Education Data Mining with China Education Panel Survey

WANG, Yiren QIU, Yirong LIU, Chang

Social Science Department

The Hong Kong University of Science and Technology

Abstract

It is interesting and meaningful to explore and analyze the educational data in social science study educational discipline. The predictive statistics can provide insight into evaluating students' performance and thus offer solutions to focus on certain factors pertinently. By using machine learning methods, we may obtain critical knowledge of educational outcomes. This study used multiple machine learning methods to evaluate Chinese students' academic performance based on the China Education Panel Survey questionnaire. We explored multiple factors in the survey that is found to be related to students' academic performance in primary and middle school education and tried multiple models to reveal significant personal and social factors that influence students' grade.

Keywords

Machine learning, CEPS, Data mining, Student Data, Performance prediction

Table of Contents

INTRODUCTION

Education has long been seen as an important component of society. Education can increase the human capital inherent in the labor force and improve labor productivity, leading to a transitional period of economic growth to a greater level of balanced output (Mankiw et al. ,1992). Moreover, education can promote economic growth by fostering the spread and distribution of the knowledge needed to understand. It leads to the successful implementation of new technologies developed by others (Benhabib and Spiegel, 1994).

Educational data mining (EDM) is a new trend in data mining and Knowledge Discovery in Databases (KDD). As an interdisciplinary field of study, it focuses on mining useful patterns from educational information systems. It is associated with applying different strategies for analyzing data generated in educational environments to discover useful knowledge. The goal is to produce models that can lead to improved learning experiences and greater institutional efficiency. Scholars in the field specialize in revealing useful insights to help educational institutions do a better job of managing their students or to help students better organize their education and outcomes to raise their performance. In the past few years, educational data mining has received a great deal of attention. When many data mining techniques have been proposed to extract hidden knowledge from educational data, the extracted knowledge starts to help the institutions to improve their teaching methods and learning processes. All these

improvements help in improving student performance and overall educational output.

This research paper will analyze students' information through machine learning methods to predict students' learning performance. The purpose of this research is to determine the relationship between students' personal and social factors and their academic performance. Predicting a student's academic performance is generally useful for educational institutions seeking to improve student performance. Given a certain data set, we can apply statistical methods to reasonably predict the results. With the predicted results, teachers can give students more personalized and targeted guidance. This newly discovered knowledge can help students and teachers better improve the quality of education by identifying students who may underperform at the beginning of the semester and giving them more attention to help their educational process and get better progress. In fact, not only the under-performing students can benefit from this research, but the excellent-performing students can also benefit from this study by putting more effort into outstanding subjects with more help and attention from teachers.

LITERATURE REVIEW

Since data mining and machine learning methods have become popular in social science disciplines, they have been widely used to predict students' academic performance. For example, Educational Testing Service (ETS) has examined (Wei, 2013) the relationships between scores and the various background variables in order to determine the extent to which background variables predict scores. One of the techniques researchers used was multilevel modeling. The researchers concluded that compared with individuals, background variables are more accurate score predictors at the management level. Therefore, the researchers believe that the application of the statistical techniques identified in this study to the background of candidates in the administrative management level can be used to continuously evaluate the performance of the Intergovernmental Test of English (TOEIC) test. The study effectively proves the necessity and validity to predict learning outcomes (scores) with background variables. This literature review aims to study the existing methods used in EDM and further review related work to show the advantages and disadvantages of these methods.

Traditional methods and limitations

A common traditional research method for predicting student performance is to conduct additional tests and use the results as a benchmark for prediction. Lindy, Gerald and Steve (2001) conducted a study of using oral reading rate to predict student performance on statewide achievement tests. A total of 77 students participated in the

study. The students were presented with an intact passage to read and prompted to read aloud for 1 minute during which the student's correct words are tallied and errors are recorded. Students usually time in three different paragraphs and record the median of these three paragraphs. Data collected from these measures of reading rate are often used to inform a broad range of educational decisions including progress monitoring, prereferral identification of students, and classification decisions. More importantly, the reading rate is used to predict students' reading and math performance on statewide achievement tests. The results indicate that the Pearson correlation coefficient between correct words per minute and statewide reading assessment is 0.66 and between correct words per minute and statewide math assessment is 0.53.

This is a typical traditional study of predicting students' performance. Based on the performance of students in reading passages, the teachers predict the performance of these students on the statewide achievement tests correspondingly. However, this kind of prediction is not rigorous and the accuracy rate is low, because the result of the prediction is based on only one variable. This simple component demonstrates the limitations of the research data since the data was created for this specific research and it can only be used for this research topic. In addition, performing an additional test is time-consuming and needs to be performed frequently to follow up, so the number of participants is very limited, which means that this method can only predict the performance of a small number of people, and cannot make large-scale predictions.

Educational Data Mining

Obviously, the oral reading rate does not apply to other aspects and the methodology works with low efficiency. Nowadays, a lot of educational institutions and social science departments restore huge volumes of data, including family background, enrollment, attendance, etc. Data mining research from these datasets yields teaching and learning processes. Traditional data mining algorithms cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a preprocessing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. In order to create a predictive model, there are several data mining techniques, which are classification, regression, and clustering. The most commonly used technique for predicting student performance is classification. There are some methods under classification such as decision tree (DT), Naive Bayes (NB), artificial neural networks (ANN), and clustering.

1. Decision Tree

A decision tree is a set of conditions arranged in a hierarchical framework. Due to its simplicity, many researchers have used this technique to transform it into a set of classification rules. C4.5 (Quadri and Kalyankar, 2010) and Classification And Regression Tree (CART) are the two popular decision tree algorithms. A study used decision tree algorithms to predict students' final grades based on data about their usage of the Moodle system (Romero et al., 2008). Moodle is an online learning platform or course management system (CMS) that is frequently used by students. The real data

was collected from seven Moodle courses at the University of Cordoba and the students were divided into two groups: 'passed' and 'fail'. The purpose of this research is to divide students with the same grades into different groups based on the activities carried out in the online course.

