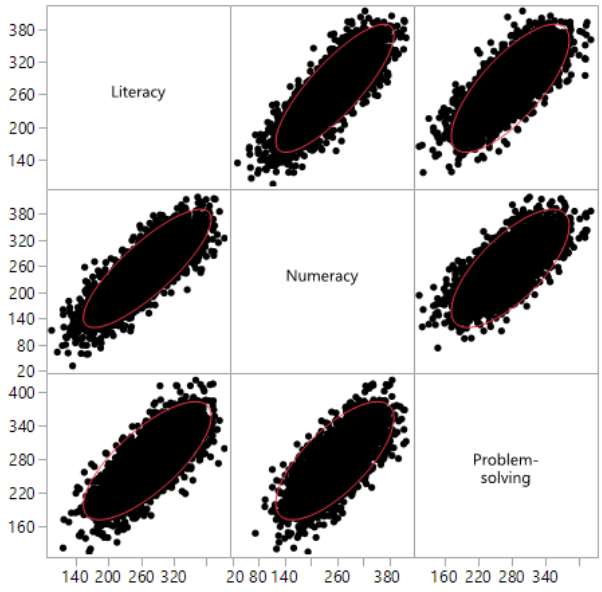

Figure 1. Correlation matrix of literacy, numeracy, and problem-solving.

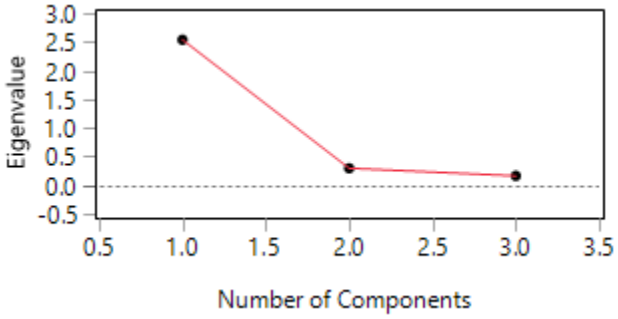

Figure 2. Scree plot of PCA of literacy, numeracy, and problem solving.

DATA ANALYSIS

Because OECD utilized multi-stage sampling, sample weights were used in all analyses. Two ensemble methods, the bootstrap forest and the boosted tree, were run with the U.S. data. The rationale of choosing the ensemble approach is simple. As mentioned before, numerous studies have confirmed that the ensemble approach outperforms any single modeling method (Dietterich, 2000; Freund & Schapire, 1997; Lemmens & Croux, 2006; Meir & Ra¨tsch, 2003; Optiz & Maclin, 1999; Schapire et al., 1998).

Bagging and boosting are the two most popular ensemble methods. Both methods are built on machine learning, in which data sets are partitioned and analyzed by different models. Each model is considered a weak learner as well as a weak classifier, and the final solution is a synthesis of all these weak learners. A weak learner is defined as a model in which the error rate is slightly better than random guessing (Hastie, Tibshirani, & Friedman, 2016). Both bagging and boosting are also resampling methods because the large sample is partitioned and re-used in a strategic fashion. When different models are generated by resampling, inevitably some are high bias model (underfit) while some are high variance model (overfit). In the end, the ensemble cancels out these errors. In addition, it can also account for sample variation. Specifically, each model carries a certain degree of sampling bias, but finally the errors also cancel out each other (Wujek, 2016).

BAGGING

Bagging, which is also known as the bootstrap forest, is a parallel method: in the first stage all resamples are generated independently by sampling with replacement and these replicates do not inform each other (Breiman, 1996). Additionally, in each bootstrap sample about 30% of the observations are set aside for later model validation. These observations are grouped as the out of bag sample (OOBS) (Zaman & Hirose, 2011). At the second stage, the computer algorithm converges these resampled results together by averaging them out. Consider this metaphor: After 100 independent researchers conducted his/her own analysis; this research assembly combines their findings as the best solution.

