Statistical Analysis of Several Factors in Predicting Student Performance

Yuyao Jiang*
School of Mathematics and Applied Mathematics, University of Reading, Reading, United Kingdom
*Corresponding author's e-mail: fg803080@student.reading.ac.uk

Abstract

Education has always been a key factor in a country's development. Improving student performance is a common goal of students, parents and teachers. Studying the factors that influence student performance can help students focus on their weak points to improve their final grades more effectively. In this paper, Random Forest Algorithm is used to extract the four most important independent variables, which are second-period grade (G2), first-period grade (G1), number of school absences (absences) and number of past class failures (failures) from a data set on student performance. Then a multiple linear regression model is then established to study the relationship between the dependent variable final grade (G3) and them. After the evaluation of the model fitting accuracy and residual test, a linear model ($y = -1.76483 + 0.97847\,X_1 + 0.14374\,X_2 + 0.03759\,X_3 - 0.25720\,X_4$) is built. It simplifies the model and can predict student performance with great accuracy.

Keywords-Multiple Linear Regression, Random Forest, K-Fold Cross-Validation

1. Introduction

Nowadays, student performance is so crucial that it can affect the next stage of students' education, like graduating or entering an ideal school. In some countries, the graduation rate of undergraduates is just over half, with a high dropout rate. It is also the main concern of their teachers and parents, who hope that students can learn rich knowledge and receive a good education in school so that they are more likely to find good jobs and better adapt to society. It tends to be late, and nothing can be changed when students get unsatisfactory final grades. So, predicting student performance in advance is necessary. It allows students to adjust themselves and teachers to guide students better.

Various machine learning algorithms, including classification techniques (e.g., Decision Tree, K-Nearest Neighbor, Naïve Bayes algorithm) and regression algorithms (e.g., Linear Regression, Support Vector Machine), have been used to predict student performance. In 2013, authors in [1] associated decision tree algorithm with data mining techniques. J48 decision tree algorithm was found to be the best suitable algorithm to construct the model. The cross-validation method and percentage split method were used to evaluate the efficiency of different algorithms. In [2] published in 2015, a model based on the Multilayer Perceptron Topology was developed and trained using data spanning five generations of graduates from the Al-Azhar University, Gaza. Test data evaluation showed that the ANN model could correctly predict the performance of more than 80% of prospective students. In 2016, Personalized Multi-regression and Matrix Factorization approach based on recommender systems was used in [3] to accurately forecast students' grades in future courses and in-class assessments. And they found that using only historical grade information coupled with available additional information such as transcript data, both linear multi-regression (PLMR) and advanced matrix factorization (MF) techniques can predict next-term grades with lower error rates than traditional methods. In 2017, both Support Vector Machine and K-Nearest Neighbor algorithms were applied to the data set, and their accuracy was compared. Authors found that the Support Vector Machine achieved slightly better results with a correlation coefficient of 0.96, while the K-Nearest Neighbor achieved a correlation coefficient of 0.95 in [4]. Naïve Bayes algorithm was also widely used by researchers to predict students' performance in [5] in 2017. The classification approach, a Naïve Bayesian classifier, was used to predict the GPA of the graduate student. It was a simple probabilistic classifier founded on Bayes theorem by naïve impartiality assumptions and was trained extremely expeditiously in supervised education. Students were clustered into collections using the K-Means clustering algorithm. In [6] published in 2017, authors observed that the linear regression-based model was well suited as it predicted the future value rather than a class label. The model was univariate, i.e., it took only one variable, but it can be extended as a multivariate model by adding more parameters to get more accurate results. Support Vector Machine was used as a supervised learning model and was a strong classifier that can identify two classes (i.e., training and test data) for prediction. In [7], by utilizing Semantic rules and SVM algorithm in 2017, authors analyzed the

International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), edited by Hong-Ming Yin, Ke Chen, Romeo Meštrović, Teresa A. Oliveira, Nan Lin, Proc. of SPIE
Vol. 12163, 121630D · © 2022 SPIE · 0277-786X · doi: 10.1117/12.2628062

students' learning and predicted their performance by conducting various tests. Artificial Neural Network is another that researchers use to predict students' performance.

Firstly, the data set about student performance used in this research paper is described in detail. The chosen data set is selected from UCI Repository. Then the strategy is discussed on how the four most important attributes are selected as independent variables by Random Forest and 10-fold Cross-Validation and the process of studying the relationship between them and student performance. After that, the statistical analysis is conducted through Multiple Linear Regression Algorithm. Finally, the model is tested from four aspects as well as the accuracy of the model is proved.