Another popular decision tree algorithm is ID3 decision tree. A study investigated 50 students who enrolled in a specific course program within 2007 to 2010 (Baradwaj and Pal, 2011). There were multiple performance indicators, such as ‘Class Test Grades’, ‘Class Test Grades’, ‘Attendance’ and etc.. ID3 is used to build the decision tree and if-then rules to eventually help teachers and students better understand and predict student performance at the end of the semester. ID3 decision tree is chosen in Baradwaj and Pal's study because it is a simple algorithm.

2. Naive Bayes

Naive Bayes is a simple probability classification technique, it is based on applying Bayes' theorem with strong independence assumptions between the features. Hien & Haddawy (2007) used Bayesian networks to predict CGPA of applicants based on their background at the time of admission. Today, educational institutions need a way to evaluate qualified applicants who graduated from various institutions. This research introduces a new method that integrates case-based components with predictive models. Case-based component searches for past students who are most similar to the assessed applicant. The challenge is to define the similarity of cases in a manner consistent with

the predictive model. This can be used in institutions with a good database of student and applicant information.

3. Regression

Another preprocessing algorithm in EDM is regression. A group of research teams in Morocco (Aissaoui et al., 2020) conducted multiple linear regression-based research to predict student's learning outcome. The researchers aimed to select among different machine learning methods to determine the most important variables that contribute to building a student performance prediction model. As this work mainly focuses on student modeling, it is one of the educational data mining objectives. The researchers concluded the various data mining techniques in former literature, such as classification and regression. He then analyzed the merits and drawbacks of each algorithm. While most of the current works tend to use all attributes to generate the prediction models, this paper only selects crucial variables. The author first decided the most important variables using multivariate adaptive regression splines (MARS) methods, and built seven different models based on the top five attributes. This study is a relatively simple but successful attempt of educational data mining.

4. Clustering

A cluster is a group of objects that are similar to each other within the group and not similar to objects belonging to other clusters. The clustering methods are usually sorted according to the approach used to implement the algorithm. Centroid-based clustering,

graph-based clustering, grid-based clustering, density-based clustering, neural network-based clustering, etc. Therefore, one can find algorithms implementing partitioning/hierarchical and hard/soft methods among these methods. Buehl & Alexander investigated students' epistemological beliefs about knowledge attainment and their learning processes (Buehl & Alexander, 2005). The aim of this study was to examine epistemological beliefs and students' achievement motivation. This study was distinctive because the authors did not examine whether individual beliefs were correlated or co-correlated to achievement and motivation. They tested different belief configurations related to students' beliefs about competence, achievement values, and learning of texts. The sample size was 482 undergraduate students, analyzed their beliefs about historical and mathematical knowledge, ability levels, and achievement values. The data were analyzed using Ward's minimum variance hierarchical clustering technique, which showed that students with different epistemological beliefs differed in their ability beliefs and achievement values. They recommend that future research could apply cluster analysis to analyze the different belief configurations related to various aspects of student learning.

Clustering can provide insights about variables that are relevant to individual clusters. Educational data are often multi-level and non-independent, so researchers must carefully select clustering algorithms that demonstrate the research question in order to obtain valid and reliable results. Scholars also must be careful when reducing the data size because when representing data in the form of fewer clusters typically loses certain fine details similar to lossy data compression.

Advantages of Using Computational Social Science (CSS) Methods

Compared with traditional quantitative research methods, computational methods show several obvious advantages. Firstly, the data used in computational social science research is the most important component. CSS data includes multi-dimensional information such as social relationships, social interactions, text, audio, and video, and most of them are recorded with the accurate and precise time. Moreover, most of the data is generated by ordinary users in the natural environment and is not deliberately created. Traditional data such as questionnaires are customized for a particular research topic, so interference and bias may exist. CSS data provides multilevel information in an analytics-friendly way.

The second advantage is the efficiency and effectiveness of data collection. The powerful functions and ease of use of calculation algorithms increase the speed of retrieving digital traces and reduce costs. Traditional research usually requires a lot of time and energy. The cost is relatively high, such as field surveys and follow-up studies, commonly used and time-consuming. However, for CSS, obtaining thousands of digital records for a single research project is not difficult. Also, digital traces can be accessed almost in real-time, making it possible to track the emergence and evolution of social phenomena or human behaviors over a longer period of time.

The third advantage of CSS is the enhanced modeling technology in data analysis. CSS

data has accurate time, and is multi-dimensional and large-scale, which helps to use enhanced modeling techniques to reveal complex mechanisms and subtle patterns in human communication. Such modeling techniques may include time series analysis, sequential modeling, mixed effects modeling, network analysis and spatial modeling (Peng, Liang & Zhu, 2019). The results and methods will also help to define future steps for the EDm research, including possible transformations of the dataset, tuning the classification algorithms' parameters, etc (kabakchie va, 2012). These enhanced modeling techniques allow CSS researchers to review classic theoretical issues through powerful research designs, or explore new research issues with new perspectives and methods. Also, CSS methods that are novel to the social sciences act as a function of the accessibility of powerful open-source tools (LSE, 2017). For instance, network analysis has long occupied social science researchers, yet a new wave of studies leverages advances in computational techniques for inferences in a degree of detail and complexity that was previously unthinkable.

Shortcomings of Using Computational Social Science (CSS) Methods

Although the CSS method has the above advantages, there are reasons to admit and discuss the CSS method's limitations. CSS data on social and mobile media was discovered casually, rather than empirical research conducted by researchers for different purposes of attention, so researchers have minimal control over research objects' representativeness related to the discovered data. A question worth pondering

is whether users who generate digital traces on the Internet can represent a social group, which means that researchers must be very cautious about the validity of the variables constructed from the discovered data.

Second, most data on the Internet are related to human behavior. They provide minimal information about human behavior's basic psychological characteristics, such as attitudes, emotions, and motivations. Although computational algorithms can infer psychological characteristics from behavioral traces, it requires a rigorous evaluation of the validity of the measurement of derived psychological characteristics based on certain facts.

