Statistical modeling of students' performance in an open-admission bachelor program in Flanders

The need to discriminate between explanatory and predictive modeling

Ramaravind Kommiya Mothilal

Thesis submitted for the degree of
Master of Science in Artificial
Intelligence, option Engineering and
Computer Science

Thesis supervisor:
Prof. dr. ir. Tinne De Laet

Assessors:
Tom Broos
Sven Charleer
Maarten Pinxten

Mentor:
Tom Broos

Academic year 2017 – 2018

Preface

I am very grateful to Prof. Tinne De Laet who provided me the freedom to pursue my own ideas and guided me whenever I needed. I would like to thank my mentor Tom Broos for his time and support, especially in the initial stage of my thesis. I would also like to thank Maarten Pinxten whose inputs played a crucial role in shaping my thesis. I also thank the jury for reading my text.

Further, I would like to thank my friends in Leuven who made me feel at home and also tolerated my eccentricities. Most importantly, I express my deepest gratitude to my parents whose love and support is what makes me who I am.

Ramaravind Kommiya Mothilal

Contents

Abstract

In universities with an open-admission system, students sometimes choose study programs that are inappropriate for their prior-knowledge, skills, and abilities and hence face huge challenges in successfully completing their education. The present study applies statistical modeling to extract insights about what is expected from the students to perform well in their first-semester exams. Good performance in the first semester reflects a smooth transition to higher education and acts as an early indicator for completing the study program. Although related works in this area are numerous, many have succumbed to the practice of indiscrimination between two distinct modeling approaches, namely explanatory and predictive, often leading to incorrect practical conclusions.

The present study discusses how appreciating the differences between explanatory and predictive modeling not only results in correct practical conclusions, but also reveals the advantages and disadvantages of these two approaches. This revelation, in turn, gives rise to three different modeling approaches to solve the problem at hand. The first approach employs interpretable statistical models to explain how different traits of students affect their first-semester performance. The second approach, still using interpretable models, tries to optimize both explanatory and predictive goals with an objective of generalizing the statistical inferences to "unseen" incoming students and strengthening the validity of explanations. Finally, the third approach employs complex algorithmic models to accurately predict students' performance by uncovering patterns and relationships that are difficult to hypothesize. The first and third approaches differ distinctly in their statistical goals, choice of models, and evaluation criteria, whereas the second approach combines aspects of the other two approaches. The present study discusses how each of the three approaches solves the problem at hand from different perspectives and further discusses their advantages and disadvantages.

The above three approaches result in three common insights: (1) prior knowledge and the effort exerted to acquire prior knowledge positively influences the first-semester performance strongly than affective and goal strategies, (2) affective strategies positively influence the first-semester performance only in an indirect way through prior knowledge and the effort exerted to acquire prior knowledge, (3) students with well-defined goal strategies tend to exert less effort in acquiring prior-knowledge, thereby perform poorly in the first semester. However, while the first two approaches show that students' preference for time pressure has no influence on the first-semester performance, the third approach shows that higher levels of pressure preference are associated, though not very strongly, with students who perform poorly in their first semester.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

math	Secondary math grade
phy	Secondary physics grade
chem	Secondary chemistry grade
hrs	Secondary math level
eff	Effort exerted to acquire secondary results
affe	Affective strategies
goal	Goal strategies
press	Preference for time pressure
wavg	Weighted average
cse	Cumulative Study Efficiency
mot	Motivation
time	Time management
conc	Concentration
anxi	Anxiety
test	Test strategy
IV	Independent Variable
DV	Dependent Variable
ECTS	European Credit Transfer and Accumulation System
PCA	Principal Component Analysis
FA	Factor Analysis
LASSI	Learning and Study Strategies Inventory
CO	Cumulative Odds (model)
LR	Likelihood Ratio (test)

Symbols

X	Students' prior-knowledge, skills, and abilities
Y	First-semester performance
$\mathcal{F}$	Theoretical relationship between X and Y
x	Operationalization of X
y	Operationalization of Y
f	Operationalization of F
$\hat{f}$	Statistical estimate of f

Chapter 1

Introduction

1.1 Setting the Context

Unlike many other Anglo-Saxon countries, neither national-level school leaving examinations nor entrance examinations to universities (except Medicine, Dentistry and Arts Education) are organized in Flanders, the northern Dutch-speaking region of Belgium. A valid Flemish Diploma is sufficient to enroll in most of the programs at Flemish universities. As the study options are numerous, the chances of choosing inappropriate programs are significant. Not surprisingly, the Organization for Economic Co-operation and Development reported that the *first time graduate rate*[1] at Bachelor level in Belgium in 2014 is only 42% [64, p. 68]. Moreover, Fonteyne et al. [34] noted that, on average, less than 40% of the university students in Belgium pass all the courses in their first year even after repeated attempts. This important issue needs to be tackled at the earliest as first-year performance is shown to be strongly related to academic retention [23, 62]. Hence, providing accurate and relevant information about what is expected from the students in order to succeed in their intended program of study is essential to help them make informed decisions.

The present study focuses on the heterogeneous incoming students (in terms of prior math/science knowledge, skills and abilities) of bachelor of engineering program at KU Leuven. The high degree of heterogeneity in incoming students is particularly challenging for STEM programs where students with less prior math/science knowledge often find it hard to live up to the expectation in a relatively challenging program. Previous research [70, 71, 72] have discussed about various characteristics that are related to incoming students' success in the engineering program at KU leuven. Some of them are academic achievement in secondary school and scores on diagnostic tests measuring learning and study strategies. The present study applies *statistical modeling* on these available data to pull out insights about what is expected from the students to perform well in the first semester of their first year. Good performance in the first semester reflects a smooth transition to higher education and acts as an early indicator for completing the study program.

[1] First-time graduate is a student who has graduated for the first time at a given level of education during the reference period [64, p. 66]

Necessarily, the data for statistical modeling is past students' records. The data for the present study is obtained from the students of Bachelor of Engineering program from the year 2015 to 2017. The data measures students' prior-knowledge, skills and abilities. It includes students' math and science grades in secondary school, the effort they put into to obtain their secondary results, their math level, scores on psychometric tests indicating their study strategies and preference for time pressure, and their final scores in different first-semester courses. In the present study, the weighted average of all the course scores is used as the dependent variable (DV) and all other variables are used as independent variables (IVs). Chapter 2 explains in detail about how the available data are used. Statistical modeling can be approached in different ways to derive different conclusions from the same data for the same problem. The next section 1.2 discusses two important distinct approaches, namely explanatory and predictive modeling that are appropriate in the context of the present study. With freedom of choice comes the responsibility to not conflate between different approaches. Section 1.3 explains the need to discriminate between these two distinct approaches for proper scientific usage and to derive correct practical conclusions. Section 1.4 discusses how many of the relevant researches in this area have succumbed to the practice of indiscrimination between these two modeling approaches. The final section 1.5 expounds how the objective of the present study unfolds in this process.

1.2 Same Problem, Different Approaches

1.2.1 Explanatory Modeling

One way to address the issue at hand is by *explaining* how students' prior-knowledge, skills and abilities affect their performance in the first semester. It should be noted that *explanation* in this context should necessarily be causal. The process of applying statistical models to test these causal explanations (or hypotheses) that are often given in terms of theoretical constructs is called as explanatory modeling [81]. Further, Shmueli [81] noted that only *association-based models* are often used for testing causal hypotheses. Hence, causality is justified by theory and the application of statistical modeling is strictly through the lens of theoretical model [81]. To put it formally, consider X and Y representing students' prior-knowledge, skills, and abilities and their first-semester performance respectively. X is hypothesized to cause Y through the function $\mathcal{F}$, such that $Y = \mathcal{F}(X)$. Let X and Y be operationalized into measurable variables x and y respectively. Similarly, let the theoretical relationship $\mathcal{F}$ be operationalized into a statistical model f. Then the focus of explanatory modeling is to use x and y to construct an accurate estimate $(\hat{f})$ of f by minimizing the bias $E(\hat{f}(x)) - f(x)$. Subsequently, statistical inference of the estimated model $\hat{f}$ is used to confirm or contradict the theoretical relationship between X and Y, all based on the assumption that f is accurately representing the proposed underlying theory provided by $\mathcal{F}$.

Once the statistical model f is validated to adequately represent $\mathcal{F}$, the fit of the estimated model $\hat{f}$ to the data measured using the in-sample metrics such as R^2 and

F statistic are typically used to indicate respectively the strength and significance of the proposed causal relationship. Finally, the statistical conclusions are converted to research conclusions that are often accompanied by introducing interventions such as policy recommendations [81]. For instance, if math grade in secondary school (an operationalization of prior math background) is shown to significantly affect the first-semester performance, then tools and instruments can be made available to encourage students with weak math background to strengthen their skill set in math.

However, in explanatory modeling, complex patterns and relationships that are hard to hypothesize in $\mathcal{F}$ are often overlooked [11, 81]. Further, only interpretable statistical models such as simple regression-type models are often used to explain the underlying causal mechanism. Any increase in the complexity of f will complicate the *interpretation* of the relationship between the constructs. Hence, the main disadvantage with explanatory modeling is the possibility of misrepresentation of the true underlying relationship.

1.2.2 Predictive Modeling

In predictive modeling, the goal is to *predict* incoming students' performance at the end of first semester based on their prior-knowledge, skills and abilities. The predicted output can be a numerical value (regression problem) like weighted average of course scores or a category (classification problem) like "at-risk"/"no-risk". Formally, the objective is to determine f directly from x and y so that f can be used to predict y for new x values. Algorithmic methods such as decision trees and neural networks are mostly preferred over data models as they could capture complex relationships between variables to provide accurate predictions [11]. This approach focuses to improve the predictive performance on a holdout set or using cross-validation or bootstrap methods [53], mostly overlooking the issue of interpretability. However, understanding the reason behind predictions is essential if decisions are made based on them. For instance, a predictive algorithm simply reporting the weighted average scores of incoming students may help in identifying students who may perform poorly in the future, but does not explain why the predictions occurred in the first place.

The above issue is better understood by drawing an interesting analogy between interpretability in statistical modeling and *persuasion* in Persuasive technology. The later is a decade old research field which refers to,

> *"The class of technologies that are intentionally designed to change a person's abilities or behaviour. Importantly, persuasion implies a voluntary change of behaviour or abilities or both. If force (coercion) or misinformation (deception) are used, these would fall outside of the realm of persuasive technology."* [48]

From the perspective of Persuasive technology [33], an interpretable model or an explainable prediction can be seen as a social actor that can persuade stakeholders through its interpretation or explanation, respectively, as social cues. In contrast, a black-box type model which just focuses on accurately predicting DV given IVs fails to persuade anyone in a social context.

Further, one cannot escape mentioning Occam's dilemma when discussing about interpretability. Simple models like Linear regression or Logistic regression are nicely interpretable but Domingos [27] prescribes that unless the target phenomenon is simple enough, it is not recommended to prefer simpler models. He comes to this conclusion after proving that the Occam's second razor (*"simpler of two models with same training set error is likely to have low generalization error"*) seldom works in practice. In such cases, Craven and Shavlik [21] and Domingos [27] suggest the use of accurate (and probably complex) models and recommend to construct comprehensible approximations to the complex models separately. A recently developed technique called LIME [76], which in my opinion, articulates their above suggestion from a different perspective. LIME can provide *explanations* (or qualitative understanding) for why a particular prediction is made for a given observation by approximating a complex model locally with an interpretable model. The authors of LIME also propose a method called SP-LIME to give a global understanding of the model by explaining a set of representative observations via submodular optimization [76]. To avoid confusion, explanation means "causal explanation" in explanatory modeling [81], but means only "qualitative understanding" in LIME, without implying causality [76].

Based on the above discussion, any statistical model which employs a technique that can provide useful information about the relationship between students' prior-knowledge, skills, and abilities and their first-semester performance can be considered "interpretable". As Breiman [11] says, *"The goal is not interpretability, but accurate information"*.

1.3 The Need for Discrimination

The two approaches discussed above have different goals through the entire modeling process, that is, from goal definition to model use and reporting [81]. In explanatory modeling, the goal is to precisely fit a model $\hat{f}$ to the data so as to reduce the statistical bias in order to accurately represent the proposed causal relationship $\mathcal{F}$. In contrast to the retrospective quality of explanatory modeling, in predictive modeling, the goal is to predict y accurately for "unseen" x by minimizing the combination of bias, $E(\hat{f}(x)) - f(x)$ and estimation variance, $E\{\hat{f}(x) - E(\hat{f}(x))\}^2$. Further, in predictive modeling, the data is partitioned into training and holdout sample and the predictive performance is evaluated on the holdout sample or by using cross-validation or bootstrap methods so as to *generalize* the predictive validity. Whereas in explanatory modeling, the data is usually not partitioned as it reduces the statistical power of significance tests. Moreover, the variables chosen in explanatory modeling are those that best operationalize the constructs in the theoretical causal structure [81]. While in predictive modeling, the variables are chosen such that they are not weakly associated with the outcome since such variables tend to reduce the predictive power. Finally, in the context of explanatory modeling, R^2 measures and

F statistic are commonly used to assess respectively the strength and significance of proposed causal relationship; but, when they are used in predictive modeling, they just show respectively the strength and significance of only the *association* between IVs and DV, and do not imply the model's predictive power. A model's predictive power is assessed only from the performance of that model on "unseen" (holdout) data.

Evidently, statistical modeling with absolute explanatory or predictive goals differ in each stage of the process. The major confusion between these two approaches in many related literature occurs during the model evaluation stage where predictive power is often implied using explanatory power [81]. Understanding the difference between between explanatory and predictive modeling is required not only for proper scientific usage of statistical methods but also to come at accurate practical conclusions. Further, the indiscrimination between these two approaches often results in ignoring the limitations of explanatory modeling and not appreciating the advantages of predictive modeling. But it is important to note that the two approaches need not be seen exclusively from one another. Many practical applications would need to possess the qualities of both the approaches. Even in such cases, appreciating the difference between these two approaches is necessary to optimize both the explanatory and predictive goals and to exploit both the explanatory and predictive power of the model as observed by Shmueli [81].

1.4 Related Works

In many previous works, mostly from *Psychology and Educational science*, the prime focus is on the justification of considered IVs that reflect students' knowledge, skills, and abilities followed by the *application* of statistical techniques to quantify their relationship with DV. But many of the well-cited studies have failed to discriminate between explanatory and predictive modeling often resulting in incorrect practical conclusions. Also, in many works where predictive models are used, the issue of interpretability as discussed in section 1.2.2 is not addressed.

1.4.1 Does it Really Predict?

Ackerman et al. [5] carried out an extensive research accounting for the usage of various factors derived from Trait-Complex approach [4], traditional factors such as prior academic achievement, and factors that are operationalization of domain knowledge with an intention to *predict* academic success and STEM retention. But their statistical goal followed the trait of explanatory modeling rather than being predictive. That is, they first proposed causal hypotheses such as *"Trait complexes would be a significant predictor of academic achievement"* grounding them in theory. Then, they used multiple linear regression and logistic regression to test various hypotheses. Further, they erroneously infer predictive power (they call it predictive validity, but both mean the same) from explanatory power of their variables (*"The trait complexes in isolation also have significant validity for predicting cumulative* *college GPA... the highest predictive validity was found for the first-year GPA ($R^2 = 0.14$)"*).

Similarly, in a study based out of Belgium [34], exploratory modeling was incorrectly performed for a predictive goal. They developed a tool called SIMON-C with an objective to *identify* those prospective students who might have a very low probability of passing. But, they instead proposed causal hypotheses and then tested them using regression methods. Subsequently, they used R^2 to incorrectly infer predictive power of input variables. In addition, the authors also used in-sample accuracy to assess their model's predictive power but using a holdout set or using cross-validation or bootstrap methods is/are preferred for correct scientific usage [53].

In another study [60], the objective was to *"harness the predictive power of LMS"* to identify at-risk students and allow for early intervention. Although the authors were aware of not implying causation through their modeling process, they incorrectly inferred predictive power from the explanatory power of a multiple regression model. They also used the in-sample accuracy of a logistic regression model to demonstrate its predictive power.

Likewise, though the titles of [29], [86], [77], [56] and [54] containing terms such as "predict" or "predicting" give a flavour of predictive modeling, their statistical goals are actually explanatory in nature. These studies employed association based statistical methods such as multiple regression to confirm their causal hypotheses and subsequently performed retrospective inferences. Some of these studies such as [86] and [77] also succumbed to the incorrect inference of predictive power from explanatory power. Furthermore, in a recent study at KU Leuven [87], predictive power is incorrectly inferred from the summary statistics of Receiver Operating Characteristic (ROC) curves. Here the authors have confused predictive modeling with another type of modeling, which Shmueli [81] calls descriptive modeling, where the goal is to only capture associations between IVs and DV rather than doing causal inference or predictions.

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.

1.4.2 How to Trust the Prediction?

Lin et al. [58] 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. The predictive performance of the models were evaluated using k-fold cross validation technique and ANN was found to outperform other models. While the authors correctly evaluated the predictive performance on a holdout sample, they concluded their research by suggesting the following: *"Model results can also be used to provide faculty and advisors with informed course selection advice to first-year engineering students"*. However, when a model's prediction is directly used as a tool for advising, developing trust on the individual prediction and on the model is important [76]. A predictive model must be interpretable from the perspective discussed in section 1.2.2, but this issue is not addressed in [58]. Similarly, in many related studies (such as [42], [84], [9]), proper usage of predictive modeling is followed but the issue of interpretability is not addressed.

1.4.3 Studies with Clear Statistical Objectives

There are also numerous studies where statistical analyses are performed properly through the lens of explanatory modeling ([28], [82], [31], [13], [10], [26], [59], [69]). Except for the incorrect usage of terms such as “predictive validity” at few places in some of these studies, the overall objective of explanatory modeling was clearly followed throughout otherwise. Moreover, in the study by Druzdze and Glymour [28], causality was not only based on theory, but also based on probabilistic methods discussed in [83]. They also verified the assumptions of the statistical techniques they employed which is seldom followed in most of the related literature. But still, they once succumbed to the improper usage of implying predictive power in terms of variance explained.

In the context of predictive modeling, studies by French et al. [38], Burtner [12], Essa and Ayad [30] investigated their model’s predictive power in predicting student success and retention while also addressing the interpretability issues. While French et al. [38], Burtner [12] used linear regression and classification techniques, Essa and Ayad [30] used more complex ensembles methods for prediction.

1.5 Thesis Objectives

Appreciating the differences between explanatory and predictive modeling not only results in proper scientific usage and correct practical conclusions, but also reveals the advantages and disadvantages of these two approaches. This revelation, in turn, gives rise to three different modeling approaches to solve the problem at hand.

First, causal explanations, in the light of data, are required to understand how students’ prior-knowledge, skills and abilities influence their performance in the first-semester exams. It leads to appropriate policy-making to ensure a smooth transition from secondary school to higher education. Hence, the initial focus of the present study is simply on explanatory modeling and is discussed in detail in Chapter 3.

Second, while explanations enhance our understanding of how different characteristics of students influence their first-semester performance, they need not necessarily hold in the future. Explanations are retrospective [81]. Hence, our model needs the ability to predict first-semester performance, in addition to explaining, to generalize our inferences to “unseen” incoming students. Further, quantifying the validity of an explanatory model is also difficult. That is, the operationalization of the proposed relationship $\mathcal{F}$ between different theoretical constructs into a statistical model f is difficult to quantify mathematically. It is mainly based on relevant theories [81].

Chapter 4 discusses to what extent the predictive validity of an explanatory model is capable of solving these problems.

Third, as mentioned previously, complex patterns and relationships that are hard to hypothesize are often overlooked when explanatory goals are considered [81]. Further, the models used in the first two approaches cannot be made complex to capture the underlying reality as it would complicate the interpretation of the underlying relationship between the constructs. An alternative approach would be to employ (algorithmic) predictive models and construct comprehensible approximations to them separately for interpretation [21, 27]. Hence, in the present study, a popular data mining algorithm called Extreme Gradient Boosting (XGBoost) is used to accurately predict students' performance. An explanation technique called LIME by Ribeiro et al. [76] is employed to reason both the individual predictions of XGBoost and to get a global understanding of all its prediction on the holdout testing set. The consistency of the results obtained through this approach and its advantages over the previous two approaches are discussed. Chapter 5 is dedicated to demonstrate this approach.

Chapter 2

Data Preparation

This chapter first provides an overview of the data available for the present study. It then discusses various necessary steps that must be carried out before proceeding with statistical modeling.

2.1 The Data

The data for the present study is obtained from the students of Bachelor of Engineering program from the year 2015 to 2017 ($N = 811$). A brief description of variables in the data is provided in this section. Further, in the context of explanatory modeling, measured variables are seen as the operationalization of constructs in the theoretical causal structure [81]. Hence, the association between different measured variables and theoretical constructs is also provided.

Math, Physics and Chemistry Grades (math, phy, chem): Students receive the grades of these subjects in percentages in Flemish schools. As there are no national-level school leaving exams in Flanders, the grades that students obtain are highly dependant on their teachers and schools. Thus, to adjust for such differences, through a self-reported questionnaire, students were asked to mention one of the following categories in which their grades fall into: 60%, 60-70%, 70-80%, 80-90%, and above 90%. Thus, *math, phy, chem* are ordinal variables with 5 categories each.

Math Level (hrs): In Flemish schools, students can choose between three different levels of mathematics they intend to study. The curriculum and the hours required per week differs between the levels. Students opting for "Low" level of mathematics spend less than 6 hours per week, while those opting for "Medium" and "High" level spend 6-7 hours and more than 7 hours per week respectively. As it can be observed, *hrs* is an ordinal variable with three levels ("Low", "Medium" and "High").

