Applying Machine Learning Techniques to Identify the Influential Factors of Students' Abilities to Apply Statistics Mathematics and Engineering Knowledge

Wen Cheng
Department of Civil Engineering
California State Polytechnic University, Pomona
Pomona, USA
wcheng@cpp.edu

Gurdiljot Gill
Department of Civil Engineering
California State Polytechnic University, Pomona
Pomona, USA
gurdiljotg@cpp.edu

***Abstract*— This Research to Practice Full Paper deals with the influential factors related to application of courses by the students. There are a large number of factors affecting students' abilities to apply knowledge of statistics, mathematics, and engineering. Even though the literature is replete with various methods to identify such factors, there is little research dedicated to exploring the influential factors using machine learning methods. Based on 200+ students' responses to senior exit survey, this study explored the influential factors using random forest approach. The variable of 'time spent at CPP (Cal Poly Pomona) towards obtaining a degree' turned out to be the most important in influencing the students' abilities at concerned subjects. This was followed by the time of sitting for EIT (Engineer in training) exam and the engagement of students in different clubs. Clearly, as per the self-assessment of students, their time devoted at CPP and their self-confidence in undertaking the EIT significantly influence their perception of statistics, mathematics and engineering knowledge. This study demonstrated that the random forest approach may be beneficial in such analysis and may be combined with the conventional methods such as multinomial logit regression to devise more informed strategies by the faculty and administration to improve student outcomes.**

Keywords*— Student outcome, senior exit survey, random forest models, machine learning

I. INTRODUCTION

The civil engineering department at California State Polytechnic University, Pomona (CPP) implements different processes for assessment and evaluation of performance: Mission and program objectives; Student outcomes; and Course objectives. These processes circulate continuous information within the key stakeholders: students, alumni, industry, and university faculty and administration. The wealth of information forms a cycle from student and alumni achievements to data collection, analysis, program changes and back to student and alumni achievements. Among the above-mentioned processes, the improvement in student' learning outcomes may be regarded as the core of other two processes as many engineering program objectives are student-centered. One of the main objectives of the majority of engineering programs is analysis of factors for students' ability to apply knowledge of mathematics, science, and engineering, which is well documented in literature [1] [2] [3]. The assessment and evaluation of students' abilities mostly adopt a subjective approach, and hence most of the education studies are based on the qualitative and or descriptive analysis. However, the quantitative assessment may uncover diverse trends which may be used in association with qualitative analysis to obtain more accurate inferences and devise more informed changes. To fill this gap, the authors proposed to identify the influential factors of aforementioned students' abilities using the popular multinomial approach for the categorical outcomes.

This Research-to-Practice Full Paper presents the application of machine learning techniques to perform an efficient evaluation of student outcomes, which plays a vital role in the continued improvement of an Engineering program. Even though the literature is replete with various methods to identify factors for students' ability to apply knowledge of mathematics, science, and engineering, there is little research dedicated to exploring the influential factors using machine learning technology. This study investigates the feasibility of using random forest model approach to analyze the senior exit survey data with the aim to rank the importance of various factors which influence the students' abilities to apply statistics, mathematics, and engineering knowledge. The senior exit survey was conducted during 2014 and 2015 and obtained the responses from 295 students.

II. DATA DESCRIPTION

The data for this study were collected from the Senior Exit Survey, which is routinely conducted by the civil engineering department involving the students who are close to completion of degree requirements. The collected information comprised of student name, years spent at CPP, admission status, the level of participation in various professional societies or clubs, association with senior project, self-assessment of various student outcomes, and so on. Since the primary focus of the current study is identification of influential factors for students' knowledge application abilities for statistics, mathematics, and engineering, this dependent variable was rated by the students on a scale of 1-5: (1) Poor; (2) Fair; (3) Average; (4) Good; (5) Excellent. As expected, majority of the students who undertook the self-assessment rated their abilities as average or above.

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

Hence, the authors regrouped the response to three categories: 1-Average or Lower; 2-Good; and (3) Excellent. The data spanning 2014-15 were utilized, and the pertinent response distributions are shown in Fig. 1.