The CSS method is easy to build models that connect different factors, but the mechanism remains unexplained. Focusing on a mere data-driven standpoint might cause ignorance of the fact that every data represents a unique human being. And when the dataset has too many observations, we might run into an overfitting issue in the regression model.

RESEARCH AIM AND METHEDOLOGY

The research aim of this study is to explore different factors related to students' cognition ability. By splitting the data into train and test datasets, we applied five machine learning models to reveal the mechanism of the most influential factors, verify their impact by Bayes probability distribution matrix, and give intuitive explanation why they might influence the score eventually.

DATA

A. Dataset

The China Education Panel Survey (CEPS) is a nationally representative large-scale tracking survey, which aims to reveal the impact of family, school, community and macro-social structures on individual educational output, and to further explore educational output. The process that takes place in the life course of an individual.

The China Education Tracking Survey (CEPS) uses the 2013-2014 school year as the baseline, the first year of junior high school (7th grade) and the third year of junior high school (9th grade) as the starting point of the survey, and randomly selects 28 county-level units (counties, districts, and cities) as the survey sites, using the average education level of the population and the proportion of migrant population as stratified variables.

The survey was conducted on a school basis, and 112 schools and 438 classes were randomly selected from the selected county-level units for the survey, with all students in the sampled classes included in the sample, making a total of about 20,000 students in the baseline survey.

B. Data Exploration

The dataset covers many aspects of students, parents and school administrators. The following table shows the sections of the questionnaires.

Table 1: Numbers of attribute in the three domains

Questionnaire	Section	Numbers of attribute
Students	Part A: Personal Background	65
	Part B: Academic Development	75
	Part C: Physical and Mental Health	71
	part D: Social Behaviors and Development	66
parents	Part A: Family Education	88
	Part B: The Relationship between Parents and School	38
	Part C: School Education	22
	Part D: Basic Information of This Child	29
	Part E: Parents, Family and Community	35
School Administrator	Part A: Basic Information of the School	48
	Part B: Basic Information of JUNIOR HIGH Students at School	41
	Part C: Basic Information of Teachers in the JUNIOR HIGH Department	40
	Part D: Enrollment of Students in the JUNIOR HIGH Department	31
	Part E: Management of the JUNIOR	82

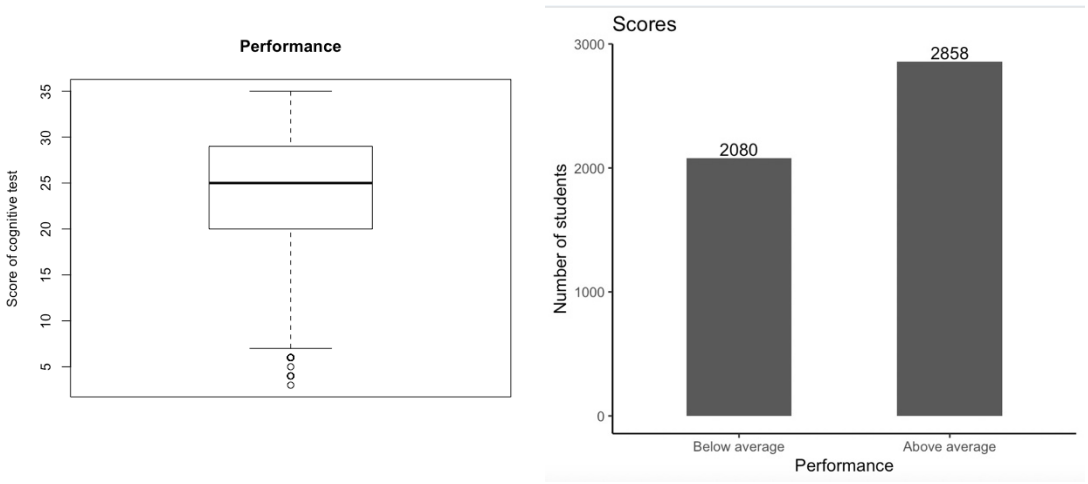
	HIGH Department	
	Part F: Personal Information	32
Total		764

In addition to the questionnaire, the CEPS designed a set of cognitive ability tests for Year 8 students, which do not cover the specific literacy knowledge taught in the school curriculum, but rather measure students' logical thinking and problem-solving skills, and are internationally comparable and nationally standardized. The test consists of 35 questions and lasts 30 minutes.

The questions are divided into 3 dimensions and 11 constructs, including: (1) language: word analogies, verbal reasoning (2) graphics: graphical pattern analysis, origami questions, geometric applications (3) calculation and logic: mathematical applications, customized rules of arithmetic, applications of series, abstract pattern analysis, probability, numerical size inverse thinking.

The following graphs shows the distribution of student's performance. The mean of students score is 23.54 and median is 24. So, we use the score of 24 as a classification standard and all students are divided into two groups. And we also use 80% of the dataset as the training set and the rest 20% as the testing set.

Figure 1. Performance boxplot and bar plot



DATA MINING PROCESS

A. Data Manipulation

1.Merging

The datasets of students, teachers and school principals are separate but have mapping relations. The student ID is shared in both students' questionnaire and the parents' one. All three datasets include the school ID. We first merged parents' data to the student's dataset by 'sid' and got *dataset_1*, then merged the school principals' data to *dataset_1* by 'schids'. This new dataset contains 876 variables and 10,280 records.

2. Data Cleaning

Firstly, the unrelated variables are deleted, such as survey time and student id, etc. When locating the NAs, we found that missing values and options that were not selected in the multiple-choice question were also recorded as null. For the latter, we replaced the null value with 0. To deal with the missing values, we deleted the attributes which lack more than ¼ of their values. For the columns that missing value is less than 10%,

we turned NA to 0. All the other missing values are deleted. After deleting 'NA's and unrelated variables, 764 variables and 4938 records remain.