In the present study, the variables *math, phy, chem* and *hrs* are seen as the operationalization of "prior academic knowledge" since these variables reflect the understanding of relevant background materials required for the engineering program.

Effort level (eff): This is an ordinal variable with 5 levels ("Very Low", "Low", "Average", "High" and "Very High") indicating the effort students put into to get their results in secondary school. Thus, *eff* is seen as the operationalization of “effort exerted to obtain prior academic knowledge”.

Learning and Study Strategies: In order to assess students’ awareness about and use of learning and study strategies, a self-reported instrument by Weinstein and Palmer [89] called Learning and Study Strategies Inventory (LASSI) is used. At KU Leuven, LASSI is administered as a 77-item questionnaire to assess and operationalize students’ standing on the following 5 areas: ***Motivation (mot), Time Management (time), Concentration (conc), Anxiety (anxi) and Test Strategies (test)***. All items are rated on a 5-point Likert type scale from 1 (“Not at all typical of me”) to 5 (“Very much typical of me”) and the final LASSI scores range from 8 to 40. In order to ensure that these scales represent coherent and reliable assessment of the underlying constructs, their internal consistency or the reliability must be evaluated [14, 22]. For the data used in the present study, the internal consistency coefficients (Cronbach’s alpha) are as follows: *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 [89] and with general standards [14], indicating that the variables *mot, time, conc, anxi* and *test* are reliable operationalization of the various psychometric constructs representing learning and study strategies. It is important to note that unlike these variables, the operationalization of variables *math, phy, chem, hrs* and *eff* are difficult to validate since there is no measures such as internal consistency coefficients.

Preference for Time Pressure (press): This scale was developed according to the procedures of Choi and Moran [17] to assess and operationalize student’s preference for time pressure in learning. This scale is expected to be important because according to Choi and Moran [17], high preference for pressure in learning among others are the main characteristics of, what they define as, active procrastinators who achieve their desired outcome by procrastinating. *press* is interpreted as a continuous variable and it ranges from 3 to 20.

All the variables discussed so far are used as independent variables which define various characteristics of students. The below two are the outcome variables, both of which are seen as the operationalization of students’ performance.

Weighted average (wavg): This is the weighted mean of scores in all the first-semester courses where each course is weighted by its ECTS credits. *wavg* is a continuous variable taking value in the range 0 to 20.

Cumulative Study Efficiency (cse): This is a discrete variable taking values equal to the number of ECTS credits obtained at the end of first semester divided by the total number of ECTS credits taken at the beginning of the semester.

The variables *wavg* and *cse* are strongly correlated with each other (pearson correlation coeff. = 0.92), thus carrying redundant information. Since it is approximately normally distributed which is a desirable property for many statistical tests, and is continuous, it is used as the dependent variable in statistical modeling to follow. The distributions of these two variables and the visualization of the correlation matrix, plotted using Seaborn Python library are shown in Appendix A.

2.2 Data Cleaning

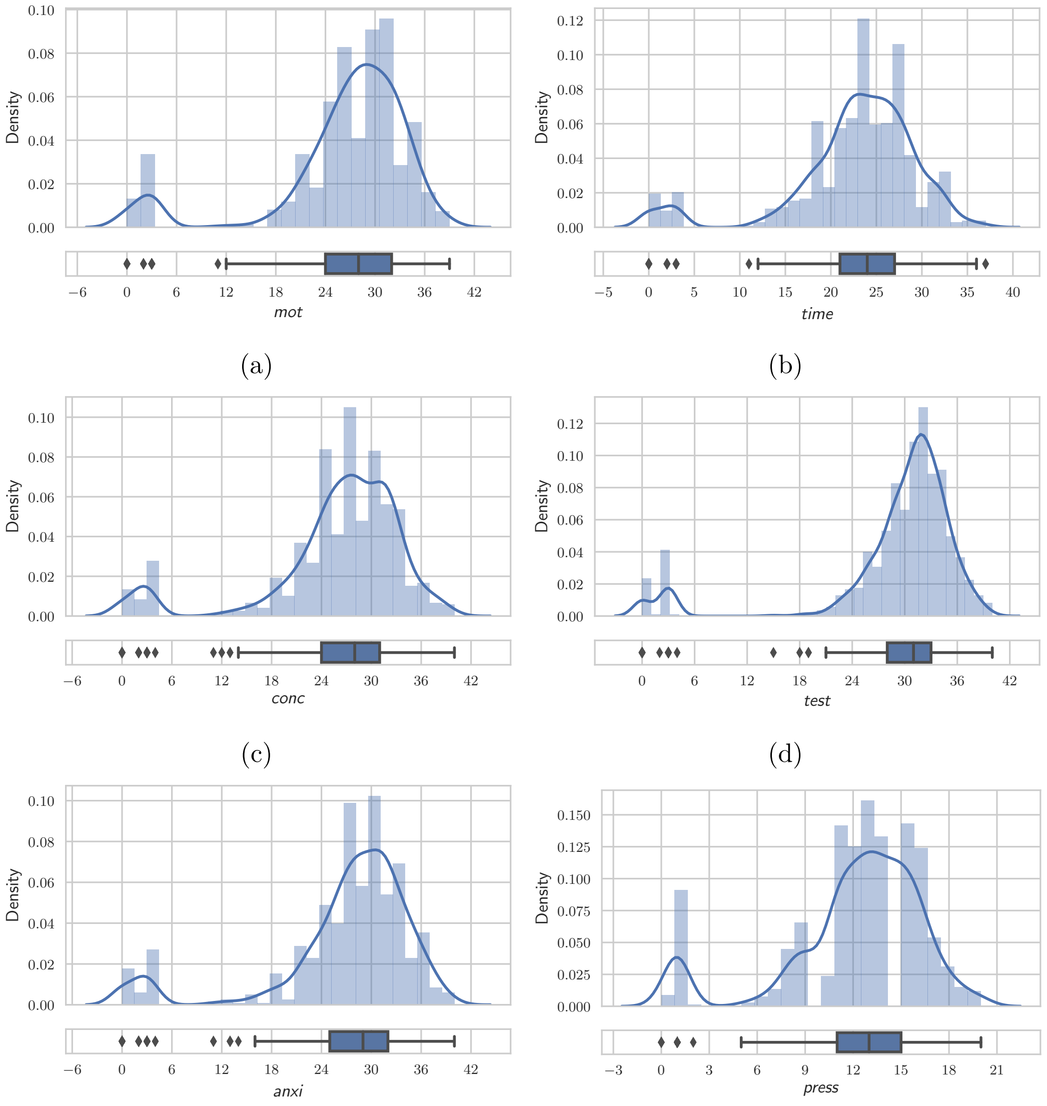
Most real world data are likely to contain *inaccurate (outliers)* and *incomplete (missing)* information. These two issues can result in misinterpretation of collected data thereby resulting in poorer statistical conclusions. The next two section aims to address these issues for the data considered in the present study.

2.2.1 Detecting Outliers

Figure 1(a-e) shows the distributions of all the LASSI scales plotted using Seaborn Python library. It was previously mentioned that the LASSI scales vary from 8 to 40. Accordingly, from Figure 1(a-e), the presence of outliers can be observed in the lower end of the distributions. To find out whether the occurrence of these outliers is random or systematic, they are compared with corresponding LASSI norm groups developed for students at KU Leuven by Olivier et al. [67]. The norm groups for each LASSI scales are computed by binning the raw scores into five categories ("Very Low", "Low", "Average", "High", "Very High") based on the guidelines described by Olivier et al. [67]. It was observed that for each LASSI scale *(mot, time, conc, test, anxi)*, every unique outlying value is associated with a fixed norm category. For instance, for variable *anxi*, all observations with value "2" correspond to "Very High" norm group, while observations with value "3" correspond to "Very Low" norm group. This indicates that the occurrence of these outliers is not random. Similarly, for variable *press*, almost all observations with value less than its minimum ("3") are equal to "1" (Figure 2.1(f)). The presence of values that systematically deviate from other observations indicates that these values could have been misrepresented during data collection. Since the outliers are systematic, simply removing them might result in a loss of information. Thus, in the present study, these values are treated as *missing* and sophisticated statistical methods, as described in the next section, are employed to replace them.

2.2.2 Handling Missing Data

Apart from the systematic outliers that are considered as missing, the data also contain other missing values that are mostly due to the unwillingness of some students to disclose their information. Values in the data are said to be *Missing Completely At Random (MCAR)* when the missingness of data is independent of any of the observed or unobserved values in the data. In the context of explanatory modeling, the data must be checked if it is MCAR in order to make unbiased statistical inferences [81]. However, if the missingness is dependent on the DV, then the model's predictive performance will be improved [25]. Hence, the data is first checked if it is MCAR, that is, if the missingness is independent of missing values, observed IVs and the DV. Another important criterion for predictive purposes is knowing whether or not the holdout sample contains missing values. It should be noted that in the present study, predictions are made on the DV only for incoming students' profiles which are *complete*, so as not to make decisions based on incomplete information. Hence,



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2.1: Univariate distributions with kernel density estimates

appropriate imputation methods are used before partitioning the data so that the model performance is evaluated on a complete holdout sample.

However, some might claim that imputing missing values before partitioning the data could lead to “leakage” of information from outside the training set in building the predictive model. But in predictive modeling, in theory, the entire data is assumed to be drawn i.i.d. from an unknown multivariate distribution [11]. Hence, imputing missing values with, for instance, mean essentially implies filling the missing values with the *most probable value* for mean from the underlying distribution [1]. Sophisticated statistical methods do a better job of modeling the underlying distribution and fill the missing values with better estimates than simple mean. Thus, the underlying theory assumption serves as the reason to disregard any information leakage from holdout sample to training set. The advanced statistical technique that is followed to identify if the data is MCAR is based on the methods proposed by Jamshidian and Jalal [49] and using the R package *MissMech* [50] which implements these methods.

Before using *MissMech*, the observations for which almost all variables have missing values are identified and removed. This reduces the number of observations from 811 to 776. Then, all continuous variables in the data are passed as input to the function “TestMCARNormality” of the package *MissMech.* The data is imputed after it is partitioned into a fixed number of groups each having identical missing data pattern. This is followed by a nonparametric test of homoscedasticity between the groups. A brief description of the procedure is given in [50]. The result of the MCAR test for our data is shown in Appendix B. The statistical test confirms that the data is not multivariate normal and uses the non-parametric method to show that there is no sufficient evidence to reject the hypothesis that the data is MCAR at 0.05 significance level. Besides, there are only 4 observations, in total, which have missing values for all the categorical IVs combined, and hence they are simply removed. Thus the final data after imputation has 720 observations for further analysis.

2.3 Feature Engineering

Once the data is clean, it must be analyzed if it can be used directly in modeling or if it needs to undergo transformations to facilitate the modeling process. In the context of explanatory modeling, the levels of measured ordinal variables must be interpretable such that they reflect the understanding of corresponding theoretical constructs. Further, for both explanation and prediction, different levels of an ordinal variable are expected to be significantly different from one another so as not to carry redundant information. Likewise, the continuous variables are also expected to be less correlated with each other. The next two subsections discuss the transformation of measured variables to address these issues.

2.3.1 Transformation of Ordinal Variables

Math and science grades are considered as the operationalization of prior academic knowledge, since they indicate the understanding of relevant background materials. Thus, the original five ordered categories, namely 60%, 60-70%, 70-80%, 80-90%, and above 90%, can be considered as the representative of "Low", "Below Average", "Average", "Above average" and "High" prior academic knowledge respectively. Similar analogy applies for the *eff* variable. But if the students falling into one of the above categories did not show any significant difference in mean performance from the students falling into another category, then these two categories can be combined into one, as they both then represent the students with same average performance. So, one-way analysis of variance (ANOVA) is used to determine if there is significant difference in mean performance between the students of different categories. As mentioned in section 2.1, *wavg* is used as the DV to represent students' performance.

One-way ANOVA test is computed using the R *stats* package [73]. The results are shown in Appendix C. They show that there is significant difference in mean *wavg* between the categories of each ordinal input variable. But a $p - value < 0.05$ does not indicate that all categories are significantly different from one another. Hence to determine if the mean difference in *wavg* between specific pairs of categories are statistically significant, *Tukey Honest Significant Differences* post hoc test is performed. The results are shown in Appendix D. Adjusted $p - value > 0.05$ for "60-70%" and "<60%" show that there is no significant difference in the mean *wavg* between students securing "<60%" and "60-70%" in math and science. Hence, these two categories are combined and renamed as "Low" while "70-80%", "80-90%" and ">90%", having significant differences among them, are renamed as "Low_medium", "High_medium" and "High" respectively. These categories are interpreted as if they represent prior academic knowledge in order.

Similarly, there is only significant differences in *wavg* between students with "Low", "Average" and "High" effort level, which originally has 5 categories. Hence, "Very Low" and "Very High" categories are combined with "Low" and "High" respectively. This results in 3 categories for the variable *eff*, namely, "Low", "Average" and "High". However, there is significant differences in the mean performance between students of every math level *(hrs)*. Hence this variable is left untouched.

To ensure that the above results are valid, the assumptions of all the one-way ANOVA tests performed above are verified (the results are not shown for brevity). First, the Normal Probability Plots of Residuals indicate that the residuals in all tests are normally distributed, thus satisfying the first important assumption. Second, Levene's test is used to test for homogeneity of variances. The results confirm the assumption of equal variances for all groups in every test. Other assumptions such as independent observations and no significant outliers are inherently satisfied for the data in the present study.

2.3.2 Transformation of Continuous Variables

Figure A.2 shows that the LASSI variables are correlated with each other carrying redundant information and hence not all of them are needed for same analysis. Multicollinearity, as Shmueli [81] explains, can lead to inflated standard errors making the inference process difficult in explanatory modeling. But as Vaughan and Berry [88] noted, though multicollinearity will not affect the ability to predict, it needs to be addressed if the influence of individual variables on prediction is required. Thus, the present study confronts the issue of multicollinearity to also obtain interpretable predictions.

Principal component analysis (PCA) and Factor analysis (FA) are two most commonly used techniques to reduce a larger number of variables to a smaller number of variables by exploiting the correlational pattern in the measured variables. The difference between these two techniques needs to be appreciated for proper scientific usage and also to not contribute to the tarnished reputation of these tools despite their scientific utility [85, p. 613]. FA is used to operationalise theoretical and intangible variables (called factors) from a set of observed variables. PCA, on the other hand, simply aggregates the correlated variables and empirically associates with a new variable (called component). They differ only in the construction of observed correlation matrix to extract factors or components and in underlying theory.

Mathematically, in PCA, all the variance of the observed variables are distributed to the components. However, in FA, only the variance that each observed variable shares with other observed variables are considered for analysis, thus excluding the unique and error variance of all the observed variables [85, p. 640]. But Arrindell and Van der Ende [8], Guadagnoli and Velicer [41] and Schonemann [80] point out that the practical difference between the results of PCA and FA is often negligible when the sample size is large, or when same number of factors or components are retained. Indeed, for the data used in the present study, the resulting values (called as scores) of the factors or the components did not differ much as expected. In addition, in the present study, the scores from PCA or FA would be discretized and converted into ordinal variables, making the results from PCA and FA virtually indistinguishable. The discretization process is similar to creating norm groups using cut-off values on the original LASSI scales for intuitive interpretation and to reflect the qualitative nature of learning and study strategies [89].

The use of both PCA and FA is prevalent in related literature. Olejnik and Nist [66] and Olaussen and Bråten [65] employed FA on the measured LASSI variables to uncover underlying theoretical factors. Whereas, Cano [13] used PCA to arrive at the final components, but reported them incorrectly as underlying factors. Although the practical results of the two methods would have been indistinguishable in [13], a conceptual clarity and an unambiguous scientific reporting is recommended [46]. Incidentally, the developers of LASSI related each of the scales to one of the three components of strategic learning: skill, will and self-regulation [89]. The association made here between the LASSI scales and the three components of learning is only based on theory and experts' decision. In the present study, PCA is preferred over FA as the criterion of interest is to find out how different LASSI scales are associated with fewer components, which in turn can be interpreted as the components of strategic learning as followed by the developers of LASSI [89].

Principal Component Analysis

The “principal” function of the R *psych* package [75] is used for applying PCA. The results provided by this function is consistent with the factor analytic tradition so as to allow comparison with other models in related literature. But before performing PCA, following practical issues need to be attended to get valid results [85, pp. 617-620]:

Sample size and missing data: The sample size of >500 in the present study is sufficient for performing PCA and the absence of any missing observations indicates that the results of PCA will be reliable.
Normality The R package *MVN* [55] indicates that the data is not multivariate normal (results not shown for brevity). However, normality is not required when PCA is used only to summarize the relationships between the variables.
Linearity and outliers: The scatter plots between the pairs of variables (plots not shown for brevity) indicate no evidence for true curvilinear relationship but indicate the presence of outliers. However, the outliers are not removed as they could reflect the uncertainties present in the real-world data used in the present study.
Factorability: The high correlation among the variables (Figure A.2) indicates the presence of factors (or components).

Following the absence of any potential issues, the decisions taken for the two main options for PCA is explained below.

Number of Components: Based on the previous studies by Olejnik and Nist [66], Olaussen and Bråten [65], and Cano [13], two principal components (PCs) are expected for the 5 LASSI scales used in the present study. While there are many solutions to the problem of finding the number of components, three widely used techniques mentioned in [79] are employed: (1) Kaiser criterion of extracting components with eigenvalue greater than 1.0, (2) Cattell’s scree plot, (3) Parallel Analysis (Figure 2.2). Each of these techniques have their own advantages and disadvantages and no technique will perfectly indicate the required number of components. However, in the present study, all these techniques suggest the use of only two factors as hypothesised.

Rotation of components: It should be remembered that the components are extracted so as to maximize the captured variance in the variables. This can result in an infinite number of orientations of components, all accounting for the correlation in the data equally well [79]. But the components can be rotated such that each of the variables relate more to (or load high on) only some of the factors. This improves the interpretability of the association between the variables and the components. The components can be rotated, either maintaining the orthogonality or allowing overlap in variance between them. The former method is orthogonal rotation and the latter method is called oblique rotation where the rotated components are correlated with one another. Oblique rotation is preferred over orthogonal as the later “strain reality”

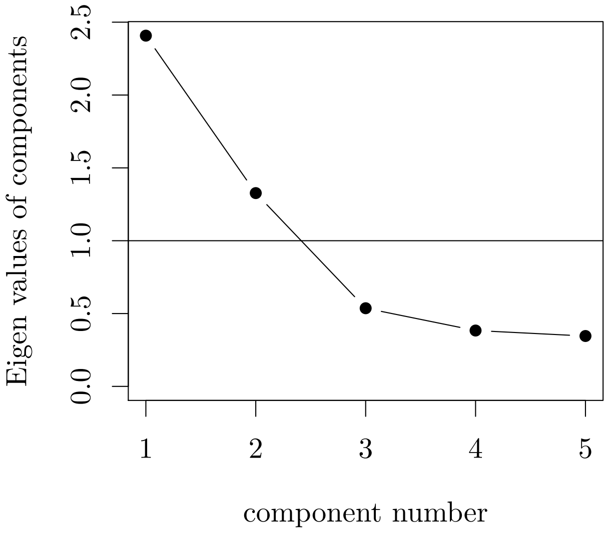


Figure 2.2: Parallel Analysis indicating that the optimal number of Principal Components are two

[85, p. 642] as some correlations between resulting components are generally expected in social sciences [20]. Although the use of rotated components is discouraged by some [74, 51], the approach followed in the present study using the R package *psych* is consistent with the factor analytic tradition.

Results Analysis: Figure 2.3 shows the results of PCA after oblique rotation. In oblique rotation, the correlation between components is provided in addition to the elements from pattern (or loadings) matrix. The pattern matrix contains the contribution (or loading) of variables on each component uncorrupted by any interaction between the components. It is expected that each of the variables are strongly associated with only one component so that a component can be easily interpretable using the associated variables. The measured variables can be considered to be strongly associated with a particular component if its loading on that particular component is greater than an acceptable cutoff. Based on the suggestion by Comrey and Lee [19], only variable loadings in excess of 0.71 are considered reliable as they contain 50% overlapping variance. It can be observed that after oblique rotation (oblimin with delta (δ)=0), the variables *mot* (motivation), *time* (time management) and *conc* (concentration) have loadings greater than 0.71 on first component while variables *anxi* (anxiety) and *test* (test Strategies) load highly (>0.71) on the second component. Further, Figure 2.4 shows the plots of components loadings before and after oblique rotation. It can be noticed from these figures that the association between the measured variables and the components is more explicit after rotation.

