To What Extend Can We Predict Students' Performance? A Case Study in Colleges in South Africa

Norman Poh
Department of Computing
University of Surrey, UK
normanpoh@ieee.org

Ian Smythe
Do-IT Solutions Ltd, UK
ian@doitprofiler.com

Abstract

Student performance depends upon factors other than intrinsic ability, such as environment, socio-economic status, personality and familial-context. Capturing these patterns of influence may enable an educator to ameliorate some of these factors, or for governments to adjust social policy accordingly. In order to understand these factors, we have undertaken the exercise of predicting student performance, using a cohort of approximately 8,000 South African college students. They all took a number of tests in English and Maths. We show that it is possible to predict English comprehension test results from (1) other test results; (2) from covariates about self-efficacy, social economic status, and specific learning difficulties there are 100 survey questions altogether; (3) from other test results + covariates (combination of (1) and (2)); and from (4) a more advanced model similar to (3) except that the covariates are subject to dimensionality reduction (via PCA). Models 1-4 can predict student performance up to a standard error of 13-15%. In comparison, a random guess would have a standard error of 17%. In short, it is possible to conditionally predict student performance based on self-efficacy, socio-economic background, learning difficulties, and related academic test results.

1. Introduction

1.1. Motivations

Context: Online educational platforms for education and assessments have generated an unprecedented volume of data that can be used to improve a student's learning experience. Today, it is common to have traditional classroom education supplemented by an online education platform. Indeed, such an online platform is often deemed beneficial to students [8]. It is fair to describe the new generation of students who live in the era of Facebook, Twitter, and YouTube, as being more receptive to the use of media for learning than more traditional media such as printed reference books. Indeed, education as a service can be delivered in the form of massively open online courses (MOOCs) such as https://www.coursera.org, https://www.udemy.com, and https://www.khanacademy.org, to name a few[1]. In this study, we would like to investigate the possibility of predicting student's performance, in the context of using an online assessment platform.

A call of urgency in South Africa's context: In South Africa, the concern focuses on the impact of the country's aim to drive its economic expansion by increasing productivity. However, that can only happen if there are enough qualified workers to fill the needs of industry. Furthermore, unemployment rates have reached over 25% [14], and over 30% for those without qualifications, rising to 35% for those who want work but are not actively seeking it (classified as expanded unemployment rate). Thus it is not only the concerns of the colleges and the employers that are in question, but also the impact on the individual and society of so many students failing to stay in education. Indeed, according to the latest available statistics [7], retention rates in Further and Education Colleges (FETs) is 61%, whilst certification rates in 41%. In the area studied in this paper (Gauteng Province) the figures were 64% and 38% respectively.

The impact of markers (variables) frequently used in studying education retention via data mining is context-dependent. Thus the traditional markers of socioeconomic status such as parental education, household income, and parent occupation [4] are not relevant in the poorer neighbourhoods of Johannesburg. Other metrics such as social integration [15] also require significant adaptation to the Gauteng context.

Motivations for performance prediction: At the state funded college level, there are at least two reasons why one should predict student's performance. First, by doing so, a teacher can identify potentially weak students or those who are at risk of falling out. Indeed, there is evidence that intervening early could improve literary skills in young children [5]. It is, therefore, highly plausible that identifying weak students and taking relevant actions could improve persistence and academic outcome. Second, an academic predictive model could help admission tutor to sift through a large number of candidate applicants to higher education. This application scenario was considered by [11] who used cognitive (grades-based) as well as non-cognitive measures for student admission.

In this paper, we shall first survey a number of theoretical models that attempt to explain factors that influence student's performance. Then, we will present a data set that can be used to study the feasibility of predicting thier academic performance.

[1] More listed on http://en.wikipedia.org/wiki/Massive_open_online_course

1.2. Literature review

A number of studies have reported on the feasibility of predicting student's performance in different contexts.