A. Data Mining Models

1) Logistic regression

Logistic regression, also known as the logit model, is used to model dichotomous outcome variables. In the logit model, the log odds of the outcome are modeled as a linear combination of predictors. This statistical methodology is frequently used in numeric predictions. We estimate a logistic regression model using the glm (generalized linear model) function. Since we convert the scores into a two-level categorical variable, 'family=binomial' was used in the model.

Figure 2. Logistic regression model

```
## Call:
## glm(formula = cogscore ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1815  -0.6028   0.1709   0.6213   3.0306
##
## Coefficients: (153 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.572e+13  4.739e+13   0.754 0.451069
## w2a01       -3.584e-01  2.579e-01  -1.389 0.164693
## w2a02       -9.350e-02  9.764e-02  -0.958 0.338276
## w2a03       -1.827e-01  3.589e-01  -0.509 0.610779
## w2a04       -9.057e-03  5.294e-02  -0.171 0.864161
## w2a05       -3.605e-01  2.188e-01  -1.648 0.099368 .
## w2a0601      6.138e-02  1.685e-01   0.364 0.715676
## w2a0602      1.267e-01  1.646e-01   0.770 0.441541
## w2a0603      2.975e-01  1.408e-01   2.114 0.034549 *
## w2a0604      1.630e-01  1.617e-01   1.008 0.313436
```

The z-statistic and the associated p-values indicates the significance of the variables,

that is, the degree of correlation between the variable and the result. For example, the most significant variable in logit model is w2b02, which is 'Do you feel struggling to learn math now', the p-value is 3.77e-16. There are comparatively more significant variables in the student questionnaire, while most of the variables in the school leadership questionnaire are not significant.

When we predict the students' scores in the test set, 'type=response' was used. The precision of the prediction is 0.7433775, the recall is 0.7714777, and the F1 score is 0.7571669.

2) LASSO Regression

LASSO regression stands for Least Absolute Shrinkage and Selection Operator, the algorithm is another variation of linear regression. In lasso, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients. We need to identify the optimal lambda value and then use that value to train the model. To achieve this, we used the glmnet function and passalpha = 1 argument. We used cv.glmnet() function to identify the optimal lambda value, which is 0.01.

Figure 3. LASSO Regression model

```
Call:  cv.glmnet(x = x_vars, y = y_var, lambda = lambda_seq, nfolds = 5,
family = "binomial", alpha = 1)

Measure: Binomial Deviance

     Lambda Measure      SE Nonzero
min 0.01000   1.071 0.01204     136
1se 0.01259   1.079 0.01070     101
```

Using the value lambda=0.01, we train the lasso model again. The precision of lasso model is 0.7472868, the recall is 0.8281787, and the F1 score is 0.7856561.

3) Random Forest (implemented by R)

Random forest is a tree-based classification and regression method. The performance of this algorithms is highly competitive with other algorithms. The following graph shows the characteristics of the random forest model.

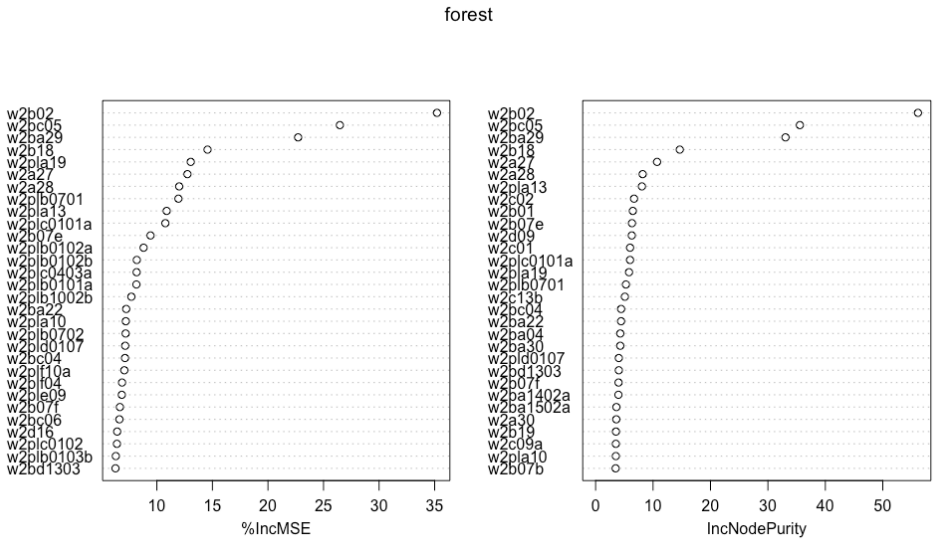
Figure 4. Random Forest model

```
Call:
 randomForest(formula = cogscore ~ ., data = train, ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 254

          Mean of squared residuals: 0.1827299
                    % Var explained: 25.17
```

Two parameters are important in the random forest algorithm: Number of trees used in the forest (ntree) and Number of random variables used in each tree (mtry). After compared the OCC and set the mtry at the best number, and slightly adjust the random forest variable, we slightly improved prediction result.

Figure 5. Important factors in random forest model



When we predict the students' scores using random forest in the test set, the precision of the prediction is 0.7592593, the recall is 0.8453608, and the F1 score is 0.8.

4) Decision Tree

Decision tree classification is very much in line with the way humans think about classification. It asks several different questions to determine a particular feature of a sample and then combines all the judgments to sample a category. A decision tree is a tree structure where each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category. Decision tree learning is a top-down recursive approach, where the basic idea is to construct a tree with the fastest decreasing entropy as a measure of information entropy. The entropy value at the leaf node is zero, at which point each instance in the leaf node belongs to the same class. We use the DecisionTreeClassifier(), and the accuracy result is 0.6417(fig 69 And the rank of feature importance is shown in the following fig(fig 7).