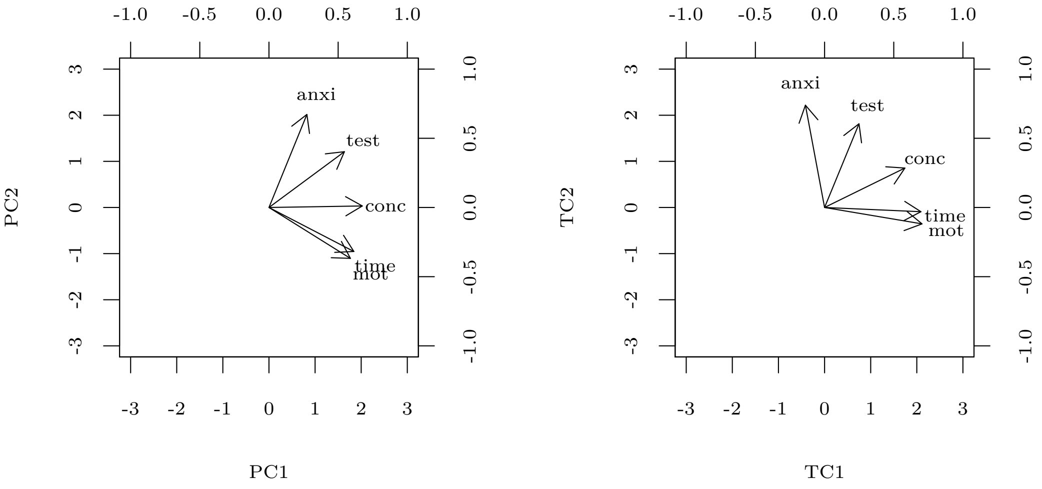
```
Principal Components Analysis
Call: principal(r = aaa[-ind, c(7:11)], nfactors = 2, rotate = "oblimin",
    n.obs = length(ind))
Standardized loadings (pattern matrix) based upon correlation matrix
       TC1   TC2   h2   u2 com
mot   0.88 -0.15 0.75 0.25 1.1
time  0.87 -0.04 0.75 0.25 1.0
conc  0.73  0.36 0.74 0.26 1.5
anxi -0.17  0.92 0.83 0.17 1.1
test  0.31  0.76 0.75 0.25 1.3

                       TC1  TC2
SS loadings           2.22 1.60
Proportion Var        0.44 0.32
Cumulative Var        0.44 0.76
Proportion Explained  0.58 0.42
Cumulative Proportion 0.58 1.00

 With component correlations of
     TC1  TC2
TC1 1.00 0.17
TC2 0.17 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.1
 with the empirical chi square  101.34  with prob <  7.8e-24

Fit based upon off diagonal values = 0.95
```

Figure 2.3: R output for PCA with oblique rotation

The results show that PCA with oblique (oblimin with delta (δ)=0) rotation yielded a two component solution and accounted for 75% of the variance in the data before rotation. The components are mildly correlated (pearson correlation coeff. = 0.17). Since the resulting association between the measured variables and the components is similar to that obtained by Cano [13], his naming convention for the components is used in the present study. Thus, component 1 is called "Affective Strategies", indicated by the variable *affe* and component 2 is called "Goal Strategies", indicated by the variable *goal*. In the context of explanatory modeling, *affe* and *goal* are seen as the operationalization of "Affective Strategie" and "Goal Strategies" respectively. It should be remembered that the components are labeled only to conveniently describe the combination of variables associated with them, and not to reflect any underlying phenomenon.

Norm Groups for Learning and Study Strategies

The LASSI scores of different students are compared for diagnostic and prescriptive purposes, commonly based on the norm groups developed to suit the local setting



(a) (b)

Figure 2.4: Plot of the component loadings before(a) and after(b) oblimin rotation. PC stands for (unrotated) Principal Component whereas TC stands for obliquely Transformed Component

[89]. Olivier et al. [67] developed the norm groups for LASSI scales for students at KU Leuven. However, norm groups for reduced components (or factors) of the LASSI scales are not yet prescribed for the population at KU Leuven. Hence, in the present study, simple cut-off values are used to transform the numerical PC scores to categorical norm groups. The split points chosen for each of the components (*affe* and *goal*) are 1 standard deviation below its mean, at its mean and 1 standard deviation above its mean. Thus, 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” similar to other IVs *math, phy* and *chem.* These categories are interpreted as if they represent different learning and study strategies in order.

Similarly, the variable *press* is also standardised and discretized into a 4-category ordinal variable. Thus, the final IVs used to represent students’ prior-knowledge, skills, and abilities are *math, phy, chem* (math and science grades), *hrs* (math level), *eff* (effort level), *affe* (affective strategies), *goal* (goal strategies) and *press* (pressure preference). The variable *wavg* (weighted average) is used as the DV to represent students’ performance.

Chapter 3

Explanatory Modeling

This chapter addresses the problem undertaken by the present study from the perspective of explanatory modeling. Specifically, causal explanations, in the light of data, are provided to understand how students' prior-knowledge, skills, and abilities affect their first-semester performance. A typical study which employs explanatory modeling starts by defining causal hypotheses based on the discussion of relevant theories [for some examples, see 86, 5, 34]. Accordingly, the next section 3.1 provides a brief discussion on how relevant theories give rise to causal hypotheses. This is followed by the application of the statistical models to test the hypotheses and is discussed in section 3.2. Finally, section 3.3 and section 3.4 discusses to what extent this modeling approach addresses the problem undertaken.

3.1 Hypotheses Formulation

The variables *math, phy, chem* and *hrs* are seen as the operationalization of prior academic knowledge since they indicate the understanding of relevant background materials. There are overwhelming empirical evidences [such as 5, 56, 82, 38] to show that higher grades and rigorous training in secondary math and science courses are strongly associated with higher performance in STEM programs in general. Previous research at KU Leuven [72] confirmed the same for our specific population of interest. Based on the above theories, the first hypothesis is as follows:

- *Prior academic knowledge (math, science grades and math level) will positively affect the academic performance at the end of first semester.*

Further, studies on population both outside [such as 5, 34, 77, 13, 31] and inside KU Leuven [72] have discussed the association between various psychological or non-cognitive factors and students' performance in higher education. Specifically, Bernold et al. [10] discussed the effect of various LASSI scales on the student outcome. In a study at KU Leuven [72], only *motivation*, *time management* and *test strategies* scales of the LASSI were shown to be significantly contributing to the variance in students' weighted average. Further, Cano [13] constructed *latent constructs* from

LASSI scales and showed that affective and goal strategies are significantly associated with the academic achievement. In another study, Choi and Moran [17] showed that in contrast to the traditional evidences, students preferring time pressure need not necessarily have poor academic achievement. Hence *press* variable is not expected to have any influence on *wavg*. Incidentally, there are *diverse* inferences made in the literature on the non-cognitive factors accounting for students' performance in higher education in general. However, no research on the use of reduced components of the LASSI scales is yet performed for the population at KU Leuven. In addition, the effect of preference for time pressure *(press)* is also not yet studied for the population at KU Leuven. The next hypothesis aims to fill the above gaps.

- *Affective strategies and goal strategies are expected to positively affect the academic performance at the end of first semester. However, preference for time pressure is expected not to affect the performance.*

Furthermore, Fonteyne et al. [34] considered the effort exerted as one of the dimensions of self-efficacy and showed that it has positive effect on academic achievement. While a previous study at KU Leuven [71] has discussed the effect of effort exerted on other psychological factors, its relation to academic performance is not yet studied. Also, Pinxten et al. [72] showed that, for the population at KU Leuven, the incremental value of non-cognitive variables (such as LASSI scales) over prior-knowledge in explaining *wavg* is very minimal. Trying to fill the above two gaps simultaneously, the final hypothesis aims to address to what extent the non-cognitive variables in the present study add incremental value over prior-knowledge in explaining the variability of academic performance.

- *Affective strategies, goal strategies, and effort level are expected to have significant incremental value over prior academic knowledge in explaining the academic performance at the end of first semester.*

3.2 Multiple Linear Regression

In all the above hypotheses, terms such as "positively affect" or "no effect" could indicate anything from simple linear to complex nonlinear relationship ($\mathcal{F}$) between the constructs. However, if nonlinear models are considered to represent the theoretical relationship $\mathcal{F}$, then the interpretation of the relationship between the constructs becomes complicated. On the other hand, the effect of each input variables on the output is easily understood if regression-type models are used. For the same reason, in most of the related literature (see section 1.4), linearity in the theoretical relationship is assumed, but often implicitly, by using a simple linear regression model. Accordingly, in the present study also, multiple linear regression is used. Table 3.1 displays the information about the variables and the types of multiple regression models required to test the hypotheses. But before performing any regression on the data, the assumptions of the regression methods are evaluated in section 3.2.1. Then, the first two hypotheses are tested using two separate standard multiple regression models and the third hypothesis is tested using a sequential multiple regression model as explained from section 3.2.2 to section 3.2.4.

Model	Response ~ Explanatory variables	Type of Multiple Regression
1	*wavg* ~ *math* + *phy* + *chem* + *hrs*	Standard
2	*wavg* ~ *affe* + *goal* + *press*	Standard
3	*wavg* ~ *math* + *phy* + *chem* + *hrs* + *affe* + *goal* + *press* + *eff*	Sequential (or Hierarchical)

Table 3.1: Information about different explanatory (multiple linear regression) models

3.2.1 Evaluation of Assumptions

This section evaluates the regression assumptions based on the guidelines provided by Tabachnick and Fidell [85, pp. 123-129] and in [3]. Fox [36] provides an elaborate discussion on various ways to diagnose and fix the violation of regression assumptions. In the present study, the assumptions are evaluated using the *analysis of residuals* method by following the proceduresl in [85, pp. 123-129]. Linear regression is performed using the "lm" function of the R *stats* package [73] and the default residual plots provided by the same package are used for the diagnosis of any violation of assumptions. The assumptions listed below are evaluated for all the three models in Table 3.1, since each one of them uses a different combination of variables.

Linearity, Normality and Homoscedasticity of Residuals: First, the expected value of the output variable in Multiple Linear Regression (MLR) is a linear function of each explanatory variable holding others constant. Also, the effects of the explanatory variables on the expected value of the output is additive [3]. To diagnose whether this assumption is violated or not, the plots between the predicted output values and the residuals are used (Figure E.1). The symmetrical spread around a horizontal line, without any distinct patterns, indicates no violation of linearity and additivity assumptions for all the models.

Second, the residuals are assumed to be normally distributed around each and every predicted output value. This assumption is required for the correct calculation of confidence intervals in the significance tests for regression coefficients. The Normal Q-Q plot (Figure E.2) for each of the models indicates the absence of any excessive skewness or kurtosis thus satisfying the normality of residuals assumption.

Third, the variance of the residuals for all predicted values are assumed to be approximately same, so that the errors in prediction do not increase over time or at larger predicted values. The plots between the square root of absolute residuals (in order to diminish skewness) against the predicted values (Figure E.3) for all the models show that the residuals are equally spread along the ranges of explanatory variables.

Statistical independence of residuals: This assumption makes sure that the errors in prediction are not associated with the order in which the observations are provided so as not to make any biased inference. The plots of residuals against the order of observations show that the errors in prediction are independent for all three models (Figure E.4).

Outliers: Although the "residuals vs leverage" plot (not shown for brevity) indicates the presence of few outliers based on Cook's distance criteria, they are not removed as their presence reflects the residual uncertainty present in this real-world data.

3.2.2 Effects of Prior Academic Knowledge

Figure 3.1 shows the results of standard multiple regression for model 1. The F statistic is $F(11, 708) = 37.97$ at $p - value :< 2.2e - 16$ and the multiple R^2 is 0.37 with 95% confidence interval (computed using R package *psychometric* [32]) from 0.32 to 0.43. The later provides an estimate of the strength of the causal relationship between the explanatory variables and the output variable, while the former determines the statistical significance of this relationship. The adjusted multiple R^2 value of 0.36 indicates that more than a third of the variability in the first-semester performance *(wavg)* is explained by the prior academic knowledge.

It is also important to compute the squared semipartial correlation (sr^2) of each variable, as they represent each variable's unique contribution to the total variance in output. sr^2 for an explanatory variable is the amount by which the R^2 is reduced if that variable is deleted from the regression equation [85, p. 144]. The unique contribution of *math, phy, chem* and *hrs* is computed as 6.6%, 2.3%, 4.5% and 3.0% respectively, and the rest 20.8% variance in *wavg* represents the shared variance which is contributed to R^2 by two or more input variables. This indicates that secondary math grade independently explains most of the variance (around 7%) among all the variables representing prior academic knowledge, showing the importance of secondary math grades for the engineering program.

All the individual regression coefficients are significant at $p = 0.05$ indicating that all variables contribute significantly in explaining the variance in *wavg*. The sign and the magnitude of the coefficients indicate that the mean estimate of *wavg* increases with increasing levels of these variables. For instance, students with high math background are expected to get a *wavg* score of 2.9 higher than those with low math background, on average, assuming that they scored equally well in *phy, chem* and took same math level *(hrs)*. It should be noted that the increasing trend in the estimated academic performance with the prior academic knowledge is what is important in explanatory modeling rather than the actual magnitude of the coefficients.

Thus the above inferences confirm that prior academic knowledge, operationalized as math and science grades, and math level, positively affects the first-semester performance.

```
Call:
lm(formula = wavg ~ math + phy + chem + hrs, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9914 -1.7617  0.3189  1.8301  5.6332 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.9439     0.7845   5.027 6.31e-07 ***
mathLow_medium    0.5453     0.2621   2.080 0.037848 *  
mathHigh_medium   1.9560     0.2864   6.830 1.83e-11 ***
mathHigh          2.8629     0.4329   6.613 7.42e-11 ***
phyLow_medium     0.8390     0.2975   2.820 0.004937 ** 
phyHigh_medium    1.1190     0.3133   3.571 0.000379 ***
phyHigh           2.0556     0.4093   5.022 6.49e-07 ***
chemLow_medium    0.7978     0.2691   2.965 0.003132 ** 
chemHigh_medium   1.7996     0.2881   6.247 7.24e-10 ***
chemHigh          2.2712     0.4137   5.490 5.61e-08 ***
hrsMedium         2.2063     0.7630   2.892 0.003949 ** 
hrsHigh           3.0747     0.7611   4.040 5.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.482 on 708 degrees of freedom
Multiple R-squared:  0.371,	Adjusted R-squared:  0.3613 
F-statistic: 37.97 on 11 and 708 DF,  p-value: < 2.2e-16
```

Figure 3.1: Multiple Regression output in R to investigate the effects of Prior Academic Knowledge

3.2.3 Effects of Principal Components of LASSI and Preference for Time Pressure

Figure 3.2 shows the results of standard multiple regression for model 2. The inferences here are made similar to how it was made in the previous section. The multiple R^2 of 0.06 indicates a weak model, although a statistically significant one ($F(9, 710) = 4.644$ at $p - value :< 5.4e - 06$). The adjusted multiple R^2 value of 0.04 indicates that only 4% of the variability in the first-semester performance *(wavg)* is explained by the affective strategies, goal strategies, and the preference for time pressure. It is found that almost all of the variance explained by model 2 is contributed by *affe*, while *goal* and *press* combinely contributes only less than 1%. Similarly, only the coefficients of *affe* are significant.

Thus, the above inference shows that only students' affective strategies influence the academic performance at the end of first semester and not their goal strategies. The model results also prove that students' preference for time pressure has no influence on their academic performance as suggested by Choi and Moran [17]. However, it is important to note that the inference made here *does not* include the effects of other variables such as students' prior-knowledge and effort level.

```
Call:
lm(formula = wavg ~ affe + goal + press, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6674 -2.1646  0.2214  2.0941  7.7583

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.503215   0.439867  19.331  < 2e-16 ***
affeLow_medium    0.645048   0.344161   1.874  0.06130 .
affeHigh_medium   1.067623   0.346564   3.081  0.00215 **
affeHigh          2.094439   0.410899   5.097 4.42e-07 ***
goalLow_medium    0.462921   0.354768   1.305  0.19236
goalHigh_medium   0.563881   0.355516   1.586  0.11316
goalHigh          0.877362   0.416511   2.106  0.03552 *
pressLow_medium   0.003486   0.358700   0.010  0.99225
pressHigh_medium -0.121927   0.382850  -0.318  0.75022
pressHigh        -0.494090   0.406321  -1.216  0.22439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 710 degrees of freedom
Multiple R-squared:  0.05559,       Adjusted R-squared:  0.04362
F-statistic: 4.644 on 9 and 710 DF,  p-value: 5.422e-06
```

Figure 3.2: Multiple Regression output in R to investigate the effects of Principal Components of LASSI and Preference for Time Pressure

3.2.4 Incremental Effects of Noncognitive Factors

The first two regression models showed that the explanatory power of prior academic knowledge in explaining the variance in *wavg* at the end of first semester (adjusted R^2 = 36%) is much higher than that of affective strategies, goal strategies and preference for time pressure combined (adjusted R^2 = 4%). In this section, a sequential (or hierarchical) multiple regression is employed to test whether the variables *affe, goal* and *press* have any significant incremental value over *math, phy, chem* and *hrs* in explaining the variance of *wavg.* Further, the effect of effort level is studied after controlling for other noncognitive variables, and hence it is entered in the regression equation after *affe, goal* and *press.* The results of the sequential multiple regression for model 3 are shown in Appendix F.

It can be observed from Table 3.2 that the addition of *affe, goal* and *press* does not make any significant contribution in explaining the variance in *wavg*, after the difference in *math, phy, chem* and *hrs* are already accounted for. For the same reason, the coefficients of these variables are not significantly different from zero (see Figures F.1, F.2, F.3). But the addition of variable effort level *(eff)* to the regression equation results in a significant increase in explaining the variability of *wavg* (the adjusted R^2 increases from 0.361 to 0.383).

Model	**Explanatory variables**	R^2	R^2 **change**	F **change**	$Pr(> F)$
1	*math + phy + chem + hrs*	0.371			
3a	*math + phy + chem + hrs + affe*	0.3723	0.0013	0.4676	0.705
3b	*math + phy + chem + hrs + affe + goal*	0.3748	0.0025	0.9644	0.409
3c	*math + phy + chem + hrs + affe + goal + press*	0.3785	0.0037	1.3561	0.2551
3d	*math + phy + chem + hrs + affe + goal + press+ eff*	0.4021	0.0236	13.783	1.35e−06

Table 3.2: Summary of Sequential Regression Model

The striking difference between model 3d and all the model preceding it is not only the significant R^2 increment due to the significance of newly added *eff* variable, but also the change in significance of the variable *goal* (Figure F.4). It can be observed that, after accounting for differences in students' effort level, the goal related strategies (but not affective strategies or pressure preference) become important in explaining the performance at the end of first semester. In order to understand the above phenomenon, the effects of *affe, goal* and *press*, in the light of other variables, are analyzed separately.

It can be observed from Table 3.3 that when *affe* is regressed separately or when it is regressed along with *goal* and *press* on *wavg*, it is contributing significantly to the variance in *wavg*. But when the variables *math, phy, chem, hrs* and *eff* are introduced in the equation, the magnitude of the relationship between *affe* and *wavg* is reduced and becomes insignificant. This indicates that the strength of the linear relationship between *affe* and *wavg* is largely influenced by the other variables in the equation. This result can be better understood by inferring the *partial correlations* between the variables instead of the regression output though both imply the same. Partial correlation is the correlation of two variables after removing the effects of a third or more other variables. As the variables in our data are ordinal, Spearman's Rank-Order Correlation r_s is used to determine the strength and direction of the monotonic relationship instead of the linear relationship between the variables. The Spearman's (partial) correlations pr_s are determined using the R package *ppcor* [52] and the results are displayed in Table 3.4. While the complete Spearman's correlation between *affe* and *wavg* determined using the "cor" function of R *stats* package [73] is significantly different from zero ($r_s = 0.22$ at $p < 0.001$), the Spearman's partial correlation of $pr_s = -0.02$ between these variables is *not* significantly different from zero (Table 3.4). Besides, *affe* is significantly partially correlated with *math, phy* and *eff*, which in turn are significantly partially correlated with *wavg*. Though *affe* is not significantly partially correlated with variables *chem* and *hrs*, the correlation becomes significant with *chem* (but not with *hrs*) if effects of *math* and *phy* are allowed. Also, all the correlations mentioned above are positive. The above observations indicate that *affe* is significantly and positively correlated with *wavg* only through the variables *math, phy, chem* and *eff*. That is, when the effects of these variables are removed/controlled, the contribution of *affe* to the variability of *wavg* becomes insignificant.

Table 3.3: Contribution by *affe* to variance in *wavg* for different models

Model	**Explanatory variables**	**contribution by *affe* (sr^2) in %**
4a	*affe*	4.7^{***}
4b#	*affe+goal+press*	3.81^{***}
4c	*affe+math+phy+chem* *hrs*	0.13
4d	*affe+math+phy+chem* *hrs+eff*	0.02

# model 4b is same as model 2
*** p $\leq$ 0.0001

Theoretically, the math and science grades, and the effort level act as *mediators* in the relationship between affective strategies and first-semester performance. The direct effect of affective strategies on the first-semester performance is reduced by the indirect or the mediated effect by math and science grades, and effort level [61]. Figure 3.3 shows the situation. In the context of explanatory modeling, affective strategies influence students in acquiring prior-knowledge and altering their effort level, which in turn affects the first-semester performance.

