Predictive Modelling of Evidence Informed Teaching

Dell Zhang and Chris Brown

Abstract In this paper, we analyse the questionnaire survey data collected from 79 English primary schools about the situation of evidence informed teaching, where the evidences could come from research journals or conferences such as EDM. Specifically, we build a predictive model to see what external factors could help to close the gap between teachers' belief and behaviour in evidence informed teaching, which is the first of its kind to our knowledge. The major challenge, from the data mining perspective, is that the Likert scale responses are neither categorical nor metric, but actually ordinal, which requires special consideration when we apply statistical analysis or machine learning algorithms. Adapting Gradient Boosted Trees (GBT), we achieve a decent prediction accuracy (MAE=0.36) and gain new insights into possible interventions for promoting evidence informed teaching.

Dell Zhang
DCSIS, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK.
✉ dell.z@ieee.org

Chris Brown
LCLL, UCL Institute of Education, 20 Bedford Way, London WC1H 0AL, UK.
✉ chris.brown@ioe.ac.uk

ARCHIVES OF DATA SCIENCE (ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. -, No. -, -
ISSN 2363-9881

1 Introduction

The research findings from educational research journals or conferences such as EDM[1] would not have much impact if they could not influence teachers' classroom practice.

Recently *evidence informed teaching*, aka evidence informed practice, has been receiving more and more attention from government policy makers, academic researchers, and school teachers (Brown, 2014; Goldacre, 2013). Despite well-documented controversy and critique, it has been shown that there are substantial benefits associated with teachers using information from research to enhance their practice. For instance, where evidences from research are used effectively as part of initial teacher education and continuing professional development, with a focus on addressing improvement priorities, it makes a positive difference in terms of teacher, school and system performance (Cordingley, 2013; Mincu, 2013; Godfrey, 2014). Furthermore, the experience of "research-engaged" schools that take a strategic and concerted approach in this area is generally positive, with studies suggesting that research engagement can shift a school from an instrumental "top tips" model of improvement to a learning culture in which staff work together to understand what appears to work, when and why (Godfrey, 2014; Handscomb and MacBeath, 2003). In addition, it is also noted that schools which have made a commitment to practitioner research report increased application for teaching posts, higher teacher work satisfaction, and better staff retention (Godfrey, 2014).

The direction of travel of recent educational policy in England has also been directed towards promoting or requiring teachers to better engage with evidence. Particularly, the significant investment in initiatives aimed at connecting practitioners with educational research undertaken by the 2007-2010 New Labour government. After New Labour, the UK's current Conservative/LibDem Coalition government, elected in 2010, changed tack and pursued a "self-improving school-led school system". Nonetheless, evidence use is still front and centre, with researchers suggesting that core characteristics of self-improvement should include: (a) teachers and schools being responsible for their own improvement; and (b) teachers and schools are required to learn from each other and from research so that effective practice spreads (Greany, 2014).

There is very much an impetus then for school leaders to ensure they and their staff seek out and engage with quality evidence from research in relation to

[1] http://www.educationaldatamining.org/

issues of teaching and learning. It has been argued by education science experts that to do so requires school leaders to focus on and address the following distinct but overlapping and interdependent areas: (i) the teachers' capacity and ability to engage in and with research; (ii) the cultures that are attuned to research use, i.e., that make research use a cultural norm; (iii) the inclusion of research use as part of an effective learning environment; and (iv) the structures, systems and resources that facilitate research use as well as the sharing of best practice. Although the areas outlined above are important to meaningful and effective research use, they are each likely to comprise a number of factors and of them some are more likely to be effective in driving evidence informed teaching than others. Correspondingly, it is crucial to understand, in relation to those factors, where school leaders should be focusing their efforts in order to establish evidence informed schools.

In this paper, we take a data mining approach to investigating the effectiveness of potential school policy levers for the promotion of evidence informed teaching. To the best of our knowledge, there has been no such study so far. The rest of this paper is organised as follows. In Section 2, we describe the data for investigation in detail. In Section 3, we carry out some exploratory analysis of the data. In Section 4, we present our predictive model learned from the data. In Section 5, we draw conclusions and discuss the future work.