Figure 6. Accuracy of the Decision Tree model

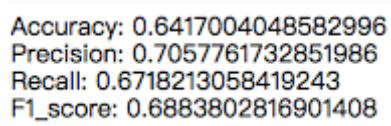
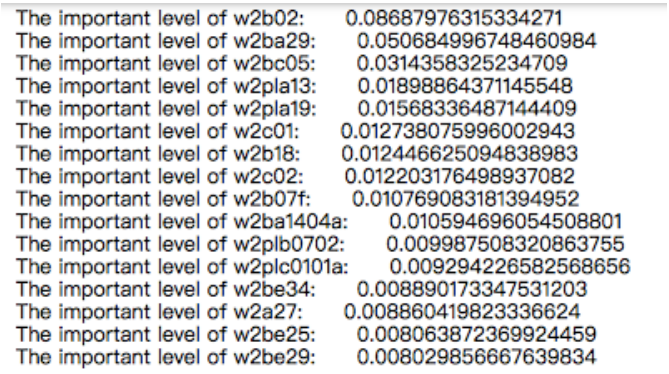
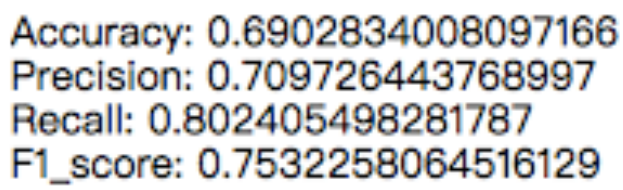


Figure 7. Important level of factors in this model



To improve the model, we find the index with the highest rating on the cross-validation dataset and get *bestdepth=5.* Applying *max_depth=5* to the DecisionTreeClassifier(), we get a better model with the accuracy of 0.6903(fig 8).

Figure 8. Adjusted accuracy of the Decision Tree model



5) SVM (Supporting Vector Machine)

Support vector machine performs the best when the boundaries are non-linear but data is limited to produce composite non-linear model. The basic idea of this method is to find the “thickest hyperplane” in order classify the different attributes. The mapping is done absolutely using kernel functions.

Figure 9 . Support Vector Machine Model

```
Call:
svm(formula = cogscore ~ ., data = train)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.001310616
    epsilon:  0.1
```

When we predict the students' scores using SVM in the test set, the precision of the prediction is 0.604333, the recall is 0.9106529, and the F1 score is 0.726525.

ANALYSIS

After handling the data and making predictions with the above models, we found the below features are frequently appeared in the crucial factors of above models. So we decide to collect all these factors and keep on research. This table shows the mutual significant variables that we find through different machine learning models

Table 2. Description and Possible Values of important attributes

Attribute	Description	Possible Values
w2a27	[Student]	1. Being one of the top five of your

	Parents' requirement on academic record	class 2. Above the average 3. About the average 4. No special requirement
w2b02	[Student] Is mathematics difficult for you?	1. Very difficult 2. A bit difficult 3. Not very difficult 4. Not difficult at all
w2b0505	[Student] My Chinese teacher always asks me to answer questions in class.	Strongly disagree Somewhat disagree Somewhat agree Strongly agree
w2b0508	[Student] My Chinese teacher always praises me.	Strongly disagree Somewhat disagree Somewhat agree Strongly agree
w2c2402	[Student] I would try my best to finish even the homework I dislike.	Strongly disagree Somewhat disagree Somewhat agree Strongly agree
w2bc05	[Parent]	1. Near the bottom

	How does this child's academic record rank in his/her class at present?	2. Below the average 3. About the average 4. Above the average 5. Around the top
w2ba29	[Parent] What is the highest level of education do you expect this child to receive?	1. Drop out now 2. Graduate from junior high school 3. Go to technical secondary school or technical school 4. Go to vocational high school 5. Go to senior high school 6. Graduate from junior college 7. Get a bachelor degree 8. Get a Master degree 9. Get a Doctor degree
w2b18	[Student] What is the highest level of education you expect yourself to receive?	1. Drop out now 2. Graduate from junior high school 3. Go to technical secondary school or technical school 4. Go to vocational high school 5. Go to senior high school 6. Graduate from junior college

		7. Get a bachelor degree 8. Get a Master degree 9. Get a Doctor degree

After going over the important factors, we construct a conditional probability table of these attributes. Intuitively, the 0.168 (which is the first number in this table) refers students who are predicted as "0" has the conditional probability of 0.168 choose the option "1". The rest can be explained in similar logic. We highlight the interesting finding (which means obvious difference between "0" and "1") in red. The function calculates the absolute difference between the classes' probabilities for each row in the confusion matrix, and only if the absolute difference between two of them is more than 0.1 (10%), it will be considered as "interesting". This method was applied by Amjad in 2016, in his paper he set the threshold at 25%, but since we have too many options in one attributes, we considerably lower this threshold and set it at 10%.

Table 3. Conditional Probability Table of important attributes

Attribute	Value	Probability	
		0 (below average)	1 (above average)
w2a27	1. Being one of the top five of your class	0.168	0.353
	2. Above the average	0.500	0.503

	3. About the average	0.237	0.083
	4. No special requirement	0.096	0.060
w2b02	1. Very difficult	0.194	0.052
	2. A bit difficult	0.461	0.316
	3. Not very difficult	0.264	0.452
	4. Not difficult at all	0.080	0.180
w2b0505	Strongly disagree	0.146	0.105
	Somewhat disagree	0.331	0.355
	Somewhat agree	0.377	0.379
	Strongly agree	0.146	0.162
w2b0508	Strongly disagree	0.183	0.151
	Somewhat disagree	0.332	0.388
	Somewhat agree	0.326	0.334
	Strongly agree	0.159	0.126
w2c2402	Strongly disagree	0.062	0.044
	Somewhat disagree	0.160	0.120
	Somewhat agree	0.411	0.425
	Strongly agree	0.367	0.410
w2bc05	1. Near the bottom	0.103	0.032
	2. Below the average	0.260	0.129

	3. About the average	0.358	0.284
	4. Above the average	0.239	0.434
	5. Around the top	0.030	0.118
w2ba29	1. Drop out now	0.002	0.001
	2. Graduate from junior high school	0.019	0.002
	3. Go to technical secondary school or technical school	0.058	0.009
	4. Go to vocational high school	0.039	0.008
	5. Go to senior high school	0.065	0.014
	6. Graduate from junior college	0.188	0.101
	7. Get a bachelor degree	0.351	0.410
	8. Get a Master degree	0.143	0.250
	9. Get a Doctor degree	0.135	0.203
w2b18	1. Drop out now	0.006	0.002