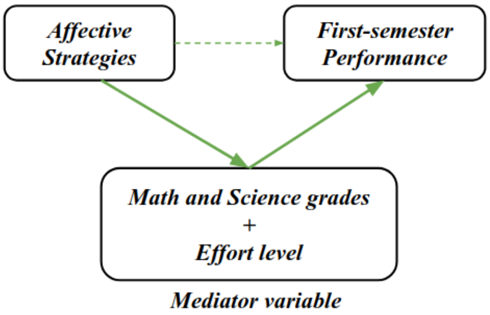


Figure 3.3: The Mediation effect between affective strategies and first-semester performance. Insignificant association is indicated by a dotted green arrow, and the strong positive association is indicated by the thick green arrows

But for *goal*, the trend is opposite to that of *affe*. Table 3.5 shows the results

	math	***phy***	***chem***	***eff***	***hrs***	***press***	***affe***	***goal***	***wavg***
math	1.00	0.26	0.17	-0.03	-0.14	0.01	0.12	0.01	0.29
phy	0.26	1.00	0.23	-0.04	0.08	0.01	0.11	-0.05	0.16
chem	0.17	0.23	1.00	0.08	-0.03	0.04	0.05	0.04	0.24
eff	-0.03	-0.04	0.08	1.00	0.00	-0.18	0.28	-0.29	0.19
hrs	-0.14	0.08	-0.03	0.00	1.00	-0.04	-0.07	-0.01	0.19
press	0.01	0.01	0.04	-0.18	-0.04	1.00	-0.08	0.16	-0.02
affe	0.12	0.11	0.05	0.28	-0.07	-0.08	1.00	0.22	-0.01
goal	0.01	-0.05	0.04	-0.29	-0.01	0.16	0.22	1.00	0.11
wavg	0.29	0.16	0.24	0.19	0.19	-0.02	-0.01	0.11	1.00

(a) Spearman's (partial) correlations pr_s between IVs

	math	***phy***	***chem***	***eff***	***hrs***	***press***	***affe***	***goal***	***wavg***
math	0.00	0.00	0.00	0.37	0.00	0.81	0.00	0.82	0.00
phy	0.00	0.00	0.00	0.31	0.03	0.79	0.00	0.19	0.00
chem	0.00	0.00	0.00	0.03	0.36	0.34	0.17	0.27	0.00
eff	0.37	0.31	0.03	0.00	0.99	0.00	0.00	0.00	0.00
hrs	0.00	0.03	0.36	0.99	0.00	0.28	0.08	0.83	0.00
press	0.81	0.79	0.34	0.00	0.28	0.00	0.04	0.00	0.56
affe	0.00	0.00	0.17	0.00	0.08	0.04	0.00	0.00	0.71
goal	0.82	0.19	0.27	0.00	0.83	0.00	0.00	0.00	0.00
wavg	0.00	0.00	0.00	0.00	0.00	0.56	0.71	0.00	0.00

(b) $p - value$ of the partial correlations test

Table 3.4: Spearman's (partial) correlations pr_s between IVs

when *goal* is regressed separately and along with other variables. The result is again better understood by inferring the partial correlations. First, It can be observed from Table 3.4 that *goal* is significantly partially correlated with *wavg*, $pr_s = 0.11$ at $p < 0.01$ and with *affe*, $pr_s = 0.23$ at $p < 0.01$. It was shown that *affe* is positively correlated with *wavg* only through *math, phy, chem* and *eff.* Accordingly, when *goal* is regressed alongside *affe* in model 2, only *affe* becomes the significant contributor, as the later is more strongly correlated with *wavg* (through other variables) than *goal.* Second, *goal* is also negatively partially correlated with *eff* ($pr_s = -0.29$ at $p < 0.001$), which in turn is positively partially correlated ($pr_s = 0.19$ at $p < 0.01$) with *wavg.* Combined, the direct positive effect of *goal* and the indirect negative effect through *eff* cancels out, resulting in a total effect of *goal* on *wavg* equal to *zero* in models 3(b-c) or in models 5(b-c). Accordingly, the complete correlation between *goal* and *wavg* determined using the "cor" function is not significantly different from zero ($r_s = 0.09$). However, when the effect of *eff* is controlled, as done in model 3d or in model 5d, the direct positive effect of *goal* results in significant contribution to the variance in *wavg.*

Theoretically, effort level acts as a *suppressor* which suppresses any unwanted

Table 3.5: Contribution by *goal* to variance in *wavg* for different models

Model	**Explanatory variables**	**contribution by *goal* (sr^2) in %**
5a	*goal*	1.0
5b	*goal+affe+press*	0.6
5c	*goal+math+phy+chem* *hrs*	0.3
5d	*goal+math+phy+chem* *hrs+eff*	0.8^*

* p $\leq$ 0.05

variance and increases the explanatory value of goal strategies. Figure 3.4 shows the situation. Goal strategies corresponds to test strategies and anxiety scales of LASSI (section 2.3.2). In the context of explanatory modeling, though students with good test-taking strategies and who manage anxiety well are expected to perform well, such students seem to exert less effort to acquire prior-knowledge required for engineering; but, effort exerted is positively associated with academic performance. Combined, these two conflicting effects cancel each other out, resulting in a total effect of goal strategies on performance equal to zero. However, when students who exerted equal effort are considered, their goal strategies become important in explaining their first-semester performance.

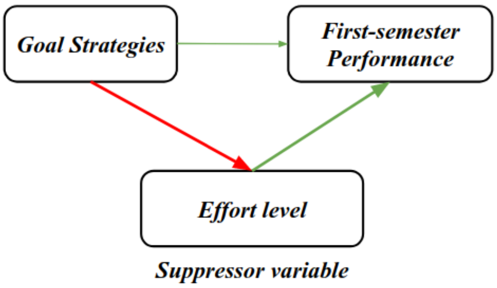


Figure 3.4: The Suppression effect between goal strategies and first-semester performance. Weakly significant association is indicated by faint green arrow, and the strong positive and strong negative association is indicated by thick green and thick red arrows respectively

Finally, the effect of *press* has to be studied. It can be observed from Table 3.4 that *press* is positively partially correlated ($pr_s = 0.16$ at $p < 0.001$) with *goal*. But as inferred previously, since *goal* is negatively correlated with *eff*, *press* is also negatively partially correlated with *eff* ($pr_s = -0.18$ at $p < 0.001$). Thus similar to the inference made for *goal*, the conflicting effects cancel out, resulting in a total effect of *press* on *wavg* equal to zero. Accordingly, the complete correlation between *press* and *wavg*, determined using the "cor" function, is not significantly different from zero ($r_s = -0.05$). However, even when the effects of *eff* and *goal* are controlled, *press* is not yet significantly related to *wavg* as indicated by its partial correlation with *wavg* ($pr_s = -0.02$) (Table 3.4). Hence, preference for time pressure does not affect the first-semester performance and consequently has no incremental value over prior academic knowledge in explaining the variance in *wavg*.

3.3 Research Conclusions

The previous section explained how students' prior-knowledge, and different skills and abilities influence their performance at the end of first semester. While some hypothesized relationships discussed in section 3.1 were proved true, some of them were contradicted. Interestingly, new patterns in the data were uncovered leading to new conclusions. The various statistical conclusions determined in the previous section are converted into research conclusions and are summarized below:

- *Prior academic knowledge operationalized as math and science grades, math level, and the effort exerted to acquire prior knowledge positively influences the first-semester performance strongly than affective and goal strategies.*
- *Affective Strategies positively influence the first-semester performance only in an indirect way through prior academic knowledge and the effort exerted to acquire prior knowledge.*
- *Though goal strategies positively influence the first-semester performance in a direct way, students with well-defined goal strategies tend to exert less effort in acquiring prior-knowledge thereby perform poorly in the first-semester exams*
- *Preference for time pressure has no influence on the first-semester performance.*

3.4 Limitations ?

Explanations are retrospective [81]. While they enhance our understanding of how different characteristics of students influence the first-semester performance, they need not necessarily hold in the future. Hence, our model needs the ability to predict first-semester performance, in addition to explaining, to generalise our inferences to "unseen" incoming students. Further, quantifying the validity of an explanatory model is also difficult. That is, the operationalization of the relationship $\mathcal{F}$ between different theoretical constructs into a statistical model f is difficult to quantify mathematically. It is mainly based on relevant theories [81]. For instance, hypothesized relationships between students' knowledge, skills and abilities is assumed linear by operationalizing F into a multiple linear regression model f, but no mathematical quantification could be provided for the same. Even if a complex nonlinear model is used as f, validating that f accurately represents $\mathcal{F}$ is mostly based on theoretical justifications. Thus, explanatory models are harder to confirm or contradict [81]. Furthermore, it is also important to note that unlike LASSI variables, the operationalization of variables *math, phy, chem, hrs* and *eff* are difficult to validate since there is no measures such as internal consistency coefficients. Hence, the best fit model $\hat{f}$ does not necessarily mean that the model f adequately represents $\mathcal{F}$.

But on the other hand, as Shmueli [81] notices, predictive validity could serve as a *"reality check"* to the relevance of theories as it can be observed whether the predictions are accurate or not. Though the explanatory model discussed in this chapter provides useful explanations through statistical inferences, determining its predictive power could reinforce our understanding of how different traits of students affect their performance. The predictive validity of an explanatory model quantifies its capabilities and could indicate whether an explanatory model f is a valid operationalization of $\mathcal{F}$. Thus statistical model aiming to optimize both explanatory and predictive power provides a better representation of the underlying reality by relating theory to practice [81]. The next chapter is aimed to exploit the uniqueness of both these approaches.

Chapter 4

Explanatory Modeling with Predictive Validity

The statistical modeling approach presented in this chapter combines both explanatory and predictive objectives. As discussed in section 3.4, the important reasons to incorporate predictive quality in an explanatory context is to generalise the statistical inferences to "unseen" observations and to reinforce the assumption that our explanatory model f is a valid operationalization of $\mathcal{F}$. The next section 4.1 presents few practical issues that must be addressed when predictive goals are combined with explanatory goals. Section 4.2 discusses the prediction techniques that are appropriate for the present study. Section 4.3 and Section 4.4 illustrates the explanatory and predictive path of this modeling approach respectively. The final section discusses to what extent this modeling approach addresses the problem undertaken.

4.1 Practical Issues in Combining Prediction with Explanation

There are several important considerations, in the context of the present study, when a statistical model is used for both explanatory and predictive purposes. First, for prediction, the data must be partitioned so that the predictive performance is evaluated on a holdout set or using cross-validation or bootstrap methods [53] to ensure that the performance is not overestimated. This implies that the model is built using training set and its predictive power is evaluated on the holdout set, but its explanatory power is evaluated based on the strength and significance of the model fit to the training set. While using cross-validation methods have many advantages over using a holdout sample in general [53], it is difficult to summarize the statistical inferences obtained in different cross-validation folds for explanatory purposes. Thus, a holdout sample is used, sacrificing some statistical power in explanatory modeling due to a reduced sample size [78].

Second, it is important to remember that PCA was performed on the complete data in section section 2.3.2 to obtain the variables *affe* and *goal*. However, partitioning the data in the current "form" could lead to "leakage" of information from outside the training set in building the model. This will result in overestimating the performance of the model on holdout/testing set. Hence, the data in the original form, containing the LASSI variables, is first partitioned and then PCA is performed only on the training set. The component scores for "unseen" testing set is computed based upon the weights obtained by applying PCA on the training set. Similarly, the transformation of categorical variables, as explained in section 2.3.1, is also performed only on the training set and the conclusion arrived based on the training set is used to transform the variables in the testing set. Fortunately, the resulting "form" of data in the training set is same as that obtained using the complete data.

Third, appropriate modeling techniques for both explanation and prediction have to be determined. For causal explanations, the coefficients of multiple regression model were used to indicate different variables' independent effects on the mean change in *wavg*. But to use this model for prediction, the predicted *wavg* scores should have a universal interpretation in the context of the present study. For instance, a *wavg* of 12 should not be interpreted as a high score by some and as a low score by others. Keeping this in mind, cut-off values have to be determined to discriminate between different *wavg* subranges such that each subrange indicates a unique level of first-semester performance. So ultimately, the goal of prediction is to classify students into different categories with meaningful directional differences [63, p. 1], so that the successive categories indicate improving performance. Thus, a statistical model capable of predicting ordinal outcomes in addition to providing interpretable coefficients (like in multiple linear regression) is preferred for the present study. Accordingly, an Ordinal Logistic Regression is used to classify students into ordered categories of first-semester performance. Further, the coefficients of ordinal logistic regression could also indicate different variables' independent effects of classifying a student into different ordered categories of first-semester performance.

The next practical issue is to determine cut-off values to transform the numerical *wavg* scores to ordinal categories. Table 4.1 displays the name of various courses that students need to take and the corresponding credits that will be obtained upon successful completion. It can be computed from that table that a *cse* in the range 0-40, 40-75 and 75-100 corresponds to "passing at most 2 of the 6 courses", "passing at most 3 or 4 courses" and "passing at most 5 or passing all the 6 courses" respectively. Students corresponding to above three ranges of *cse* can be conveniently interpreted as "at-risk", "moderate-risk" and "no-risk" students, respectively. But it was mentioned that *cse* and *wavg* are highly correlated and Figure 4.1 shows that *cse* of 40 and 75 approximately corresponds to *wavg* of 8.5 and 11.5 based on the linear (regression) fit between them. Thus, *cse* in the ranges 0-40, 40-75 and 75-100 corresponds to *wavg* in the ranges 0-8.5, 8.5-11.5 and 11.5-20 respectively. Hence the original continuous variable *wavg* is cut at 8.5 and 11.5 so as to transform it to a 3-category ordinal variable *wavg_ord* with categories "0-8.5", "8.5-11.5" and "11.5-20" in order.

Further, it can be observed from Figure 4.2 that the frequencies of different categories of *wavg_ord* are approximately equal, which is a desirable property to

Course	Maximum Credits
Analyse, deel 1	6
Toegepaste algebra	5
Algemene en technische scheikunde	7
Toegepaste mechanica, deel 1	5
Probleemoplossen en ontwerpen, deel 1	4
Wijsbegeerte	3

Table 4.1: First semester courses of bachelor of engineering program at KU Leuven and the corresponding number of credits that will be obtained upon successful completion.

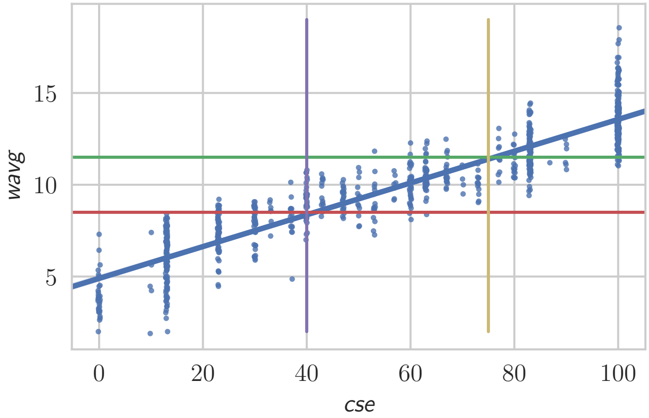


Figure 4.1: Relationship between weighted average *(wavg)* and cumulative study efficiency *(cse)*. The red and the blue line corresponds to a *wavg* of 8.5 and 11.5 respectively. The purple and the yellow line corresponds to a *cse* of 40 and 75 respectively.

avoid deceptive performance results [45]. Hence, while partitioning, the relative frequencies of different categories of *wavg_ord* are approximately preserved in both the training and the testing set so as not to incur any performance loss due to class-imbalance problem [35].

4.2 Ordinal Logistic Regression

Logistic regression relates the independent effects of IVs on the natural log of odds of an event's occurrence (π). It is formally put as $\ln(\frac{\pi}{1-\pi}) = b_0 + b_1p_1 + b_2p_2 + ... + b_np_n$ where $p_1, p_2, ..., p_n$ are IVs and $b_1, b_2, ..., b_n$ are coefficients representing the expected change in the log odds for different values of the IVs; b_0 represents the intercept. When the outcome (J) is binary (if seen as $J = 1$ or $J = 0$), π represents the occurrence of $J = 1$, that is, $\pi = P(J = 1)$. However, different representations of π are possible for ordinal outcomes, each giving rise to different logistic model. Agresti

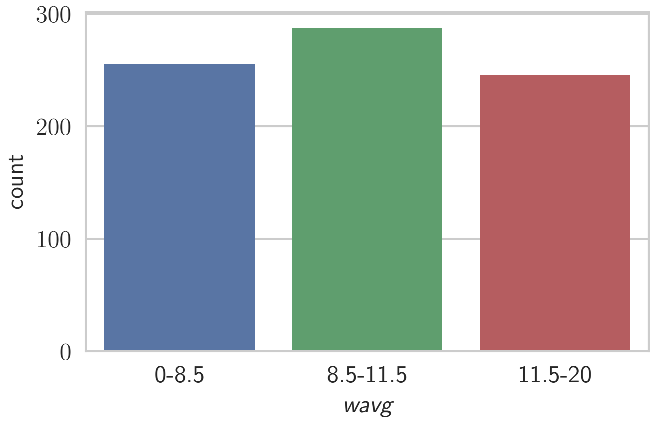


Figure 4.2: Frequency distribution of *wavg_ord*

[6, pp. 173-189], Hosmer et al. [47, pp. 288-305] and O'Connell [63] provide an elaborate discussion on the different types of logistic regression models for ordered outcomes. In sum, in every model, the K-level ordinal outcome variable is used to transform the ordinal classification problem into *K-1* ordinary binary classification problems, each with a different representation of π.

In the present study, "cumulative odds" (CO) ordinal regression model is used where π represents the cumulative probability for J at different levels. Thus the logits (or log odds) now represent the probability of the outcome falling below or at a category compared to the probability of the outcome falling above that particular category. Formally, the logits of the cumulative odds model are $\ln(\frac{P(J \leq k)}{P(J > k)})$ for $k = 1, 2, ...K-1$. In the present study, two logits have to be used since $K = 3$ for *wavg_ord*, and they are $\ln(\frac{P(J=\text{“}0-8.5''\text{)}}{P(J=\text{“}8.5-11.5'')+P(J=\text{“}11.5-20'')})$ and $\ln(\frac{P(J=\text{“}0-8.5'')+P(J=\text{“}8.5-11.5'')}{P(J=\text{“}11.5-20'')})$. It can also be noticed that the cumulative probabilities reflect the ordering of the outcome since $P(J \leq \text{“}0 - 8.5'') \leq P(J \leq \text{“}8.5 - 11.5'') \leq P(J \leq \text{“}11.5 - 20'')$. Further, the effect of each of the IVs is assumed constant for all $K - 1$ cumulative logits to result in a parsimonious model and ease the interpretation of the effects of IVs. This property is called as the proportional (or parallel) odds assumption of the model [6, p. 189]. Cumulative odds model with proportional odds property is used in the present study because this is what is expected to result when an underlying continuous variable is discretized to derive an ordinal outcome variable [47, p. 298]. Further, the interpretation of input variables' effects on both the ordinal and underlying continuous variables are expected to be the same irrespective of the choice of categories for J, as both the ordinal and continuous variables represent the same underlying phenomenon. Thus an implication of using this model is that the effects of IVs should roughly be the same if future works aim to use cut-off values different from that followed in this study to discretize *wavg* [6, p. 189].

4.3 Explanatory Path

4.3.1 Models Formulation

Table 4.2 displays the information about the different models that are required to test various hypotheses discussed in section 3.1 using the cumulative odds model. The resulting inferences are expected to be the same as the key inferences made in section 3.3. This is because, as discussed in the last section, the interpretation of results of a cumulative odds model is expected to be consistent with its equivalent multiple regression model due to the proportional odds property.

Course	**Response ~ Explanatory variables**
6a	$wavg \sim math + phy + chem + hrs$
6b	$wavg \sim math + phy + chem + hrs + eff$
7	$wavg \sim affe + goal + press$
8	$wavg \sim math + phy + chem + hrs + affe + goal + press + eff$

Table 4.2: Information about different explanatory (ordinal logistic regression) models

4.3.2 Analyses of Results

CO ordinal logistic regression is performed using the function “clm” of the R package *ordinal* [18]. Multiple criterias are recommended in assessing and comparing the fit of different logistic regression models [47, pp. 143-203]. In the present study, Deviance statistic (D) and different pseudo R^2 metrics are used to assess and compare how well these models explain the data. The D statistic measures how poorly the fitted model represents the data by comparing the likelihood of the fitted model to that of the saturated model. Pseudo R^2 measures the proportional reduction of error (in terms of log-likelihood) relative to the null model by comparing the likelihood of the fitted model to that of the intercept-only model. Table 4.3 compares the fit of different models to the data, based on D statistic and pseudo R^2 metrics.

Model	**McFadden R^2**	**Cox_Snell R^2**	**Nagelkerke R^2**	**D statistic**
6a	0.16	0.23	0.28	1002.318
6b	0.17	0.25	0.31	981.215
7	0.02	0.04	0.05	1158.989
8	0.19	0.27	0.33	966.200

Table 4.3: Comparision of different CO models using Deviance statistic (D) and different pseudo R^2 metrics

It can be observed that the deviance D from saturated model is much lower for models 6(a-b) compared to model 7 showing that the models 6(a-b) are fitted less poorly than model 7. Similarly, the different R^2 metrics indicate that the proportional reduction of error by the variables of models 6(a-b) is much higher than that of model 7. Next, Figure 4.3 shows the effect of variable *eff* over the variables *math, phy, chem* and *hrs* using the likelihood-ratio (LR) test implemented in the “add1” function[1] of R package *stats* [73]. The LR statistic of 21.1 with $df = 2$ indicates that *eff* contributes significantly ($p < 0.001$) to model 6a in differentiating the students of different performance levels. Further, the independent effects of different variables of models 6(a-b) using the “drop1” function[1] of R package *stats* [73] show that all variables are significant (Figure 4.4). It is to be noted that the inferences are made using the “drop1” and “add1” functions rather than through the coefficients in the model outputs. This is because the former methods employ likelihood-ratio test while the later employs Wald test, for significance testing. Since, Wald test is shown to be less robust than likelihood-ratio test, the later is preferred for significance tests [47, p. 16]. In sum, these results, similar to that obtained in the last chapter, show that prior academic knowledge and the effort level differentiate the students of different performance level more strongly than students’ affective strategies, goal strategies and preference of time pressure.

```
Call: add1(model6a, scope = ~.+eff, test = "Chi")

Single term additions

Model:
wavg_ord ~ math + phy + chem + hrs
       Df    AIC    LRT  Pr(>Chi)
<none>    1028.3
eff     2 1011.2 21.104 2.615e-05 ***
```

Figure 4.3: Results of the *R* function “add1” to investigate the effect of variable *eff* over the variables *math, phy, chem* and *hrs*

Similarly, Figure 4.5 shows the independent effects of different variables of model 7 using the “drop1” function and Figure 4.6 shows the effect of adding variable *affe* to models 6(a-b) using the “add1” function. It can be observed that *affe* is a significant explanatory variable when considering only *goal* and *press*. However, when variables *math, phy, chem, hrs* and *eff* are considered in the equation, *affe* becomes insignificant in discriminating the students of different performing levels. Thus the interpretation of the effect of *affe* is same as that observed in the last chapter. That is, affective strategies influence the first-semester performance only indirectly through prior knowledge and effort level.