2 Data

The data of investigation come from a questionnaire survey conducted from 2 October to 19 October 2014 in 79 English primary schools. It was developed using SurveyMonkey[2] and distributed electronically to all involved schools via their principal/headteacher. The response rate was above 80% in half of the schools, and above 60% in most (four-fifth) of the schools.

The questionnaire itself aims to provide an indication of the base state of each school with respect to evidence informed teaching. The design of the survey was undertaken in conjunction with Prof Alan Daly (University of California, San Diego) who is experienced in examining the movement of evidence within and between schools in Californian school districts (Daly, 2010). Before it was distributed, the survey was also piloted with teachers from the primary sector

[2] https://www.surveymonkey.com/

Table 1 The questionnaire survey for evidence informed teaching.

	Variable	Question
Effects	R-support	I do not support implementing a school-wide change without research to support it.
	R-practice	Information from research plays an important role in my teaching practice.
	R-approach	I have found information from research useful in applying new approaches in classroom.
	R-discussion	I have discussed relevant research findings with my colleagues in the last year.
Causes	Strategies	Research and evidence are used to inform staff about potential improvement strategies.
	Conversation	Research and evidence are used to stimulate conversation or dialogue around an issue.
	Encouragement	My school encourages me to use research findings to improve my practice.
	Overall-trust	Staff in my school trust each other.
	SeniorL-trust	When senior leadership in my school tell me something I can believe it.
	MiddleL-trust	When middle leadership in my school tell me something I can believe it.
	Teacher-trust	When teachers in my school tell me something I can believe it.
	Respect	Staff in my school respect each other.
	Depending	Staff in my school can depend on each other even in difficult situations.
	Sharing	Staff in my school are eager to share information about what does and doesn't work.
	Key-assumptions	Staff in my school frequently discuss underlying assumptions that might affect key decisions.
	New-ideas	Staff in my school value new ideas.
	Training	My school has made time available for education or training activities for school staff.
	Forums	My school has forums for sharing information among staff.
	Experimentation	My school experiments with new ways of working.
	Evaluation	My school has a formal process for evaluating programs or practices.

(not involved in the project) in order to test "face" and "construct" validity. Feedback from the pilot was then incorporated into the final questionnaire.

Table 1 lists the questions each of which corresponds to a variable of interest to us for this investigation. The first four variables are about a teacher's own experience with evidence informed teaching, while the other sixteen variables refer to the external factors that may affect his or her experience. Therefore we call the former group of variables *effects* and the latter *causes*.

The answer to each question would be in a five-level *Likert scale* (Likert, 1932): "strongly disagree", "disagree", "neither disagree nor agree", "agree", and "strongly agree". For the convenience of further data analysis and modelling, we represent the above Likert scale points as numerical integers values $[-2,-1,0,+1,+2]$.

One teacher can submit one and only one response to the questionnaire. Each distinct teacher's response is considered as an example (in the context of statistical *machine learning* (Hastie et al, 2009)). Since a teacher was not required to answer all questions, some values could be missing in the data. After discarding those examples with too many (more than a quarter) missing values, we have in total 696 examples left. For the remaining examples, we simply fill the missing values with the neutral value 0. It is possible to use more sophisticated filling methods (such as matrix factorisation), but that would not make much difference as the missing values that need to be filled are only of a very small percentage.

