Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Benchmarking and Tuning Regression Algorithms on Predicting Students' Academic Performance

Owen H.T. Lu[a], Anna Y.Q. Huang[b] & Stephen. J.H. Yang[c*]
[abc]*Department of Computer Science and Information Engineering, National Central University, Taiwan*
*jhyang@csie.ncu.edu.tw

Abstract: With the adaption of online learning environment, students' learning behavior can be recorded as digital data. In order to implement the conceptual framework of learning analytics, many researchers applied machine learning methodologies and used data which collected from digital learning environment to predict students' academic performance for targeting at-risk population. However, along with the characteristic of machine learning methodologies, it presents diversity prediction performance due to the statistical property of educational data and these caused the difficulty to applied machine learning technology to classroom. In this study, we collected the state-of-the-art on regression algorithms and used an E-book-based learning dataset within 53 students for benchmarking the suitable algorithm for targeting at-risk students. In addition, we address the issues from learning environment, including over-concentration score, dropout students and data instance insufficiently, for improving prediction performance. The results revealed that the proposed performance tuning process could obtain optimal performance metrics and avoid over-fitting problem.

Keywords: Learning Analytics, At-risk Student Identification, Regression

1. Introduction

Learning analytics is a conceptual framework to help students to get higher achievement in classroom. In 2011, Horizon Report, a report of educational trends, defined learning analytics as a method based on educational data collection and information exploration that enables teachers to understand students' learning behavior and identify learning risk population at an early stage(L Johnson, Smith, Willis, Levine, & Haywood, 2012). Under the framework of learning analytics, teachers will improve teaching strategy, design of learning activities, design of pedagogic or even teaching material based on results of educational data mining. One of the advanced definition of learning analytics was mentioned by Horizon Report 2016, it defined that at-risk student can be identified at early stage of semester by applying machine learning technology and give timely intervention based on result of machine learning(Larry Johnson et al., 2016). In practice, many researchers have tried to prove the benefits of learning analytics, such as Lu, Huang, Huang, and Yang (2017) measured students' clickstream from a digital learning environment and intervened risk student according to the level of engagement. On the other hands, Van Leeuwen, Janssen, Erkens, and Brekelmans (2013) collected student's online discussion behavior and asked the assistant to participate in when the discussion was deviation. The result also shows that by using data analysis to decide when to intervene students' learning, which can effectively improve the learning outcome.

In order to use machine learning to identify at-risk students, the research field of learning analysis has begun to use students' final grades, scores and online learning behavior as the starting point to establish the at-risk student identification model. In instance, Hu, Lo, and Shih (2014) developed an early warning system by using a decision tree classifier. The model was constructed from data on 300 students and contained 13 online variables, including for how long each student had used the system and how many documents had been read by each student in the preceding week. The results revealed a 95% accuracy in predicting whether students would pass or fail based on 1–4 weeks of data from a skewed data set. Moreover, Romero, López, Luna, and Ventura (2013) collected data on 114 students from an online discussion forum and separated them into several data subsets on a weekly basis before evaluating each data set's predictive accuracy through several data-mining methods. Romero et al. (2013) used the sequential minimal optimization classification algorithm and student interaction data before a midterm exam to achieve the highest accuracy for predicting student learning performance.

On the other hands, another group of researchers used regression algorithm to try to accurately predict the student's final score. In instance, in order to reduction dimension for optimize the regression formula, Yang et al. (2018) developed a methodology which combined multiple linear regression and principle component analysis to predict students' final score, They prove that by adopting appropriate feature extraction and data pre-processing in a MOOCs and online assessment system enabled calculus course, student's final score can be predicted in one third of the semester, and the evaluator of root mean square error can reach about 12 (Lu et al., 2018), which means 88% of accuracy. Moreover, Huang and Fang (2013) used students' final grades as prediction targets. To evaluate the prediction results, the researchers designed two quantitative indicators to transfer the regression mean square error into prediction accuracy. The final results showed that the students' final exam scores were predictable to 88% accuracy based on eight variables collected from a learning management system. Previous studies have explained 4 that "at-risk" can generally be used to describe students who dropout, fail, or achieve low grades on courses.