Chemer, Hu, and Garcia's assertion on academic self-efficacy Chemer, Hu, and Garcia [1] conducted a longitudinal study of first year university students. They subsequently proposed a predictive model that includes high school performance (measured in terms of grade-point average, GPA), academic self-efficacy, and optimism as the main predictor variables, whilst adjusting for a number of moderator variables (covariates) that represent the specific characters of each student. The authors' model includes three predictor variables and a number of adjusted variables or covariates. The three predictor variables are self-efficacy, optimism, and student's past performance, as measured by his/her high school grade-point average (GPA). The covariates include academic expectations and student's perceived ability to cope with problems and challenges (termed challenge and threat evaluation). They found that academic self-efficacy and optimism were strongly related to performance and adjustment, both directly on academic performance and indirectly through expectations and coping perceptions (challenge-threat evaluations) on classroom performance, stress, health, and overall satisfaction and commitment to remain in school.

Rossi and Montgomerys alternative model of student's performance: Rossi and Montgomery [10] conjectured that the contextual environments such as student's personal, home, community, and school characteristics may also influence student's performance. They argued that a student's personal, home, community, and school characteristics should not be studied in isolation; all these variables contribute to student's performance in an interdependent manner. Recognising these interactive dynamics, they proposed a theoretical model that explains the variety of reasons that some students fail whilst others succeed. Rossi and Montgomery argued that academic progress is primarily an ongoing function of two factors. The first one is the quality of student's resources. These include abilities, family support, and educational opportunities. The second one is the students perceived incentives and pressures to invest these resources in academic achievement. In this view, past "returns" on educational investments will influence on student's competency and aptitude to succeed academically and to persist in school.

Schmitt's noncognitive measures: Schmitt *et al.* [11] postulated that a number of personal measures that are not related to ability henceforth called non-ability measures may also influence the success of four-year undergraduate academic performance of college applicants. Examples of non-ability measures considered are personality, motivation, and experience measures.

Bayesian knowledge tracing and contextual guess and slip model: Reye *et al.* [9] proposed to use a Bayesian Network in order to continuously assess a student's current understanding of each skill. When a student first attempts a question, their model will assess the initial knowledge of the student. For the subsequent attempts of problems related to the same skill, their model is able to assess whether or not a student has guessed the correct answer or the student has acquired the correct understanding in terms of probability. When the student makes a mistake despite having already learnt the skill, their model characterises this by "slippage". Reye *et al.* called their model has Bayesian Knowledge Tracing. They reported that the ability to assess the students latent knowledge enables the teacher to tailor the amount of practice each student receives. They claimed that this approach can significantly improve student's learning outcomes.

Narrow traits versus the big-five traits: Lounsbury *et al.* examined narrow traits in addition to the Big Five in predicting academic success among adolescents [6]. They investigated individual grade point average (GPA) and scores from the Adolescent Personal Style Inventory among 220 seventh-graders and 290 tenth-graders. They measured aspects such as agreeableness, conscientiousness, emotional stability, extraversion, and openness. In addition, they measured four narrow traits, namely, aggression, optimism, tough-mindedness, and work drive. All traits correlated significantly ($P < .01$) with GPA among both 7th- and 10th-graders. The Big Five traits together accounted for 15% and 10% of variance in GPA among 7th and 10th graders, respectively. In comparison, and consistent with prior research, narrow traits accounted for 8% and 12% of the variance in GPA, respectively, over and above those that are predicted by the Big Five. Based on these results, Lounsbury *et al.* concluded that there is a clear relationship between personality and academic success among adolescents.

On cultural difference: Motivated by the fact that most previous studies on college students' performance have focused on conventional learning environments in Western cultures, Cheung and Kan [2] evaluated factors related to student's performance in the open and distance-learning environment in Hong Kong. Using two-way cross-tabulations with chi-square testing and equality of academic performance by proposed factors, the authors examined 168 students in a distance-learning business communication course. Results show that tutorial attendance, gender, relevant academic background, previous academic achievement, and relevant learning experience are related to student's performance. The results are mostly similar to those of prior studies despite differences in culture, teaching mode, and subject.

