Predicting first-year engineering student success:
from traditional statistics to machine learning

Ramaravind Kommiya Mothilal
Research Intern
Microsoft Research Lab, Bangalore, India
raam.arvind93@gmail.com

Tinne De Laet
Head Tutorial Services Engineering Science
Leuven Engineering and Science Education Center (LESEC), KU Leuven
Leuven, Belgium
tinne.delaet@kuleuven.be

Tom Broos
PhD student
Leuven Engineering and Science Education Center (LESEC), KU Leuven
Leuven, Belgium
tom.broos@kuleuven.be

Maarten Pinxten
Research Manager
Leuven Engineering and Science Education Center (LESEC), KU Leuven
Leuven, Belgium
maarten.pinxten@kuleuven.be

Conference Key Areas: Retention of students, Engineering Skills, Recruitment

Keywords: student success, retention, machine learning, first-year experience

INTRODUCTION

First-year student success in Engineering Bachelor programs is well-studied. Both traditional statistical modelling and machine learning approaches have been used to study what makes students successful. While statistical modelling helps to obtain population-wide patterns, they often fail to create accurate predictions for individual students. Predictive machine learning algorithms can create accurate predictions but often fail to create interpretable insights. This paper compares a statistical modelling and machine learning approach for predicting first-year student success. The case study focuses on first-year Bachelor of Engineering Science students from KU Leuven between 2015-2017 and relates first-semester academic achievement to prior education, learning and study strategies, effort level, and preference for time pressure.

1 GENERAL

First-year student success and retention in STEM and Engineering bachelor programs are well-studied. The relation between prior academic achievement, learning and study skills and strategies, self-efficacy, self-regulatory skills but also gender and socio-economic factors have been the subject of many studies. Against the background of the increasing need for skilled scientists and engineers, the heterogeneous inflow of incoming students in science and engineering programmes is particularly challenging in universities with an open-admission system. These universities are therefore looking for predictors of first-year student success that might be used in advising on the one hand, and remediation on the other hand. While population-wide conclusions provide insight in the factors that are key, predicting individual student success allows to take the next step to advising individual incoming students.

The present study first applies statistical modelling to obtain population-wide insights on what is expected from the students to perform well in the first semester of their first year of an open-admission university KU Leuven in Belgium. Good performance in the first semester reflects a smooth transition to higher education and acts as an early indicator for completing the study program. Secondly, the predictive validity of the obtained statistical models was assessed. Finally, a dedicated machine learning approach was used for prediction. Such algorithms can fit complex non-linear relations between the independent variables (here: prior academic achievement and learning skills) and the dependent variable (first-year student success).They often result in high predictive performance but challenge the interpretability of the predictions: i.e. it is not easy to see which independent variables are mainly contributing to the predicted class. Getting interpretable insights from the prediction is however key for both individual predictions and population-wide analyses. Firstly, when the predictions are used for academic advising, one should understand the factors that cause a particular prediction: e.g. what causes this particular student to be at risk? Secondly, the machine learning approaches might discover non-linear relationships at the population level that were hard to hypothesize, but that should be interpretable in order to be usable. Different methodologies are available to create local interpretable approximations of the complex non-linear models.

This paper focuses on three research questions: RQ1: Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success?; RQ2: Can boosted trees more accurately predict first-year student success than logistic regression? RQ3: Can Local Interpretable Model-agnostic Explanations (LIME) [1] generate interpretable insights in the factors important for predicting first-year student success?

The paper of Pinxten et al. [2] is key prior work as it used statistical modelling, and multiple linear regression in particular, to investigate the relation between first-year student success in the same study program as targeted in this paper and using a similar dataset. Specifically, the influence of secondary school math level, math and science secondary school GPA, diagnostic test score, study strategies, and advice of the secondary school teacher board was researched. While this paper focuses on a similar problem and dataset, it focuses on predicting rather than explaining first-year student success. We refer to this paper for an elaborate literature survey on the factors that are important for first-year STEM student success.

Like the study of Ackerman et al. [3], [2] used explanatory statistical modelling, including linear and logistic regression, to study the factors that impact STEM retention and academic success. Another recent study on predicting academic success in Belgium used linear and logistic regression but as [3], they erroneously infer predictive power from explanatory power of their variables and use an in-sample accuracy to assess the predictive validity rather than a test set or cross-validation. While no doubt is raised on the theoretical validity of the studies discussed above, only their statistical objectives are criticised, so as not to come at incorrect scientific and practical conclusions.