The previous studies used data which collected from students' online learning behavior to train a classification or regression model for targeting at-risk students. The result shows around 80% of prediction accuracy on students' grade or final score. However, in the field of machine learning, it is necessary to consider the characteristics of regression and classification algorithms and also data statistical properties. The risk identification model need to consider as more metrics as possible during the training and evaluation process, for example, is the model over-fit? In this study, we will sort out the various situations which will be encountered in model training process, and expose prediction performance on several regression algorithms. Moreover, we will use the characteristics of the data and regression algorithm to try to optimize the model training performance by using an online learning dataset. The following research questions were proposed:

- **RQ1:** Benchmarking regression algorithms for predicting students' academic performance in E-Book-based Learning.
- **RQ2:** Tuning the regression algorithm to address (1) dropout students, (1) final score over-concentration and (3) prediction model over-fitting issues.

2. Literature Review

2.1 Design of Benchmarking Experimental

To identify at-risk student, the strategy we adopted was to use the regression model to predict the student's final score, and then to intervention students whose score are lower than expected, so we need to benchmark the performance of regression algorithms. In order to design the benchmark experimental, we first refer to Romero et al. (2013)'s article, they collected data from an online discussion forum and defined several features from the collected dataset. Several statistical methods were adopted to select features for the dimension reduction propose at incoming step. At the end, they selected 20 classification algorithms and two metrics to benchmark performance, and finally selected Expectation–maximizations as the best prediction algorithm.

However, Romero et al. (2013) started with the classification problem, and performance metrics could not be applied to benchmark regression algorithms. Therefore, we refer to Loterman, Brown, Martens, Mues, and Baesens (2012)'s article, they used metrics including: RMSE (root mean square error), R^2 (determination of coefficient), AUC (area under curve) and MAPE (mean absolute percentage error) to benchmark performance of regression algorithms on bank loss given problem. Finally, we were reexamined several studies in this field of educational data mining to define whether the range of metrics are acceptable, the RMSE in Lu et al.'s study revealed about 12 in the proposed of calculus course; in MAPE, the values of Huang and Fang (2013) and Lu et al.

(2018) fell between 0.82-0.90; The part of the R^2 in model which trained by a course which included 114 students in Çevik (2015)'s study reached 0.3.

2.2 Model Performance Tuning

There are many factors affect the optimized of machine learning, one of the major effect is feature extraction (as known as feature selection, attribute selection), which is a process of selecting a subset of relevant features. The aim of feature extraction is for simply the model, makes training time shorter, and also reduce over-fitting. In Romero et al. (2013)'s research, they applied 10 feature selection algorithms to rank the importance of the feature. The result demonstrated that classification accuracy can be improved by these feature selection process. Moreover, in Hall and Holmes (2003)'s research, they benchmarked sixteen feature selection and extraction algorithms. The result also shows feature extraction process is statistically significant improvement machine learning accuracy.

The other problem is data imbalanced, which means the number of data instance along with one class significantly outnumbering than others. If classification or regressions results are all easily align to that outnumbering class, the accuracy can be presented well even if the model was not optimized. This problem occurs in the context of educational data mining frequently because in most courses, most students pass exam, and low score tend to be only a minority in the group. In Thammasiri, Delen, Meesad, and Kasap (2014)'s research, they benchmarked sixteen resample algorithms on educational data and the result shows the prediction accuracy for the minority class can be improved. On the other hands, Chawla, Bowyer, Hall, and Kegelmeyer (2002) take a noise injection approach which named SMOTE(synthetic minority over-sampling technique), they demonstrated the SMOTE can improve not only prediction accuracy, but also solve the problem of data instance insufficient.