Context in South Africa: However, all these models are developed through research and analysis of mature learning environments. In South Africa, the socioeconomic factors are more likely to include if the student studied by candlelight – which explains 8.3% of variance [13] – whilst social integration could include how far one has to walk home and how many dependants one cares for.

1.3. Our Approach and Contributions

In the above studies, there is near-exclusive use of statistical models for causal explanation. Implicit to these studies is the assumption that the models with high explanatory power are inherently of high predictive power. Shmueli [12] argued that while "conflation between explanation and prediction is common, the distinction must be understood for progressing scientific knowledge". We have therefore opted for a direct prediction of student results, using variables that are deemed important from the above studies. By prediction, we mean that we shall train our model on one cohort of students and use the model to predict the academic performance of a held-out set of students. The arguments for predictive analysis are debated and substantiated at length by Shumueli [12].

In order to strike a balance between explanatory modelling and predictive modelling, we have chosen to use white-box models that allow us to interpret the results. However, rather than choosing only a handful of variables or computing proxy variables, e.g., deriving an intermediate variable that represents socio-economic status, or one variable that represents academic self-efficacy, we have chosen to fit our statistical model with all the available raw variables, which is fairly large in dimension. In order to reduce the number of variables, we use regularisation, also known as lasso or L1 penalisation. We can then interpret the variables with non-zero coefficients because they are actually used in predicting examination results.

In summary, our contributions can be summarised as follow: First, we show that it is possible to predict student's performance from a relatively large cohort of approximately 8,000 college students. The predicted student's results are found to be a lot better than random guessing. Second, we find that by using white-box (interpretable) models with regularisation, it is possible to verify that the retained variables with non-zero coefficients are reasonable and conform to the literature.

2. Methodology

In order to find out how well one can predict a student's academic performance, we shall analyse the examination results from Further Education and Training (FET) College students from Gauteng, South Africa as a case study.

2.1. Context

The students were those studying at Gauteng FET colleges in 2012-2013. They were invited to be assessed using the Do-IT Profiler, an online student profiling tool that collects data with respect to demographics, socio-economics as well as performing academic and cognitive evaluations [13]. Students were not preselected, but those who had completed the task on the date of data collection. There are eleven national languages in South Africa, and the cohort represented all of them. However, the reliability of that data is questionable since many claim that English is their first language when it is not for reasons of social status. Hence at this stage, language is not a variable analysed.

Covariates: The data collected can be categorised into three parts:

1. Demographics – including college location, age, and dependants;
2. socioeconomic – containing information such as employment status of care-giver, quality of housing, and nutrition; and,
3. Study progress – capturing information such as feeling about success/failure, and study trends

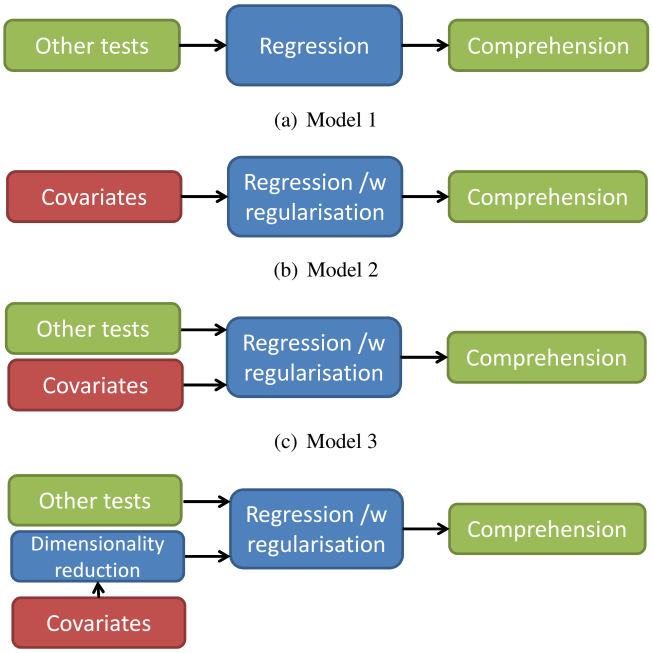
Academic and cognitive assessments: The following series of online academic and cognitive assessments were used.

