Journal of Information Processing Vol.26 170–176 (Feb. 2018)

[DOI: 10.2197/ipsjjip.26.170]

Invited Paper

Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis

Stephen J.H. Yang[1,†1,a)] Owen H.T. Lu[1] Anna Y.Q. Huang[1] Jeff C.H. Huang[2]
Hiroaki Ogata[3] Albert J.Q. Lin[1]

Received: September 5, 2017, Accepted: November 24, 2017

Abstract: With the rise of big data analytics, learning analytics has become a major trend for improving the quality of education. Learning analytics is a methodology for helping students to succeed in the classroom; the principle is to predict student's academic performance at an early stage and thus provide them with timely assistance. Accordingly, this study used multiple linear regression (MLR), a popular method of predicting students' academic performance, to establish a prediction model. Moreover, we combined MLR with principal component analysis (PCA) to improve the predictive accuracy of the model. Traditional MLR has certain drawbacks; specifically, the coefficient of determination (R^2) and mean square error (MSE) measures and the quantile-quantile plot (Q-Q plot) technique cannot evaluate the predictive performance and accuracy of MLR. Therefore, we propose predictive MSE (pMSE) and predictive mean absolute percentage correction (pMAPC) measures for determining the predictive performance and accuracy of the regression model, respectively. Analysis results revealed that the proposed model for predicting students' academic performance could obtain optimal pMSE and pMAPC values by using six components obtained from PCA.

Keywords: learning analytics, multiple linear regression, principal component analysis

1. Introduction

In recent years, educators have applied learning analytics to improve the quality of teaching and learning. In Europe and the United States, the Horizon Report has annually investigated the benefits and methods of learning analytics since 2011. The Horizon Report: Edition 2011 proposed that the goal of learning analytics is to enable human tailoring of students' responses through adapting learning content and assisting at-risk students at the right time [1]. With the prominence of big data analytics, the Horizon Report: Edition 2016 proposed learning analytics to become the future trend in education [2]. Learning analytics is a process of measuring and analyzing learning data collected from learning environments [3], [4], [5]. Predicting students' learning performance is one of the main research topics in learning analytics. For example, Hu et al. collected data from 300 students and established a student risk prediction model. Experimental results revealed a 95% accuracy in predicting students' passing or failure rates based on 1–4 weeks of data [6]. Meier et al. designed a neighborhood selection process to predict students' grades. They claimed that the proposed algorithm achieved 76% accuracy [7]. After predicting the students' final performance, Purdue University designed and implemented "Course Signals" [8], an early warning solution to increase students' success by early risk identification. Moreover, learning analytics were combined with various strategies such as computer-supported collaborative learning (CSCL). For example, Van Leeuwen et al. developed a chat tool wherein the instructor can decide when to intervene in the group discussion based on the results of text emotion analysis [9]. Lu et al. developed a pair programming tool, wherein the instructors can provide timely intervention according to the engagement measurement results [10].

Several researchers have applied multiple linear regression (MLR) to predict students' learning performance [11], [12], [13], [14], [15], [16] or to identify at-risk students by predicting the course pass or fail [17], [18], [19]. With the rapid growth of information technology, the number of collected data variables from blended learning environments has also increased considerably. However, the number of used variables considerably affects the goodness of fit of the prediction model obtained using MLR. For predicting students' learning performance by using MLR, several researchers have reduced the number of variables through selecting some variables with higher predictive power [17], [18], [19]. Therefore, the aim of the current study was to investigate whether MLR is suitable for predicting students' academic performance by using a multivariable learning profile collected from the proposed blended calculus course.

The traditional measures used in MLR include mean square

[1] Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
[2] Department of Computer Science and Information Engineering, Hwa Hsia University of Technology, Xinbei, Taiwan
[3] Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[†1] Presently with Asia University, Taiwan
[a)] stephen.yang.ac@gmail.com

error (MSE), coefficient of determination (R^2), and quantile-quantile plot (Q-Q plot). These measures can only measure the goodness of fit of a regression model but cannot evaluate the prediction performance of MLR. In the field of education, it is difficult for teachers to determine whether the prediction results of MLR are credible through these measures. Therefore, it is necessary to define performance measures when using MLR to build a prediction model. Accordingly, this study focused on designing performance measures to measure the prediction performance of a regression model.

To reduce teachers' intervention, providing higher predictive accuracy to teachers is necessary. For predicting students' learning performance, several researchers have focused on how to improve the predictive accuracy. Therefore, this study investigated methods of improving the predictive accuracy of regression models, and attempted to answer the following research questions:

- **RQ1:** Is MLR suitable for predicting students' academic performance by using a multivariable learning profile collected from the proposed blended calculus course?
- **RQ2:** Is it possible to improve the predictive accuracy of the proposed MLR process?