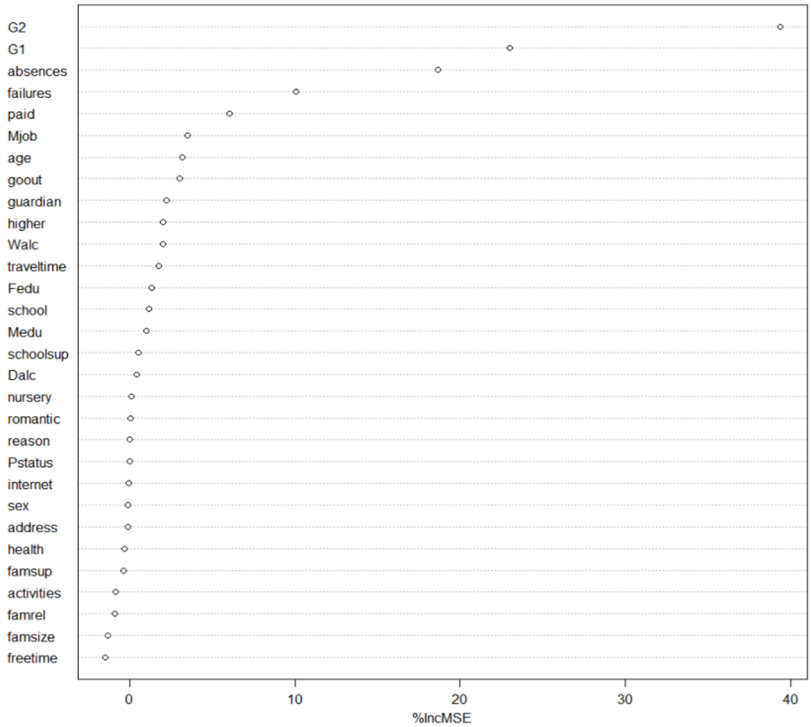
2. Material and Research Method

2.1 Data Source

The UCI Machine Learning Repository was founded in 1987 by David Aha and his colleagues at UC Irvine. It features complete data and reasonable classification. So far, the archive has been used as the primary source for researchers' empirical analysis of machine learning algorithms and has been cited more than 1000 times. In our study, a data set called Student Performance [8] is collected from the UCI Repository. Paulo Cortez collected the data from two public schools in Portugal between 2005 and 2006. School mark reports (i.e., number of absences and three-period grades) and questionnaires are two main sources of the database. Questionnaires are some additional attributes that may affect students' performance related to demographic (e.g., family income, quality of family relationships), social (e.g., going out with friends) and school (e.g., weekly study time) related features. It was reviewed by school professionals and tested on a small set of 15 students to get feedback at first, the formal ones contained 37 questions in a single A4 sheet, and 788 students in the class answered it. One hundred eleven answers were discarded due to a lack of identification details (necessary for merging with the school reports). Finally, the data was integrated into a Student Performance data set with 395 examples and 33 attributes.

2.2 Research Variables Selection

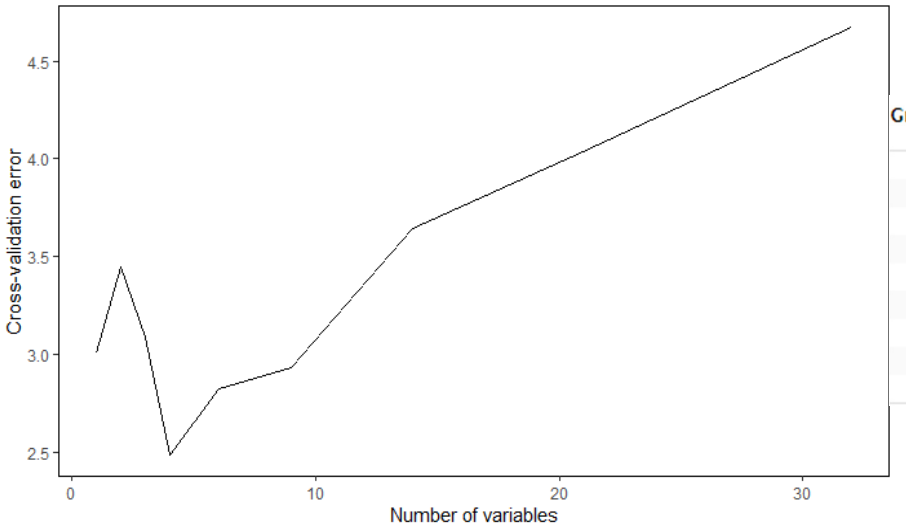
The final grade (G3) is chosen as the dependent variable. The relationship between other variables and the final grade (G3) is studied. Considering that the number of variables in the data set which is 32 is too large, it is necessary to select the most important ones.