In addition, the effect of adding variable *goal* to models 6(a-b) is shown in Figure 4.7. It can be observed that the effects of goal becomes significant ($p = 0.05$) only after controlling for the effects of *eff*, as inferred in the previous chapter. Likewise, it

[1] Interpretation of significance codes in the *R* output are as follows: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
Call: drop1(model6a,test=‘‘Chi’’)

Single term deletions

Model:
wavg_ord ~ math + phy + chem + hrs
       Df    AIC    LRT  Pr(>Chi)
<none>    1028.3
math    3 1062.5 40.174 9.786e-09 ***
phy     3 1032.1  9.763  0.020695 *
chem    3 1053.4 31.111 8.056e-07 ***
hrs     2 1033.6  9.310  0.009513 **
```

```
Call: drop1(model6b,test=‘‘Chi’’)

Single term deletions

Model:
wavg_ord ~ math + phy + chem + hrs + eff
       Df    AIC    LRT  Pr(>Chi)
<none>    1011.2
math    3 1045.3 40.065 1.032e-08 ***
phy     3 1015.6 10.404   0.01543 *
chem    3 1031.0 25.786 1.057e-05 ***
hrs     2 1016.1  8.842   0.01202 *
eff     2 1028.3 21.104 2.615e-05 ***
```

Figure 4.4: Results of the *R* function “drop1” to investigate the independent effects of different variables of models 6(a-b)

```
Call: drop1(model7,test=‘‘Chi’’)

Single term deletions

Model:
wavg_ord ~ affe + goal + press
       Df    AIC     LRT Pr(>Chi)
<none>    1181.0
affe    3 1190.8 15.7911 0.001251 **
goal    3 1181.1  6.1253 0.105669
press   3 1178.6  3.6317 0.304073
```

Figure 4.5: Results of the *R* function “drop1” to investigate the independent effects of different variables of models 7

```
Call: add1(model6a,
          scope = ~.+affe, test = ‘‘Chi’’)

Single term additions

Model:
wavg_ord ~ math + phy + chem + hrs
       Df    AIC    LRT Pr(>Chi)
<none>    1028.3
affe    3 1032.8 1.5192   0.6779
```

```
Call: add1(model6b,
          scope = ~.+affe, test = ‘‘Chi’’)

Single term additions

Model:
wavg_ord ~ math + phy + chem + hrs + eff
       Df    AIC    LRT Pr(>Chi)
<none>    1011.2
affe    3 1015.5 1.7448    0.627
```

Figure 4.6: Results of the *R* function “add1” to investigate the effect of adding variable *affe* to models 6(a-b)

can be observed from Figure 4.8 that press has no influence on the first-semester performance.

```
Call: add1(model6a,                            Call: add1(model6b,
          scope = ~.+goal, test = ‘‘Chi’’)              scope = ~.+goal, test = ‘‘Chi’’)

Single term additions                          Single term additions

Model:                                         Model:
wavg_ord ~ math + phy + chem + hrs             wavg_ord ~ math + phy + chem + hrs + eff
       Df    AIC    LRT Pr(>Chi)                      Df    AIC   LRT Pr(>Chi)
<none>    1028.3                               <none>    1011.2
goal    3 1029.8 4.4883   0.2133               goal    3 1007.5 9.749  0.02083 *
```

Figure 4.7: Results of the *R* function “add1” to investigate the effect of adding variable *goal* to models 6(a-b)

```
Call: add1(model6a,                            Call: add1(model6b,
          scope = ~.+press, test = ‘‘Chi’’)             scope = ~.+press, test = ‘‘Chi’’)

Single term additions                          Single term additions

Model:                                         Model:
wavg_ord ~ math + phy + chem + hrs             wavg_ord ~ math + phy + chem + hrs + eff
       Df    AIC    LRT Pr(>Chi)                      Df    AIC    LRT Pr(>Chi)
<none>    1028.3                               <none>    1011.2
press   3 1030.3 4.0036   0.2611               press   3 1015.8 1.3703   0.7125
```

Figure 4.8: Results of the *R* function “add1” to investigate the effect of adding variable *press* to models 6(a-b)

Further, while the coefficients of model 8 (Figure G.2) and model 3d (multiple linear regression) imply different meanings, the interpretation of effects of all the IVs are exactly the same. For instance, according to the model 3d (Figure F.4), students with high math background are expected to get a wavg score of 2.9 higher than those with low math background on average, assuming that their other traits are same. Whereas, according to the odds ratios of model 8 (Figure G.1), the odds of being in the higher performing levels is, on average, around 8 times higher for students with high math background than those with low math background, assuming that their other traits are same. Evidently, both convey the same effect in different ways. Thus, in the context of explanatory modelling, the multiple linear regression model and the ordinal logistic regression model are nothing but two different operationalizations of proposed theoretical model $\mathcal{F}$, thus yielding similar results. Further, the above results also indicate that data partitioning has not lead to any significant reduction in the power of statistical tests.

4.3.3 Evaluation of Assumptions

The inferences made hitherto are under the assumption of proportional odds. Though the assumption of proportional odds is motivated by the underlying continuous outcome variable, it still needs to be investigated to ensure accurate modelling. Harrell Jr [43, p. 313,315] notices that statistical tests such as score test often reject proportional odds assumption even when they hold and hence is unreliable in many practical applications. However, he suggests the use of a graphical method to test the proportional odds assumption. When the assumption holds, the coefficients of different levels of an IV for different binary logits at varying cut points are the same. Hence, when an IV is regressed separately, the difference between the predicted values of the logits, at different cut points of the DV, for varying levels of that explanatory variable should also be the same [43, p.315].

Figure 4.9 displays the predicted values of different binary logistic regressions at varying cut points on the DV (each of the cut points 1,2 and 3 represents the categories "0-8.5", "8.5-11.5" and "11.5-20" respectively), when IVs are regressed one at a time. If the proportional odds assumption holds, then the difference between the predicted values of both the logits must be approximately same at all levels of IVs. To facilitate visual inspection, the first set of coefficients (column $Y \geq 2$ in Figure 4.9) are normalized by subtracting them from the second set of coefficients (column $Y \geq 3$ in Figure 4.9) so that there is a common reference point [2]. Hence now, all the modified predicted values must approximately be the same. The "plot" function from R package *Hmisc* [37] is used to plot the modified predicted values and is displayed in Figure 4.10. It can be observed that the predicted values of the logit are approximately around -1.5 for most levels of different IVs. Agresti [6, p. 185] recommends not to reject the proportional odds assumption unless the lack of fit of the ordinal model is very poor in a practical sense. Also, from Figure 4.10, it can be observed that the proportional odds assumption could be violated due to the variables *math, phy, chem* and *affe*, as some of their levels (or categories) show large deviance from -1.5. Hence, LR test is performed between the original model 8 with proportional odds assumption for all variables and the modified model (model8_nom) with the proportional odds assumption relaxed for variables *math, phy, chem* and *affe*. Hence, these variables are treated as nominal rather than ordinal. Figure 4.11 shows the results. As it can be observed that making some variables nominal in model8_nom did not result in any significant difference in the output, the assumption of proportional odds is considered satisfied for all the variables.

4.4 Predictive Validity

This section evaluates the predictive power of different ordinal logistic regression (CO) models on the holdout sample. Table 4.4 shows the predictive performance of the models in terms of recall, precision and F1-score. Recall for a class answers the question: what fraction of the observations that are predicted as belonging to that class truly belong to that class, and precision for a class answers the question: what fraction of the observations that are truly belonging to that class are predicted. $F1$

```
Call: s <- with(dat, summary(as.numeric(wavg_ord) ~ math + phy + chem + hrs + eff +
                                                  affe + goal + press, fun=sf))

as.numeric(wavg_ord)     N= 542

+-------+-----------+---+----+----------+----------+
|       |           |N  |Y>=1|Y>=2      |Y>=3      |
+-------+-----------+---+----+----------+----------+
|math   |Low        |118|Inf |-0.2384110|-2.2749336|
|       |Low_medium |195|Inf | 0.4700036|-1.4186596|
|       |High_medium|181|Inf | 1.3252585|-0.2780202|
|       |High       | 48|Inf | 2.7080502| 1.3350011|
+-------+-----------+---+----+----------+----------+
|phy    |Low        | 90|Inf |-0.5947071|-2.8332133|
|       |Low_medium |184|Inf | 0.5340825|-1.1574528|
|       |High_medium|202|Inf | 1.1932775|-0.5266990|
|       |High       | 66|Inf | 1.8458267| 0.4946962|
+-------+-----------+---+----+----------+----------+
|chem   |Low        |116|Inf |-0.4560174|-2.7454377|
|       |Low_medium |196|Inf | 0.6551198|-1.1819939|
|       |High_medium|174|Inf | 1.2750687|-0.3246611|
|       |High       | 56|Inf | 2.1202635| 0.8303483|
+-------+-----------+---+----+----------+----------+
|hrs    |Low        |  9|Inf |-0.6931472|-2.0794415|
|       |Medium     |232|Inf | 0.5477813|-0.9867640|
|       |High       |301|Inf | 0.8362480|-0.6831968|
+-------+-----------+---+----+----------+----------+
|eff    |Low        |218|Inf | 0.2025243|-1.3466291|
|       |Average    |194|Inf | 0.8034952|-0.8034952|
|       |High       |130|Inf | 1.4853853|-0.1541507|
+-------+-----------+---+----+----------+----------+
|affe   |Low        | 86|Inf | 0.1865860|-1.9195928|
|       |Low_medium |190|Inf | 0.6074917|-0.9496890|
|       |High_medium|183|Inf | 0.7932306|-0.5962974|
|       |High       | 83|Inf | 1.2144441|-0.2666287|
+-------+-----------+---+----+----------+----------+
|goal   |Low        | 84|Inf | 0.2392297|-1.2992830|
|       |Low_medium |171|Inf | 0.7462570|-0.6931472|
|       |High_medium|207|Inf | 0.6714857|-0.8495994|
|       |High       | 80|Inf | 1.0986123|-0.6190392|
+-------+-----------+---+----+----------+----------+
|press  |Low        | 79|Inf | 0.8292794|-0.4372138|
|       |Low_medium |219|Inf | 0.6931472|-0.6726686|
|       |High_medium|135|Inf | 0.5306283|-1.0498221|
|       |High       |109|Inf | 0.7487170|-1.2119410|
+-------+-----------+---+----+----------+----------+
|Overall|           |542|Inf | 0.6820973|-0.8262997|
+-------+-----------+---+----+----------+----------+
```

Figure 4.9: Predicted values of different binary logits at varying cut points. The cut points 1,2 and 3 represent the categories "0-8.5", "8.5-11.5" and "11.5-20", respectively, of *wavg_ord*

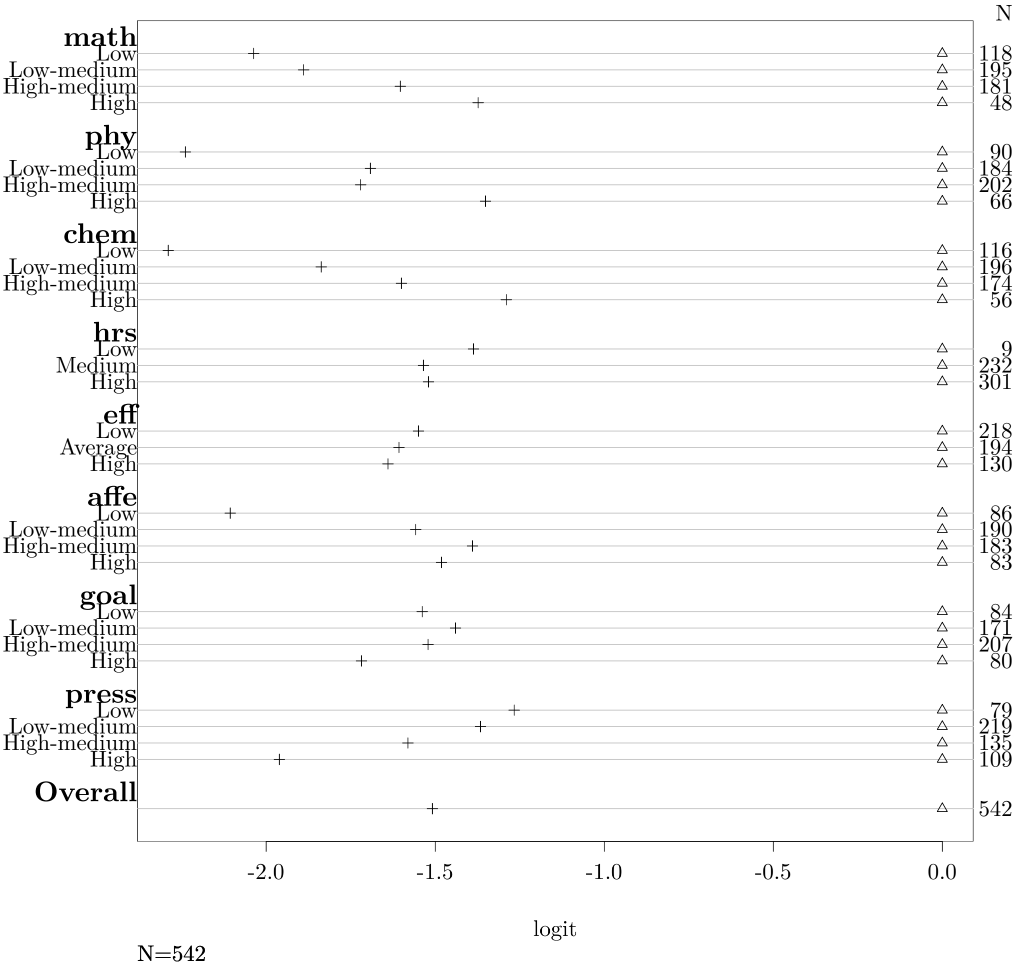


Figure 4.10: Graphical method for testing Proportional Odds assumption. The "+" marker indicates the modified predicted values. If the markers are diffused for most level of an IV, it indicates that the proportional odds assumption may not hold for that particular IV.

```
Call: anova(model8, model8_nom)

Likelihood ratio tests of cumulative link models:

           formula:                                     nominal:
model8     wavg_ord ~ math + phy + chem + hrs +                 ~1
                        eff + affe + goal + press
model8_nom wavg_ord ~ math + phy + chem + hrs +        ~math + phy + chem + affe
                        eff + affe + goal + press