2. Literature Review

MLR is a predictive analysis method based on the multivariate statistical technique, and it has been widely used in education. The number of variables has a considerable influence on the performance of the MLR process. The MSE and R^2 measures and the Q-Q plot technique have been used for evaluating the goodness of fit of regression models [20]. However, these measures cannot evaluate predictive performance of regression model. Therefore, providing more robust performance measures to teachers in the field of education is necessary.

For measuring prediction error, the mean absolute percentage error (MAPE) is calculated to measure the prediction error percentage of a prediction model [21], [22]. The MAPE is one of the most commonly used methods for evaluating prediction error. The lower the MAPE value is, the lower is the prediction error of the prediction model. Therefore, this study proposes the predictive mean absolute percentage correction (pMAPC) based on the calculus concept of MAPE.

On the basis of the covariance matrix of the data, principal component analysis (PCA) is typically used to determine uncorrelated eigenvectors through singular value decomposition and set the eigenvectors as the principal components of the data [23], [24]. The determined components can be used as a new variable set with higher discriminative power in a linear regression. Some researchers have proposed that predictive accuracy of MLR can be improved using PCA [25], [26], [27]. Therefore, this study combined PCA with MLR to improve the predictive accuracy. In addition, this study applied MAPC to measure the predictive accuracy of the regression model.

3. Blended Calculus Course

3.1 Participants

Fifty-eight university freshmen from Northern Taiwan participated in this study, which was conducted from September 2015 to February 2016. This experiment was conducted in a course named United Classes of Calculus. The participants comprised 33 male and 25 female students. Students learned calculus in the proposed blended learning course.

3.2 Learning Activities in Blended Calculus Course

To improve the quality of teaching and learning, the proposed blended calculus course combined an online learning environment and an online practice environment with classroom teaching of calculus. The learning activities of the proposed blended calculus course comprised previewing online learning materials, instructing calculus, practicing online exercises, practicing homework, and quizzes. The learning data in the proposed course were collected by recording students' clickstreams in the online course platform and online calculus practice environment. In addition to enriching the dataset, we collected the obtained learning grades in the quizzes and homework. The detailed information about the collected learning data is described in the next section.

The proposed blended calculus course was aimed at developing students' mathematics ability through the proposed learning activities. In this study, a Chinese version of Open edX [*1] was built to enable students to preview the online learning materials before class, after which the teacher instructed the subject in the class. To continue the learning behavior after class, the students performed online exercises and homework as part of their learning activities. A calculus online learning environment, namely Maple T.A. [*2], was built for students to engage in online calculus learning activity. Moreover, the teacher assigned homework to the students to continue the learning behavior after class. For measuring students' learning performance, the teacher administered a quiz to the students every 2 weeks. We collected learning data from the applied online learning environments to analyze the learning behavior. Furthermore, we built a model for students' academic performance prediction by combining PCA and MLR.

4. Methodology

4.1 Data Collection

In the proposed blended calculus course, we collected learning data from Open edX and Maple T.A. We collected the students' video-viewing behavior and exercise grades from Open edX and Maple T.A., respectively. To predict students' academic performance, we built a model for predicting students' final grades.

4.2 Datasets of Learning Activity and Learning Variables

For the datasets of learning activity, the students' learning data collected in this study comprised video-viewing behavior in Open edX, exercises in Maple T.A., homework completion, and the quiz grades. For guiding students to continue learning calculus, the instructor assigned paper-format homework exercises to the students every 2 weeks. To measure the learning performance for each learning topic, the instructor administered a quiz in the class every 2 weeks. The proposed blended calculus course lasted 18 weeks; that is, there were nine homework assignments and

[*1] https://open.edx.org/
[*2] https://www.maplesoft.com/products/mapleta/

```
{
   "username":"■■■■■■",
   "event_type":"pause_video",
   "ip":"123.110.40.112",
   "agent":"Mozilla/5.0",
   "host":"courses.openedu.tw",
   "session":"4c0801d5ce19ce4e9485bcf5ad647a7e",
   "event":"{
    \"id\":\"i4x-NKUHTx-TC101-video-\
        437045b9661a40609e4fff0a8ef0e24d\",
    \"currentTime\":923.549472,
    \"code\":\"html5\"}",
   "event_source":"browser",
   "context":{
      "user_id":14514,
      "org_id":"NKUHTx",
      "course_id":"NKUHTx/TC101/201511",
      "path":"/event"
   },
   "time":"2016-01-01T13:22:12.181487+00:00",
   "page":"https://courses.openedu.tw/courses/NKUHTx/TC101\
    /201511/courseware/4e5d487d59ac460890f71edbd37d7f1c\
    /b25a4f18519643a8b651a0c55af04ffa/"
}
```


Fig. 1 Example of tracking logs for a pause video in JSON format.