	%IncMSE <dbl>
G2	39.395021
G1	22.991368
absences	18.663228
failures	10.087547
Fjob	-1.585402
age	3.189720

Figure 1: Top 30 variables importance

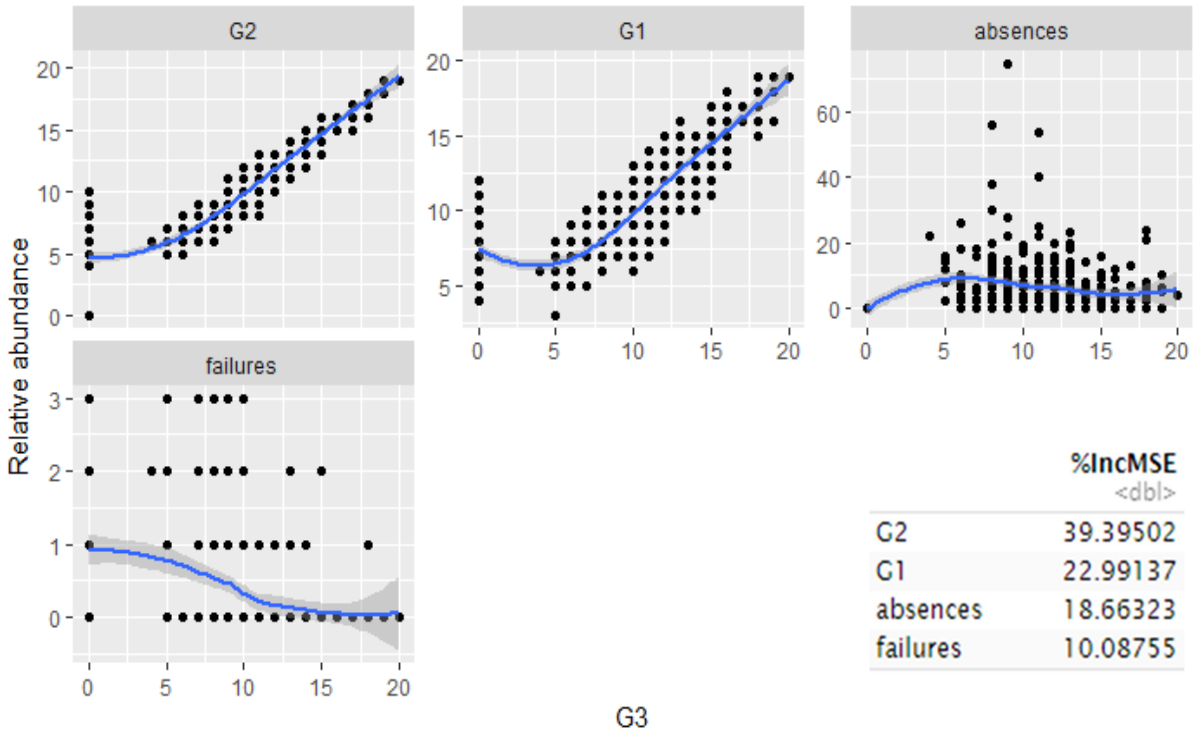
Variables are chosen in terms of values of mean squared error (%IncMSE). The larger %IncMSE, the more important the variable. The top 30 important variables from the lowest to the highest are plotted, and some specific values are listed above.



Group.1 <dbl>	x <dbl>
1	3.010648
2	3.445202
3	3.087011
4	2.489154
6	2.826811
9	2.935255
14	3.642242
21	4.036282
32	4.671210

Figure 2: Cross-validation error about the number of variables

Then the appropriate number of predictive variables need to be determined. Repeated 10-fold cross-validation in the training set is performed, and the relationship between the number and cross-validation error is presented using ggplot [9]. From the graph, the error is minimized when the number of variables is about 4.