The last problem is model over-fitting, which means prediction model that corresponds too exactly to training data. The prediction accuracy or MSE will present low performance once if we applied non-training data into the over-fitting model. After we reviewed the above-mentioned multiple versions, especially in the field of educational data mining, which means that there may be no real appraisal of the risk prediction model to the real curriculum. Therefore, in this paper, we will add one more performance evaluator: training loss to measure if over-fitting happened on selected regression algorithms. Moreover, we will use the method of dropout or early-stop which proposed by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014) to avoid problem of over-fitting and expose actual prediction performance for each selected regression algorithms.

3. Methodology

3.1 Dataset characteristics

In this study, we selected an opened and de-identified data which collected from Kyoto University e-book learning system: **BookRoll** (Flanagan & Ogata, 2017; Ogata et al., 2017; Ogata et al., 2015), hereinafter referred to as **KEL** (Kyoto E-Book-based Learning) dataset.

We first plot each student's weekly activities and final score as a heatmap from clickstream, as shown in Figure 1, the learning activities of this course is very similar as Self-learning, most learning activities happened on the first week, such as student *ds121* and *ds122*. In addition, the level of engagement on **BookRoll** seems not related to students' final score, for example: *ds124* and *ds128* has actives only in the first week, but they get score of 100 points at the end of course, therefore this would increase the difficulty of model training.

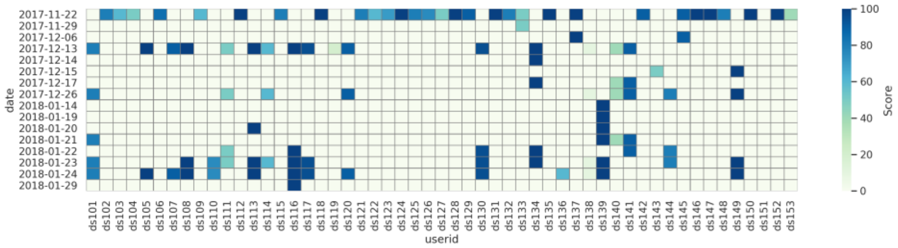

Figure 1. Students' Weekly Activity versus Final Score in **KEL** Dataset by Heatmap Plot

Since original data is a clickstream from **BookRoll**, we surveyed the relevance study from Yamada, Oi, and Konomi (2017) who discuss the correlation between E-Book learning behavior and students' learning outcomes, they defined 15 identical features. As shown in Figure 2, which is count distribution on 15 features from **KEL** dataset along with the definition from Yamada et al. (2017), we will use this figure to discuss how we reduce dimension in the following session. In addition, as shown in Figure 3, we plot students' score distribution in **KEL** dataset. From this figure, we can know that most students concentrate on 80-100 point, only a few students concentrate on less than 80 points. This is an imbalanced problem which we mentioned in session of literature review, therefore, we will discuss how to address this problem at the stage of data preprocess.