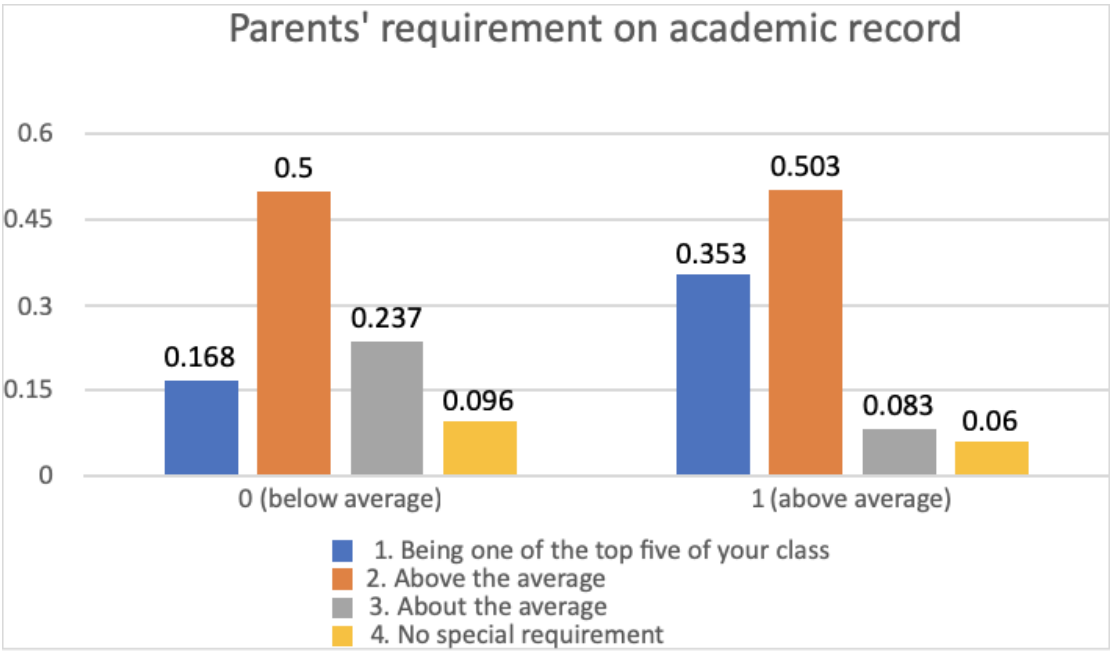
	2. Graduate from junior high school	0.034	0.003
	3. Go to technical secondary school or technical school	0.044	0.007
	4. Go to vocational high school	0.040	0.013
	5. Go to senior high school	0.100	0.031
	6. Graduate from junior college	0.161	0.122
	7. Get a bachelor degree	0.349	0.384
	8. Get a Master degree	0.128	0.237
	9. Get a Doctor degree	0.010	0.167

After finding the interesting contrast above, we visualize them and make reasonable interpretation.

A) [Student] Parents' requirement on academic record

As shown in the figure, the blue bar and the grey bar are significantly different between the below average and the group of above average. The probability of students who are carrying high parental expectations to get high scores are significantly higher. In contrast, students who carried fewer parents' requirements may be more likely to achieve a score lower than the average.

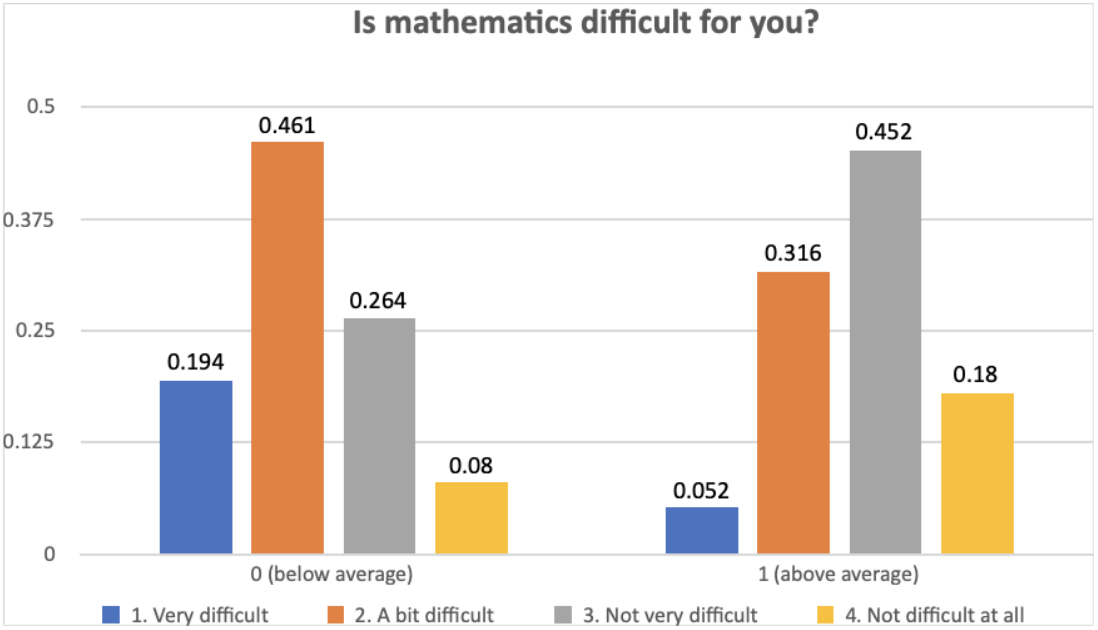
Figure 9. Parents' Requirement



B） [Student] Is mathematics difficult for you?

How students feel about the subject of mathematics has a significant impact on their scores for cognitive skills. Students who did not perceive mathematics as particularly difficult were more likely to score highly. If students perceived mathematics as a little difficult or hard, they were more likely to score below average.