Lin et al. [4] compared four different predictive models, namely, artificial neural network (ANN), logistic regression, discriminant analysis, and structural equation in predicting student retention using cognitive and non-cognitive data, where ANN was found to outperform other models. While the authors correctly evaluated the predictive performance on a test set, they concluded their research by suggesting the following: “The model results can also be used to provide faculty and advisors with informed course selection advice to first-year engineering students”, without however handling the interpretability of their models. The studies [5] & [6] (linear regression) and [7] (ensemble methods) soundly assess their models’ predictive power in predicting student success and retention while also addressing the interpretability issues.

2 MATERIALS AND METHODS

2.1 Available data

Data was collected of first-year Bachelor of Engineering Science students of KU Leuven in two academic years: 2015-2016 and 2016-2017 (N=811). The independent variables (IV) were collected using a paper-and-pencil questionnaire administered in the first week of the academic year. Two sets of IV operationalize “prior academic experience”: (1) *math, phy, chem:* math, physics, and chemistry grades obtained in the last year of secondary education self-reported using the following categories: 60%, 60-70%, 70-80%, 80-90%, and above 90%. (2) *hrs:* numbers of hours of mathematics per week in the curriculum of the last year of secondary education self-reported using the following categories: low (<6 hours), (6 or 7 hours), and high (8 hours). Regarding soft-skills the following data was collected: (1) *eff:* The effort level in secondary education is self-reported using the question “How frequently did you study to obtain your math and science results in the final year of secondary education?” using five categories (very low, low, average, high, and very high); (2) *mot, time, conc, anx, and test:* Five scales (motivation (mot), time management (time), concentration (conc), performance anxiety (anx), and the use of test strategies (test)), were used from the Learning and Study Strategies Inventory (LASSI) to assess the students’ learning skills (resulting in 30 questions). The internal consistency coefficients (Cronbach’s alpha) were: mot 0.77, time 0.76, conc 0.84, anxi 0.84, and test 0.71. These values are in accordance with both the standards provided in the user’s manual and with general standards [8]; (3) *press:* The preference for time scale of Choi and Moran [9] was used. Press was standardised and discretized into a 4-category ordinal variable with categories “Low”, “Low_medium”, “High_medium”, and “High.

The dependent variable (DV) *GPA* operationalizing first-semester academic achievement (AA) was collected from the universities data warehouse. The GPA, between 0 and 20, is the weighted average of the grades on the first semester courses obtained before the resit, weighted with the ECTS credits of each course.

2.2 Methodology

The goal of **explanatory modelling** is to discover patterns IVs (student characteristics) and the DV (AA). Based on literature, three hypotheses were formulated:1) “Prior academic experience positively AA.”, 2) “Affective and goal strategies positively affect AA.”; and 3) “Preference for time pressure does not affect AA.” Multiple linear regression was used to investigate these hypotheses.

Table 1 provides an overview of the explanatory models built. A sequential (or hierarchical) multiple regression (model 3) was built to test whether the variables affe, goal, press, and eff have a significant incremental explanatory power in explaining AA. To assess the **predictive validity** of the explanatory model for predicting AA, ordinal logistic regression was used to classify students in three groups at-risk (GPA≤8.5), middle group, and no-risk (GPA>11.5). For logistic regression a cumulative odds model with proportional odds property was used.

Table 1: Explanatory models

	model	**regression type**
1	wavg ~ math+phy+chem+hrs	standard
2	wavg ~aff + goal + press	standard
3	wavg ~ math+phy+chem+hrs + aff + goal + press + eff	sequential

Gradient boosting was used for **predictive modelling.** It is an ensemble technique where new decision trees are sequentially added to compensate the errors made by the already existing trees, using gradient descent. The final prediction score is the sum of the prediction scores of each individual tree. XGBoost, which is based on the gradient boosting, incorporates various additional improvements to avoid overfitting, to handle sparsity pattern in data, and parallelization [10]. Instead of learning an multi-class classifier, two independent classifiers were used, as multiple binary classifiers are often easy to train and optimize than a single multinomial classifier. The first classifier, XGBoost_1, was used to distinguish students of class “≤8.5” (“at-risk”) from students of class “>8.5” (“moderate-risk” or “no-risk”) while the second classifier, XGBoost_2, was used to distinguish students “>11.5” (“no-risk”) from students of class “≤11.5” (“at-risk” or “moderate-risk”). The output of the two binary-classifiers model were combined into a multi-class classification (*Table 2*). In this paper XGBoost_1 and XGBoost_2 are optimized for high recall of ≤8.5 and ≤11.5 respectively (high identification rate of the “at risk”- students) and high precision of >8.5 and >11.5 respectively (few misclassification as “no-risk”). Once the two binary classifiers were learnt, LIME was used to construct interpretable approximations of the classifiers. LIME constructs “textual or visual artefacts that provide qualitative understanding” of the prediction. LIME, in short, approximates a complex classifier in the neighbourhood of an observation (for which prediction is required) with a simple interpretable model like linear regression.