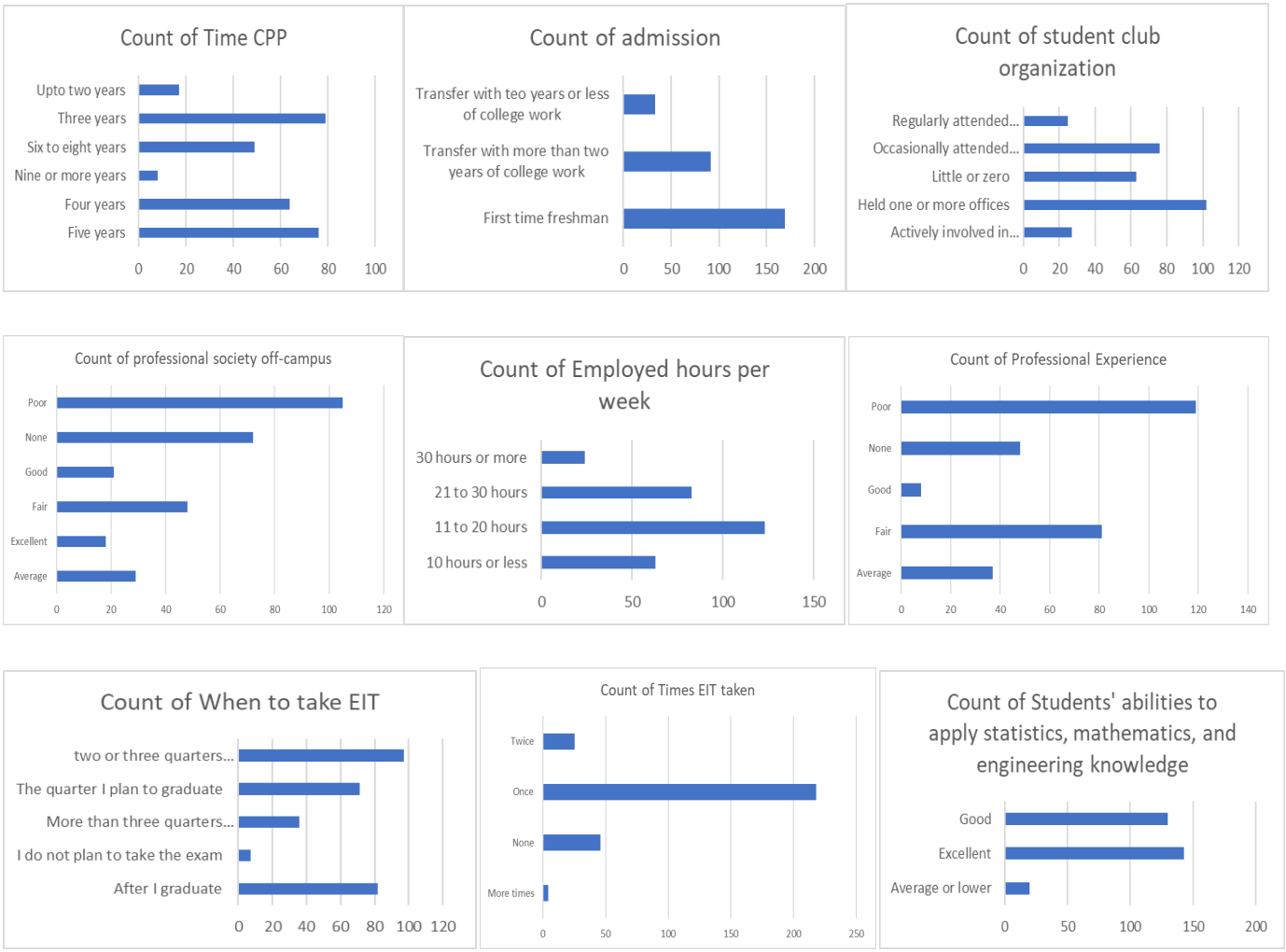

Fig. 1. Visual representation of dependent and independent variables

III. Methodology

In comparison with the traditional statistical regression models, the machine learning technologies seem to have some advantages in that they are robust to outliers and are capable of detecting complex interactions without the need to make specific assumptions of data distributions. Among the large multitude of machine learning techniques, the random forest approach was selected in the study due to its reputation for transparency and relative simplicity in design which is a popular and very efficient algorithm, based on model aggregation ideas, for both classification and regression problems, introduced by Breiman [4]. The principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the learning sample and choosing randomly at each node a subset of explanatory variables.

The Generalized Boosted Regression Models (gbm) [5] package was utilized within the software R. Within the package, gbm function was employed to fit the boosted regression trees to the Senior Exit Survey data. As stated before, in this study the response variable showing students' self-evaluation of capabilities of applying Engineering and Mathematics knowledge has more than two levels:1-Average or Lower; 2-Good; and (3) Excellent. Hence, this converts to a multiple classification problem where the "multinomial" distribution is employed. This may be regarded as the alternative to the conventional Multinomial Logistic Regression Model (MLR), which is a type of Generalized Linear Model (GLM), a Bayesian and non-Bayesian approach to the analysis of static regression problems given by Nelder and Wedderburn [6] that explains how a dependent variable can be described by a set of explanatory variables [7] [8]. The quantification of the variable importance is a crucial issue not

only for ranking the variables before a stepwise estimation model but also to interpret data and understand the underlying phenomenon in many applied problems. The random forest approach ranks the variable importance according to three different issues: the first one deals with the sensitivity to the sample size and the number of variables, the second examines the sensitivity to method parameters and, the third one deals with the variable importance in the presence of groups of highly correlated [9] [10].