3 Analysis

Likert scale values are not exactly nominal (categorical), because they have a rank order by which they can be sorted — "strongly agree" is usually regarded better than "agree". Likert scale values are not exactly metric (interval or ratio) either though they are encoded as integers in this paper, because they do not allow for relative degree of difference between them — the distances between "strongly agree" and "agree" may not be the same as that between "agree" and "neither disagree nor agree", for example; people actually often think that there is a bigger difference between items at the extremes of the range than in the middle. Likert scale data are in fact ordinal (Agresti, 2010), therefore one should make use of *nonparametric* statistical methods (Field and Hole, 2003) for data analysis and modelling rather than standard parametric techniques such as Stu-

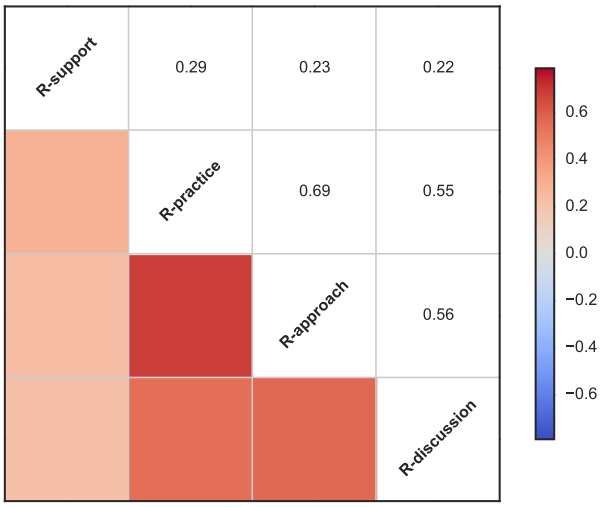

Fig. 1 The correlation plot of effect variables.

dent t-test which assumes a normal distribution of data. Although Likert scale questionnaires are widely used in research fields such as human-computer interaction and computer science education, the sad truth is that most researchers wrongly use parametric techniques to deal with them. For example, it has been found that in the CHI proceedings prior to 2010, 45.6% of the papers reported on Likert-type data, but only 8.3% of them applied nonparametric techniques (Kaptein et al, 2010).

For the effect variables, we measure their pairwise associations using a nonparametric method Kendall's τ rank correlation coefficient (the tau-b version which accounts for ties). As shown in Figure 1, the correlation between `R-support` that reflects teachers' belief and each other effect variable (`R-practice`, `R-approach`, `R-discussion`) that reflect teachers' behaviour is quite low (< 0.30), which suggests that there is a gap between supporting the idea of evidence informed teaching and putting it into practice. In particular, according to the nonparametric Wilcoxon signed-rank test, the discrepancy between `R-support` and `R-practice` is significant (p-value $1^{-25} \ll 0.01$). The contingency table (crosstab) of them is shown in Figure 2.

For the cause variables, we perform hierarchical agglomerative clustering (the average-linkage version) based on their pairwise Kendall's τ rank correlation coefficient. As shown in Figure 3, the cause variables are clearly grouped

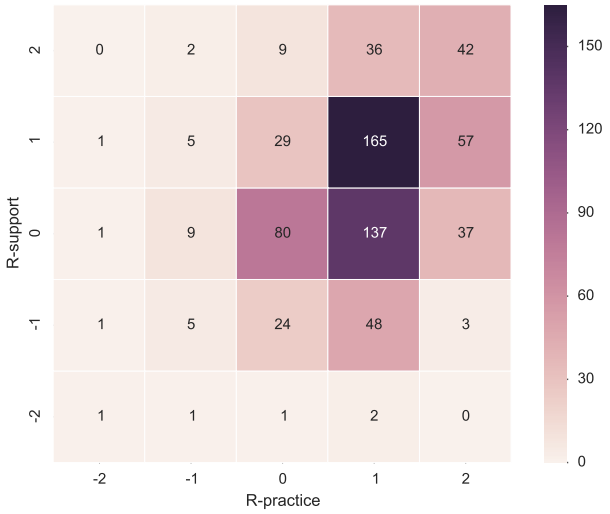

Fig. 2 The crosstab of `R-support` and `R-practice`.

into three clusters. For example, one cluster contains three cause variables: `Strategies`, `Conversation`, and `Encouragement`.

4 Modelling

It is of particular interest to school leaders to find out why some teachers give *support* to the idea of evidence informed teaching but do not put it into *practice*, and how the situation could be made better. For this purpose, we build a predictive model for teachers who gave positive answers ("agree" or "strongly agree") to the `R-support` question to see what their answers would be to the `R-practice` question based on their external factors. More specifically, we formulate the research problem as follows: given an example (with `R-support`¿0) represented as a vector of those sixteen cause variables (i.e., features), predict the corresponding value of the effect variable `R-practice` (i.e., target or label). Since we have labelled data, we can address this problem through *supervised learning* (Hastie et al, 2009).