All classifiers in this study are trained using a training set of 542 out of 720 subjects and evaluated using a test set of 178 out of 720 created using stratified sampling.

Table 2: Combining outcome of two binary classifiers into multi-class classification. (other combinations, which rarely occur, are classified as moderate at-risk)

XGBoost_1 outcome	XGBoost_2 outcome	interpretation
≤8.5	≤11.5	at-risk
>8.5	<11.5	moderate at-risk
>8.5	>11.5	no-risk

3 RESULTS

3.1 Data preparation

As the five measured learning and studying skills are correlated, PCA was performed (after checking for the assumptions of linearity and removing outliers) to reduce the dimensionality of the data. PCA with oblique (oblimin with δ=0) rotation yielded a two component solution accounting for 75% of the variance. Since the resulting association between the measured variables and the components (*Table 3*) is similar to that obtained by Cano [8], the same naming convention is used: “Affective Strategies” (*aff)* “Goal Strategies”(*goal)*. The variables *affe* and *goal* are discretized into a 4-category ordinal variable and the categories are named as “Low”, “Low_medium”, “High_medium”, and “High.

Table 3: PCA component loadings. Proportion variance explained by TC1 is 0.44 and by TC2 is 0.32. Pearson correlation: 0.17. RMS of residuals was 0.1.

	TC1	TC2	h^2	
mot	*0.88*	-0.15	0.75	aff
time	*0.87*	-0.04	0.75	
conc	*0.73*	0.36	0.74	
anxi	-0.17	*0.92*	0.83	goal
test	0.31	*0.76*	0.75	

3.2 Explanatory modelling

For ***model 1*** a significant regression equation was found (F(11,708) = 37.97 at p< $2.2e^{-16}$) , with an R^2 of 0.37. (adjusted multiple R^2 value 0.36). All included independent variables were significant contributors (p<0.05). The unique contribution of math, phy, chem and hrs is computed as 6.6%, 2.3%, 4.5%, and 3.0% respectively. The remaining 20.8% variance is contributed by two or more input variables. These results confirm that prior academic experience, operationalized as math and science grades, and math level, positively affect the first-semester academic achievement.

For ***model 2*** a significant regression equation was found (F(9,710) = 4.644 at p< $5.4e^{-16}$), with an R^2 of 0.06. The adjusted multiple R^2 value of 0.04 shows that soft skills are only a weak predictor of AA. Most variance explained by model 2 is contributed by affe, while goal and press together contribute less than 1%. Similarly, only the coefficients of affe are significant. These results show that students’ affective strategies influence AA, while goal strategies don’t.

The results of the sequential model (***model 3***) show that after accounting for differences in students’ effort level, the goal related strategies (but not affective strategies or pressure preference) become important in explaining AA.

3.3 Explanatory modelling with predictive validity

Logistic regression according to ***model 1*** showed that prior academic achievement is a strong predictor of AA (Nagelkerke R^2=0.28). Preference for time pressure and affective and goal strategies on the contrary are only week predictors (***model 2***, Nagelkerke R^2=0.05), they are however of added predictive value on top of prior academic achievement (***model 3***, Nagelkerke R^2=0.33). *Table 4* presents the detailed prediction outcome of model 3. This model, involving all IV, has a recall of 60% for the class "≤8.5", which implies a high risk of not identifying "at-risk" students. Similarly, the precision of 41% for the class ">11.5" implies the high risk of misclassifying many students as "no-risk". The F1 score, which is an overall measure of the predictive performance, is only 62% for the "at-risk" class and 43% if the "no-risk" class.