The random forest approach ranks the variables based on importance while the MLR approach gives a coefficient to quantify the relationship and provides the level of statistical significance. This study is primarily focused on demonstrating the application of random forest but also provides the MLR model for reference. The model can be expressed by the following equation:

$$Log(\Pr(Y_i)/1 - \Pr(Y_i)) = f(x) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \quad (1)$$

The intercept β_0 is the value of Y when all of the independent variables are equal to zero. β_1, β_2, ..., β_k are the regression coefficients of X_1, X_2,…, X_k . The probability of performance level 1 (Average or lower) is regarded as the reference.

IV. RESULTS

Table 1 demonstrates the results from the random forest model. The importance of each variable has been quantified based on the relative influence (RI), where the higher value of RI represents the greater importance of that specific variable for the dependent variable of students' abilities. As evident from the table, the time spent at CPP turned out to be the most important factor, followed by the time when students plan to take EIT, and participation in student club organizations. Conversely, the number of times for taking EIT ranked the lowest. These results suggest that the as per the exit survey based on the students' self-assessment of their abilities to apply statistics, mathematics, and engineering knowledge, the time spent towards obtaining their degree significantly influences their level of confidence. Understandably, the decision of a student to undertake the EIT is influenced by his/her confidence level. The importance of timing of EIT in the results depicts that the timing of EIT may be correlated with the confidence level for the concerned abilities of students. Fig. 2 depicts the partial dependence plots for the top three influential variables. These plots are the graphical visualizations of the marginal effect of each of three variables on the dependent variable after integrating out the other variables. In other words, this method essentially averages out rest of the predictors, except one, to understand the importance of that specific predictor. The changing relationship between different values of predictors and dependent variable is evident in the three graphs. The importance of these three predictors is represented by the largely non-zero relationship with the dependent variable (represented by the y-axis for all three cases).

The results of random forest may be compared to the findings of MLR model, which are shown in Table 2. The MLR approach quantifies the relationship by generating the coefficient and also provides the statistical significance of the dependency. Only the predictor of time spent at CPP turned out to be statistically significant in case of excellent performance level (dependent variable). In such a scenario, the random forest approach may prove to be beneficial in the analysis of such data since the regression approach is limited in identifying the most important variables which should be monitored to devise more informed strategies to improve student outcomes.

TABLE I. INFLUENTIAL FACTORS FROM RANDOM FOREST

Variable	Relative Influence
Time CPP years	22.67199
When EIT	17.76063
Student club organization	16.16541
Professional society off-campus overall	12.61906
Professional Experience	10.60059
Employed hours per week	9.646789
Admission	8.682377
EIT Times Taken	1.853168