Apart from the above assessment results, students also answered a number of survey questions about their specific learning difficulties, study skills, and socio-economic backgrounds. We refer to these variables as covariates or exogenous variables because they are not assessment results, i.e., they are not in the unit of marks; but are in binary representation.

2.2. Modelling

We shall define the underlying regression problem to be one that predicts the English Comprehension examination result due to its all inclusive nature. There are two categories of predictor (independent) variables, namely the seven remaining examination results; and a list of covariates mentioned above. With these two categories of variables, we can employ regression models, as shown in Figure 1.

While each of models 1 and 2 takes two categories of variables separately, models 3 and 4 consider these variables jointly. Because the number of covariates is very large (100 variables), we have to employ regression with regularisation for model 2 whilst this is not necessary for model 1. Tables 2 and 3 list some of the variables found by regularisation. Note that we model the value of the survey responses directly, i.e., each variable that is fed to the regression model is one of the possible response values. Therefore, one survey question may have several responses which are represented by several binary variables that are given as input to the regression model.



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 1. Four different models. "Other tests" is a real-valued vector of 7 dimensions whereas "Covariates" is a vector of 100 binary elements.

Compared to Model 3, Model 4 is different mainly due to the use of Principal Component Analysis as a means to reduce the large dimension of the covariates. As a result of this dimensionality reduction step, Model 4 is expected to generalise better (can do better prediction) but doing so at the risk of not being able to interpret the coefficients.

2.3. Inclusion/exclusion criteria

While there are a total of 11,104 students taking at least one test, not all students have completed all the tests. Our inclusion criteria are that (1) all students have to have completed all the tests at the time of data extraction; and (2) the results must not be zero. The first criterion results in 7,995 students being retained whereas the second further excluded 6 students. Therefore, a total of 7,989 are used for subsequent analyses.

For a more rigorous comparison of predictive performances, it is common practice to divide the data randomly into a training and a test set of equal proportion. While the training set is used to train a regression model, the test set is used uniquely to test the generalisation performance of the trained model. This results in a training set containing 3,994 students whereas the test set containing 3,995 students. Of course after such comparison, the trained model is qualitatively validated for its explanatory power.

Table 1. The eight academic and cognitive tests used for assessing the FET college students.

Reading	**Vocabulary test** – Students have to match a picture with one word chosen from a list. This a form of reading vocabulary test **Cloze procedure** – Students have to choose the correct word. An example of question is "Fill in the missing word: There are seven days in a {week / well / weak / wheel}. **Comprehension test** – students are presented with short ten stories, each of which has two questions and four multiple choice answers. **Word choice** – Students are presented with visually similar words and have to choose the correct word in the right context, e.g., Monk versus Munk.
Spelling	**Spelling test** – A brief test of spelling of words and non-words
Sound discrimination	**Sound discrimination test** – Students have to listen to two words and identify if they are the same words or different, e.g., "and" versus "end"
Maths	**Arithmetic test** – Students have to use the four operands **Maths Problems** – Students are presented with word-based maths questions