No double counting on this type of collective wisdom is better than relying on one-person decision. However, it is important to note that the bootstrap method works best when each model yielded from resampling is independent and thus these models are truly diverse. If all researchers in the assembly think in the same way, then no one is thinking. By the same token, if the bootstrap replicates are not diverse, the result might not be as accurate as expected. Putting it bluntly, if there is a systematic bias and the classifiers are bad, bagging these bad classifiers can make the end model worse (Hastie, Tibshirani, & Friedman, 2016). As mentioned before, in theory, an ensemble method should suppress both bias and variance by merging overfitted and underfitted models. However, Kotsiantis (2013) found that bagging tends to generate less heterogeneous models than its boosting counterpart. Additionally, Fumera, Roli, and Serrau (2005) found that the misclassification rate of bagging has the same bias as a single bootstrap though the variance is reduced by increasing the number of resamples. This can be explained by the disposition of overfitting in bagging. When these overfitted models are averaged, the same bias is retained while the variance is cancelled out.

BOOSTING

Boosting, also known as the boosted tree, is a sequential and adaptive method because the previous model informs the next model so that improvement can be made in subsequent modeling (Breiman, 1998; Freund & Schapire, 1997; Optiz &Maclin, 1999). Initially, all observations are assigned the same weight. If the model fails to classify certain observations correctly, then these cases are assigned a heavier weight so that they are more likely to be selected in the next model. In the subsequent steps, each model is constantly revised in an attempt to classify those observations successfully. Boosting is so named because of gradient improvement by learning mistakes in previous steps. Ultimately, the final model is created by a majority vote as the best solutions are kept and the worst ones are eliminated. While bagging requires many independent models for convergence, boosting reaches a final solution after a few iterations. Hence, boosting is much less computing-intensive than bagging. The differences between bagging and boosting is summarized in Table 1.

Characteristics	**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

Table 1. Comparison of bagging and boosting

In JMP Pro bootstrap forest, boosted tree, and model comparison can be assessed at Predictive Modeling, as shown in Figure 3.

Figure 3. Predictive modeling in JMP.

In the SAS programming environment the procedures to run a random forest and a gradient boost tree are PROC HPFOREST and PROC TREEBOOST, respectively. HP stands for high performance, meaning that this set of procedures is built for running extremely large data sets (count in terabytes). For this data set it is an overkill. Nonetheless, all tasks are run with recycled electrons and thus there is no waste.

```
PROC HPFOREST DATA= <DATA>;
TARGET <DEPENDENT VARIABLE> /
LEVEL= <BINARY, NOMINAL, INTERVAL>;
INPUT <VARIABLE 1, VARIABLE 2, VARIABLE 3…ETC.>/
LEVEL= <BINARY, NOMINAL, ORDINAL, INTERVAL>;
RUN;

PROC TREEBOOST DATA= <DATA>;
INPUT <PREDICTOR 1, PREDICTOR 2…>/ LEVEL=;
TARGET <DEPENDENT VARIABLE> /;
ASSESS VALIDATA= <partition_for_training/validation>;
RUN;
```

In SAS Enterprise Miner, the analyst can go one step further by creating an ensemble of models yielded from different modeling techniques, such as regression, neural networks, decision tree, boosting, bagging…etc. Figure 4 shows that high performance forest, gradient boosting, and conventional regression are run with the PIAAC data set. Their results are merged into the control point node. This node does not perform any computation or produce any result; rather, it simply stores the results for the ensemble node. Figure 5 shows the result of the ensemble model. At the end, all results are evaluated by model comparison.