           link: threshold:
model8     logit flexible
model8_nom logit flexible

           no.par    AIC  logLik LR.stat df Pr(>Chisq)
model8         24 1014.2 -483.10
model8_nom     36 1027.5 -477.74  10.714 12     0.5536
```

Figure 4.11: *R* output for ANOVA test between model 8 with and without proportional odds assumption for variables *math, phy, chem* and *affe*

score represents how good the combination of recall and precision is, by computing their harmonic mean. These measures are preferred here because they offer more information about the model's predictive performance than classification accuracy. The objective of performing predictive validity of explanatory models is two-fold. First, to strengthen the validatity of our statistical (explanatory) models so as to indicate that they adequately represent the theoretical causal structure $\mathcal{F}$. Second, to generalise the inferences to "unseen" observations.

Table 4.4(a-b) shows that model 7 with variables *affe, goal* and *press* with an average $F1$ score of only 0.39 has very poor predictive performance compared to model 6b with variables *math, phy, chem, hrs* and *eff* whose average $F1$ score is 0.54. Not surprisingly, model 7 is also the one with worst performance of all the models. Further, Table 4.4(c-d) shows that model 8 with and without variables *affe* and *press* have almost same predictive performance. In fact, a closer look at Table 4.4(c-d) indicates that even variable *goal* is not contributing much to the predictive power. It can be noted that all the above observations are in consistent with the inferences made earlier in the context of explanatory modeling. However, it can be observed that though *goal* was shown as a significant explanatory variable in the context of explanatory modeling, it does not contribute anything to the predictive power. It should be remembered that significance of a variable implies only the presence of a relationship with DV and not its strength. The partial contribution to the variance indicates the strength of the relationship and it was observed earlier that *goal* contributes only less than 1% (Table 3.5). This is the reason why *goal* does not contribute anything to the predictive power. Theoretically, the above observations collectively imply that prior academic knowledge predicts the first-semester performance much better than the affective strategies, the goal strategies and the preference for time pressure combined.

	Precision	**Recall**	$F1$ **score**	**support**
0-8.5	0.64	0.57	0.60	60
8.5-11.5	0.60	0.52	0.55	54
11.5-20	0.42	0.52	0.46	64
avg / total	0.55	0.53	0.54	178

(a) model 6b: *wavg* ~ *math* + *phy* + *chem* + *hrs* + *eff*

	Precision	**Recall**	$F1$ **score**	**support**
0-8.5	0.47	0.35	0.40	60
8.5-11.5	0.41	0.31	0.36	54
11.5-20	0.35	0.50	0.41	64
avg / total	0.41	0.39	0.39	178

(b) model 7: *wavg* ~ *affe* + *goal* + *press*

	Precision	**Recall**	$F1$ **score**	**support**
0-8.5	0.63	0.60	0.62	60
8.5-11.5	0.63	0.59	0.61	54
11.5-20	0.41	0.45	0.43	64
avg / total	0.54	0.55	0.55	178

(c) model 8: *wavg* ~ *math* + *phy* + *chem* + *hrs* + *eff* + *affe* + *goal* + *press*

	Precision	**Recall**	$F1$ **score**	**support**
0-8.5	0.63	0.57	0.60	60
8.5-11.5	0.61	0.57	0.59	54
11.5-20	0.40	0.45	0.42	64
avg / total	0.54	0.55	0.55	178

(d) model 8 without *affe* and *press*: *wavg* ~ *math* + *phy* + *chem* + *hrs* + *eff* + *goal*

Table 4.4: Predictive performance of different Cumulative Odds models

Since the predictive performance of different models are in consistent with the statistical inferences made hitherto, our explanatory models can now be seen as valid operationalizations of the theoretical causal structure $\mathcal{F}$. However, these CO models cannot be used to predict future students' first-semester performance levels as the predictions are not accurate enough. That is, the generalising ability of these models are not satisfactory. For instance, for model 8 with all variables, the low recall of only 0.60 for the outcome class "0-8.5" implies the high risk of not identifying too many "at-risk" students, and similarly, the low precision of only 0.41 for the outcome class "11.5-20" implies the high risk of misspecifying too many students as "no-risk".

4.5 Limitations ?

The CO model is a straightforward extension of ordinary logistic regression model and thus it classifies students into different performance levels only based on the linear combination of the IVs. But an overall poor predictive power (for model 8) observed in the last section indicates the presence of complex patterns and relationships between the IVs and DV or among the IVs or both, that are hard to hypothesize. The predictive power displayed for the CO model was only useful to justify the simple hypotheses formulated in section 3.1, but not sufficient enough to generalise the inferences to "unseen" observations. However, the models trying to optimize both explanatory and predictive goals cannot be made complex since it will complicate the interpretation of the underlying relationship between the constructs. Thus, this leads to a chicken and egg situation!

However, using predictive (algorithmic) models in such contexts can help capture the reality better as it can be made complex by focusing only on the predictions of *wavg_ord* classes, without considering the need for explaining the association between the variables. But such models often lack interpretability which is inappropriate for our problem as discussed in section 1.2.2. In order to avoid such problems, instead of combining explanatory and predictive goals as recommended by Shmueli [81], another approach would be to construct accurate predictive models and comprehensible approximations to them separately for interpretation as suggested by Craven and Shavlik [21] and Domingos [27]. The next chapter demonstrates this approach in detail and discusses its advantages over the methods described so far.

Chapter 5

Predictive Modeling and Comprehensible Approximations

This chapter constructs an accurate predictive model to classify incoming students into different performance levels (“at-risk”/ “moderate-risk”/ “no-risk”) and to explain the reasons behind the predictions. As mentioned earlier, explaining here refers to “qualitative understanding” of the individual predictions in contrast to “causal explanation” in the explanatory modeling context. A popular data mining algorithm called Extreme Gradient Boosting (XGBoost), which is based on the principles of Gradient Boosting Trees [39], is used for prediction. An explanation technique called Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro et al. [76] is employed to reason the individual predictions of the model. Section 5.1 describes how these two algorithms can be used to tackle the problem undertaken by the present study and section 5.2 discusses the tuning of the model parameters. The predictive performance of the model is discussed in section 5.3, and section 5.4 discusses about assessing trust in our model’s performance by understanding the mechanics behind the predictions. The last section section 5.5 discusses to what extent this modeling approach has addressed the problem undertaken.

5.1 Model Design using XGBoost with LIME

Gradient Boosting algorithm, in short, 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 decision tree algorithm, incorporates various additional improvements such as better regularized formulation and column subsampling to avoid overfitting, ability to handle sparsity pattern in data, and parallelization [15]. It is also shown to outperform many other data mining algorithms on real-world datasets in both speed and performance

[16, 68]. A detailed discussion on the working of this algorithm is given by Chen and He [16].

But before using this algorithm, the classification setting defined in the previous chapter is modified in one aspect with an objective to get more accurate predictions: two binary classifiers, defined in the style of CO model without proportional odds property, is preferred over a single multinomial classifier. That is, one classifier is used to distinguish students of class "≤8.5" ("at-risk") against students of class ">8.5" ("moderate-risk" or "no-risk") while another classifier is used to distinguish students of class "≤11.5" ("at-risk" or "moderate-risk") against students of class ">11.5" ("no-risk"). An illustration of the two-binary-classifiers model and the interpretation of its outcome is given in Figure 5.1 and Table 5.1 respectively . The reason behind this design of model is that, in practice, it is often difficult to identify students in the "moderate-risk" zone than those in either of the other two categories. Hence, in the current design of the model, "moderate-risk" ("8.5-11.5") students are not explicitly identified, but those students who are identified as belonging to classes ">8.5" and "<11.5" by the first and the second classifier respectively are interpreted implicitly as "moderate-risk" students. Further, multiple binary classifiers are often easy to train and optimize than a single multinomial classifier as shown by some studies such as [7] and [40].

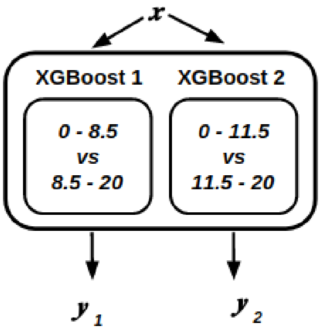


Figure 5.1: Illustration of two-binary-classifiers predictive model

y_1	y_1	**Interpretation**
0 - 8.5	0 - 11.5	at-risk
8.5 - 20	0 - 11.5	moderate-risk
8.5 - 20	11.5 - 20	no-risk

Table 5.1: The interpretation of different outcomes

Next, it was discussed in section 1.2.2 why the reasons behind the predictions need to be investigated when decisions are made based on them in a social context. In the context of the present study, LIME is used to explain the prediction of the two binary classifiers ("≤8.5" vs ">8.5" and "≤11.5" vs ">11.5"). Local Interpretable Model-agnostic Explanations (LIME) by Ribeiro et al. [76] is an algorithm that can present explanations that refer to *"textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components....and the model's prediction"*. LIME, in short, approximates a complex algorithm in the neighbourhood of an observation (for which prediction is required) with a simple interpretable model like linear regression. The authors of LIME also show that their technique is superior to other related techniques by assessing "trust" in individual predictions and by demonstrating how explanations are faithful to the models. It is interesting to note that, the above process of approximating a complex model locally to explain the individual predictions reflects the idea suggested by Craven and Shavlik [21] and Domingos [27], who recommended the use of complex models in practice and the construction of comprehensible approximations to these complex models separately. Further, in the present study, the LIME explanations are also used to get a global understanding of our model's predictive performance and to check its consistency with the observations made in the previous chapters.

5.2 Parameters Tuning

The python implementation of XGBoost is used in the present study and it comes with a multitude of parameters that must be tuned to achieve good performance. Grid Search with 5-fold cross-validation function, available in scikit-learn python library, is used to tune the parameters of two XGBoost classifiers separately. These parameters are tuned such that they optimize two criterias that are important in the context of the present study.

First, the parameters that are responsible to get the best trade-off between bias and variance are tuned to get good generalization ability. The XGBoost parameters documentation provides a detailed account on these parameters. Appendix H shows the final values of these parameters that are associated with the best performance of our model. All parameters except the number of boosting trees take equal values for both classifiers. An interesting observation that can be made here is, while an ensemble of 90 trees is required to get the least generalisation (or testing set) error for the first classifier ("≤8.5" vs ">8.5"), an ensemble of only 9 trees is sufficient for the second classifier ("≤11.5" vs ">11.5"). This difference in model complexity hints that differentiating between "at-risk" and "moderate-risk" students is difficult than differentiating between "no-risk" and "moderate-risk" students, under a reasonable assumption that "at-risk" and "no-risk" students can mostly be easily distinguished.

Second, the parameters that rebalance the dataset to address the class-imbalance problem is tuned, as it can be observed from Figure 5.2 that the output classes are imbalanced in both of our binary classification problems. However, it is important to note that rebalancing of the dataset will be a problem when the true probability of classifying an observation into one of the classes is expected to be much higher than others. But, in our case, the students have equal odds of being classified into ">8.5" against "≤8.5" or "≤11.5" against ">11.5", and the probability of them falling into one of the two classes depends only on their characteristics x. Hence, the presence of imbalanced classes in our dataset does not indicate that the odds of being classified into one of the two classes are higher than the other. The classes are imbalanced in our case only because the frequencies of original three categories ("0-8.5", "8.5-11.5", "11.5-20") of *wavg_ord* are approximately equal (Figure 4.2), and hence, when any two of these three categories are combined (to get ">8.5" or "≤11.5"), the frequency of one of the resulting categories will always be double of the other.

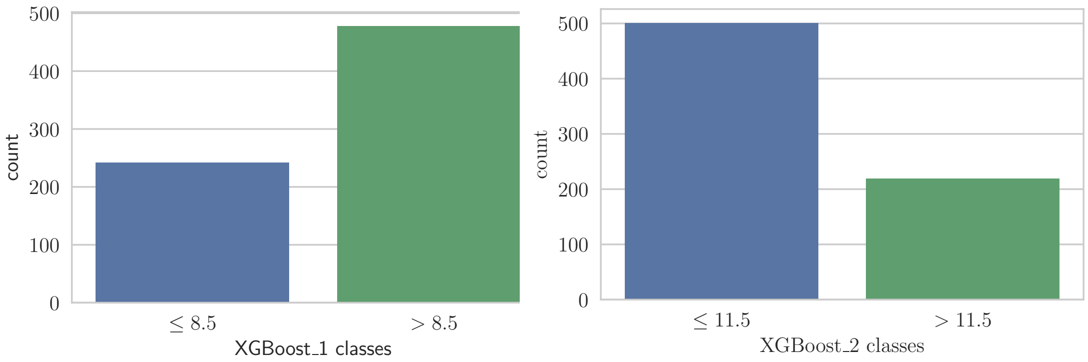


Figure 5.2: Class distribution of two-binary-classifiers predictive model

5.3 Model Evaluation using Predictive Performance

In this section, our two-binary-classifiers model is evaluated based on its predictive performance on the holdout/testing set. Different metrics are used for each of the two classifiers. This is because, for the first classification problem ("≤8.5" vs ">8.5"), the cost of not correctly identifying true "at-risk" ("≤8.5") students will be much higher than the cost of misclassifying students as "at-risk" in a practical sense. Hence, for the class "≤8.5", its recall could be maximised even at the cost of low precision. In addition, the precision of correctly classifying students who are not truly "at-risk" (">8.5") is more important than correctly identifying all of them so as to minimize the misclassification of truly "at-risk" students as being not "at-risk". Therefore, the first classifier is evaluated based on the recall of class "≤8.5" and the precision of class ">8.5". Table 5.2 shows that both of these measures take high values (≥0.80) for the first classifier.

	Precision	**Recall**	$F1$ **score**	**support**
≤8.5	0.64	0.80	0.71	60
>8.5	0.88	0.77	0.82	54
avg / total	0.80	0.78	0.79	178

Table 5.2: Predictive performance of XGBoost_1

Similarly, for the second classifier, the recall of class "≤11.5" is important as the practical cost of not correctly identifying truly "at-risk" or "moderate-risk" students is high. Further, the precision of class ">11.5" is also important so as not to misclassify students as "no-risk" who truly should be identified as either "at-risk" or "moderate-risk". It can be observed from Table 5.3 that the recall of class "≤11.5" is high at 0.85. The high recall of both the classes "≤8.5" (0.80) and "≤11.5" (0.85) indicate that true "at-risk" and "moderate-risk" students can be identified accurately. In contrast, the precision of class ">11.5" is relatively low at 0.68 indicating the significant number of students who truly are in the "at-risk" or "moderate-risk" zone are misclassified into "no-risk" zone. However, it was observed earlier that very less proportion of truly "at-risk" students will be classified as being not "at-risk" since the precision of class ">8.5" is high at 0.88. Thus, the low precision of class ">11.5" is largely due to the misclassification of students in the "moderate-risk" zone than those in the "no-risk" zone.