Table 4: Prediction outcome of logistic regression with wavg ~math + phy + chem + hrs + eff + affe + goal + press (model 3)

	precision	**recall**	**F1-score**	**support**
≤8.5	0.63	*0.60*	0.62	60
8.5-11.5	0.63	0.59	0.61	54
>11.5	0.41	*0.45*	0.43	64

3.4 Predictive modelling with comprehensible approximations

Table 5 shows the predictive performance of both XGBoost_1 and XGBoost_2 using all IV. The high recall of XGBoost_1 ≤8.5 shows that most at-risk students are correctly identified. The overall F1-score of 71% for the "at-risk" group (XGBoost_1 ≤8.5) and 69% for the "no-risk" group (XGBoost_2) shows that the overall predictive performance is adequate.

Figure 1 shows an example of the LIME output. Firstly, the prediction probabilities are shown. Next, a bar chart shows the importance (or weights) of each of the IV in the prediction in descending order. The "weight" in the bar chart indicates that if an IV does not take the displayed value, then the probability of the displayed class to which the IV is contributing will be on average be reduced in value equal to the weight. As such, the LIME explanations allow to assess to which level a IV contributes to the predication made: e.g. what are the aspects that make the student at-risk. Population-wide patterns can be discovered by listing how often the IVs occur in the prediction of students belonging to the different classes. Similar to the explanatory modelling this shows that IVs math, phy, chem and eff contribute more to the predicted probabilities than IVs affe, goal, or press. Additionally, some patterns, previously not discovered in explanatory modelling, can be observed here. For instance (*Figure 2*), (1) while hrs contributes to distinguishing at-risk students (≤8.5) from the rest, it is not key in distinguishing "no-risk" (>11.5) students from the rest; and (2) affe considerably contributes to distinguishing "no-risk" students from the rest, while it does not significantly contribute to distinguishing "at-risk" students from the rest.

Table 5: Prediction outcome of predictive performance with boosted trees.

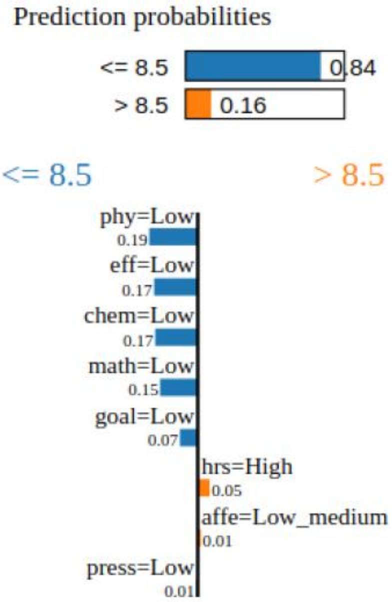
		precision	**recall**	**F1-score**	**support**
XGBoost_1	**≤8.5**	0.64	*0.80*	0.71	60
	>8.5	*0.88*	0.77	0.82	118
XGBoost_2	**≤11.5**	0.87	*0.85*	0.86	124
	>11.5	*0.68*	0.70	0.69	54

4 DISCUSSION & CONCLUSION

The discussion is organized around the three research questions.

Regarding RQ1, we found that the both statistical modelling (linear and logistic regression) and boosted trees create the same insights regarding factors important for first-year student success: prior academic achievement, affective strategies, and goal strategies affect first-semester academic performance. This confirms the findings of [2], which showed that the students' motivation/persistence, concentration, and time management skills significantly influenced first-year student achievement although the incremental value over prior achievement was small.

The results confirm that students' preference for time pressure has no influence on their academic performance as suggested by Choi and Moran [9].