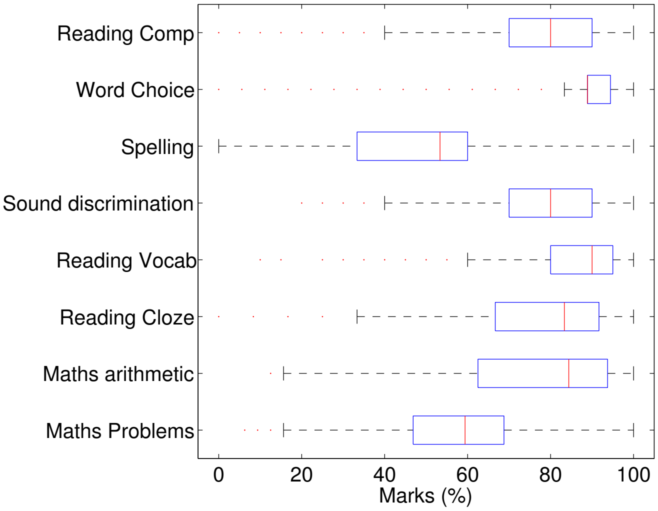


Figure 2. The distribution of subject marks

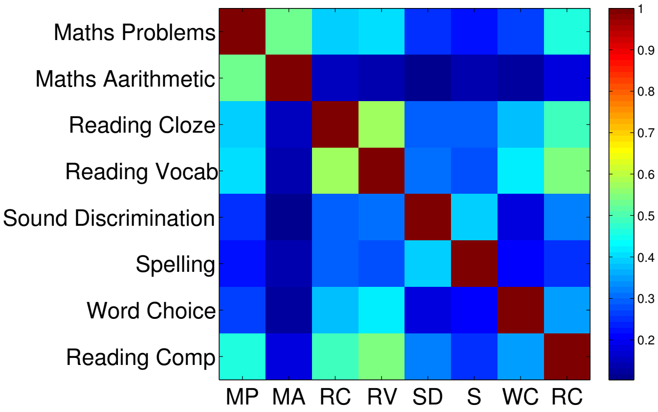


Figure 3. The pair-wise correlation of various test marks. An element in the X-axis has a corresponding element in the Y-axis, i.e., MP=Math Problems, MA=Maths Arithmetic, and so on until RC=Reading Comprehension.

3. Experimental results

We shall report three experiments: (1) preliminary analyses in order to characterise the data distribution and correlation; (2) prediction performance; and (3) coefficient-level analysis.

3.1. Data characteristics

In order to describe the examination marks, we show two diagrams: (1) the marginal distribution of the examination marks, and (2) their correlation. The results are shown in Figures 2 and 3. Since the exercise being undertaken here does not involve prediction, all the 7,989 observations are used.

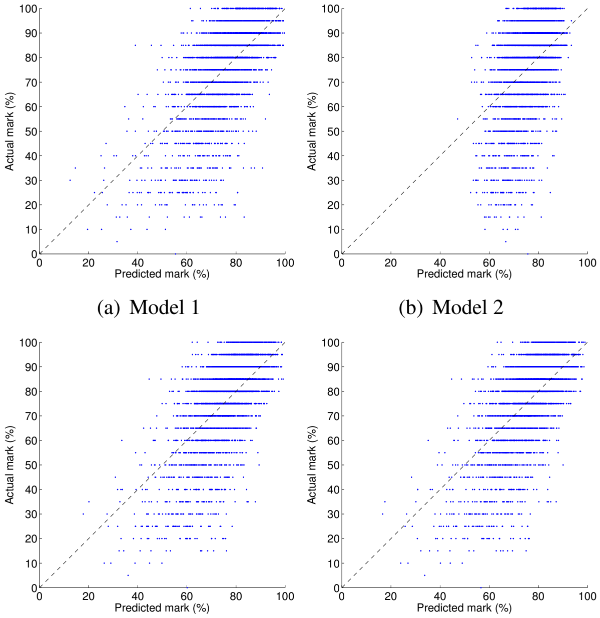
Each box in Figure 2 is a "boxplot". The bounding box surrounding a boxplot shows the first and the third quartiles of the examination marks, with the middle line showing the median mark. The dashed lines show the extend of the 5-th and 95-th percentiles of the marks. Finally, each dot shows an outlier beyond the interval spanned by 5-95 percentiles.

The correlation in Figure 3 corresponds well to our expectation. For instance, students who are good at arithmetic are also good at maths problems (and vice-versa). Another observation is that the examination marks of Reading Cloze, Reading Vocabulary, and Reading Comprehension are all correlated.

3.2. Prediction results

The next experiment examines the generalisation performance of the four regression models. The predicted versus the actual English Comprehension examination marks are shown in Figure 4. A diagonal line shows that identity line where the predicted and the actual examination marks are equal. A model that is better should follow the diagonal line closely. One can observe that Model 2 does not predict the examination mark as well as the three other models. This is reflected by their predicted error residuals in terms of expected standard deviations, which are ±13.15%, ±14.99%, ±12.80%, and ±12.74 respectively of each of the four models. The last model, i.e., Model 4 achieves the best generalisation performance but this is only slightly better than Model 3. While Model 3 can be interpreted, Model 4 cannot. Therefore, the better academic prediction may come at the loss of model interpretability. However, the performance gain is not significant in this case.