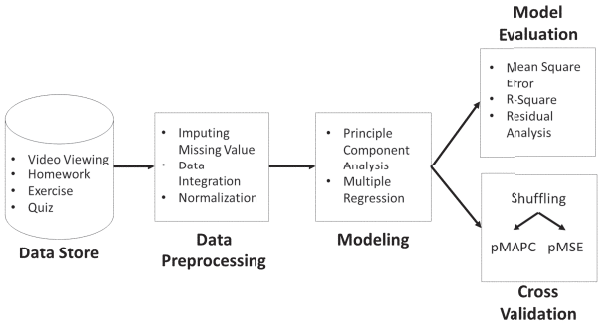


Fig. 2 Procedures involved in proposed student academic performance prediction model.

nine quizzes. For collecting learning data from homework and quizzes, we recorded the average grades obtained by the students in the homework assignments and quizzes. For collecting the exercise practice data, we recorded the average grade obtained in the online exercises in Maple T.A.

For collecting learning data from Open edX, we used Open edX Pipeline to retrieve the students' numerous learning actions from the tracking logs in Open edX. Open edX Pipeline is an open-source project that is fully integrated with analysis tools such as Apache HDFS, Jenkins, and MySQL. **Figure 1** shows an example of tracking logs in JSON format. This study applied 27 variables extracted from Open edX and Maple T.A., homework and quiz.

4.3 Procedures Involved in Student Academic Performance Prediction Model

The procedures involved in developing the proposed student academic performance prediction model entailed a data-preprocessing phase, modeling phase, and evaluation phase (**Fig. 2**). The data preprocessing phase involved missing value imputation, data integration, and data normalization. The modeling phase entailed the execution of PCA and MLR. Finally, the evaluation phase comprised model evaluation and cross validation.

4.3.1 Data Preprocessing Phase

The data preprocessing phase entailed extracting and transforming unstructured data to structured data to simplify the analysis. The data integration process focused on integrating the learning data derived from Open edX, Maple T.A., homework, and quizzes to generate the proposed 27 learning variables. The data normalization process was applied to redefine the range of data values in a smaller and specific range, because the range of various data values may be excessively wide. We normalized the range of the proposed 27 variables from 1 to 10.

4.3.2 Modeling Phase

The modeling phase entailed building the student academic performance prediction model. Accordingly, we combined MLR and PCA. First, PCA was applied to reduce the number of independent variables by extracting a new variable set from the original variable set. After performing PCA, MLR could be executed to build the student academic performance prediction model by using the factor scores of the components extracted through PCA.

4.3.3 Evaluation Phase

The evaluation phase involved measuring the performance of the proposed student academic performance prediction model. This phase involved model evaluation and cross validation. The model evaluation and cross-validation processes are described as follows:

- Model evaluation: By using MLR to build the students' academic performance prediction model, we could use the MSE and R^2 measures and the Q-Q plot technique to evaluate the goodness of fit of the regression model. A smaller MSE indicates a higher model goodness of fit. A closer R^2 value to 1.0 indicates a higher model goodness of fit.
- Cross evaluation: Cross validation is a model validation technology that combines the average measured values to derive the estimated value of prediction performance and accuracy of the prediction model. In cross validation, 10-fold cross validation with shuffling is performed to measure the prediction performance and accuracy of the prediction model. In 10-fold cross validation with shuffling, the original data are first shuffled, after which the original dataset is partitioned into 10 equal-sized subsets. Among the 10 subsets, 1 is selected as the testing set and the remaining 9 are selected as the training sets. The prediction regression model can be built using the training set. The prediction performance and accuracy of the prediction model can be calculated using the testing set. Each of the 10 subsets must be set exactly once as the test set. The average of the 10 results for the prediction model can be considered as the estimated value of the prediction performance and accuracy.