As we have explained earlier in Section 3, both the features and the labels for this supervised learning problem are of ordinal nature. Therefore, we choose to employ the Decision Tree learning algorithm, to be more precise, CART

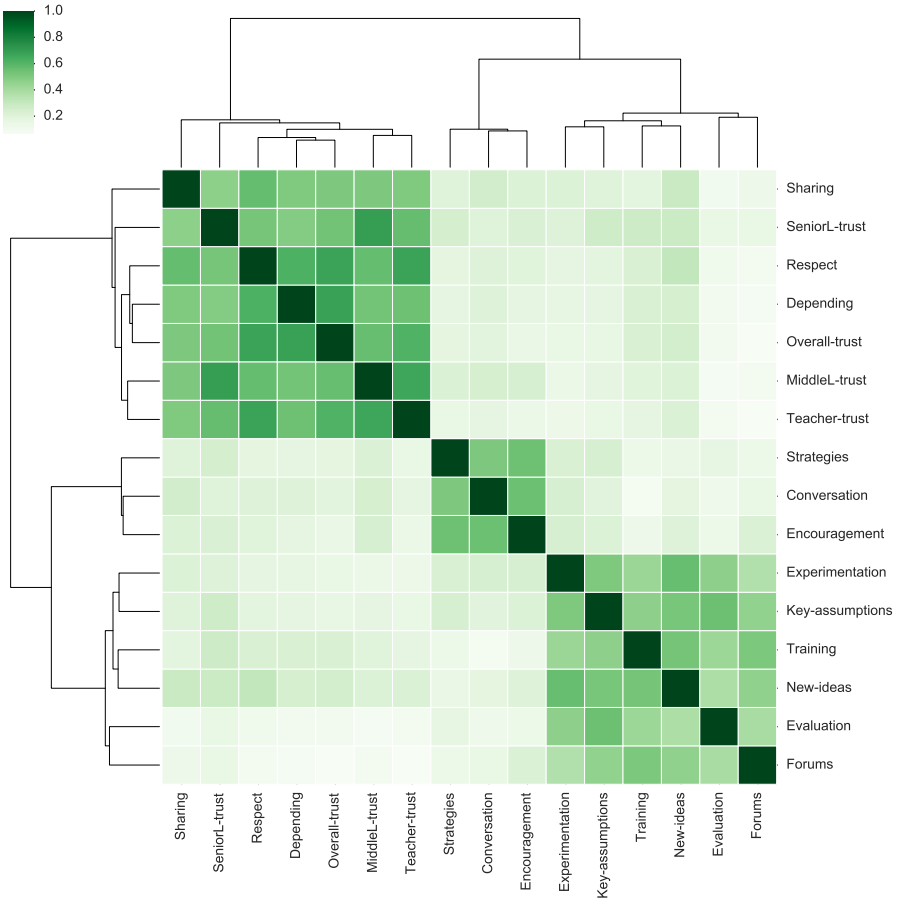

Fig. 3 The cluster map of cause variables.

(Breiman et al, 1984). Unlike commonly used predictive models (such as Linear Regression, Naive Bayes, and Support Vector Machines), Decision Tree is a nonparametric supervised learning algorithm that can handle ordinal value features for classification and regression. However, the standard version of Decision Tree learning algorithm is still not able to handle ordinal targets directly. One simple solution is to treat the ordinal targets as numerical metric values, run the Decision Tree learning algorithm for regression, and then translate the regression output back into a discrete class label in a post-processing step (e.g., rounding to the nearest ordinal target value). Another possible solution is to