	%IncMSE <dbl>
G2	39.39502
G1	22.99137
absences	18.66323
failures	10.08755

Figure 3: The four most important variables

Considering the increase in mean squared error (%IncMSE), the four largest values are 39.395 (second-period grade G2), 22.991 (first-period grade G1), 18.663 (number of school absences) and 10.088 (number of past class failures), respectively. They are more important than other attributes. Therefore, they are singled out and chosen as independent variables. The trend of the relationship of them with the variable final grade G3 is obvious, indicating that they are highly correlated with final grade G3.

Then compare the random forest [10] regression model using all 32 independent variables and four selected predictive variables. Denote the former method 1 and the latter method 2.

"% Var explained" of method 1 is 84.75 compared with 82.43 of method 2. In "Mean of squared residuals", method 1 gets 2.965 and method 2 gets 3.414. Method 2 explains less of the total variance than method 1 and is with larger error, but their differences are little.

Call: randomForest(formula = G3 ~ ., data = d1_train, importance = TRUE)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 10

Mean of squared residuals: 2.965162

% Var explained: 84.75

Call: randomForest(formula = G3 ~ ., data = d1_train.select, importance = TRUE)

Type of random forest: regression

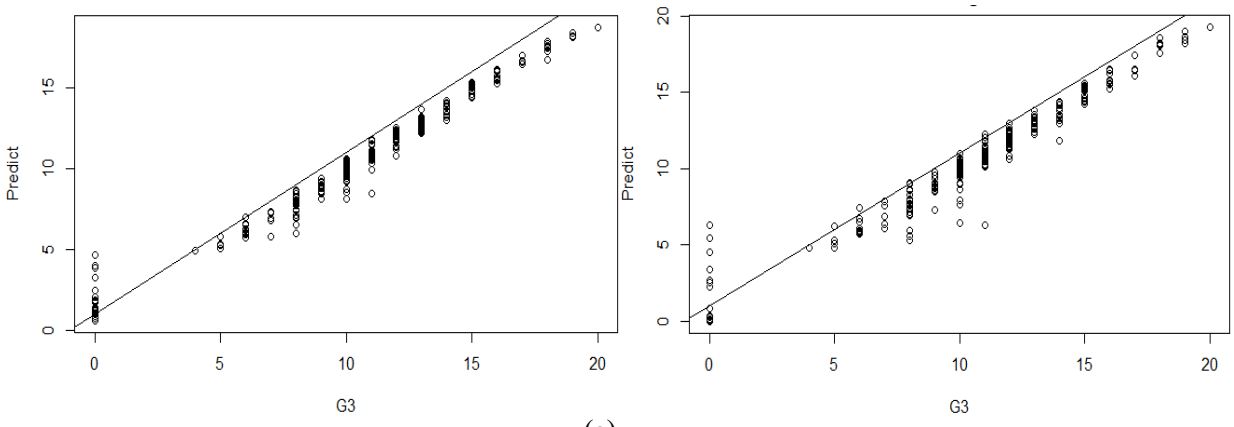
Number of trees: 500

No. of variables tried at each split: 1

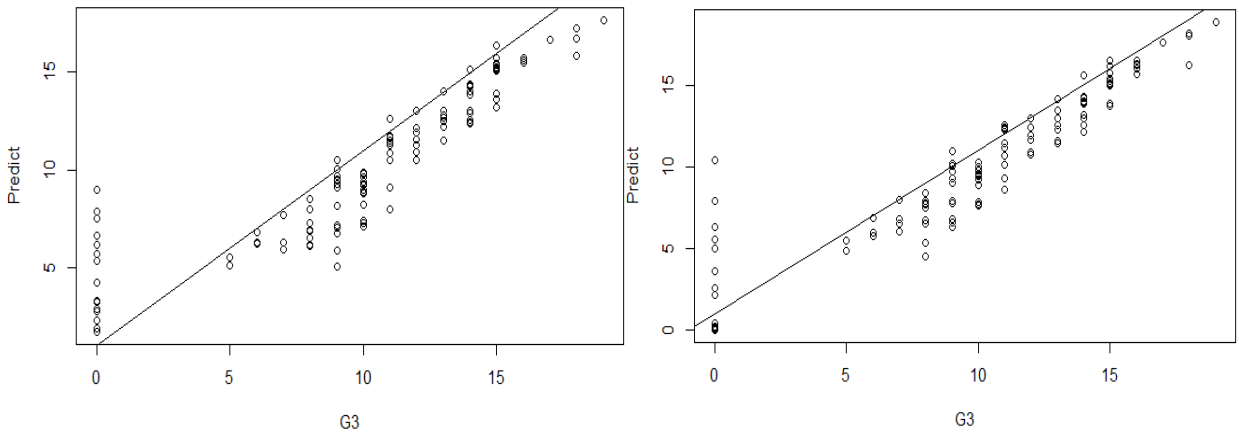
Mean of squared residuals: 3.414538

% Var explained: 82.43

Two methods perform almost the same on the training set and testing set:



(a)



(b)

Figure 4: Comparison of prediction accuracy, training set (left) and testing set (right) of method 1 (a), training set (left) and testing set (right) of method 2 (b)

The AUC value of method 1 is 0.531, close to that (0.594) of method 2, indicating that the two methods have similar reliability.

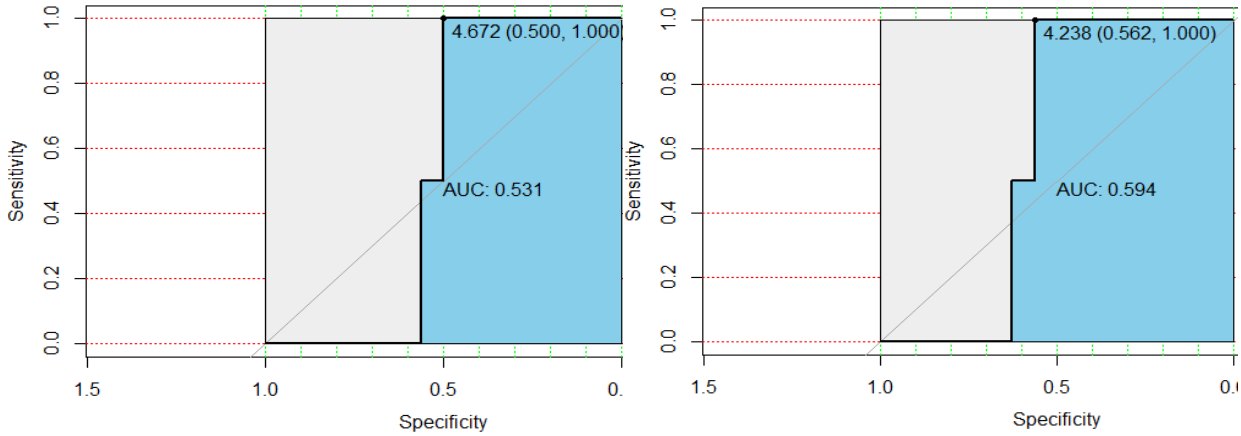


Figure 5: ROC curve and AUC value of method 1 (left) and method 2 (right), mtry=3, ntree=500

Through the error analysis and ROC curve comparison on the training set and testing set, we find that the four variables with the most important characteristics can well replace the original 32 independent variables, and the study of the relationship between these four variables and the final grade can greatly simplify the model on the premise of ensuring accuracy.

Second-period grade (G2), first-period grade (G1), number of school absences (absences) and number of past class failures (failures) are all numerical variables with some specific values. Therefore, we use efficient machine learning algorithms to predict student performance concerning four independent variables, G2, G1, absences and failures.

Table 1: Information of research variables

Research Variables	Type	Value
G2	numerical	from 0 to 20
G1	numerical	from 0 to 20
absences	numerical	from 0 to 93
failures	numerical	n if 1<=n<3, else 4
G3	numerical	from 0 to 20

2.3 Statistical Analysis and Model Selection

An R file about student performance in Mathematics is downloaded from the UCI Repository. It is uploaded to RStudio for research and analysis. The multiple linear regression is selected as a statistical model in our research. Our methodology consists of three steps. The first step is to build a multiple linear regression model based on second-period grade (G2), first-period grade (G1), the number of school absences (absences) and the number of past class failures (failures) and derive coefficients of the regression equation. Next, calculate and evaluate the linear model's residuals, standard error, R-squared, p-value, etc. The third step aims to verify whether the model satisfies linear hypothesis, normally distributed residuals, residual independence, and multicollinearity.