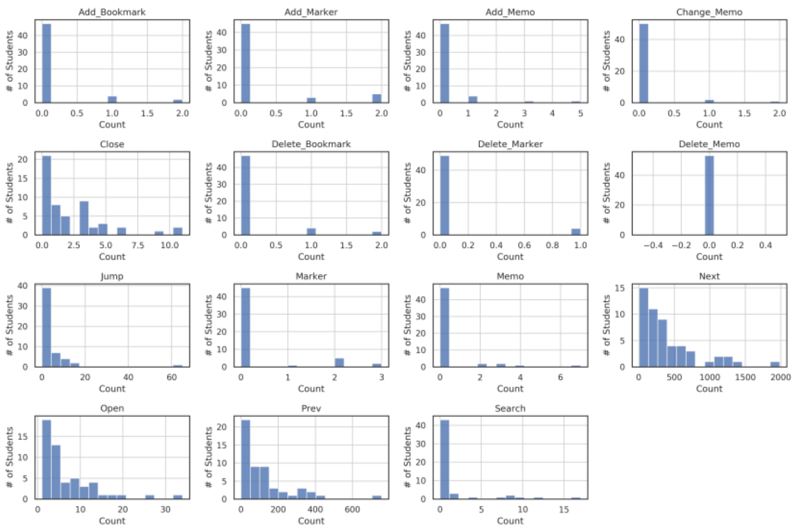

Figure 2. Distribution of Students' Learning Activities from **KEL** Dataset

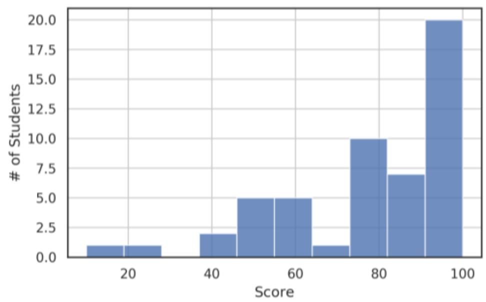

Figure 3. Distribution of Students' Academic Performance from **KEL** Dataset

3.2 Regression Algorithms and Evaluation Metrics

As listed in Table 1, we selected five common regression algorithms for the benchmark experimental, from the most basic MLR (Multiple Linear Regression), non-linear CART (Classification and Regression Tree), to the top algorithm on Kaggle: XGBoost (Extreme Gradient Boosting) and SVR (Support Vector Regression), and finally the most flexible ANN (Artificial Neural Network). Moreover, we refer to Loterman et al. (2012)'s work to list several performance metrics to evaluate which regression algorithm produce accurate predictions. There are five metrics listed in Table 2 and each of them has its own worst and best range to quantify the algorithm performance.

Table 1
Regression Models Employed for Benchmark Experimental

Regression Algorithms	Description
MLR	Multiple Linear Regression (Draper & Smith, 2014) is a linear algorithm which uses several explanatory variables to train a model for predicting target. The goal of training procedural is to produces a regression models that with the minimal difference between values observed in training set and values which predicted by regression model.
CART	Classification and Regression tree (Steinberg & Colla, 2009) uses a tree-like graph of decisions and possible outcomes, and has capability to take continuous values.
SVR	Support Vector Regression (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) project data point into a hyperplane and separate them with the maximum margin.
XGBoost	Extreme Gradient Boosting(Chen & Guestrin, 2016) uses sparsity-aware algorithm for sparse data and weighted for decision tree learning procedural.
ANN	Artificial neural networks (Bishop & Bishop, 1995) is based on a collection of connected units called artificial neurons. Each connection can transmit a signal between neurons, and signal can be processed when an artificial neuron received it. Moreover, signal can be add weight or basis after passing through the artificial neurons. ANN offers several flexible tuning methodologies such as Dropout (Srivastava et al., 2014) and Early-stop (Prechelt, 1998) to prevent over-fitting problem.

Table 2
Performance Metrics for Benchmarking Regression Algorithms

Metrics	Description	Worst	Best
RMSE	RMSE gives information about prediction error, which is the difference between value* observed in testing set and value which predicted by regression model. The metrics is presents by square root after average difference.	∞	0
R^2	R^2 gives information about how goodness of fit of the regression model, which is the difference between value observed in testing set and value which predicted by regression model. The metrics is presents by square of the difference.	$-\infty$	1
AUC	AUC is imply information about data imbalanced, which is area under ROC (Receiver Operator Characteristic), which show how the number of correct classified positive examples varies with incorrect classified negative examples. In general, AUC metrics is for classification algorithm, in this paper, we refer to (Bi & Bennett, 2003) to calculate AUC in regression algorithm.	0.5	1
MAPE	MAPE gives information about prediction error, which is the difference between value* observed and value which predicted by regression model. The metrics is present percentage after average difference.	∞	0
Training Loss	Training Loss is imply information about over-fitting (Srivastava et al., 2014), which is similar as RMSE but is in difference between value observed in training set and value which predicted by regression model.	Not equal to RMSE	Equal to RMSE

* Value here will be students' final score

3.3 Dataset processing and variable selection

Since we visualized **KEL** in previous session, we can clearly observe several obvious problems: First, data instance is insufficient; the number of students is 53, far less than the number of 15 features, these makes regression algorithm difficult to find the optimal solution of the equation. Therefore, we will remove the feature without any value: *Delete_Memo*, then extract features through PCA for the dimension reduction propose, and then inject the necessary noise by normal distribution for increase data instances propose. In order to solve the problems of Self-Learning and imbalanced scores, resample will be adopted for improving the performance of the regression algorithm.