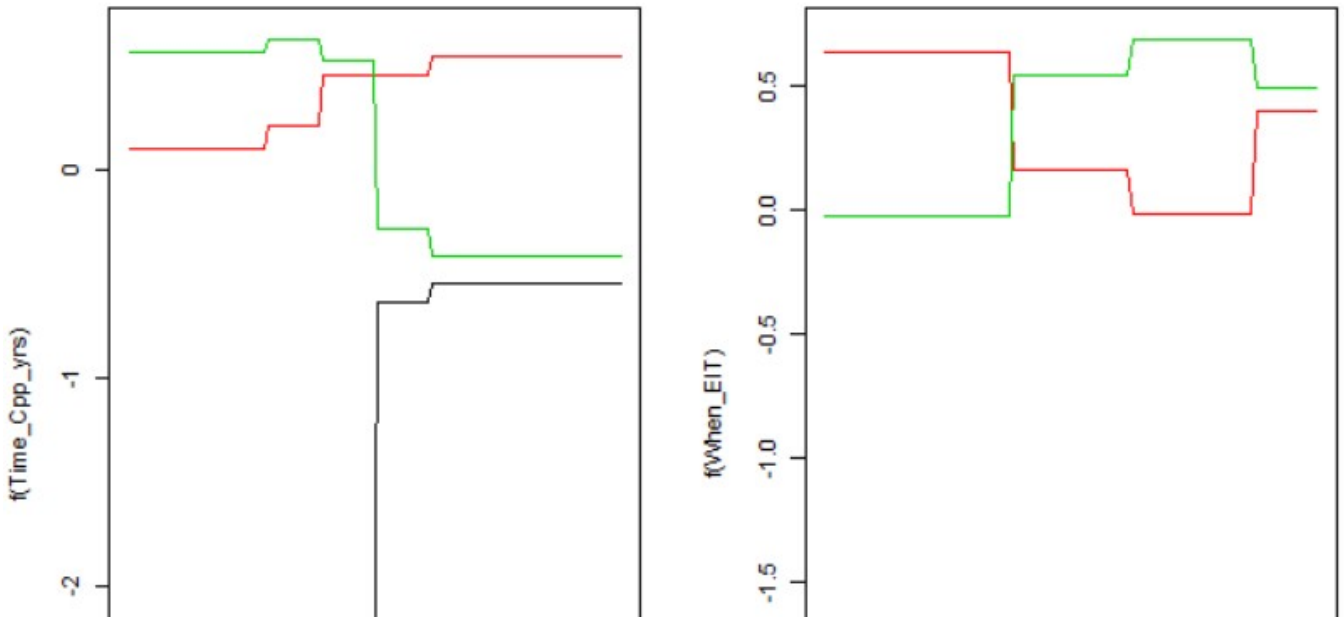

Fig. 2. Partial dependence plots for top three important variables

TABLE II. INFLUENTIAL FACTORS FROM MULTINOMIAL LOGIT REGRESSION MODEL

Variable	Performance level=good		Performance level=excellent	
	Coefficient	p-value	Coefficient	p-value
Intercept	1.917023	0.122193	2.236468	0.080706
Time CPP years	-0.22996	0.092298	-0.58646	0.000431
When EIT	0.061753	0.802575	0.233962	0.357148
Student club organization	-0.09372	0.589141	-0.03898	0.826771
Professional society off-campus overall	0.161196	0.488612	0.208923	0.37564
Professional Experience	0.568262	0.105923	0.605836	0.087757
Employed hours per week	0.167113	0.547988	0.263423	0.359636
Admission	-0.15871	0.652715	-0.07971	0.829376
EIT Times Taken	0.526723	0.272664	0.581394	0.236713

V. SUMMARY

Based on 200+ students' responses to senior exit survey, the study explored the influential factors of students' abilities to apply knowledge of statistics, mathematics, and engineering. The machine learning approach of random forest was employed in this problem of multiple classification. The variables of time spent at CPP towards obtaining a degree turned out to be the most important in influencing the students' abilities at concerned subjects. This was followed by the time of sitting for EIT exam and the engagement of students in different clubs. Clearly, as per the self-assessment of students, their time devoted at CPP and their self-confidence in undertaking the EIT significantly influence their perception of statistics, mathematics and engineering knowledge. This study demonstrated that the random forest approach may be beneficial in such analysis and may be combined with the conventional methods such as multinomial logit regression to devise more informed strategies by the faculty and administration to improve student outcomes.

ACKNOWLEDGMENT

The authors are indebted to all participating students' responses to senior exit survey conducted by the Civil Engineering Department of CPP.

REFERENCES

[1] I. Lawrence, M. Meiers, and A. Beavis, "Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes & efficacy." Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, 13, 2005.

[2] S. B. Robbins, K. Lauver, H. Le, D. Davis, R. Langley, and A. Carlstrom. "Do psychosocial and study skill factors predict college outcomes? A meta-analysis." Psychological bulletin, 130(2), p.261, 2004.

[3] Benz, Michael R., L. Lindstrom, and P. Yovanoff. "Improving graduation and employment outcomes of students with disabilities: Predictive factors and student perspectives." Exceptional Children 66, no. 4, 509-529, 2000.

[4] Breiman, L. "Random forests." Machine learning, 45(1), 5-32, 2001.

[5] Ridgeway, G. "Generalized Boosted Models: A guide to the gbm package." Update, 1(1), 2007.

[6] Nelder, J. A., and Baker, R. J. "Generalized linear models" John Wiley & Sons, Inc, 1972.

[7] C. Kwak, and A. Clayton-Matthews, "Multinomial logistic regression" Nursing research, 51(6), 404-410, 2002.

[8] R. D. Retherford, and M. K. Choe, "Multinomial logit regression." Statistical Models for Causal Analysis, 151-165, 2011.

[9] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests." Pattern Recognition Letters, 31(14), 2225-2236, 2010.

[10] R. V. Diaz-Uriarte, R. Diaz-Uriarte, R. Diaz-Uriarte, and R. Diaz-Uriarte. "Variable selection using random forests." R Package Version 0.7-3, 2014.