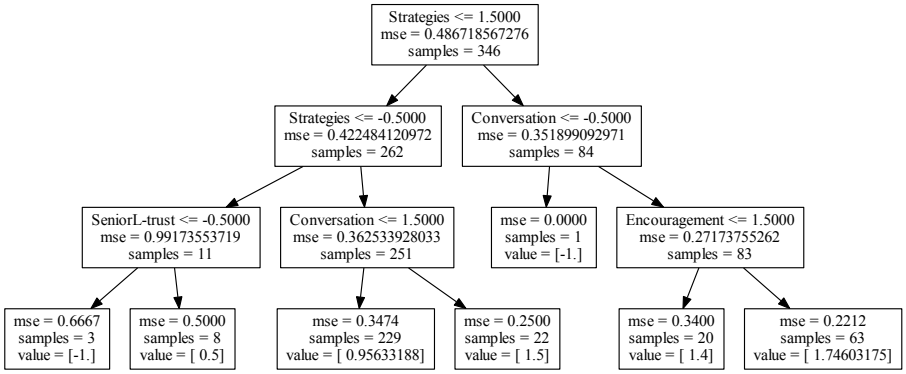

Fig. 4 A Decision Tree learned from our data.

carry out a sequence of binary classification using the Decision Tree learning algorithm (Frank and Hall, 2001), e.g., in our case, for five-level Likert scale targets represented as integers ranging from -2 to $+2$, we need four binary classifiers: (i) $\{-2\}$ vs $\{-1,0,+1,+2\}$, (ii) $\{-2,-1\}$ vs $\{0,+1,+2\}$, (iii) $\{-2,-1,0\}$ vs $\{+1,+2\}$, (iv) $\{-2,-1,0,+1\}$ vs $\{+2\}$. On our data, these two approaches worked similarly well, so we only report the results of the former in this paper. Furthermore, we could enhance the prediction accuracy by using not just one single Decision Tree, but an *ensemble* of Decision Trees (Hastie et al, 2009; Seni and Elder, 2010). There are two typical ensemble learning methods, bagging and boosting. When applied to Decision Trees, they lead to two popular learning algorithms Random Forest (RF) (Breiman, 2001a) and Gradient Boosted Trees (GBT) (Friedman, 1999a,b) respectively. Both RF and GBT have kept showing the best performances on a variety of real-world data mining problems (Caruana and Niculescu-Mizil, 2006). In this work, we have adapted the implementation of these tree-based learning algorithms from an open-source Python machine learning library, scikit-learn[3], for the prediction of Likert scale values.

An advantage of using tree-based learning algorithms is that the generated model is simple to understand and to interpret. Figure 4 shows a Decision Tree learned from our data.

Only two main parameters of GBT have been tuned in the experiments: one is the the number of trees in the ensemble (`n_estimators`), and the other is the size of the random subsets of features to consider when splitting a

[3] `http://scikit-learn.org/`

Table 2 The prediction performance of tree-based learning algorithms.

	RMSE	MAE
Decision Tree	0.82 ± 0.11	0.62 ± 0.05
Random Forest	0.66 ± 0.10	0.41 ± 0.07
Gradient Boosted Trees	0.64 ± 0.07	0.36 ± 0.08

node (`max_features`) in each tree. For `n_estimators`, usually the larger it is the better performance, but also the longer time the computation will take. In the end, we settled at a GBT with 50 trees which yields a decent prediction performance. For `max_features`, usually the lower it is the greater the reduction of variance, but also the greater the increase in bias. The rule of thumb is to use $\texttt{max\_features} = \texttt{n\_features}$ for regression problems and $\texttt{max\_features} = \sqrt{\texttt{n\_features}}$ for classification problems, where `n_features` is the number of features in the data. Our task is more similar to to a classification problem rather than a regression problem, because the target variable only has five possible values. So we set this parameter to $\sqrt{\texttt{n\_features}}$ instead of its default value `n_features`.