4. Results and Discussion

4.1 Benchmarking Regression Algorithms

After the regression algorithms, dataset and performance metrics are defined, we will feed **KEL** into five regression algorithms for training prediction models. During the training process, 70% of the data instances will be randomly sampled each time, and the rest will be used as cross validation. In the first round of experimental, the validation results of the five regression algorithms are shown in Table 3.

First, we observed the RMSE obtained after the first round of verification, each regression algorithm applied default parameter. The MLR got the worst RMSE, which up to 80.69; it means that each prediction result will have an error of 80 points. The situation will be considered in the case of a student's score of 100, this RMSE cannot be accepted in real-life situations. On the contrary, we observe the RMSE of the remaining four regression algorithms, which is between 23.32 in the SVR and 38.62 in the ANN. Although it is better than the MLR, there is still a gap behind 12 points from pervious study. Therefore, as expected, the next round of validation will perform data preprocessing for increasing evaluation metrics propose.

Table 3

*Performance Metrics on **KEL** Dataset for each Regression Models*

	Algorithm	RMSE	AUC	R^2	MAPE	Training Loss
	MLR	80.69	0.36	-13.23	102.81	83.17
1st Round:	CART	31.30	0.82	-0.12	40.27	18.14
Default	SVR	23.32	0.89	0.02	32.89	23.85
Parameters	XGBoost	30.84	0.86	-0.25	40.92	11.07
	ANN	38.62	0.81	-0.50	43.44	24.21
	MLR	85.52	0.33	-13.08	101.21	80.95
2nd Round:	CART	23.44	0.89	-0.05	47.68	18.55
Noise	SVR	23.23	0.89	-0.03	48.82	22.27
Injection	XGBoost	25.45	0.86	-0.24	45.88	11.12
	ANN	28.69	0.86	-0.58	50.69	21.35
3rd Round:	MLR	85.56	0.33	-13.37	100.23	80.76
Noise	CART	23.44	0.88	-0.08	45.56	19.00
Injection +	SVR	22.43	0.89	0.01	45.60	22.44
Feature	XGBoost	24.26	0.88	-0.16	46.70	11.99
Extraction	ANN	23.57	0.88	-0.09	44.88	19.73

As we mentioned in the previous chapter, **KEL** dataset has 15 features with 53 data points, it is difficult for the other regression algorithm to optimal solution during training, therefore, we referred to Thammasiri et al. (2014)'s study to apply an oversample methodology by inject then same amount of noise meaning that each will multiply the point by a standard deviation of the normal distribution and then merge it into the original data set. As shown in Figure 4, we re-projected the original data by t-SNE to two dimensions, and the original loose data points were injected after the noise. It can be visually observed that the distribution format of data tends to be significant. As shown in Table 3, in the second round of testing after the injection of noise, most of the indicators have improved. Taking ANN as an example, the original RMSE improved from 38.62 to 28.69, and the AUC improved from 0.81 to 0.86, which can be attributed to the effects of noise injection.