	Precision	**Recall**	**$F1$ score**	**support**
≤8.5	0.87	0.85	0.86	60
>8.5	0.68	0.70	0.69	54
avg / total	0.81	0.81	0.81	178

Table 5.3: Predictive performance of XGBoost_2

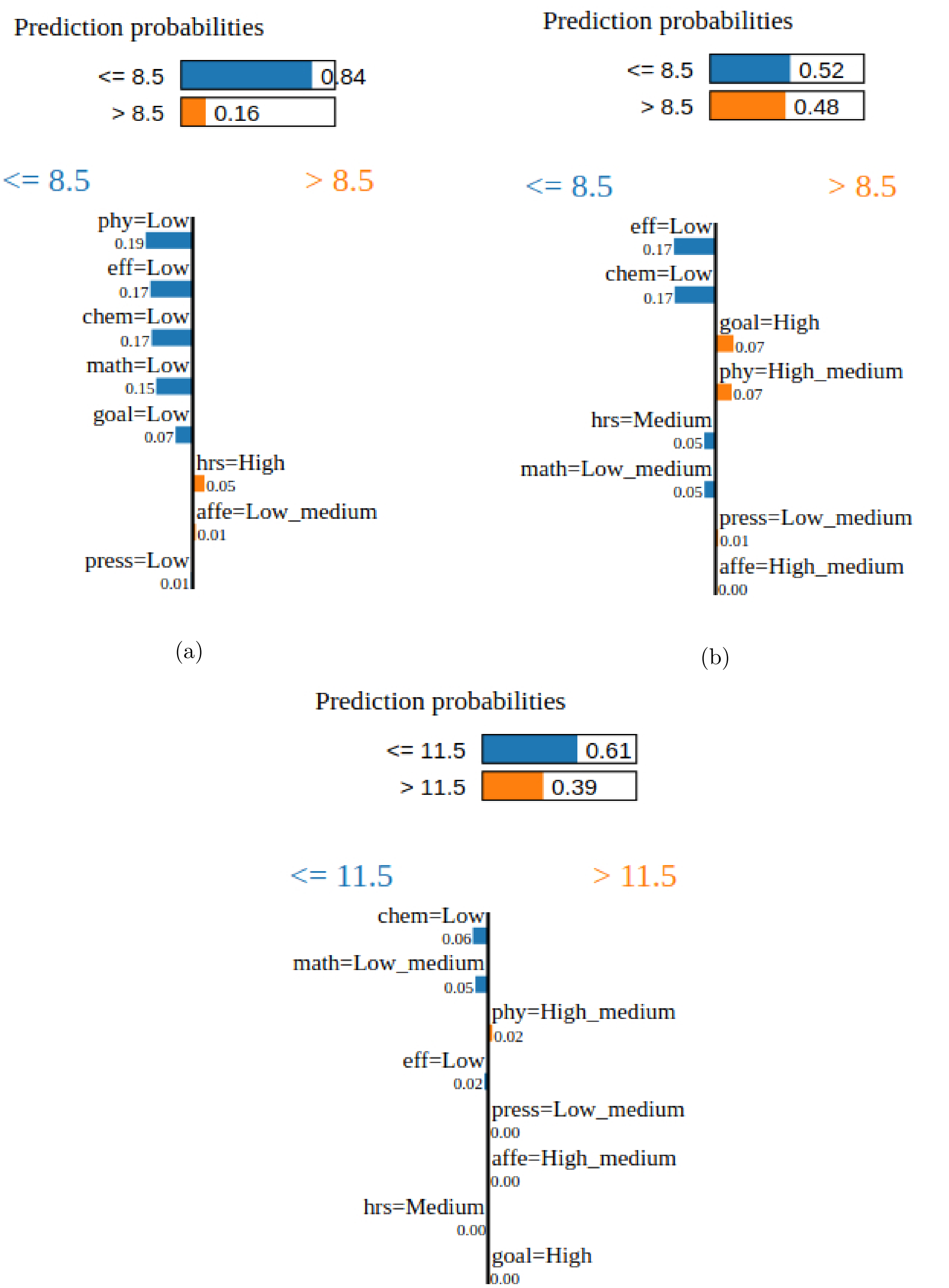
In sum, based on the results obtained, our model is capable of correctly identifying true "at-risk" and "moderate-risk" students with a high recall. Further, while true "at-risk" students have a very less chance of being classified into other classes, true "moderate-risk" students stand a relatively high chance to be classified as "no-risk". In other words, it is difficult to distinguish "no-risk" students from "moderate-risk" students than from "at-risk" students.

5.4 Trusting Model's Predictive Performance

This section assesses the trust in our model's performance by understanding the mechanics behind the predictions using the explanations provided by LIME. The output of LIME for some of the observations whose true and predicted classes are both "≤8.5" is shown in Figure 5.3 . The bar chart in each explanation represents the importance (or weights) given to the each of the IV-value pair. The predicted probability of each class is also displayed. The "weight" in the bar chart indicates that if a IV does not take the displayed value, then the probability of the displayed class to which the feature is contributing to will be reduced in value, on average, equal to the weight. The blue and orange color indicates that a particular IV-value pair contributes to class "≤8.5" and to class ">8.5" respectively. First, section 5.4.1 demonstrates how LIME explanations could help to understand the reasons behind the high recall of students in the "at-risk" and "moderate-risk" zones. Next, section 5.4.2 investigates whether the misclassification of students in the "at-risk" or "moderate-risk" zone into the "no-risk" zone is due to some irrational reasons or not. Finally, section 5.4.3 discusses the contribution of different IVs to the predictions using LIME explanations.

5.4.1 Understanding the High Recall of "at-risk" and "moderate-risk" Students

In addition to the default LIME outputs that explain individual predictions, the frequencies of different values taken by each IV are plotted for different set of



(a)

(b)

(c)

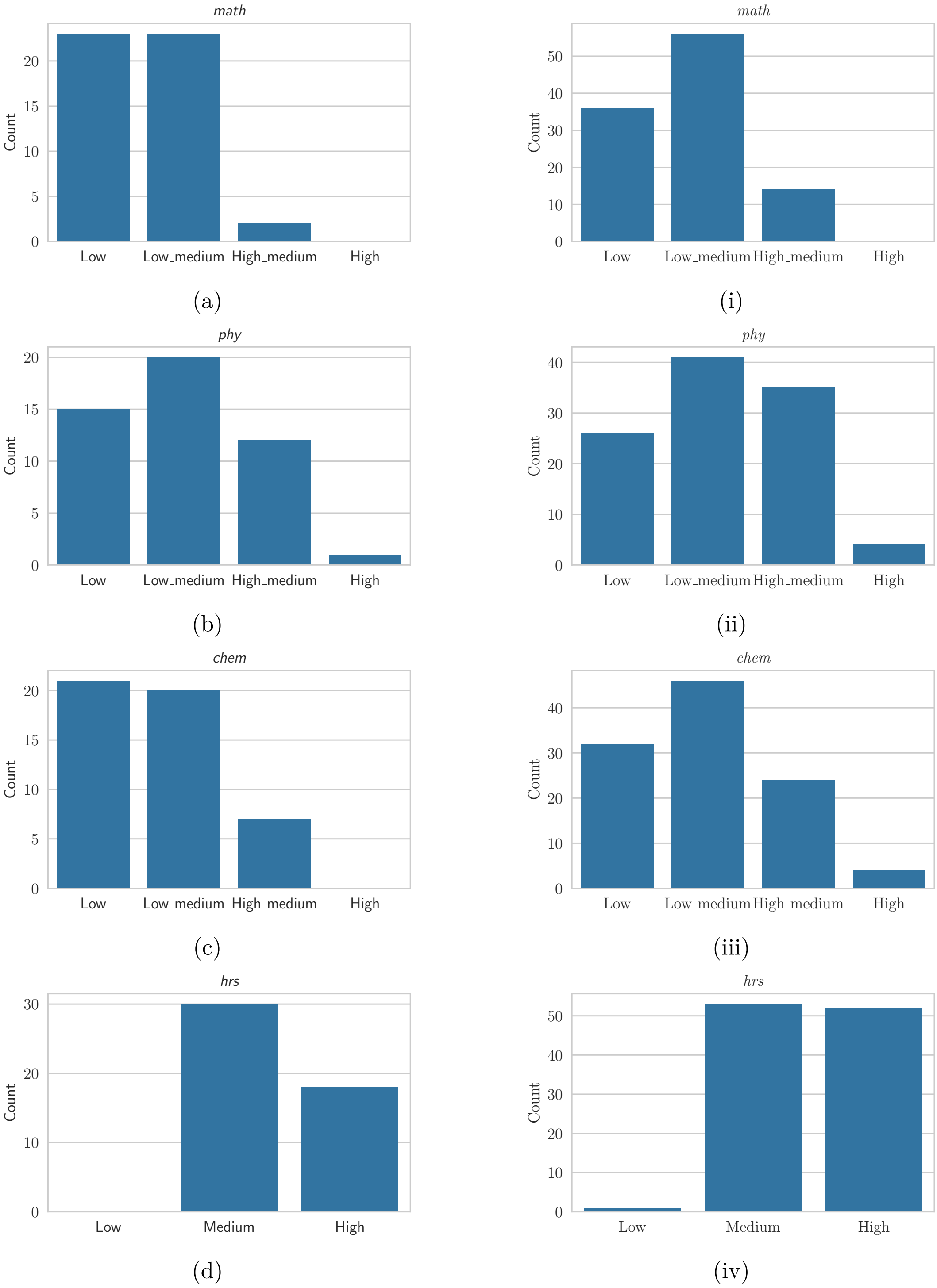
Figure 5.3: LIME outputs to understand high recall of students in the "at-risk" and "moderate-risk" zones

observations. Figure 5.4 shows the frequency plots for IVs of two sets of observations. For the first set of observations (Figure 5.4(a-h)), true and predicted classes are both "≤8.5", while for the second set of observations (Figure 5.4(i-viii)), the true and predicted classes are both "≤11.5". This plot gives a global understanding of how different values taken by the IVs contribute to the correct identification of students in the "at-risk" and "moderate-risk" zones thereby increasing the recall.

It can be observed from Figure 5.4(a-e) that "Low" and "Low_medium" levels of *math, phy,* and *chem*, "Medium" level of *hrs*, and "Low" level of *eff* are strongly associated with "at-risk" students, with the lower levels of *math* being the strongest. However, it can also be observed that, after considering the students in the "moderate-risk" zone Figure 5.4(i-v), the frequencies of levels higher than level "Low" for these IVs have significantly increased. However, the frequencies remain almost constant for *affe* even after considering the students in the "moderate-risk" zone (Figure 5.4e). These inferences indicate that moving towards the higher levels of *math, phy, chem, hrs* and *eff* (but not *affe*) increases the chances of being *not* classified as "at-risk". It is important to note that the above inferences are consistent with that obtained from the perspective of explanatory modelling.

Further, it can be observed from Figure 5.4(g) and Figure5.4(vii) that students with average levels ("Low_medium", "High_medium") of *goal* have higher chance of being classified into "at-risk" and "moderate-risk" zones than "Low" or "High" levels of *goal.* Juxtaposing the above observation with that obtained in explanatory modeling indicates that students with average levels of *goal* are more likely to exert less effort, and hence are identified as being either in the "at-risk" or "moderate-risk" zones. However, while the first-semester performance was shown to be invariant of preference for time pressure *(press)* in explanatory modelling, the explanations by LIME indicate that increasing levels of *press* are associated, though not very strongly, with the "at-risk" students. This is because, Figure 5.4(h) and Figure5.4(viii) shows that higher levels of *press* are strongly associated with the students in the "at-risk" and "moderate-risk" zones.

Figure 5.3(a-b) displays the LIME output for two observations whose true and predicted classes are both "≤8.5". It can be observed that a "Low" value for most of the IVs contributes heavily to class "≤8.5". For instance, the weights of IV-value pairs in Figure 5.3a indicate that if the IVs *math, phy, chem* had not taken "Low" values, then the probability of that particular observation being classified into class "≤8.5" would, on average, be only 84 - (19+17+15) = 33. Likewise, moving towards the higher levels of an IV increases the chances of being not classified as "at-risk" (">8.5"). For instance, a "High_medium" on *phy* and a "High" on *goal* for the observation represented in Figure 5.3b contributes significantly to the increase in the probability of class ">8.5". Figure 5.3c displays the LIME output for the second observation, but here it is classified into class "11.5" by the second classifier.

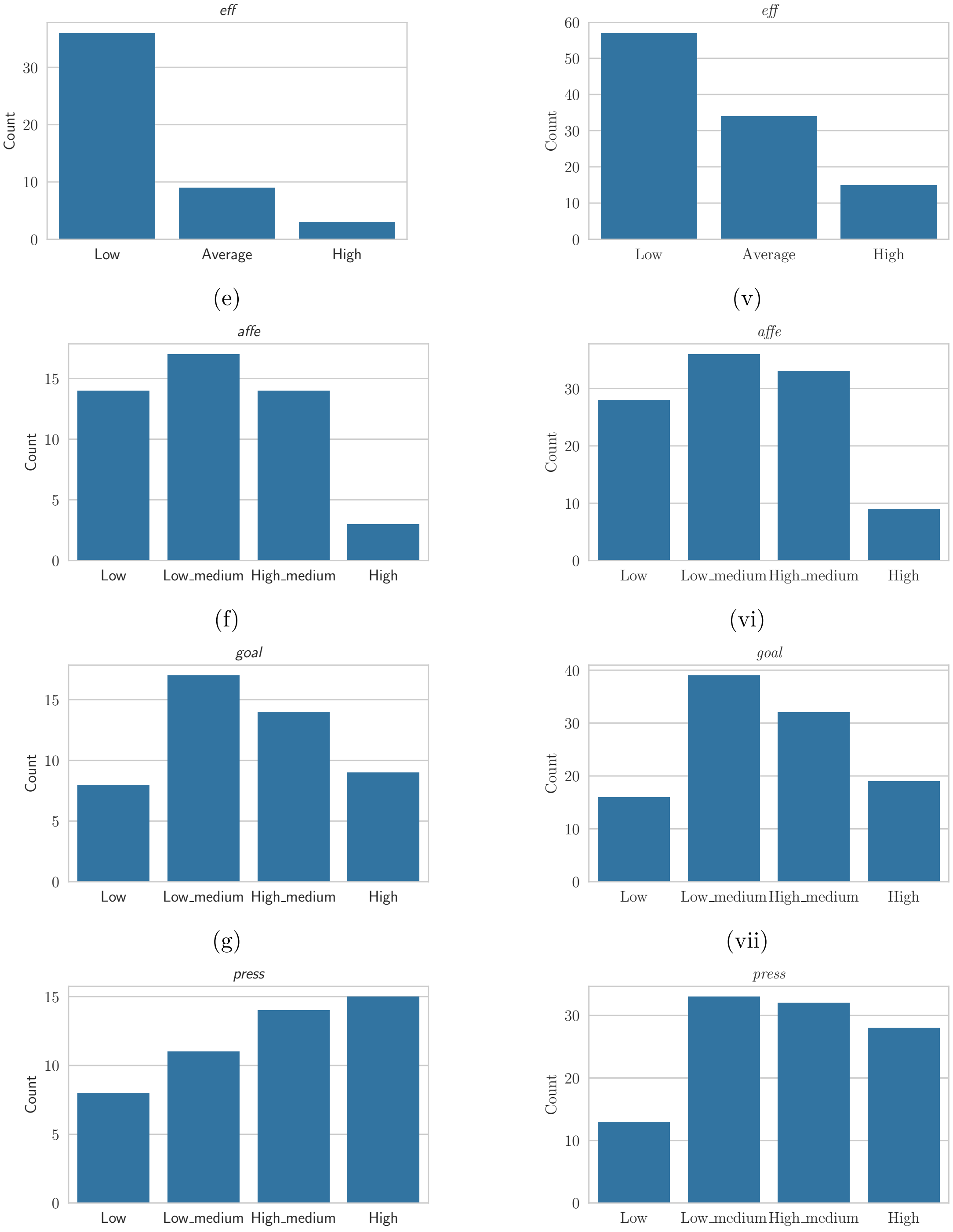


(a) (i)

(b) (ii)

(c) (iii)

(d) (iv)



(e) (v)

(f) (vi)

(g) (vii)

(h) (viii)

Figure 5.4: Figures (a-h) show the frequency plots for IVs of those observations whose true and predicted classes are both "≤8.5". Figures (i-viii) show the frequency plots for IVs of those observations whose true and predicted classes are both "≤11.5"

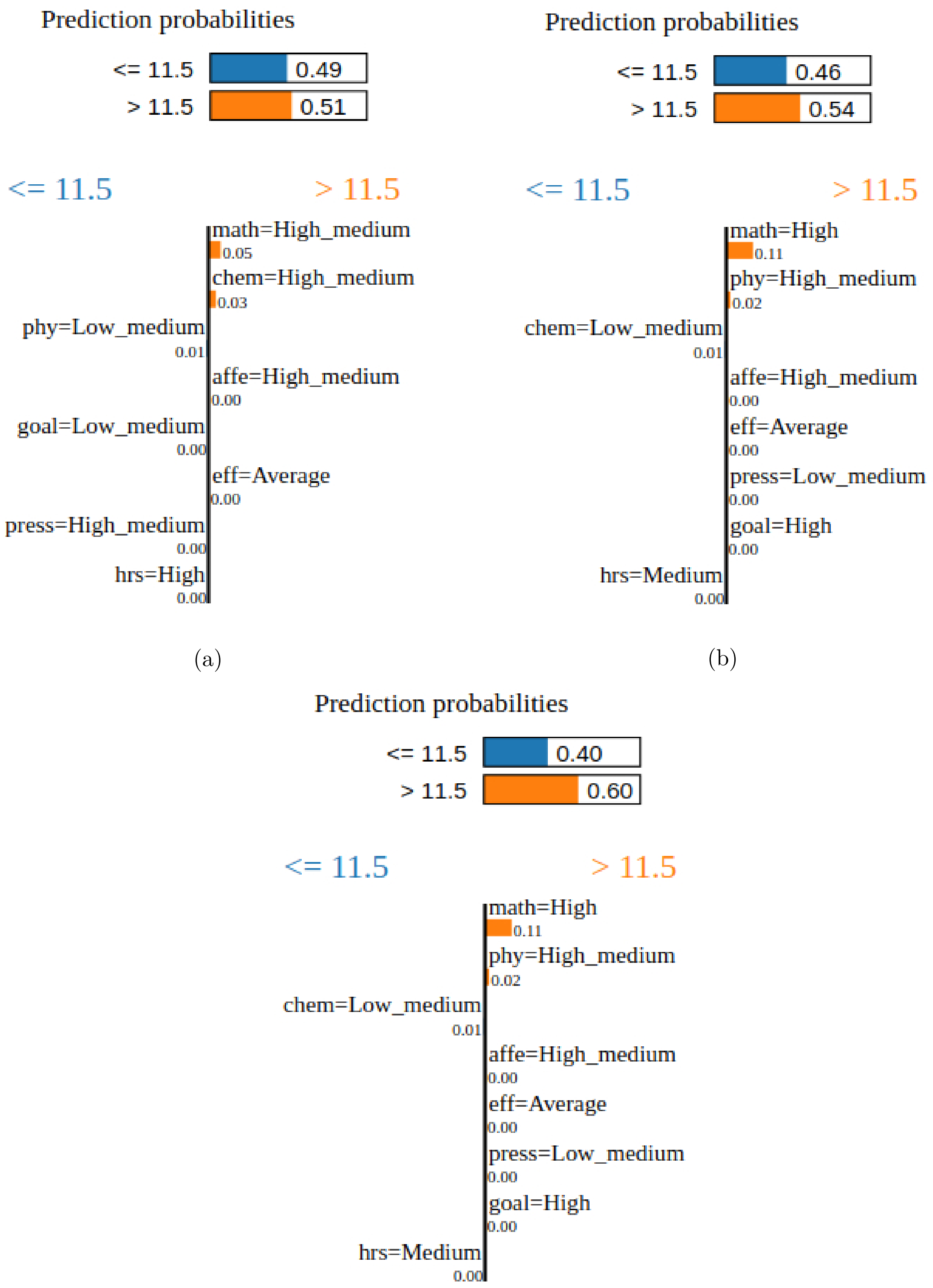
5.4.2 Understanding the Misclassification of "at-risk" and "moderate-risk" Students

Figure 5.5 displays the LIME output for some observations whose true and predicted classes are "≤11.5" and ">11.5" respectively. It can be observed that though these observations are misclassified into class ">11.5", the predicted probabilities of both the classes are almost equal. To see whether this trend is followed for all such misclassified observations in the testing set, the distribution of the "differences in the predicted probabilities" of these observations are plotted (Figure 5.6). The plot clearly indicates that the differences in the probabilities of almost all the misclassified observations are less than 0.075 implying that the observations have almost equal predicted probabilities on both the classes. Hence, the classifier is unable to distinguish between the two classes for some observations. This is perhaps because these misclassified observations take average values ("Low_medium", "High_medium", "Average", "Medium") on most of the IVs as shown in Figure 5.5 for some observations. Alternatively, it is also possible that some subtle differences in the different combinations of the IVs are not identified by the classifier. Leaving such explorations to future work, the important observation to note at this point is, the classifier is not confidently (due to the low differences in probabilities) or erratically (due to the average values on most of the IVs) making wrong predictions. This indicates that the misclassification of students in the "at-risk" or "moderate-risk" zone into the "no-risk" zone is not irrational.

5.4.3 Understanding the Importance of Different Independent Variables

The output of LIME conveys the importance of different IVs by listing the contribution of IVs to the predicted probabilities in descending order. It can be observed from the LIME explanations in Figure 5.3 and Figure 5.5 that IVs such as *math, phy, chem* and *eff* often contribute more to the predicted probabilities than IVs *affe, goal* or *press.* To see if this pattern holds for both the classifiers, for each IV, the number of times it is displayed at different positions in the lists of contribution for all the observations in the testing set are computed and plotted as shown in Figure 5.7. It can be observed that for the first classification problem ("≤8.5" vs ">8.5"), the IVs *math, chem* and *eff* occur more frequently in the beginning of the descending list of contributions (Figure 5.7(a-h)), implying that these IVs contribute more to the predicted probabilities than other IVs. Similarly, for the second classification problem ("≤11.5" vs ">11.5"), the IVs *math, phy, chem* and *eff* contribute more to the predicted probabilities than other IVs (Figure 5.7(i-viii)).

Further, it was earlier inferred in explanatory modeling that IVs *math, phy, chem, hrs* and *eff* contribute more in explaining the variance in the DV. A quick glance at Figure 5.7 shows that these IVs also contribute more than other IVs to the prediction. However, some interesting patterns that were not observed in explanatory modeling can be observed here. For instance, (1) the IV *hrs* is *not* contributing considerably to the predictions that distinguish the students in the "no-risk" zone from those who



(a)

(b)

(c)

Figure 5.5: LIME outputs to understand misclassification of "at-risk" and "moderate-risk" students

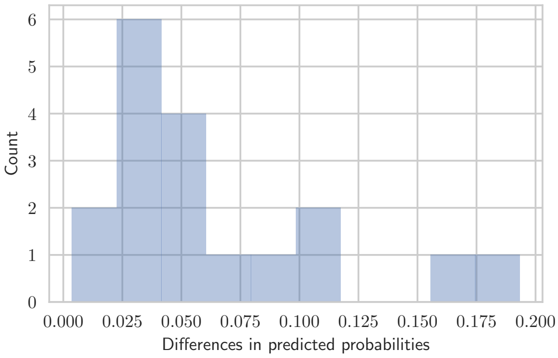


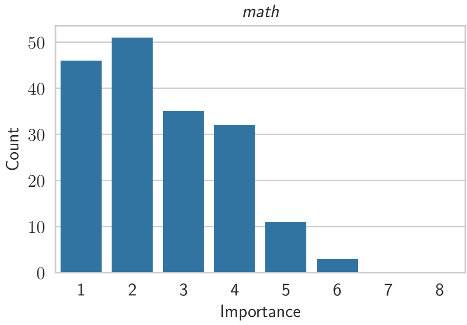
Figure 5.6: Distribution of the "differences in the predicted probabilities" of misclassified observations

are are not, as it is contributing to distinguish the students in the "at-risk" zone from those who are not; (2) the IV *affe* is one of the least contributing variables to the predictions for the first classification problem while it is contributing considerably in distinguishing the two classes "≤11.5" vs ">11.5".

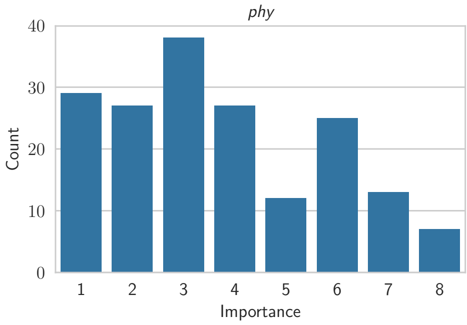
Finally, the above inferences indicate that the complex XGBoost classifiers have uncovered some patterns that are slightly different from the average-case statistical patterns obtained in explanatory modeling and also have uncovered some patterns that were not previously identified.

5.5 Limitations ?

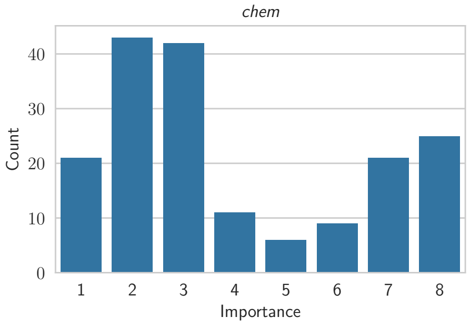
In the previous chapter where both explanatory and predictive goals were tried to optimize, the model was evaluated on a holdout sample because it is difficult to summarize the statistical inferences obtained in different cross-validation folds for explanatory purposes. The same training and holdout testing set is used for the predictive approach described in this chapter to mainly show how complex, algorithmic and black-box type model like XGBoost can produce far more accurate and precise predictions, and uncover complex relationships compared to the simple and interpretable statistical models. However, Hawkins et al. [44] observed that model evaluation using cross-validation methods, in general, provide more robust estimate of the model's performance than evaluating on a holdout sample. But using LIME explanations in a cross-validation context makes the model evaluation process tedious because the explanations for observations in each of the folds need to be investigated to assess the model as a whole. Hence, the model is evaluated only on a holdout sample sacrificing some robustness. However, to make the results obtained in the present study more practically relevant, it is necessary to investigate in the future about how cross-validation techniques can be used in approaches where both explanatory and predictive goals are optimized, and in approaches where predictive modeling is used in conjugation with explanations techniques like LIME.



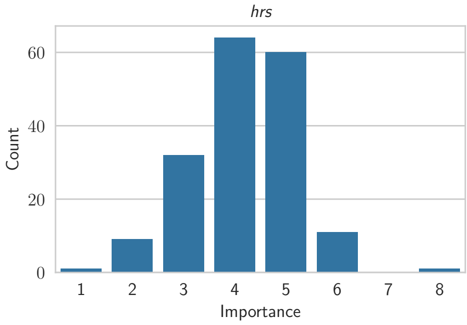
(a)



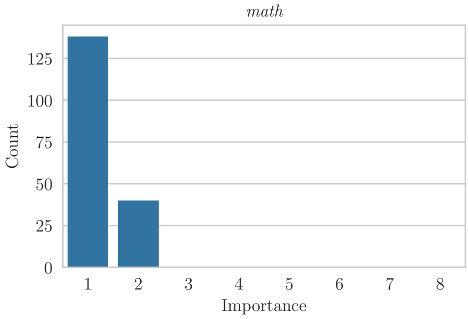
(b)



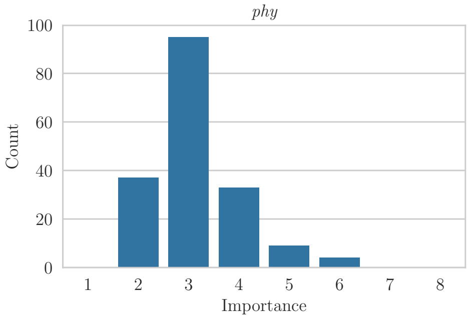
(c)



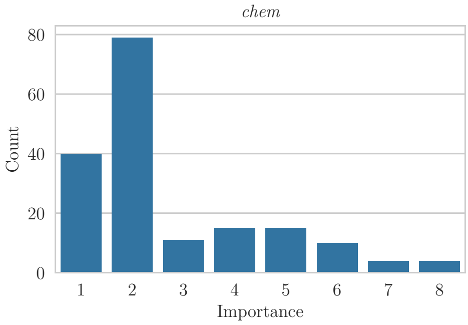
(d)



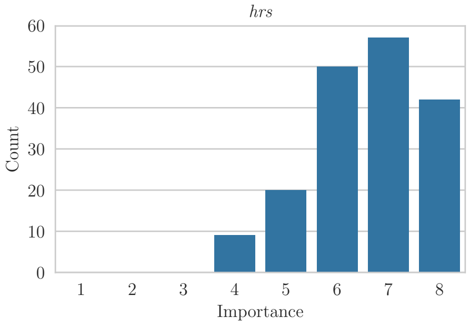
(i)



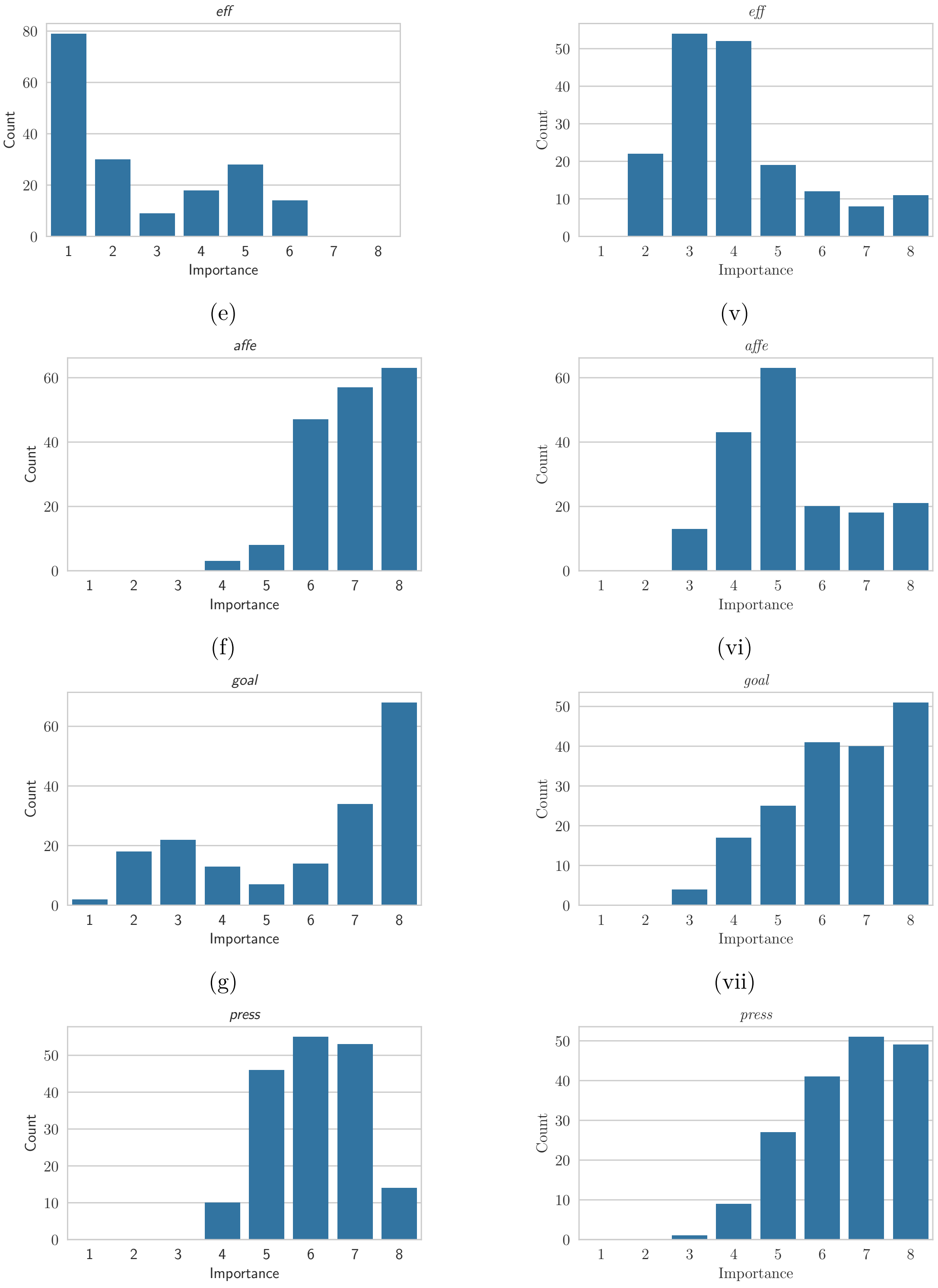
(ii)



(iii)



(iv)



(e)

(v)

(f)

(vi)

(g)

(vii)

(h)

(viii)

Figure 5.7: Figures (a-h) and Figures (i-viii) show the Importance of IVs in the predictions by XGBoost_1 and XGBoost_2 respectively

Chapter 6

Final Remarks

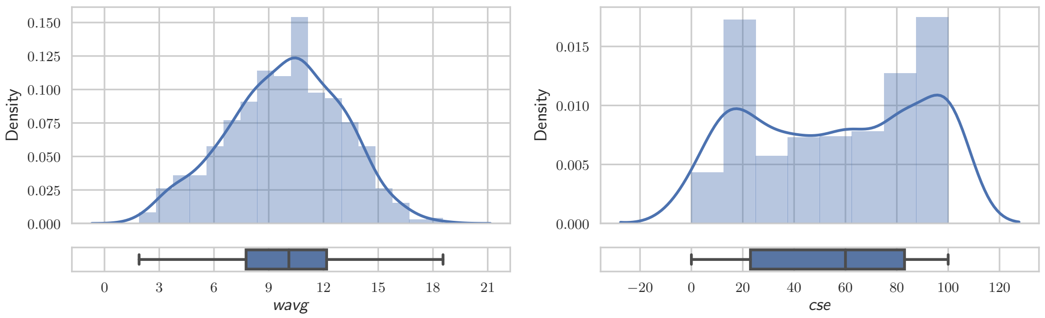
The present study provided various insights about *"what is expected from the students to perform well in their first-semester exam?"* from the perspective of three different modeling approaches that arose as a result of simply appreciating the differences between explanatory and predictive modeling. Along the lines of observations made by Shmueli [81], I argued that many of the related literature in the domain of educational research fail to discriminate between explanatory and predictive modeling. This not only leads to incorrect practical conclusions but also results in ignoring the limitations of explanatory modeling and not appreciating the advantages of predictive modeling. For instance, the subtle effects of some IVs such as *press* on the first-semester performance was observed only in predictive modeling using LIME explanations, but not in explanatory modeling. Further, it is also important to note that instead of seeing the three approaches discussed in the present study independently, using all the three together helps in reinforcing our understanding of the underlying phenomenon far better than using any one of them.

Furthermore, there are many areas that could be explored in the future from the trail left behind the present study. Some of them are discussed below. First, only five LASSI scales were used in the present study because students at KU Leuven are administered only for these scales while the original inventory measures students' learning and study strategies using ten scales [89]. Hence, it would be interesting to use all the ten LASSI scales and see how they interact with other existing variables. Second, In explaining predictions, frequency plots for IVs of the different set of observations were used to get a global understanding of how different IVs contribute to the predictions. However, for the same purpose, a method called SP-LIME proposed by the developers of LIME themselves is recommended by them because it is more robust and concise. Hence, using SP-LIME instead of the frequency plots method used in the present study is an important arena to explore. Finally, in the predictive modeling approach discussed in the last chapter, the ordinal multinomial classification problem was modified into two binary classification problems. An interesting task for future work is to propose a customized loss function for XGBoost classifier similar to the those carried out in some of the studies [24, 57], in order to approach the prediction problem as an ordinal classification problem.

Appendices

Appendix A

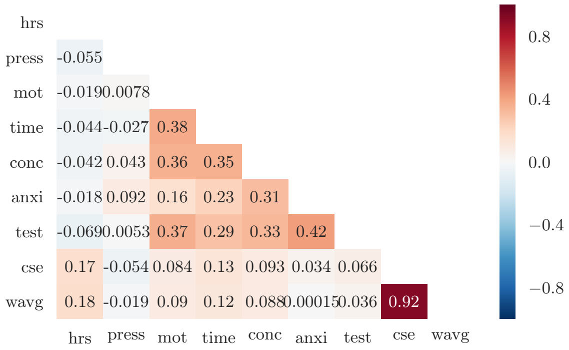
Distributions of *wavg* and *cse*, and Visualization of the Correlation Matrix



(a)

(b)

Figure A.1: Univariate distributions of *wavg* and *cse* with kernel density estimates



	hrs	press	mot	time	conc	anxi	test	cse	wavg
hrs									
press	-0.055								
mot	-0.019	0.0078							
time	-0.044	-0.027	0.38						
conc	-0.042	0.043	0.36	0.35					
anxi	-0.018	0.092	0.16	0.23	0.31				
test	-0.069	0.0053	0.37	0.29	0.33	0.42			
cse	0.17	-0.054	0.084	0.13	0.093	0.034	0.066		
wavg	0.18	-0.019	0.09	0.12	0.088	0.00015	0.036	0.92	

Figure A.2: Heatmap of correlation matrix

Appendix B

Results of MCAR Test