Since a perfect regressor should have an expected standard error of zero, it is arguable that the four models have not fared very well. However, these errors should be compared with the best random guess; and the best guess is to opt for the mean value, which is



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure 4. The predicted versus the actual English Comprehension test results using four different models, as measured on the held-out, unseen test set. The error residuals of these methods – from figures (a) to (d) – are ±13.15%, ±14.99%, ±12.80%, and ±12.74%, respectively. In comparison, the expected error residuals due to the best guess, i.e., always giving the mean value of the result, has an error residual of ±17.15%. Therefore, the conditionally predicted English Comprehension test outcome is better than the best random guess.

76%. This is to say, for any student, one systematically guesses that he/she will invariably obtain 76%. The expected error residual due to this best guess attempt is ±17.15%. This is much higher than those predicted by the models (i.e., ±13.15%, ±14.99%, ±12.80%, and ±12.74 for each of the four models). Therefore, we can conclude that our regression model performed better than the best random guess.

3.3. Coefficients analysis

As a final experiment, we would like to investigate how the covariates or factors affect students' performance (in terms of English comprehension) without the additional influence of the results of "the other tests". Among the four models investigated, only Model 2 satisfies this criterion. We should mention that Model 3 is best used for predicting students' performance because by considering the students' other tests, it produces a more accurate prediction than Model 2. However, our exercise here is not one of prediction; but one of understanding the influencing factors.

Coefficients with negative values are listed in Table 2 whereas those with positive values are shown in Table 3.

In these two tables, the magnitude of large coefficients are significant because they can positively or negatively influence the resultant predicted examination marks. For instance, the first coefficient in Table 2 has a negative value of 0.08092. This means that the corresponding variable – studying by street light – leads to a reduction of 8.09% marks. The second row shows that if a student feels that he is likely to fail, i.e., low academic self-efficacy, 3.8% marks are penalised from the predicted result. Similarly, the third row shows that if a student feels that his/her duties at home interfere with his/her studies, than 3.03% marks are penalised from the predicted mark. The responses appear to be reasonable.

As the coefficient values become smaller and smaller, the effect on the predicted mark is also becoming less and less important. For instance, it would be difficult to explain why eating potato chips before taking an exam or test can lead to a 0.79% of penalised mark[2]. However, since the magnitude of the coefficient is very small, this clearly indicates that this factor is not important.

The table with positive coefficients follow a similar trend. However, the most significant variables should be read from the bottom of the list. The first significant variable shows that having daily access to running water could add 3.08% marks to the predicted result. Asking teachers in order to clear up one's confusion also adds 2.37% to the predicted mark. Feeling positive about the future adds 2.17% to the predicted mark. This is consistent with academic self-efficacy. Having access to electricity adds 1.64%; and this is a good indicator of a good socio-economic status; so as the running water. Going up the list, the influence of the variables on the predicted mark becomes less and less important.

4. Discussions and conclusions

Although a number of literature has shown the feasibility of explaining student performance, few work has actually attempted to predict student performance. As a result, it remains unknown how well one can predict student performance. We use machine-learning techniques that are able to automatically identify features that are relevant. With the right model, it is possible to predict student's academic performance as well as explain the variables that are likely to be useful in the prediction. We achieved this by using linear regression models with regularisation.

The ability to predict student's academic performance will entail a number of potential implications. Such a predictive analytics can be integrated into an online assessment system so that educators can prioritise teaching for students whose performances are predicted to be low. In addition, it is also possible to actively introduce interventions that reduce the risk factors that cause poor academic performance. Although colleges can have little influence in certain socio-economic factors (e.g. employment of carers and those being cared, running water at home, access to electricity at home etc), this research suggests that simple, low cost interventions such as opening colleges at night for self-guided studies ("homework"), or running study skills courses (only 55% had any experience of study skills) could have a significant impact on the retention and qualification rates.