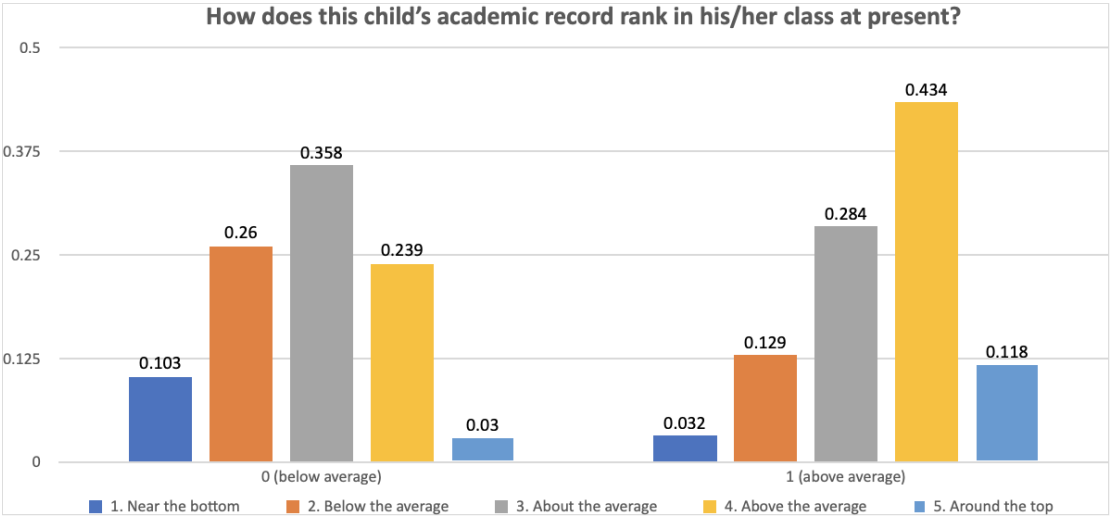
Figure 9. Difficulty of Math



C） [Parent] How does this child's academic record rank in his/her class at present?

It is easy to see how students' academic performance can be used to predict their level of cognitive ability. The fact that this question came from a parent questionnaire rather than a student questionnaire, suggests that parents' perceptions of secondary school students may be more in line with reality than their children's own perceptions. When parents perceive their child's grades to be at the class average, the child's cognitive score is likely to be lower than the average score. When parents perceive their child's academic performance to be above the class average, the child is much more likely to receive an above average cognitive score.

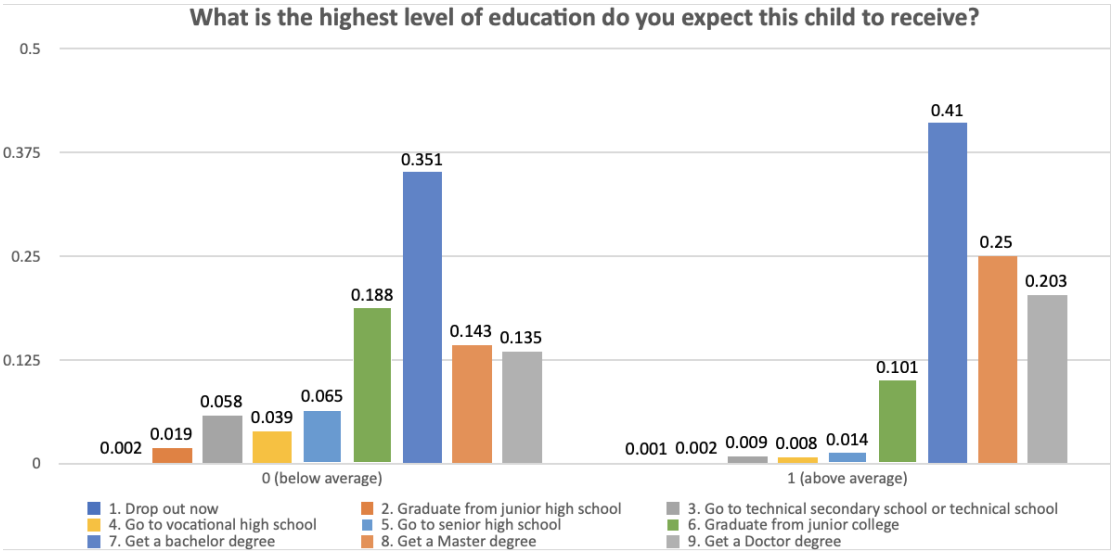
Figure 10. Academic Rank



D） [Parent] What is the high level of education you expect your child to receive?

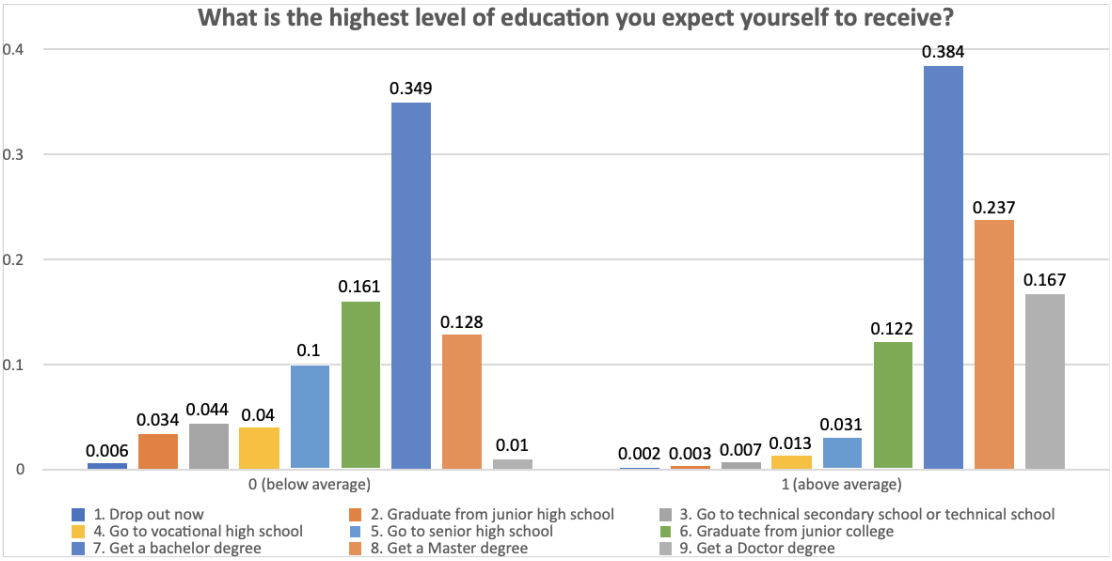
Parents' expectations of their children's educational attainment can also be used as a predictor of a student's high or low cognitive level. When they expect their child to earn a master's degree, the child is more likely to score above average on cognitive tests. When expectations were only high school level, the child was more likely to have a lower level of cognition.