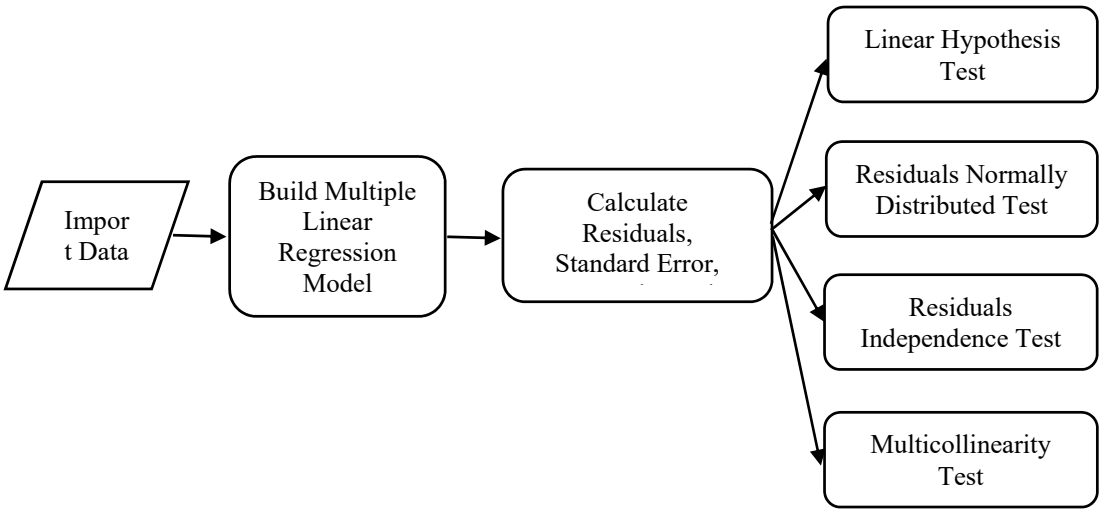


Figure 6: Flowchart of statistical analysis

In predicting student performance, machine learning techniques like multiple linear regression [11], a supervised machine learning technique, are implemented in teaching and learning considering past student grades, demographic, social and school-related features [12]. In life, a phenomenon is often associated with multiple factors, and the optimal combination of multiple independent variables is used to predict or estimate the dependent variable effectively and in line with reality. It can also indicate the influence intensity of multiple independent variables on a dependent variable, accurately measure each factor's correlation degree and fitting degree and improve the effect of prediction. Therefore, multiple linear regression is appropriate to analyze the data set for predicting student performance based on multiple variables.

In this research, the R of version 4.1.0 (2021-05-18) is employed to analyze the secondary data.

3. Results

Multiple linear regression is implied to predict the dependent variable final grade G3 based on four independent variables second-period grade (G2), first-period grade (G1), number of school absences (absences) and number of past class failures (failures). A significant regression equation is found ($F(4, 390) = 468.3$, $p<2.2e-16$), with an R-squared of 0.8277. Denote student's final grade (G3) as y, second-period grade (G2) as X1, first-period grade (G1) as X2, number of school absences (absences) as X3, number of past class failures (failures) as X4. The multiple linear regression model is:

$$y = -1.76483 + 0.97847\,X_1 + 0.14374\,X_2 + 0.03759\,X_3 - 0.25720\,X_4 \qquad (1)$$

where G2 and G1 are measured in points, absences and failures are measured in times. Student's final grade increases 0.97847 points for each point of second-period grade G2 and G1 weighed 0.14374 points less than G2. However, final grades G3 decreased 0.25720 points for each time of failure. G2, G1, absences and failures were significant predictors of final grades G3.

Residuals:

Min	1Q	Median	3Q	Max
-9.3136	-0.3296	0.2701	0.9521	3.7871

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.76483	0.37769	-4.673	4.1e-06 ***
G2	0.97847	0.04921	19.882	< 2e-16 ***
G1	0.14374	0.05575	2.579	0.01029 *
absences	0.03759	0.01206	3.117	0.00196 **
failures	-0.25720	0.13956	-1.843	0.06609 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.911 on 390 degrees of freedom

Multiple R-squared: 0.8277, Adjusted R-squared: 0.8259

F-statistic: 468.3 on 4 and 390 DF, p-value: < 2.2e-16

There are no points that deviate from the straight line, indicating a good linear relationship between four independent (G2, G1, absences and failures) variables and Component + Residuals.