Acknowledgement

The authors would like to thank Dr Simon Lygo-Baker and Dr David Windridge for fruitful discussions.

References

[1] M. M. Chemers, L.-t. Hu, and B. F. Garcia. Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, 93(1):55, 2001.

[2] This line is not shown here, since the table lists only the top-ten variables sorted by the magnitude of coefficients.

Table 2. The top-ten most influential Model-2 variables with negative coefficients, sorted in ascending order.

Coef	Question	Description
-0.08092	Do you study by:	Street Light
-0.03779	How do you feel you are doing in your studies?	Failing
-0.03033	Do you feel your duties at home interfere with your studies:	All the time
-0.02704	What is the employment status of your parent or caregiver?	Student / learner
-0.02459	Do you feel your duties at home interfere with your studies:	Too much
-0.02163	Do you study:	With Family
-0.01615	How do you feel you are doing in your studies?	Feel like you are drowning
-0.01532	Where do you study for a test or exam?	I don't study for exams or tests
-0.01528	Do you take care of:	Parents
-0.01401	Do you feel:	That you will fail this year

Table 3. The top-ten most influential Model-2 variables with positive coefficients, sorted in ascending order.

Coef	Question	Description
0.00264	What time of the day do you study?	Later Evening (10pm - 12pm)
0.00332	Do you have daily access to:	Sufficient street lighting
0.00379	Do you have daily access to:	Toilet inside house
0.00418	Do you study by:	Sunlight
0.00433	When you study, do you:	Struggle to concentrate
0.00437	Do you have daily access to:	National / local newspaper
0.00440	When preparing for tests or exams, do you ask teachers / lecturers for help:	To work out past paper questions
0.00444	Is your area classed as a formal or informal settlement?	Formal settlement
0.00458	What time of the day do you study?	Evening (8pm - 10pm)
0.00464	Do you do any of the following before taking an exam or test:	Eat something small

[2] L. L. Cheung and A. C. Kan. Evaluation of factors related to student performance in a distance-learning business communication course. *Journal of Education for Business*, 77(5):257–263, 2002.

[3] M. S. DeBerard, G. Spielmans, and D. Julka. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal*, 38(1):66–80, 2004.

[4] E. R. Dickinson and J. L. Adelson. Exploring the limitations of measures of students' socioeconomic status (ses). *Practical Assessment, Research and Evaluation*, 19(1):1–14, 2014.

[5] R. H. Good III and R. A. Kaminski. Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly*, 11(4):326, 1996.

[6] J. W. Lounsbury, E. Sundstrom, J. L. Loveland, and L. W. Gibson. Broad versus narrow personality traits in predicting academic performance of adolescents. *Learning and Individual Differences*, 14(1):65–75, 2002.

[7] Mokwena. 2012 national examinations provisional report. In *Workshop*, Tshwane North FET College, 2013.

[8] J. O'Malley and H. McCraw. Students perceptions of distance learning, online learning and the traditional classroom. *Online journal of distance learning administration*, 2(4), 1999.

[9] J. Reye. Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, 14(1):63–96, 2004.

[10] R. Rossi and A. Montgomery. Educational reforms and students at risk: A review of the current state of the art. *Washington, DC: US Department of Education, Office of Educational Research and Improvement, Office of Research. Retrieved May*, 25:2007, 1994.

[11] N. Schmitt, J. Keeney, F. L. Oswald, T. J. Pleskac, A. Q. Billington, R. Sinha, and M. Zorzie. Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, 94(6):1479, 2009.

[12] G. Shmueli. To explain or to predict? *Statistical Science*, pages 289–310, 2010.

[13] I. Smythe. Data driven decision-making in further education. In *Presentation at the FET Conference*, Gauteng, South Africa, 2013.

[14] Statistics South Africa. Quarterly labour force survey quarter, 2014.

[15] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8(2):307–325, 2010.