  The traditional MSE and R^2 measures and Q-Q plot technique cannot measure the prediction performance and accuracy of regression models. Therefore, we propose predictive MSE (pMSE) and predictive mean absolute percentage correction (pMAPC) for measuring the model prediction performance and accuracy, respectively. We applied 10-fold cross validation with shuffling to calculate the pMSE and pMAPC values. We modified the MSE and thus obtained pMSE to calculate the prediction performance by using the testing data in cross validation. MAPE was used to measure the prediction error percentage of the prediction model. By

modifying the MAPE, we derived the pMAPC measure to determine the accuracy of the prediction model. The definitions of pMSE and pMAPC are shown in Eqs. (1) and (2), respectively.

$$pMSE = \frac{1}{n_{test}} \sum_{n_{test}}^{i=1} (p_i - a_i)^2 , p_i \in p^{test}, a_i \in A \quad (1)$$

$$pMAPC = 1 - \frac{1}{n_{test}} \sum_{n_{test}}^{i=1} \left| \frac{p_i - a_i}{a_i} \right|, p_i \in p^{test}, a_i \in A \quad (2)$$

The set $A = \{a_1, a_2, ..., a_n\}$ comprises the actual academic grades of students. The symbol n_{test} indicates the number of data items in the test set. The set $p^{test} = \{p_1, p_2, ..., p_{n_{test}}\}$ comprises the predicted academic grades in the testing data. We can calculate the pMSE and pMAPC values using Eqs. (1) and (2). A lower pMSE value indicates higher predictive performance. Moreover, a higher pMAPC value indicates higher model accuracy.

5. Results and Discussion

To obtain the best explanatory power of the regression model, we extracted principal components from the original data after the data preprocessing step in the proposed student academic performance prediction model. In the scree plot shown in **Fig. 3**, each bar of the bar chart and each point of the line chart represent the explanatory power of each component and the accumulated explanatory power, respectively. The explanatory power of the first component for the regression model was higher than 81%. By contrast, the accumulated explanatory power levels of six components were higher than 96%, and the predictive performance of the regression model for each component will be discussed in Section 5.2.

5.1 MLR Model Evaluation of Goodness of Fit

In general, MLR is used to predict the value of dependent variables according to historical information. For selecting independent variables, the causal relationship between independent and dependent variables must be considered. For evaluating MLR, traditional measures such as R^2 and MSE are used to examine the goodness of fit of regression models. In this study, the goodness of fit of the regression model was first examined using the traditional measures. However, these traditional measures cannot evaluate the predictive performance of regression models. Consequently, teachers cannot obtain predictive accuracy by using traditional measures in the actual teaching environment. Therefore, we propose additional measures for determining the predictive performance of regression models, thus enabling teachers to evaluate predictive accuracy.

The MSE measure is used to evaluate how close a prediction regression line is to a set of actual values of dependent variable. This measure is used to calculate error variance by using the residual sum of squares divided by the number of predicted data. In a regression model, the residual is defined as the predictive value of data minus the actual value of the data. A lower MSE value indicates a higher model goodness of fit. In the proposed student academic performance prediction model, the academic score in calculus serves as the predicted dependent variable. Therefore, in this study, PCA was first performed, followed by MLR using the extracted principal components. The ranges of the predicted dependent variables and MSE were 0–100 and 0–10,000, respectively. The MSE values obtained in the regression model after PCA are presented in **Fig. 4**, indicating the MSE range to be 108.27–248.1. This thus implies that the range of the predictive error for each student was 10.4–15.8. Figure 4 shows that when the number of applied principal components is 4, the value of MSE is dramatically decreased to 152.83, after that, the value of MSE is continue decreasing incrementally.

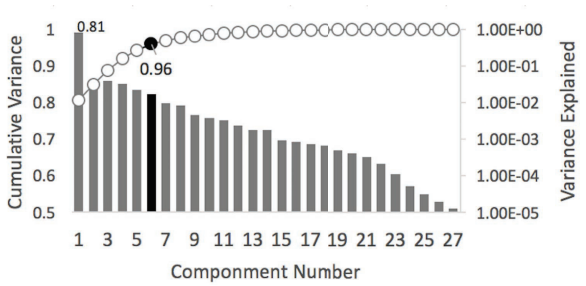


Fig. 3 The explanatory power and accumulated explanatory power values for each component.