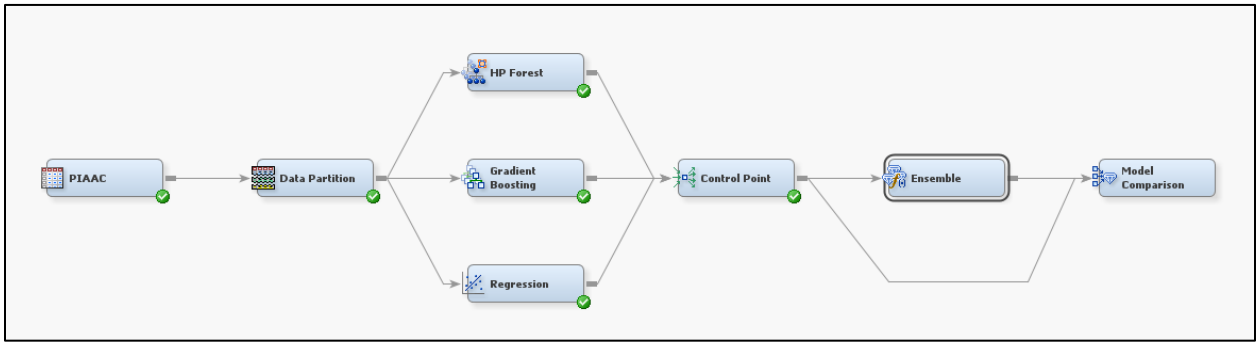

Figure 4. Flowchart in Enterprise Miner.

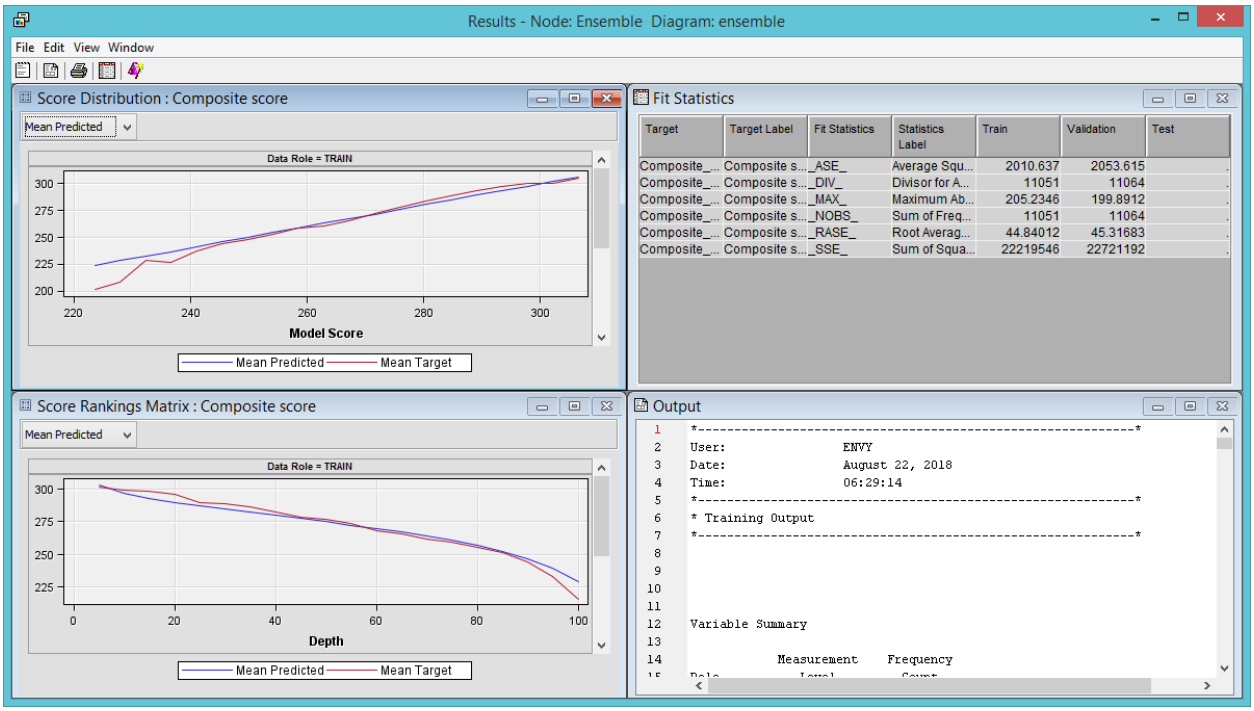

Figure 5. Result of ensemble mode in Enterprise Miner.