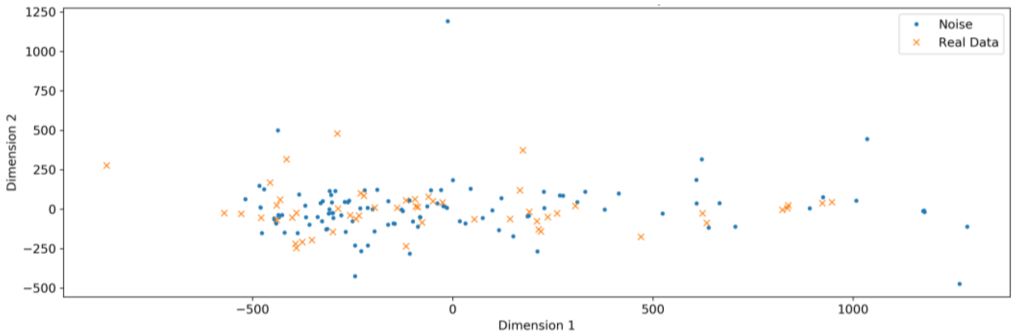

Figure 4. **KEL** Datasets Distribution after Noise Injection by t-SNE Plot

The third round of verification is to reduce the dimension of the data, so we take the **KEL** dataset which injected noise to PCA. The results are shown in Figure 5. We can obtain the variance explained in 99.8% of original 14 features by using 9 dimensions after PCA decomposition. Therefore, we use 9 dimensions for regression model training, and the verification results are also shown in Table 3. Observed from the metrics, after noise injection and feature extraction, the prediction performance of the SVR is still the best, which RMSE value is 22.43, and with the only positive R^2. The most improved algorithm is ANN, whose RMSE improved from 28.69 to 23.57, and AUC improved from 0.86 to 0.88.

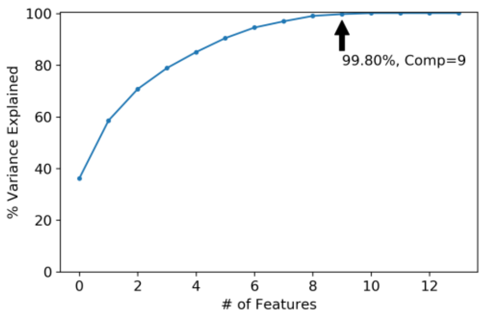

Figure 5. **KEL** PCA Decomposition Result

Finally, as show in Figure 6, we feed training data and testing data into model which trained by selected regression model, the result is projects with students’ final score as x-axis and prediction score as y-axis. First, observe the SVR and find that regardless of the input data, the output is always 80 points. Compared with the characteristics of **KEL**, the students' scores are almost distributed over 80 points. In the meantime, students who distributed under score of 80 are just a minority in the group; these would not affect performance metrics a lot. Therefore, we thought SVR is in a situation of overestimation. Moreover, CART and XGBoost have similar performance with SVR, however, we can observe from the Figure 6that the training data is almost accurately predicted, but there is no way to test the data, so we look back at Table 3 and we observed that two algorithms’ Training Loss is far below then RMSE in the three rounds of validation, so it can be concluded that the over-fitting problem has occurred in these two algorithms. Finally, expect for under-estimation of the MLR, we will use the ANN for the next stage.

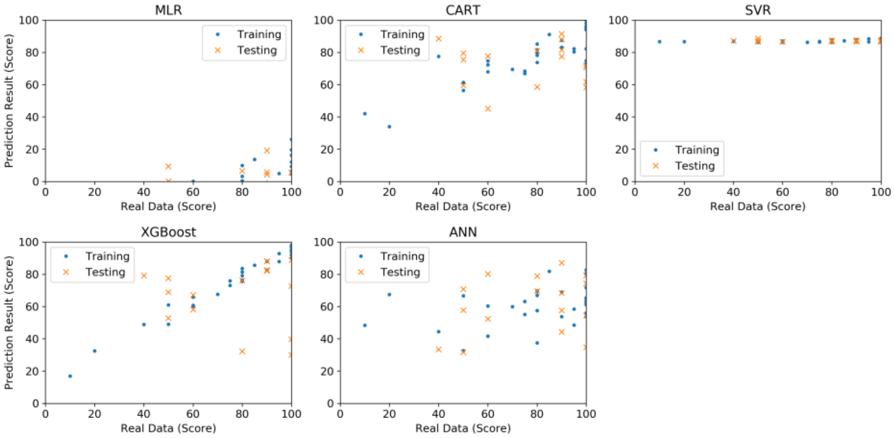

Figure 6. Prediction Students' Final Score by Different Regression Algorithm.