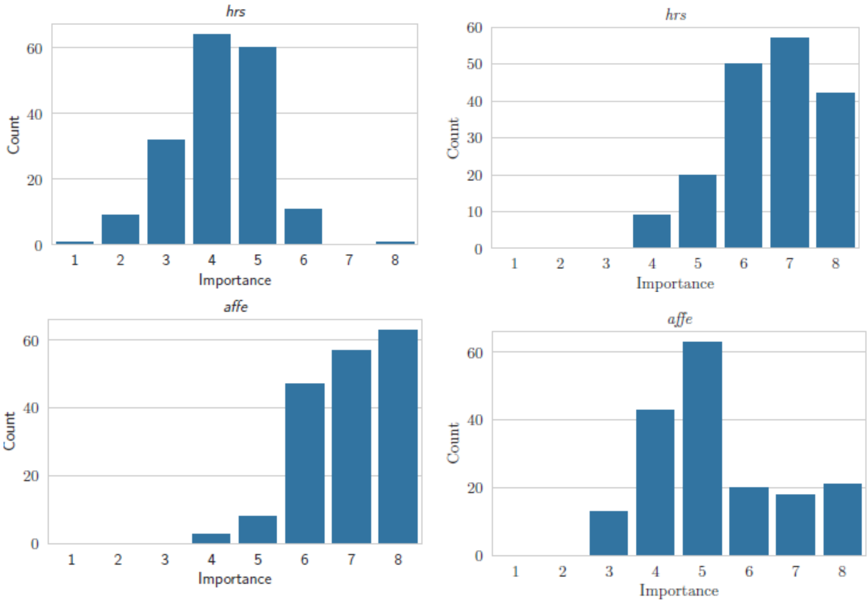


Figure 1: LIME output for at-risk student that was also predicted as at-risk.

Figure 2: Population-wide analysis of importance of IV hrs and aff for predicting at-risk students (left, XGBoost_1, ≤8.5) and no-risk students (right, XGBoost_2, >11.5).

RQ2: Can boosted trees more accurately predict first-year student success than logistic regression?

The results show that boosted trees can more accurately predict AA, providing a positive answer to RQ2. Boosted trees outperform more-commonly used logistic regression, as both precision and recall increased by more than 20%. Additionally, we have shown that LIME can be used to get interpretable insights from the boosted trees model both for predicting individual student success as for studying population-wide patterns.

Future work should focus in including more IV such as recommendation by the secondary school teacher board [2], study effort [2], and positioning test score [11]. Furthermore, future work should not only focus on predicting first-year students success but also on long-term success, and explore different machine learning approaches to this end. The approach should be tested to engineering programs of other universities. Finally and most importantly, the usability of the interpretable models (LIME) for advising students should be assessed.

ACKNOWLEDGMENTS

This research is co-funded by the Erasmus+ program of the European Union (562167-EPP-1-2015-1-BE-EPPKA3-PI-FORWARD).

REFERENCES

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, “&quot;Why Should I Trust You?&quot; Explaining the Predictions of Any Classifier,” in *KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[2] M. Pinxten, C. van Soom, C. Peeters, T. de Laet, and G. Langie, “At-risk at the gate: prediction of study success of first-year science and engineering students in an open-admission university in Flanders—any incremental validity of study strategies?,” *Eur. J. Psychol. Educ.*, 2017.

[3] P. L. Ackerman, R. Kanfer, and M. E. Beier, “Trait complex, cognitive ability, and domain knowledge predictors of baccalaureate success, STEM persistence, and gender differences.,” *J. Educ. Psychol.*, vol. 105, no. 3, pp. 911–927, 2013.

[4] J. J J Lin, P. K. Imbrie, and K. Reid, “Student Retention Modelling: An Evaluation of Different Methods and their Impact on Prediction Results,” *2009 Res. Eng. Educ. Symp. REES 2009*, 2009.

[5] J. Burtner, “The Use of Discriminant Analysis to Investigate the Influence of Non-Cognitive Factors on Engineering School Persistence,” *J. Eng. Educ.*, vol. 94, no. 3, pp. 335–338, Jul. 2005.

[6] B. F. French, J. C. Immekus, and W. C. Oakes, “An Examination of Indicators of Engineering Students’ Success and Persistence,” *J. Eng. Educ.*, vol. 94, no. 4, pp. 419–425, Oct. 2005.

[7] A. Essa and H. Ayad, “Student success system: risk analytics and data visualization using ensembles of predictive models.,” in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK ’12*, 2012, p. 158.

[8] F. Cano, “An In-Depth Analysis of the Learning and Study Strategies Inventory (LASSI),” *Educ. Psychol. Meas.*, vol. 66, no. 6, pp. 1023–1038, Dec. 2006.

[9] J. N. Choi and S. V. Moran, “Why Not Procrastinate? Development and Validation of a New Active Procrastination Scale,” *J. Soc. Psychol.*, vol. 149, no. 2, pp. 195–212, Apr. 2009.

[10] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*, 2016, pp. 785–794.

[11] J. Vanderoost, R. Callens, J. Vandewalle, and T. De Laet, “Engineering positioning test in Flanders : a powerful predictor for study success ? Conference Topic : The Attractiveness of Engineering ; Education al Research Methods INTRODUCTION,” in *Proceedings of the 42nd Annual SEFI conference*, 2014, pp. 1–8.