DEBATE ON BAGGING AND BOOSTING

Whether bagging or boosting is better has been an ongoing debate for nearly two decades. It is not surprising to see that in some situations, bagging outperforms boosting whereas in others the outcomes are reversed (Chandrahasan, Christobel, Sridhar, & Arockiam, 2011; Dietterich, 2000; Khoshgoftaar, van Hulse, & Napolitano, 2011; Kotsiantis, 2013; Wang, Zhang, & Guo, 2015; Zaman & Hirose, 2011). Many studies concluded that boosting outperforms bagging in most cases, specifically when the analyst works with a noisy data set. On the other hand, bagging is a suitable option in data environment with less noise (Dietterich, 2000; Khoshgoftaar et al., 2011). Nonetheless, it is impractical for the researcher to analyze how noisy the data set is before choosing a particular ensemble approach. In addition, the bias–variance tradeoff is a central but insurmountable problem in machine learning. Ideally, the analyst hopes to obtain a model that can accurately detect the patterns in the data set and also generalize the finding to unseen data. As aforementioned, bagging is good at minimizing variance whereas boosting is capable of reducing bias, but none can accomplish both simultaneously.

The authors are convinced that there is no single best ensemble method applicable to analyze all situations. A strategy is to run both analysis and select the better fitting one by model comparison. In model comparison, there are several criteria for assessing the goodness of a model, namely, the R2, the Root Average Squared Error (RASE), and the Average Absolute Error (ASE). The R^2 is the variance explained whereas ASE is the average error rate of the model. RASE is the same as RMSE except that RMSE adjusts for degrees of freedom but RASE does not. The values in the final model (i.e. the validation model), instead of the training model, were evaluated because the training model is always overfitted. Unlike classical hypothesis testing, which relies on a cut-off for decision-making, the data mining method aims to recognize the data pattern, without a rigid cut-off for variable selection.

After identifying best model and the most important predictors, median smoothing was utilized to examine the relationship of the predictors and the learning outcomes. In this large-scale assessment, the sample size of each OECD member nation was around 5,000. When thousands of data points generate a noisy scatterplot, detecting a pattern within the sample becomes challenging. This problem, called overplotting,

is resolved by dividing the data into several portions along the x–dimension, computing the median of y in each portion, and looking at the trend after connecting the medians (Tukey, 1977; Yu, 2014).

RESULT

BAGGING, BOOSTING, AND MODEL COMPARISON

Table 2 shows the descriptive statistics of the US test scores. For inferential statistics, variables related to readiness to learn, cultural engagement, political efficacy, and social trust were input into bagging and boosting as predictors of composite learning outcomes, respectively.

Gender	Literacy		Numeracy		Problem-solving		Composite	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Female (n=2,323)	269.18	47.73	245.38	53.68	275.00	42.20	259.82	46.86
Male (n=2,687)	270.39	49.15	260.48	56.78	280.26	44.40	266.58	49.01

Table 2. Descriptive statistics of test scores of the USA.

The OLS regression, bagging, and boosting results were evaluated by model comparison criteria and the best one was retained (Table 3).

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

Table 3. Model comparison.

It is important to point out that there are different criteria for evaluating the goodness of a model. There is no single best model. Determining which one is more appropriate depends on the goal of the modeling: decision (e.g. yes/no, pass/fail, accept/reject), rankings (e.g. assign a score to each case), or estimates (e.g. least error) (SAS Institute, 2018). In this study, the weight is put on estimates.

Based on this criterion, it is obvious that both bagging and boosting outperformed OLS regression in terms of variance explained and the error rate. More importantly, in OLS regression almost every predictor is found to be significant in a two-tailed test ($P < .05$). If a one-tailed test is used, then every predictor is significant (see Table 4). It is important to re-emphasize that no cross-validation (CV) by subsetting the data was done for regression modeling and thus stability of this "good" result is in question. On the other hand, subsetting was used in both bagging and boosting. In training the bootstrap method yielded overfitted models because the R^2 is unreasonably high. Therefore, a proper comparison should be based on the validation results only. Using the criteria of R-square, RASE, and AAE, the boosted tree model slightly outperformed the bagging approach (higher variance explained and lower error).

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*

Table 4. OLS regression result.

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	

Table 5. The final boosted tree model for the USA sample..

Table 5 shows the ranking of predictors in relation to the overall learning outcomes. The top three predictors were cultural engagement (voluntary work for non-profit organizations), social trust (other people take advantage of you), and readiness to learn (like learning new things).

The relationship between readiness to learn and learning outcomes were positive and linear. However, non-linear patterns were detected when social trust and cultural engagement regressed against learning outcomes. Because the sample size was extremely large, median smoothing was employed for each level of the X variable. By doing so, the X-Y association could be detected by the trend of the medians (see Figure 6-8).