The prediction performance of our model is measured by Root Mean Squared Error (RMSE) $\sqrt{(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2)/n}$ and also Mean Absolute Error (MAE) $(\sum_{i=1}^{n}|\hat{y}_i - y_i|)/n$, where $\hat{y}_i$ is the perdition of the i-th true target value y_i, and n is the total number of predictions. Although they are not really made for ordinal target variables, they are the two most widely used performance measures for recommender systems (Jannach et al, 2010) which also make predictions about users' Likert-type scale ratings of items (e.g., from 1-star to 5-star). To be consistent with existing research literature and facilitate performance comparison, we also would like to use them here. Table 2 shows the prediction performance of those three tree-based learning algorithms evaluated using *stratified 10-fold cross-validation*. It can be seen that GBT achieves the best prediction performance (i.e., the smallest RMSE and MAE), so we will focus on it in the rest of our discussion.

A feature's relative importance in the Decision Tree, with respect to the predictability of the target variable, could be roughly assessed by the expected fraction of examples split by the node using that feature. In ensemble algorithms like GBT, those expected fractions are averaged over all the (randomised) trees in the ensemble to give a fairly accurate estimate of feature importance. The current importance scores of the features in our final GBT predictive model are shown in Table 3. The three most important features, or cause variables,

Table 3 The current importance score and potential (cumulative) improvement contribution of each feature in the final predictive model.

feature	importance	improvement
Strategies	0.318	+0.162
Conversation	0.271	+0.488
Encouragement	0.165	+0.763
SeniorL-trust	0.052	+0.841
MiddleL-trust	0.049	+0.873
Overall-trust	0.036	+0.873
Depending	0.025	+0.873
New-ideas	0.014	+0.873
Experimentation	0.013	+0.873
Sharing	0.012	+0.873
Respect	0.012	+0.873
Forums	0.009	+0.873
Evaluation	0.008	+0.873
Teacher-trust	0.007	+0.873
Training	0.006	+0.873
Key-assumptions	0.003	+0.873

are Strategies, Conversation, and Encouragement. They actually correspond to one cluster of cause variables that we have discovered before (see Section 3). So it is fair to say that interventions in these three aspects would be most effective for promoting evidence informed teaching.

One benefit of having a predictive model is the ability to forecast the amount of improvement that can be brought by any change to the input cause variables. One by one, in the descending order of their importance scores, we push the features' values to their maximum ("strongly agree") and check how much cumulative increase we can see in the target variable R-practice on average. It is demonstrated in Table 3 that by just optimising the three most important cause variables, Strategies, Conversation, and Encouragement, the effect variable R-practice is likely to be increased by 0.763, which means that its average score could be improved from the initial value 1.127 (i.e., mostly "agree") to 1.890 (i.e., almost all "strongly agree").

5 Conclusions

The major contribution of this paper is an accurate predictive model built on the data from a large-scale questionnaire survey which connects teachers' real

practice in evidence informed teaching to various external factors affecting whether they can and will engage in meaningful evidence use. To our knowledge, this is the first of its kind. In particular, interventions in three aspects — `Strategies`, `Conversation`, and `Encouragement` — are found to be most influential in promoting evidence informed teaching. Hopefully such findings can help teachers and other stakeholders close the loop between educational research from EDM etc. and educational outcomes.

The results outlined naturally come with a number of caveats in relation to how they should be interpreted. First, the 79 schools surveyed are all primary schools, correspondingly no relationship can be made between this analysis and England's 3200+ secondary schools. Second, it is likely that the schools involved are more predisposed to research engagement than the majority of England's primary schools: of the schools involved in the survey, 20 were in a formal Teaching School Alliance and a further 20 in a similar relationship (but had not applied or were in the process of applying to be Teaching School Alliance). Teaching School Alliances form a key driver of England's self-improving school system and there are clear expectations that they act as leaders in relation to evidence use. Nonetheless, our analysis does provide useful indicators as to where school leaders might focus their efforts should they wish to establish their school as one engaged with and in evidence.

An anticipated by-product of this paper is our advocacy of using the right nonparametric statistical methods for analysing and modelling Likert-type scale data that are prevalent in educational research.