4.2 Tuning Algorithm based on Learning Activities and Characteristics of Algorithm

Continuing from the previous session, we noticed that ANN is the only one regression algorithm that does not underestimate or overestimate, but slightly over-fitting. Therefore, in this section, we extent the **KEL** visualization observation from session 3.1 and defining two issues that should be solved: (1) Over-concentration of students' score from teacher and (2) dropout phenomenon caused by assumption of self-learning strategy. In this session, we continue use the dataset with feature extraction and noise injection, and then perform the outlier detection and data resampling, and enable the dropout function of the neural network.

In the step of detection the outliers, we select directly from Figure 1 and remove students who only have activities in the first week, for example: *ds102* and *ds103*. On the other hand, we randomly sample and remove students whose scores are above 80 points. The data distribution after resample is shown in Figure 7. Comparing Figure 3, it can be observed that data has been shifted to the right has gradually aligned to the center.

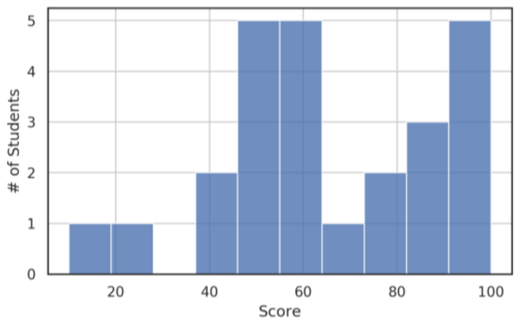

Figure 7. Distribution of Students' Academic Performance after Re-sample

Before us solving over-fitting problem, we injected the resampled data into the ANN again, and the results are shown in the Table 4 and Figure 8. The first metrics we are concerned about: RMSE is improved again from 23.57 to 19.00. On the other hands, although R^2 performs worthy, but it improves from the negative to positive. However, observing Figure 8, we can notice that with the increase of training epochs, the training loss gradually becomes lower and lower, but prediction loss (RMSE in Table 4) keep flat, the distance between them is getting farther and farther. This is another proof that our current training meets over-fitting problem.

Table 4
*Performance Metrics on **KEL** Dataset for each Regression Models*

	RMSE	AUC	R^2	MAPE	Training Loss

Remove outlier and resample	19.00	0.86	0.01	42.53	9.16
Enable dropout	20.00	0.88	0.01	27.84	17.66

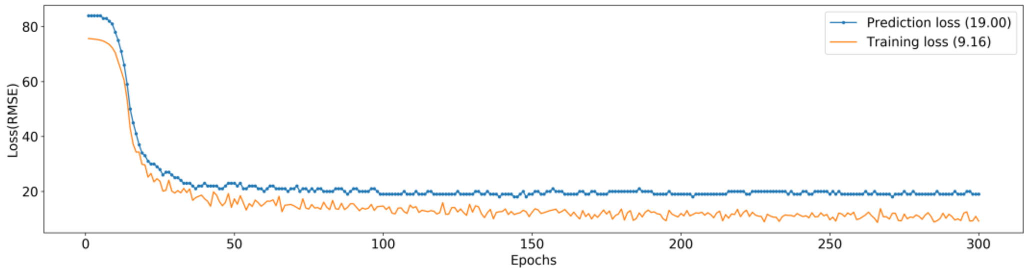

Figure 8. Training epochs after removing outlier and re-sampling

After removing the outliers and resampling, we enabled the Dropout parameters in the neural network. The final result is shown in Table 4 and Figure 9. It can be observed that training loss is close to prediction loss under the same number of training epochs with previous round, the distance between the two values is not farther and farther. At the end, we can claim to use the **KEL** dataset to train a student risk prediction model with an RMSE of 20.00, AUC of 0.88 and without over-fitting. However, comparing to our previous work at prediction students' final score in a blended Casuals course, we have obtained RMSE around 16.9 and similar MAPE, which means the prediction result on **KEL** dataset might have chance to improve continuously due to the parameter tuning or feature selection. The working items for improving prediction result will be addressed in the future works.