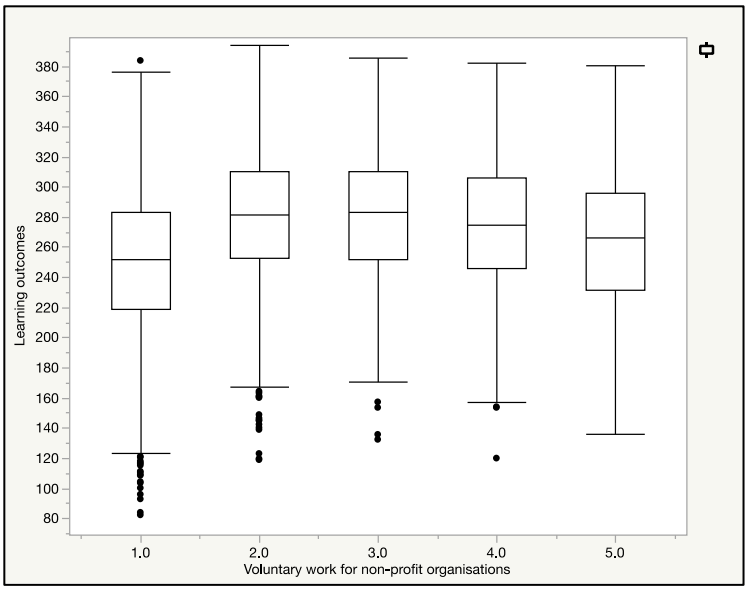

Figure 6. Median smoothing plot of learning outcomes and cultural engagement in the US sample.

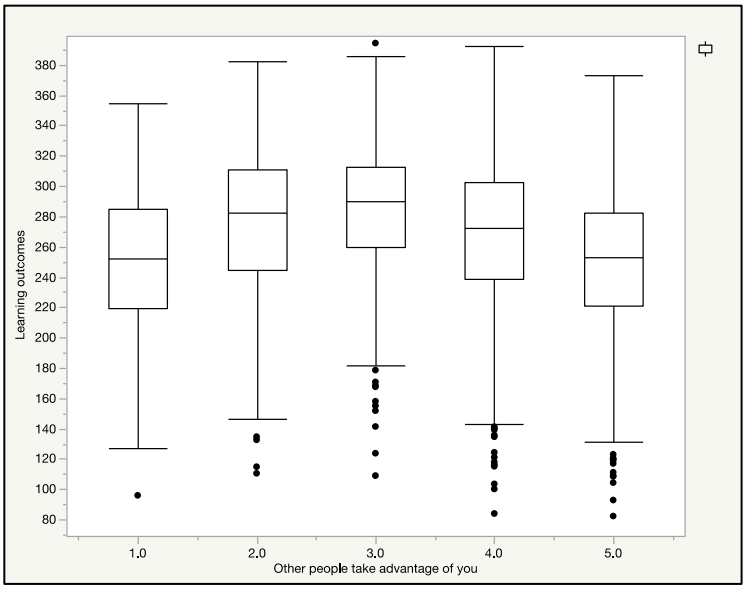

Figure 7. Median smoothing plot of learning outcomes and social trust in the US sample.