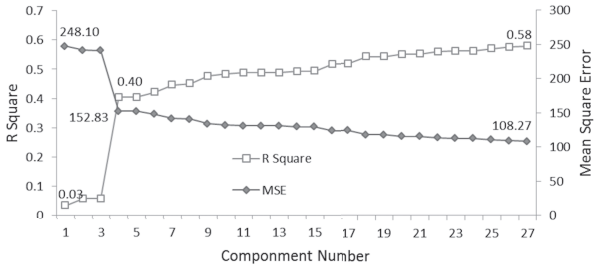
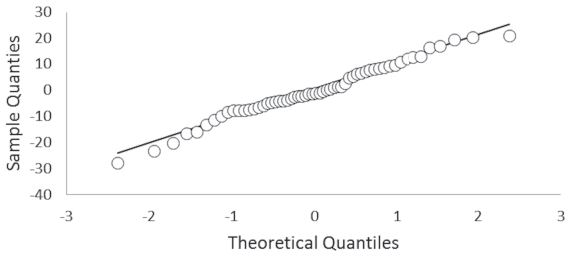


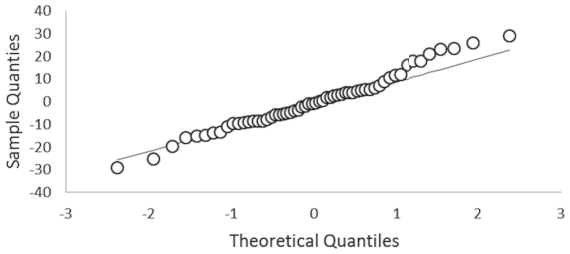
Fig. 4 MSE value for each component in the regression model after PCA.

R^2 is used to measure the explanatory power of a regression model by using variances of percentage between the independent and dependent variables. Moreover, R^2 is one of the goodness-of-fit measures for a regression model. A higher R^2 value indicates a higher explanatory power for the regression model. The R^2 values for MLR performed in this study by using the components extracted through PCA are presented in Fig. 4. According to the first principal component in Fig. 3, the value of the explanatory power for first component of the proposed regression model was 0.81, and this can be attributed to some missing information from the original data because of PCA. In addition, for the first component, the R^2 value was only 0.03 (Fig. 4). The explanatory power of the regression model increased when more components were applied. Subsequently, the R^2 value increased gradually, from 0.40 for 4 components to 0.58 for 27 components. According to the increasing trend of R^2 value after 4 components, the number of applied components should be more than 4.

In addition to examining the goodness of fit of a regression model by using R^2 and MSE, the distribution of residuals for the regression model must be examined to determine whether the hypothesis of normal distribution is supported. In residual analysis, the Q-Q plot is the most widely used technique to verify this hypothesis. The results of residual analysis presented using the Q-Q plot are shown in **Fig. 5**. As indicated in Figs. 5 (a) and (b), the



(a) Result of residual analysis for the MLR without PCA



(b) Result of residual analysis for the MLR with the six components extracted from PCA

Fig. 5 Results of residual analysis by using Q-Q plot.

distributions of residuals for the regression model involving MLR without and with PCA are both similar to a straight line. The p value of the test for the regression model involving MLR without PCA was 0.35 and that for the regression model involving MLR with PCA was 0.23. Thus, both the aforementioned results support the hypothesis of normal distribution.

To address **RQ1**, according to the described results of the goodness-of-fit test and residual analysis of the regression model, the results of these measures obtained using MLR with and without PCA are satisfactory. However, the explanatory power of the regression model determined by using MLR with PCA was lower than that determined by using MLR without PCA. Through the use of MLR with PCA, the values of measures such as MSE and R^2 can be accepted in the field of education. For example, the MSE value determined using MLR without PCA was 10.23 for each student, and the MSE value determined using MLR with PCA increased to 12.33 for each student, thus indicating that the gap between the method of MLR without and with PCA is 1.9. This gap is acceptable when the range of student scores from 0 to 100 is considered.

5.2 Improving MLR Predictive Accuracy Using PCA

The R^2 and MSE measures can evaluate only the goodness of fit of a regression model but cannot evaluate the accuracy of the model. However, in practice, teachers need to know the performance accuracy in order to reduce the risk of wasting resources through incorrect interventions. Therefore, this study introduced 10-fold cross validation with shuffling to partition the original dataset into a training dataset and testing dataset. The shuffling mechanism enables overcoming the problem of higher residual errors influenced by outlier data caused by a single round of 10-fold cross validation. Moreover, we applied the pMAPC measure to measure the accuracy of the regression model. According to **Fig. 6**, the pMSE values in the first 4 rounds dropped from 503.4

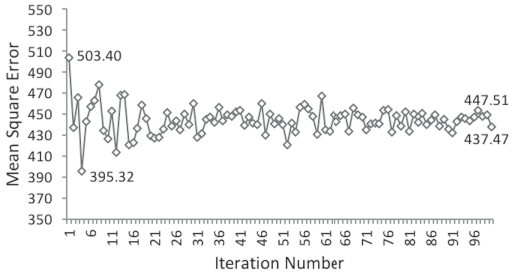


Fig. 6 Influence of shuffling times on the value of pMSE for MLR without PCA.