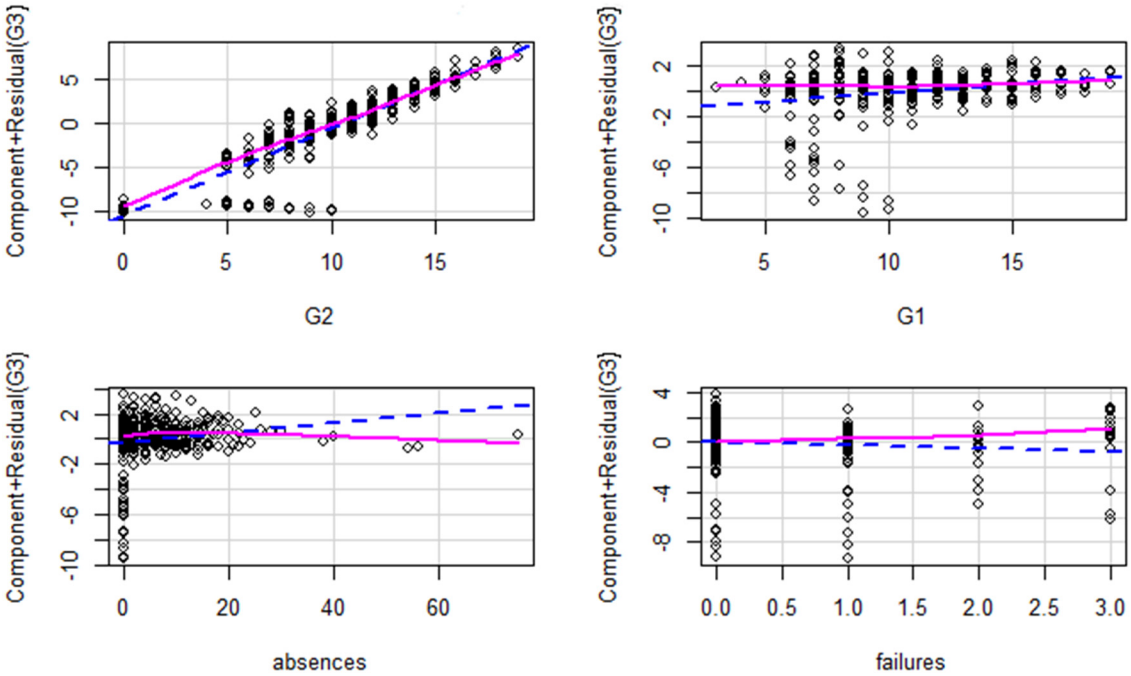


Figure 7: Partial-residual plots between four factors and Component+Residuals

After plotting the quantile-comparison plot of the linear regression model (1), most of the points are near the blue line except for some outliers like 265 and 342. It indicates the model has good normality. The outliers may result from error of data and the small sample size.

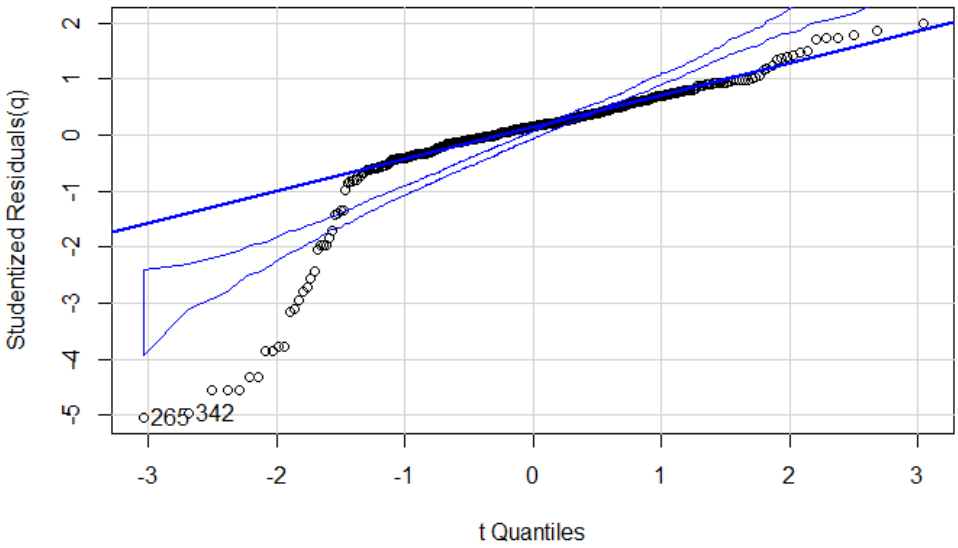


Figure 8: Q-Q plot of the linear regression model

P-value is larger than 0.05, and the errors can be considered independent of each other.

lag	Autocorrelation	D-W Statistic	p-value
1	0.08230762	1.834481	0.098

Alternative hypothesis: rho != 0

In an ideal linear model, each independent variable should be linearly independent. If there exist collinearity among independent variables, the accuracy of regression coefficients will be reduced. In Statistical Learning, collinearity is measured by Variance Inflation Factor (VIF) and it is generally believed that collinearity exists when VIF exceeds 4. VIF of four independent variables of our model are all relatively small, which indicates that the multicollinearity problem does not need to be considered.