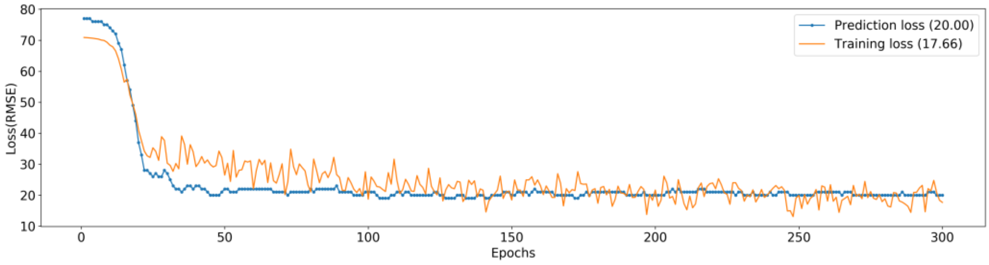

Figure 9. Training epochs after enabling Dropout

Conclusion

In this study, we used Kyoto University's E-book-based Learning dataset to benchmark the performance of various regression algorithms for targeting at-students, and also addressed several situations in educational area, which including: over-concentration of students' score, and dropout students in self-learning, the number of students insufficient, and continuously optimize the performance of the model in the process of solving the problem. In the future, there are still many parameters adjustment work that has not been completed during the experiment, such as the proportion of injected noise, degree of Dropout, methods of resample, and even the number of layers of the neural network. This will rely on the method of hyperparameter tuning to continue to deepen.

Acknowledgements

This work was supported by Ministry of Science and Technology, Taiwan under grants MOST-105-2511-S-008-003-MY3, and MOST-106-2511-S-008 -004 -MY3.

References

Bi, J., & Bennett, K. P. (2003). *Regression error characteristic curves.* Paper presented at the Proceedings of the 20th international conference on machine learning (ICML-03).

Bishop, C., & Bishop, C. M. (1995). *Neural networks for pattern recognition*: Oxford university press.

Çevik, Y. D. (2015). Predicting college students' online information searching strategies based on epistemological, motivational, decision-related, and demographic variables. *Computers & Education, 90*, 54-63.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (Vol. 326): John Wiley & Sons.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). *Support vector regression machines.* Paper presented at the Advances in neural information processing systems.

Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering, 15*(6), 1437-1447.

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in human behavior, 36*, 469-478.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education, 61*, 133-145.

Johnson, L., Becker, S. A., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC horizon report: 2016 higher education edition*: The New Media Consortium.

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2012). *NMC Horizon Report: 2011 Edition*. Retrieved from

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting, 28*(1), 161-170.

Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning. *Journal of Educational Technology & Society, 21*(2), 220-232.

Lu, O. H., Huang, J. C., Huang, A. Y., & Yang, S. J. (2017). Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course. *Interactive Learning Environments, 25*(2), 220-234.

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., . . . Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education *Smart Sensors at the IoT Frontier* (pp. 327-350): Springer.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). *E-Book-based learning analytics in university education.* Paper presented at the International Conference on Computer in Education (ICCE 2015).

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks, 11*(4), 761-767.

Romero, C., López, M.-I., Luna, J.-M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458-472.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929-1958.

Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. *The top ten algorithms in data mining, 9*, 179.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321-330.

Van Leeuwen, A., Janssen, J., Erkens, G., & Brekelmans, M. (2013). Teacher interventions in a synchronous, co-located CSCL setting: Analyzing focus, means, and temporality. *Computers in human behavior, 29*(4), 1377-1386.

Yamada, M., Oi, M., & Konomi, S. i. (2017). Are Learning Logs Related to Procrastination? From the Viewpoint of Self-Regulated Learning. *International Association for Development of the Information Society*.

Yang, S. J., Lu, O. H., Huang, A. Y., Huang, J. C., Ogata, H., & Lin, A. J. (2018). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. *Journal of Information Processing, 26*, 170-176.