Table 1 Comparison of pMSE and pMAPC between MLR without PCA and MLR with PCA (comp = 6).

	pMSE	pMAPC
MLR	455.87	0.71
MLR+PCA(comp=6)	198.62	0.81
p	<0.05	<0.05

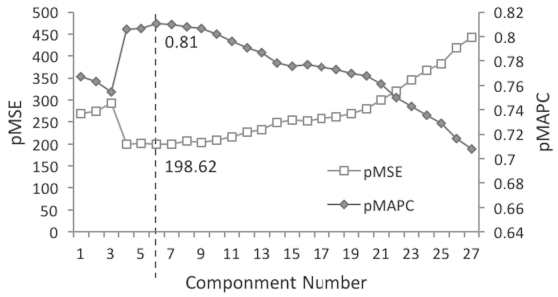


Fig. 7 Value of pMSE and pMAPC for each component after performing PCA.

to 395.32. The maximal difference among the first 4 rounds was as high as 27%; this was engendered by the outliers in the training or testing datasets. To reduce the difference among the rounds, we shuffled all data items after each round of 10-fold cross validation. After 100 rounds of shuffling, the average pMSE value could be considered as the pMSE value for the regression model. The difference range of pMSE for the latter rounds was 447.51–437.47, signifying that the difference range could be effectively reduced to 2%.

The pMSE values for the regression model after cross validation are presented in **Table 1**. The pMSE and pMAPC values for the regression model determined by using MLR without PCA were 455.87 and 0.71, respectively. In the field of education, the predictive error of students' academic performance was close to 21, according to the pMSE value. Moreover, the academic scores of 3 out of 10 students were inaccurately predicted.

According to the pMAPC equation in Section 4.3.3, we can easily quantify the correct prediction rate, thus facilitating prediction tasks in the real field of education. Incorrect prediction of at-risk students not only increases the teaching costs after school but also considerably influences the students' psychologically.

For the regression model involving MLR with PCA, the pMSE and pMAPC values for each component are presented in **Fig. 7**, revealing that the optimal pMSE and pMAPC values could be obtained by using six components. The optimal pMSE and pMAPC values were 198.62 and 0.81, respectively. Therefore, the predictive error for each student's academic score was close to 14; furthermore, the academic scores of 8 out of 10 students were

accurately predicted.

To address **RQ2**, a t-test was performed to examine the difference in the values of the predictive performance measures between MLR with and without PCA. The p values were less than 0.05 for pMSE and pMAPC (Table 1). Therefore, the predictive performance of MLR can be improved considerably by using six components of PCA. This result indicates that the original dataset had two properties: first, strong correlations existed among the independent variables. Therefore, the predictive performance could be improved using PCA, and six components could be used to obtain the best predictive performance. Second, the outliers had influenced the seventh to the twenty-seventh components. This is thus the main reason for the optimal pMSE and pMAPC values of the regression model obtained using six components.

5.3 Limitation

In this study, we proposed a methodology to establish a model for predicting students' academic performance. The model was built from a dataset collected from Open edX and Maple T.A. Moreover, we designed an 18-week learning activity that included homework, quizzes, and video-based learning, and was integrated with the aforementioned learning environment. In particular, the prediction model is associated with this learning activity and these particular data attributes; thus, the model is not applicable to other courses with different learning activities and data attributes.

6. Conclusion

The aim of learning analytics is to improve learning performance by predicting at-risk students and providing them with the necessary intervention. With the increasing complexity of the learning environment and diversity of available learning tools, traditional prediction methods have some limitations. In this study, we collected learning data from video-viewing, exercises, quizzes, and homework in a blended calculus course to predict student performance. First, we investigated whether MLR is suitable for building a model for predicting students' academic scores. Subsequently, we combined PCA and MLR to improve the predictive accuracy of the model.

According to goodness-of-fit and residual analysis results for the established regression model, MLR is suitable for building a student academic performance prediction model for the blended calculus course with many variables. For providing the predictive performance of teachers, we also propose the pMSE and pMAPC measures by applying cross validation. According the analysis results, the predictive performance of MLR with PCA was higher than that of MLR without PCA. In the future, the original dataset will be separated by middle exam to predict at-risk students at an early stage. To validate the predicted students' academic performance, we also intend to provide the predicted information of the at-risk students to the university.

Acknowledgments This work is supported by Ministry of Science and Technology, Taiwan under grants MOST-104-2511-S-008-006-MY2, MOST-105-2511-S-008-003-MY3, MOST-106-2511-S-008 -004 -MY3, MOST-105-2622-S-008 -002-CC2.