G2	G1	absences	failures
3.695497	3.691906	1.004176	1.161441

4. Discussion

Given the advantages of a multiple linear regression model, such as the influence intensity of multiple independent variables being indicated on a dependent variable, each factor's correlation degree and fitting degree can be accurately measured. A multiple linear regression model is constructed. After calculating and analyzing some parameters like residuals, standard error, R-squared, p-value etc., the performance of the built model is tested from four aspects: linear hypothesis, residuals normally distributed, residuals independent, and multicollinearity, respectively. Graphs and data intuitively illustrate that the significant regression equation ($F(4, 390) = 468.3$, $p<2.2e-16$), with an R-squared of 0.8277, can replace the original 32 variables to predict student performance effectively. In general, the multiple linear model represented by formula (1) can be used to calculate a student's final grade relatively accurately as it fits the data well. However, there exists a flaw in the result. The estimated coefficient of the variable final grade G3 is positive, which means that the number of absences of students is positively correlated with their final scores, but this is not consistent with the actual situation. There are two reasons to consider: the data in the dataset may not be accurate enough, and the sample size selected may be too small to be representative. We can use some other machine learning algorithms to tackle this problem.

5. Conclusion

This study used Random Forest Algorithm and 10-fold Cross-Validation to extract the four most important variables from the original 32 independent variables. In other words, students' final grades (G3) are highly correlated with second-period

grade (G2), first-period grade (G1), number of school absences (absences) and number of past class failures (failures). Then through Multiple Linear Regression Algorithm, a linear model represented by formula (1) with higher accuracy is built and its prediction accuracy and bias are tested. This model greatly reduces the number of independent variables while ensuring accuracy, which will help predict student performance.

In future work, we would like to use other criteria to select important variables and make the model can be tested for better analysis and accuracy to be more applications such as finding some methods to handle outliers. Also, some other regression and classification techniques would be tried to predict student performance.

References

[1] Mrinal Pandey and Vivek Kumar Sharma. A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications, 61(13), 2013.

[2] Samy S Abu-Naser, Ihab S Zaqout, Mahmoud Abu Ghosh, Rasha R Atallah, and Eman Alajrami. Predicting student performance using artificial neural network: In the faculty of engineering and information technology. 2015.

[3] Asmaa Elbadrawy, Agoritsa Polyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, and Huzefa Rangwala. Predicting student performance using personalized analytics. Computer, 49(4):61–69, 2016.

[4] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, and Sunday O Olatunji. Student performance prediction using support vector machine and k-nearest neighbor. In 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE), pages 1–4. IEEE, 2017.

[5] Fahad Razaque, Nareena Soomro, Shoaib Ahmed Shaikh, Safeeullah Soomro, Javed Ahmed Samo, Natesh Kumar, and Huma Dharejo. Using naïve bayes algorithm to students' bachelor 2 academic performances analysis. In 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), pages 1–5. IEEE, 2017.

[6] Mahesh Gadhavi and Chirag Patel. Student final grade prediction based on linear regression. Indian J. Comput. Sci. Eng, 8(3):274–279, 2017.

[7] Ankita Kadambande, Snehal Thakur, Akshata Mohol, and AM Ingole. Predict student performance by utilizing data mining technique and support vector machine. International Research Journal of Engineering and Technology, 4:2818–2821, 2017.

[8] Paulo Cortez, Student Performance Data Set, the first version, 2014.11, Retrieved from http://archive.ics.uci.edu/ml/datasets/Student+Performance

[9] Hengl T, Nussbaum M, Wright M N, et al. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables[J]. PeerJ, 2018, 6: e5518.

[10] Breiman L, Cutler A. Random forest-manual[J]. Online: http://www. stat. berkeley. edu/~ breiman/RandomForests/cc_manual. htm, 2004.

[11] Tranmer M, Elliot M. Multiple linear regression[J]. The Cathie Marsh Centre for Census and Survey Research (CCSR), 2008, 5(5): 1-5.

[12] Oyerinde O D, Chia P A. Predicting students' academic performances–A learning analytics approach using multiple linear regression[J]. 2017.