Figure 11. Highest expected level of education from parents



E） [Student] What is the highest level of education you expect yourself to receive? Unlike the previous question, this one comes from the student's perspective. However, the two present very similar results, with above-average awareness levels more likely than below-average when the expectation is a bachelor's degree or higher. One more feature than the previous question is the considerably higher than average level of cognition that students have when they expect themselves to earn a PhD.

Figure 12. Highest expected level of education from students



CONCLUSION

In this research paper, we applied multiple data mining methods to find the best fit predictive model using the CEPS survey data. Based on the result from previous model construction and prediction, the random forest model was found to be the most accurate one in this research. Random forests can handle very high dimensional data and do not have to do feature selection. For unbalanced datasets, Random Forests can balance the

error. When there is a classification imbalance, Random Forests can provide an effective way to balance the error in the dataset. Random forests are more resistant to overfitting, introducing randomness and are less prone to overfitting. Our dataset has over seven hundred variables and using the random forest approach really demonstrates its benefits.

Although this is an immature attempt to understand the survey data from Chinese student and generalize potential learning outcome from the answer we obtained from these students, we still successfully raised the accuracy of prediction to a relative high level. First, we retrieved our dataset from the panel data and after processing and tried different machine learning models and make predictions, we generalize a table of the most influential features. The questionnaire may not directly reflect the true learning ability, but our analysis actually finds some interesting facts, such as learning math well may lead to higher cognition ability, which is logical and stay in line with our common sense. In conclusion, this study explores the students' data and suggested some important cognition patterns behind the questionnaires, which can help parents and teachers to better understand their children' performance at school in many ways.

LIMITATIONS & FUTURE RESEARCH

Although the prediction accuracy turns out to be fairly satisfying, some limitations exist in this study. There are three main limitations of this research that can be addressed by later research.

Firstly, the training set after data preprocessing is slightly smaller than we expected.

This study contains 764 variables (after converting to dummy variables), which is quite slow for the computer to process. Some of the variables are very similar to each other; we can use the feature engineering process to combine these variables or delete some of the irrelevant features. (For example, questions in the survey asking for similar questions, and kids in real life usually give the same answers to both questions).

Secondly, resolving data quality might be another challenge. The data we used were mainly questionnaire results from young kids. While most of the questions intended to discover the kids' background information, some important factors that directly influence the learning outcome were omitted (eg. Academic achievement, scholarship). This might affect the functionality of decision trees. To improve the model, we may add more attributes closely related to the study behavior to reduce classification errors. Also, some of the survey questions in this study were ambiguous, so that the responses might be inaccurate in the first place.

Moreover, there are more advanced algorithms to be tested. The most suitable candidate algorithms for this study are LASSO and random forest. It is possible that other mining tasks that can achieve higher prediction accuracy were left out of this study; it would be interesting to examine how well other algorithms perform in this dataset.

REFERENCES

Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of students' performance in E-learning environment using random Forest. International Journal of Innovative Computing, 7(2), 1–5.

Aissaoui, O, E. et al. (2020). A Multiple Linear Regression-Based Approach to Predict Student Performance.

Cortez, P., Silva, A. (2008). Using data mining to predict secondary school student performance. In: Proceedings of 5th Annual Future Business Technology Conference, pp. 5–12.

Baradwaj, B.K. and Pal, S., (2011). Mining Educational Data to Analyze Students' Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

Benhabib, J. and Spiegel, M. M. (1994). The role of human capital in economic development: Evidence from aggregate cross-country data. Journal of Monetary Economics 34(2), 143–174.

Dutt, A., Ismail, M, A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya.

Francis, B. K., & Babu, S. S. (2019). Predicting academic performance of students using a hybrid data mining approach. Journal of Medical Systems, 43(6). https://doi.org/10.1007/s10916-019-1295-4.

Hien, N.T., & Haddawy, P. (2007). A decision support system for evaluating international student applications. 2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, F2A-1-F2A-6.

Iatrellis, O. et al. (2020). A two-phase machine learning approach for predicting student outcomes. Education and Information Technologies. https://doi.org/10.1007/s10639-020-10260-x

Lee, K. (2018). Machine learning approaches for learning analytics: Collaborative filtering or regression with experts ? Korea, 1–11.

Lindy Crawford, Gerald Tindal & Steve> Stieber (2001) Using Oral Reading Rate to Predict Student Performance on Statewide Achievement Tests, Educational Assessment, 7:4, 303-323, DOI: 10.1207/S15326977EA0704_04

Mankiw, N. G., Romer, D., and Weil, D. (1992). A contribution to the empirics of economic growth. Quarterly Journal of Economics 107(2), 407–437.

Quadri, M., & Kalyankar, N. (2010). Drop Out Feature of Student Daa for Academic Performance Using Decision Tree Techniques. Global journal of computer science and technology, 10.

Rai, S. et al. (2020). Machine Learning Approach for Student Academic Performance Prediction. Evolution in Computational Intelligence.

Romero, C., Ventura, S. & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384. Elsevier Ltd. Retrieved November 8, 2020 from https://www.learntechlib.org/p/66484/.