References

[1] Consortium, N.M. et al.: The 2011 horizon report (2011).
[2] Becker, S.A., Cummins, M., Davis, A., Freeman, A., Giesinger, C.H., and Ananthanarayanan, V.: NMC Horizon Report: 2017 Higher Education Edition, *The New Media Consortium* (2017).
[3] Baker, R.S. and Inventado, P.S.: Educational data mining and learning analytics, *Learning analytics*, pp.61–75, Springer (2014).
[4] Papamitsiou, Z. and Economides, A.A.: Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence, *Journal of Educational Technology & Society*, Vol.17, No.4, p.49 (2014).
[5] Peña-Ayala, A.: *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*, Vol.94, Springer (2017).
[6] Hu, Y.-H., Lo, C.-L. and Shih, S.-P.: Developing early warning systems to predict students' online learning performance, *Computers in Human Behavior*, Vol.36, pp.469–478 (2014).
[7] Meier, Y., Xu, J., Atan, O. and van der Schaar, M.: Predicting grades, *IEEE Trans. Signal Processing*, Vol.64, No.4, pp.959–972 (2016).
[8] Arnold, K.E. and Pistilli, M.D.: Course signals at purdue: Using learning analytics to increase student success, *Proc. 2nd International Conference on Learning Analytics and Knowledge*, pp.267–270, ACM (2012).
[9] Van Leeuwen, A., Janssen, J., Erkens, G. and Brekelmans, M.: Supporting teachers in guiding collaborating students: Effects of learning analytics in cscl, *Computers & Education*, Vol.79, pp.28–39 (2014).
[10] Lu, O.H., Huang, J.C., Huang, A.Y. and Yang, S.J.: Applying learning analytics for improving students engagement and learning outcomes in an moocs enabled collaborative programming course, *Interactive Learning Environments*, Vol.25, No.2, pp.220–234 (2017).
[11] Huang, S. and Fang, N.: Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, *Computers & Education*, Vol.61, pp.133–145 (2013).
[12] Tempelaar, D.T., Rienties, B. and Giesbers, B.: In search for the most informative data for feedback generation: Learning analytics in a data-rich context, *Computers in Human Behavior*, Vol.47, pp.157–167 (2015).
[13] Zacharis, N.Z.: A multivariate approach to predicting student outcomes in web-enabled blended learning courses, *The Internet and Higher Education*, Vol.27, pp.44–53 (2015).
[14] Morris, L.V., Finnegan, C. and Wu, S.-S.: Tracking student behavior, persistence, and achievement in online courses, *The Internet and Higher Education*, Vol.8, No.3, pp.221–231 (2005).
[15] Sorour, S.E., Mine, T., Goda, K. and Hirokawa, S.: A predictive model to evaluate student performance, *Journal of Information Processing*, Vol.23, No.2, pp.192–201 (2015).
[16] Yoo, J. and Kim, J.: Predicting learner's project performance with dialogue features in online q&a discussions, *International Conference on Intelligent Tutoring Systems*, pp.570–575, Springer (2012).
[17] Marbouti, F., Diefes-Dux, H.A. and Madhavan, K.: Models for early prediction of at-risk students in a course using standards-based grading, *Computers & Education*, Vol.103, pp.1–15 (2016).
[18] Macfadyen, L.P. and Dawson, S.: Mining lms data to develop an "early warning system" for educators: A proof of concept, *Computers & education*, Vol.54, No.2, pp.588–599 (2010).
[19] Agudo-Peregrina, Á.F., Iglesias-Pradas, S., Conde-González, M.Á. and Hernández-García, Á.: Can we predict success from log data in vles? classification of interactions for learning analytics and their relation with performance in vle-supported f2f and online learning, *Computers in Human Behavior*, Vol.31, pp.542–550 (2014).
[20] Taneja, A. and Chauhan, R.: A performance study of data mining techniques: Multiple linear regression vs. factor analysis, arXiv preprint arXiv:1108.5592 (2011).
[21] O'Connell, R.T. and Koehler, A.B.: *Forecasting, time series, and regression: An applied approach*, Vol.4, South-Western Pub (2005).
[22] Hyndman, R.J. and Koehler, A.B.: Another look at measures of forecast accuracy, *International Journal of Forecasting*, Vol.22, No.4, pp.679–688 (2006).
[23] Jolliffe, I.T.: Principal component analysis and factor analysis, *Principal component analysis*, pp.115–128, Springer (1986).
[24] Hira, Z.M. and Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data, *Advances in Bioinformatics*, Vol.2015 (2015).
[25] Ul-Saufie, A., Yahya, A. and Ramli, N.: Improving multiple linear regression model using principal component analysis for predicting PM_{10} concentration in seberang prai, pulau pinang, *International Journal of Environmental Sciences*, Vol.2, No.2, p.403 (2011).
[26] Qiuhua, L., Lihai, S., Tingjing, G., Lei, Z., Teng, O., Guojia, H., Chuan, C. and Cunxiong, L.: Use of principal component scores in