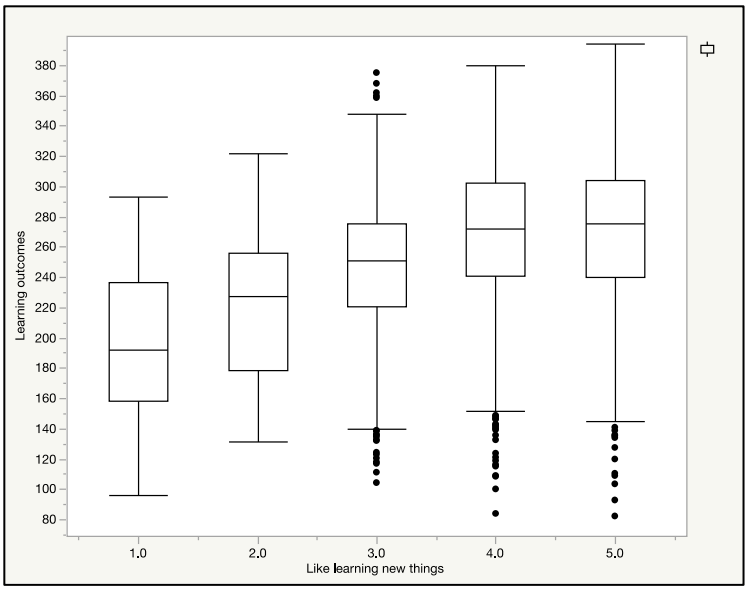

Figure 8. Median smoothing plot of learning outcomes and readiness to learn in the US sample.

CONCLUSION

Like classical procedures, there are pros and cons in different data mining techniques, and as a result sometimes it is difficult to determine which method is more appropriate. Some researchers count on simulation methods to examine the robustness of various techniques based on the assumption that real-world data are usually messy. However, it is unlikely for simulators to generate all possible scenarios and therefore advice like "in most cases" one particular method is superior to another is not helpful. Hence, it is the conviction of the authors that method choice and model goodness should be assessed on a case-by-case basis. Despite the fact that bagging is relatively resource-demanding, most mid-range computers are capable of performing a bootstrap forest in a short time. It is advisable to run both bagging and boosting, and then choose the best result according to the criteria of model comparison. In addition, this illustration focuses on the ensemble of the same modeling techniques.

Migrating from classical procedures to big data analytics is no doubt a paradigm shift. In hypothesis testing, decisions are based on certain cut-off points (e.g. $p < .05$, RMSEA < .1) whereas data mining emphasizes pattern recognition (Bishop, 2006; Kosinski, Wang, Lakkaraju, & Leskovec, 2016). Researchers who are trained in classical methods might be puzzled by the omission of decision points in big data analytics. As shown in the results section, the output tables show the rank order of predictors yielded by bagging or boosting, but none of the predictors is marked as "significant" or "not significant." At first glance, it is a shortcoming, but indeed it is a blessing in disguise. Rosnow and Rosenthal (1989) pointed out the problem of the cut-off logic by saying, "Dichotomous significance testing has no ontological basis. That is, we want to underscore that, surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?" (p.1277). The meaning of the p value is about the chance of observing the test statistics if the same study is repeated over and over in the long run given that the null hypothesis is true. However, when the analyst has a great amount of observations at hand, why does s/he bother to ask a hypothetical question about big data accumulated by repeated studies? Big data analytics remediates the problem of hypothesis testing by utilizing both model building and data visualization. Very often, data visualization could unveil patterns that might go undetected by statistical figures. For example, in conventional data analysis non-linear regression is seldom used, but running OLS regression modeling on this data set could result in misleading conclusions because both cultural engagement and social trust are not linearly associated with learning outcomes. Consider this sample. Figure 9 shows an overplotted scattergram and a linear regression line of learning outcomes and social trust. Without pattern recognition by median, smoothing the regression line and the regression coefficient could fool analysts.

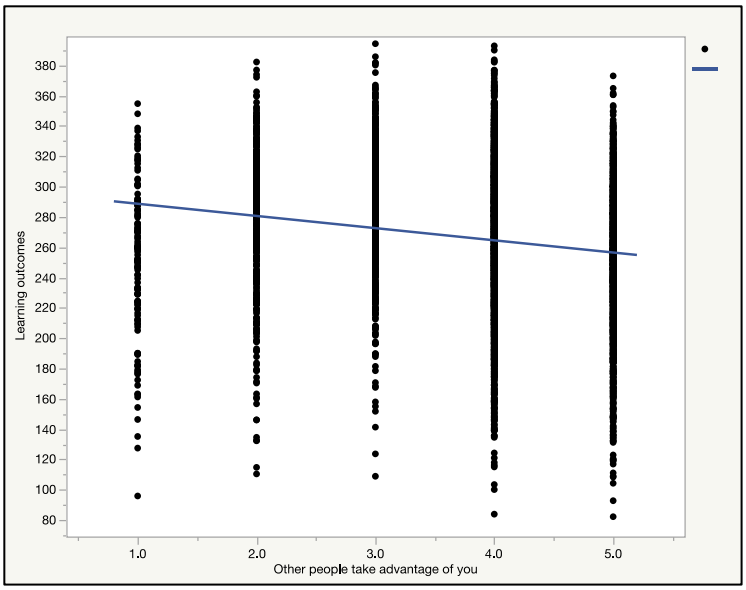

Figure 9. Overplotted scattergram of learning outcomes and social trust without median smoothing.