Moreover, as Breiman has noted (Breiman, 2001b), there exist two cultures in the use of statistical modelling to reach conclusions from data. "One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown." This paper obviously lies in the latter category which in our opinion does provide a more accurate and informative alternative in comparison to the former category on small data sets like ours, especially when there is little known about the underlying stochastic mechanisms.

Regarding the future work, it is promising to adapt and apply other learning algorithms such as Factorization Machines (Rendle, 2012) which have been very successful in recommender systems for this data mining problem. In addition, we can of course build a predictive model for each of the four effect variables separately. However, because the four effect variables — dependent on the same input of cause variables — are themselves correlated (see Section 3), it is probably better to build a single joint model capable of predicting all the

four effect variables simultaneously through multi-task learning (Chapelle et al, 2010; Dumont et al, 2009). First, it could be more efficient as only one single model would need to be trained and used. Second, it could be more effective as the model's generalisation ability would be increased.

Acknowledgements We thank the fellow researchers who provided useful feedback on this paper at the ECDA-2015 conference.

References

Agresti A (2010) Analysis of Ordinal Categorical Data, 2nd edn. Wiley-Blackwell

Breiman L (2001a) Random forests. Machine Learning 45(1):5–32

Breiman L (2001b) Statistical modeling: The two cultures. Statistical Science 16(3):199–231

Breiman L, Friedman J, Stone C, Olshen R (1984) Classification and Regression Trees. Chapman & Hall/CRC, Belmont, CA, USA

Brown C (2014) Evidence-Informed Policy and Practice in Education: A Sociological Grounding. Bloomsbury Academic

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: ICML, Pittsburgh, PA, USA, pp 161–168

Chapelle O, Shivaswamy PK, Vadrevu S, Weinberger KQ, Zhang Y, Tseng BL (2010) Multi-task learning for boosting with application to web search ranking. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC, USA, pp 1189–1198

Cordingley P (2013) The contribution of research to teachers' professional learning and development. In: Research and Teacher Education: the BERA-RSA Inquiry, BERA, London, UK

Daly AJ (ed) (2010) Social Network Theory and Educational Change. Harvard Education Press

Dumont M, Marée R, Wehenkel L, Geurts P (2009) Fast multi-class image annotation with random subwindows and multiple output randomized trees. In: Proceedings of the 4th International Conference on Computer Vision Theory and Applications (VISAPP), Lisboa, Portugal, pp 196–203

Field A, Hole G (2003) How to Design and Report Experiments. Sage

Frank E, Hall MA (2001) A simple approach to ordinal classification. In: Proceedings of the 12th European Conference on Machine Learning (EMCL), Freiburg, Germany, pp 145–156

Friedman J (1999a) Greedy function approximation: A gradient boosting machine. Tech. rep., IMS 1999 Reitz Lecture

Friedman J (1999b) Stochastic gradient boosting. Tech. rep., Stanford University

Godfrey D (2014) Leadership of schools as research-led organisations in the english educational environment: Cultivating a research-engaged school culture. Educational Management Administration & Leadership p 1741143213508294

Goldacre B (2013) Building evidence into education. Tech. rep., Department for Education, UK

Greany T (2014) Are We Nearly There Yet?: Progress, Issues, and Possible Next Steps for a Self-Improving School System. IOE Press

Handscomb G, MacBeath J (2003) Professional development through teacher enquiry. Professional Development Today 7:6–12

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer

Jannach D, Zanker M, Felfernig A, Friedrich G (2010) Recommender Systems: An Introduction. Cambridge University Press

Kaptein MC, Nass C, Markopoulos P (2010) Powerful and consistent analysis of likert-type rating scales. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI), Atlanta, GA, USA, pp 2391–2394

Likert R (1932) A technique for the measurement of attitudes. Archives of Psychology 5:228–238

Mincu M (2013) Teacher quality and school improvement: What is the role of research? In: Research and Teacher Education: the BERA-RSA Inquiry, BERA, London, UK

Rendle S (2012) Factorization machines with libFM. ACM Transactions on Intelligent Systems and Technology (TIST) 3(3):57:1–57:22

Seni G, Elder JF (2010) Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan & Claypool Publishers