multiple linear regression models for simulation of chlorophyll-a and phytoplankton abundance at a karst deep reservoir, southwest of china, *Acta Ecologica Sinica*, Vol.34, No.1, pp.72–78 (2014).
[27] Pires, J., Martins, F., Sousa, S., Alvim-Ferraz, M. and Pereira, M.: Selection and validation of parameters in multiple linear and principal component regressions, *Environmental Modelling & Software*, Vol.23, No.1, pp.50–55 (2008).
[28] Goossens, M., Mittelbach, F. and Samarin, A.: *The LaTeX Companion*, Addison Wesley (1993).
[29] Lamport, L.: *A Document Preparation System LaTeX User's Guide & Reference Manual*, Addison Wesley (1986).

Stephen J.H. Yang is now the Vice President of Asia University, Taiwan. He is also associated with the National Central University as the Distinguished Professor of Department of Computer Science & Information Engineering. Dr. Yang was the Director of Department of Information and Technology Education, Ministry of Education, Taiwan (2013–2014), during the two years of service in Taiwan government, Dr. Yang was responsible for the national information & technology education, he also launched Taiwan's national digital learning initiative which includes the construction of 100 G Taiwan Academic Network for national network infrastructure, the construction of Education Cloud for national data infrastructure, and innovation programs such as Taiwan MOOCs and mobile learning. Dr. Yang also served as the Convener of Information Education Discipline, Ministry of Science & Technology. Dr. Yang received his Ph.D. degree in Electrical Engineering & Computer Science from the University of Illinois at Chicago in 1995. Dr. Yang has published over 70 SSCI/SCI journal papers, his research interests include Big Data, learning analytics, Artificial Intelligence in education, educational data mining, and MOOCs. As shown on Google Scholar, Dr. Yang's publication citation indices has been over 8,800, especially on the main research themes, Education data mining is ranked #3, MOOCs is ranked #3, Artificial Intelligence in education is ranked #7, Learning analytics is ranked #8. Dr. Yang received the Outstanding Research Award from Ministry of Science & Technology (2010) and Distinguished Service Medal from Ministry of Education (2015). Dr. Yang is currently the Co-Editors-in-Chief of the International Journal of Knowledge Management & e-Learning.

Owen H.T. Lu is a Student of Computer Science & Information Engineering, National Central University, and also the Section Manager of Smart System Institute, Institute for Information Industry, Taiwan. Mr. Lu received his M.S. degree in Department of Electronics and Communications Engineering and National Chung Hsing University at Taiwan in 2009. His research interests include Cloud Computing, Big Data Technology, Data Security and Learning Analytics.

Anna Yu-Qing Huang is a Postdoctoral Researcher of Computer Science & Information Engineering, National Central University, Taiwan. Dr. Huang received her Ph.D. degree in Institute of Engineering Science and Technology from National Kaohsiung First University of Science and Technology at Taiwan in 2011. Her research interests include Big Data technology, Learning Analytics, mobile learning, Massive Open Online Courses (MOOCs), Computer Supported Collaborative Learning (CSCL).

Jeff Cheng-Hsu. Huang is an Associate Professor of Computer Science & Information Engineering, Hwa Hsia University of Technology, Taiwan. Dr. Huang received his Ph.D. degree in Computer Science & Information Engineering from National Central University at Taiwan in 2009. His research interests include e-learning, mobile learning, social networking, social computing, 3D virtual world, creative design, Computer Supported Collaborative Learning (CSCL), Big Data technology, Learning Analytics.

Hiroaki Ogata is a Professor at Learning and Educational Technologies Research Unit, the Academic Center for Computing and Media Studies, and the Graduate School of Informatics at Kyoto University, Japan, and also, a honorary professor at the Education University of Hong Kong, and a Visiting Chair Professor of Asia University in Taiwan as well.

Albert J.Q. Lin is an Engineer of Taiwan Semiconductor Manufacturing Company. Mr. Lin received his M.S. degree in Computer Science & Information Engineering from National Central University at Taiwan in 2017. His research interests include Big Data Technology and Learning Analytics.