In conclusion, it is the conviction of the authors that while the ensemble method, model comparison, and data visualization are employed side by side, interesting patterns and meaningful conclusions could be extracted from a big data set.

REFERENCES

Aparicio, M., & Costa, C. J. (2015). Data visualization. *Communication Design Quarterly Review, 3*, 7-11.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics, 26*, 801–849.

Büchlmann, P., & Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics, 30*, 927-961.

Chandrahasan, R. K., Christobel, A., Sridhar, U. R., & Arockiam, L. (2011). An empirical comparison of boosting and bagging algorithms. *International Journal of Computer Science and Information Security, 9*(11), 147-152.

Chen, S. C., Lin, S. W., & Chou, S. Y. (2011). Enhancing the classification accuracy by scatter-search-based ensemble approach. *Applied soft computing*, *11*, 1021-1028.

Cheung M. L., & Jak, S. (2016) Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7,* Article 738. doi: 10.3389/fpsyg.2016.00738

Cutler, R. (2017). What statisticians should know about machine learning? Proceedings of 2017 SAS Global Forum. Retrieved from http://support.sas.com/resources/papers/proceedings17/0883-2017.pdf

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning,* 40(2), 139–157.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Fumera, G., Roli, F. & Serrau, A. (2005). Dynamics of variance reduction in bagging and other techniques based on randomisation. *Lecture Notes in Computer Science, 3541*, 316–325.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. (2011). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 41*, 552–568.

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec. J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods, 21*, 493–506.

Kotsiantis, S. (2013). Bagging and boosting variants for handling classifications problems: A survey. *Knowledge Engineering Review, 29*(1), 78–100. doi:10.1017/S0269888913000313.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 63,* 276–286.

Meir, R., & Ra˙tsch, G. (2003). An introduction to boosting and leveraging. In: Advanced Lectures on Machine Learning. *Lecture Notes in Computer Science, 2600*, 118–183.

Miller, P., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods, 21*, 583–602.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11*, 5-11.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi: 10.1126/science.aac4716. Retrieved from http://science.sciencemag.org/content/349/6251/aac4716

Optiz, D., Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169–198.

Organization for Economic Co-operation and Development [OECD]. (2016). *Technical report of the survey of adult skills (PIAAC)*. Retrieved from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1-39.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

SAS Institue. (2018). *Applied analytics using SAS Enterprise Miner*. Cary, NC: Author.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics, 26*, 1651–1686.

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice, 35*(3), 38-54.

Skurichina, M., & Duin, R. P. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, *5*, 121-135.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison–Wesley Publishing Company.

Tukey, J. W. (1986). *The collected works of John W. Tukey, Volume III: Philosophy and principles of data analysis: 1965–1986*. L. V. Jones (Ed.). Pacific Grove, CA: Wadsworth.

Wang, G. W., Zhang, C. X., & Guo, G. (2015). Investigating the effect of randomly selected feature subsets on bagging and boosting. *Communications in Statistics—Simulation and Computation*, *44*, 636–646.

Wujek, B. (2016, September). *Practical guidance for machine learning applications*. Paper presented at SAS Analytics Experience Conference, Las Vegas, NV.

Yu, C. H. (2014). *Dancing with the data: The art and science of data visualization.* Saarbrucken, Germany: LAP.

Zaman, M. F., & Hirose, H. (2011). Classification performance of bagging and boosting type ensemble methods with small training sets. *New Generation Computing, 29*, 277-292.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chong Ho Yu
Department of Psychology; Department of Math and Physics
Azusa Pacific University
cyu@apu.edu; chonghoyu@gmail.com
http://www.creative-wisdom.com/pub/pub.html
https://www.researchgate.net/profile/Chong_Ho_Yu

.