```
Call:
TestMCARNormality(data = newdata[, c(5:13)])

Number of Patterns:  10

Total number of cases used in the analysis:  724

 Pattern(s) used:
          hrs   press   mot   time   conc   anxi   test   cse   wavg   Number of cases
group.1     1       1     1      1      1      1     NA     1      1                19
group.2     1       1     1      1      1     NA      1     1      1                30
group.3     1       1     1      1      1      1      1     1      1               531
group.4     1      NA    NA      1      1      1      1     1      1                 8
group.5     1       1     1      1     NA      1      1     1      1                23
group.6     1       1     1     NA      1      1      1     1      1                20
group.7     1       1    NA      1      1      1      1     1      1                25
group.8     1      NA     1      1      1      1      1     1      1                41
group.9    NA       1     1      1      1      1      1     1      1                 7
group.10    1       1     1      1      1      1      1     1     NA                20


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity:  0.04894875

    Either the test of multivariate normality or homoscedasticity (or both) is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity:  0.4207877

    Reject Normality at 0.05 significance level.
    There is not sufficient evidence to reject MCAR at 0.05 significance level.
```

Appendix C

Results of One-way ANOVA for Categorical Variables

	df	Sum Sq	Mean Sq	F value	Pr(>F)
math	4	1664.92	416.23	56.45	0.0000
Residuals	715	5272.08	7.37		

Table C.1: Results of `aov(wavg ~ math, data = dat)`

	df	Sum Sq	Mean Sq	F value	Pr(>F)
phy	4	1374.63	343.66	44.17	0.0000
Residuals	715	5562.36	7.78		

Table C.2: Results of `aov(wavg ~ phy, data = dat)`

	df	Sum Sq	Mean Sq	F value	Pr(>F)
chem	4	1486.02	371.51	48.73	0.0000
Residuals	715	5450.97	7.62		

Table C.3: Results of `aov(wavg ~ chem, data = dat)`

	df	Sum Sq	Mean Sq	F value	Pr(>F)
eff	4	524.37	131.09	14.62	0.0000
Residuals	715	6412.62	8.97		

Table C.4: Results of `aov(wavg ~ eff, data = dat)`

Appendix D

Results of Tukey HSD post hoc test for Categorical Variables

	diff	**lwr**	**upr**	**p adj**
60-70%-<60%	0.75	-0.80	2.29	0.68
70-80%-<60%	1.78	0.31	3.26	0.01
80-90%-<60%	3.72	2.24	5.21	0.00
>90%-<60%	5.56	3.88	7.25	0.00
70-80%-60-70%	1.04	0.24	1.84	0.00
80-90%-60-70%	2.98	2.17	3.79	0.00
>90%-60-70%	4.82	3.68	5.95	0.00
80-90%-70-80%	1.94	1.28	2.61	0.00
>90%-70-80%	3.78	2.75	4.82	0.00
>90%-80-90%	1.84	0.80	2.89	0.00

Table D.1: Results of `TukeyHSD(aov(wavg ~ math, data = dat))`

	diff	**lwr**	**upr**	**p adj**
60-70%-<60%	0.21	-1.84	2.26	1.00
70-80%-<60%	2.10	0.13	4.07	0.03
80-90%-<60%	3.22	1.26	5.18	0.00
>90%-<60%	5.01	2.94	7.07	0.00
70-80%-60-70%	1.89	0.99	2.80	0.00
80-90%-60-70%	3.01	2.12	3.90	0.00
>90%-60-70%	4.80	3.69	5.90	0.00
80-90%-70-80%	1.12	0.44	1.79	0.00
>90%-70-80%	2.90	1.96	3.84	0.00
>90%-80-90%	1.79	0.86	2.71	0.00

Table D.2: Results of `TukeyHSD(aov(wavg ~ phy, data = dat))`

	diff	**lwr**	**upr**	**p adj**
60-70%-<60%	1.00	-0.60	2.60	0.43
70-80%-<60%	2.45	0.92	3.98	0.00
80-90%-<60%	3.99	2.46	5.52	0.00
>90%-<60%	5.48	3.77	7.18	0.00
70-80%-60-70%	1.45	0.62	2.28	0.00
80-90%-60-70%	2.99	2.16	3.82	0.00
>90%-60-70%	4.48	3.35	5.60	0.00
80-90%-70-80%	1.54	0.86	2.22	0.00
>90%-70-80%	3.03	2.02	4.04	0.00
>90%-80-90%	1.49	0.47	2.51	0.00

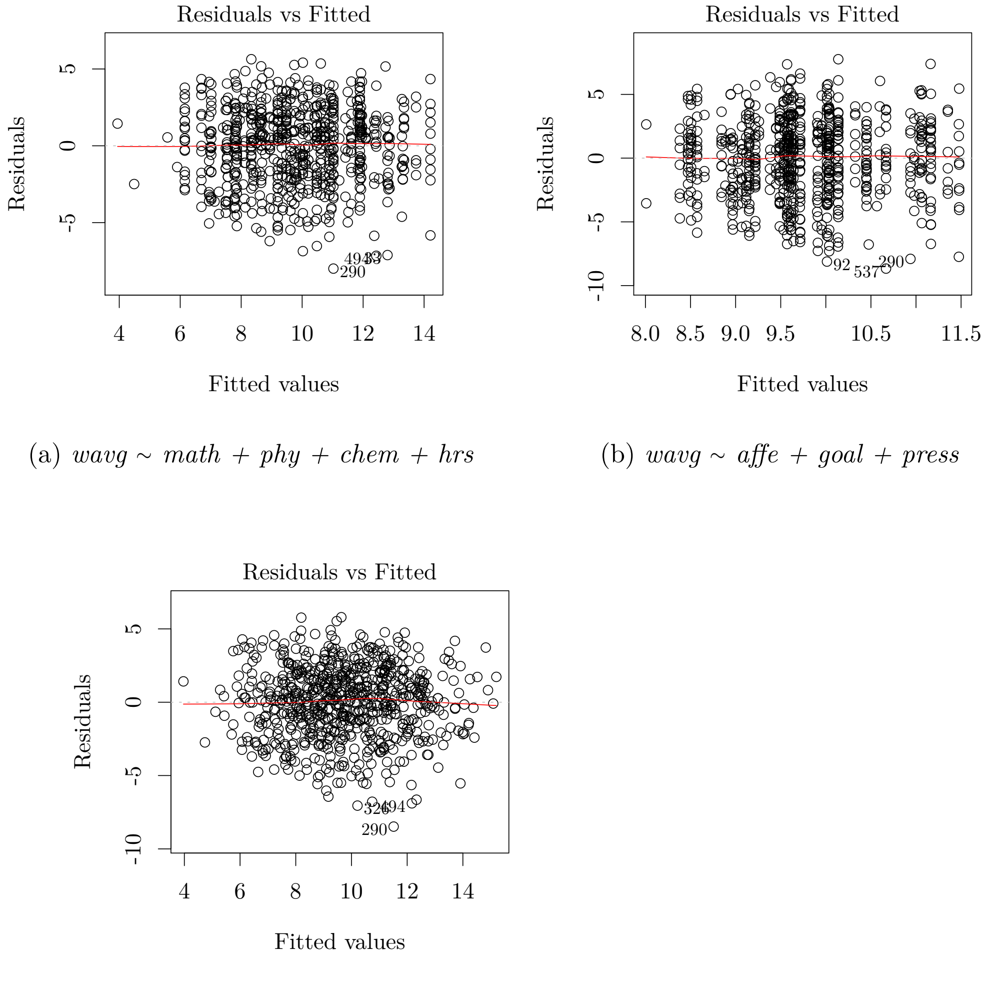
Table D.3: Results of `TukeyHSD(aov(wavg ~ chem, data = dat))`

	diff	**lwr**	**upr**	**p adj**
High-Average	0.93	0.10	1.76	0.02
Low-Average	-0.94	-1.66	-0.22	0.00
Very High-Average	0.29	-2.85	3.43	1.00
Very Low-Average	-2.41	-3.88	-0.94	0.00
Low-High	-1.87	-2.70	-1.04	0.00
Very High-High	-0.64	-3.81	2.52	0.98
Very Low-High	-3.34	-4.88	-1.81	0.00
Very High-Low	1.23	-1.91	4.37	0.82
Very Low-Low	-1.47	-2.94	0.00	0.05
Very Low-Very High	-2.70	-6.09	0.69	0.19

Table D.4: Results of `TukeyHSD(aov(wavg ~ eff, data = dat))`

Appendix E

Analysis of Residuals

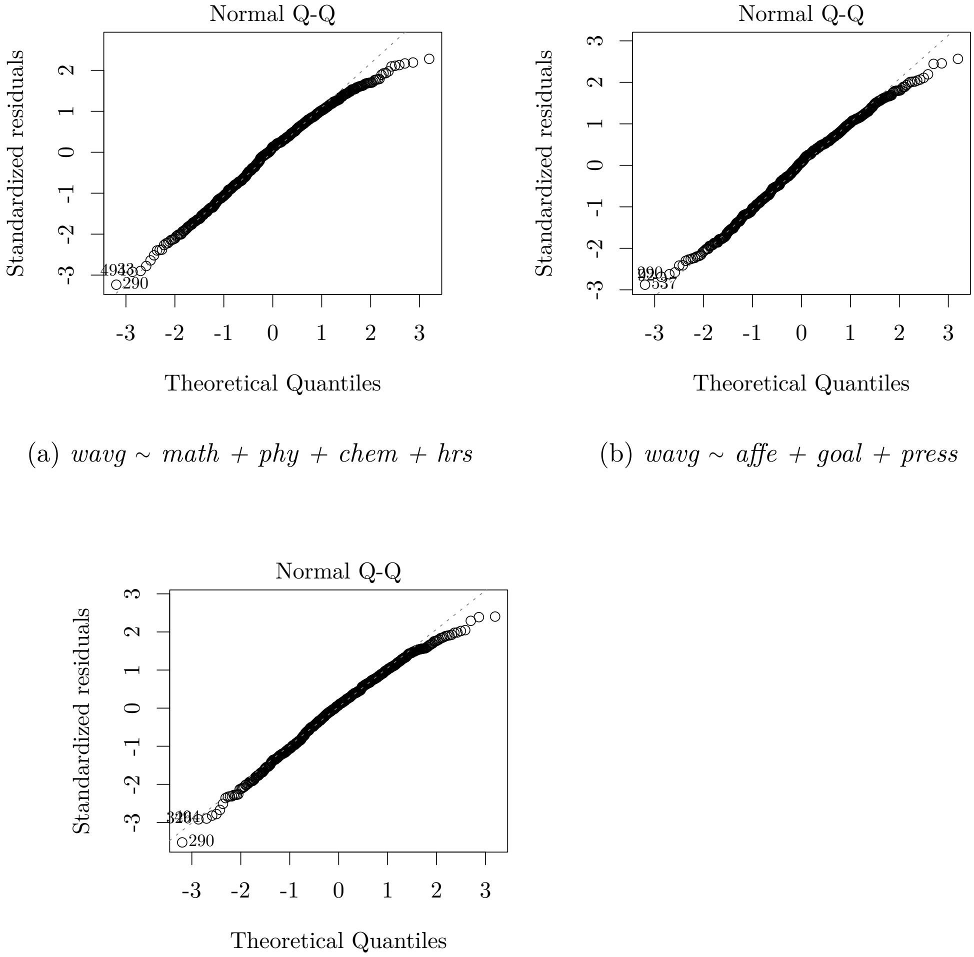


(a) *wavg ∼ math + phy + chem + hrs*

(b) *wavg ∼ affe + goal + press*

(c) *wavg∼math+phy+chem+hrs+eff+affe+goal+press*

Figure E.1: Plots for testing Linearity of Residuals

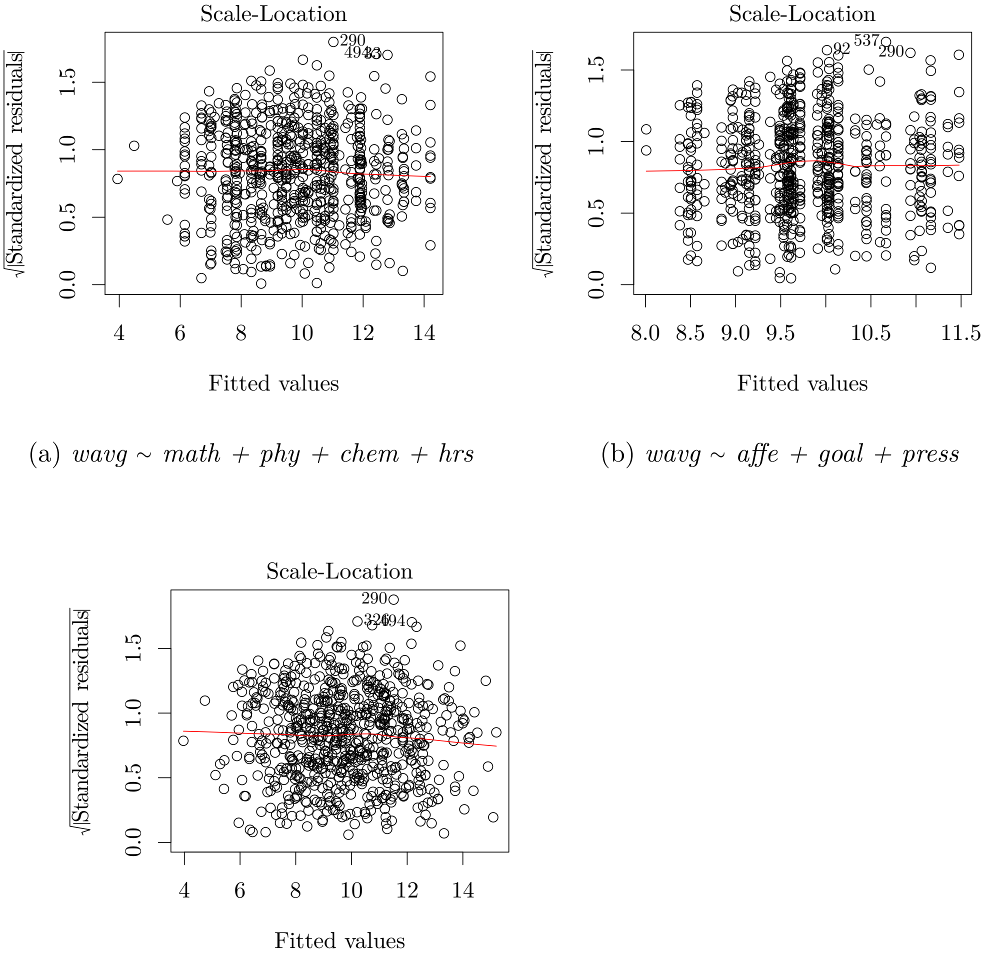


(a) *wavg* ~ *math* + *phy* + *chem* + *hrs*

(b) *wavg* ~ *affe* + *goal* + *press*

(c) *wavg~math+phy+chem+hrs+eff+affe+goal+press*

Figure E.2: Plots for testing Normality of Residuals

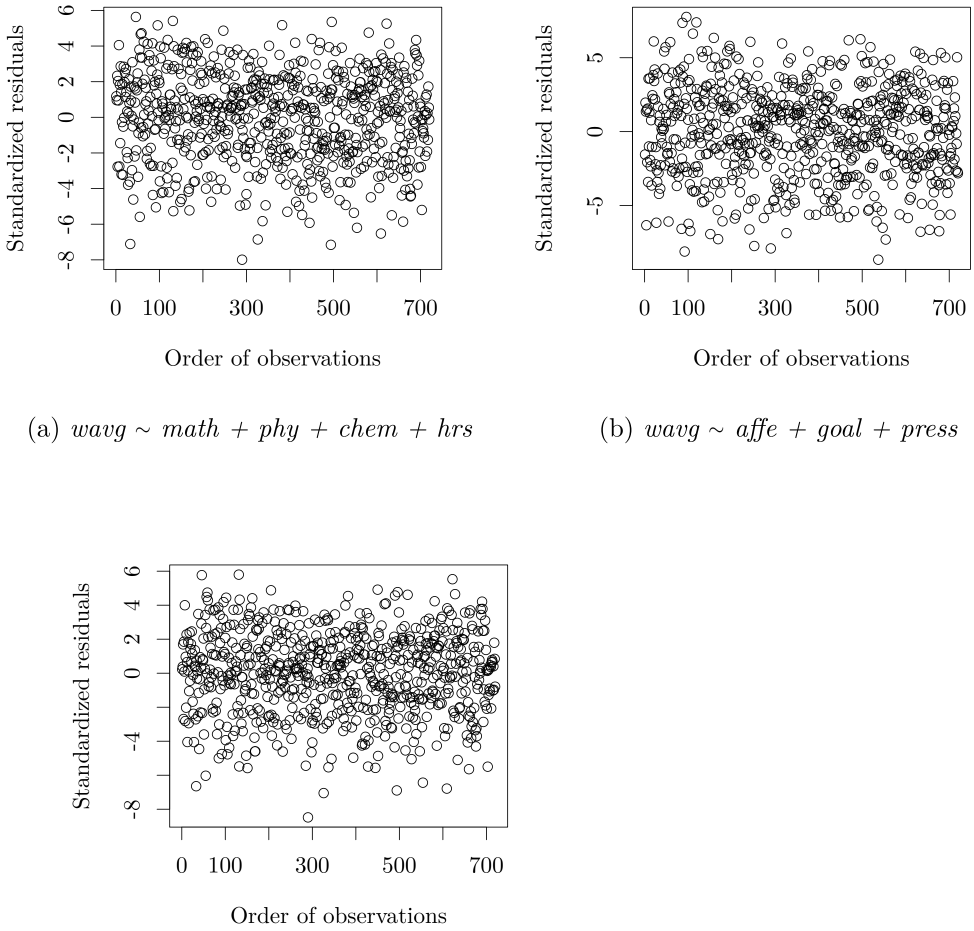


(a) *wavg ~ math + phy + chem + hrs*

(b) *wavg ~ affe + goal + press*

(c) *wavg~math+phy+chem+hrs+eff+affe+goal+press*

Figure E.3: Plots for testing Homoscedasticity of Residuals



(a) *wavg ~ math + phy + chem + hrs*

(b) *wavg ~ affe + goal + press*

(c) *wavg~math+phy+chem+hrs+eff+affe+goal+press*

Figure E.4: Plots for testing Statistical Independence of Residuals

Appendix F

Results of Incremental Effects of Noncognitive Factors

```
Call:
lm(formula = wavg ~ math + phy + chem + hrs + affe, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1755 -1.7878  0.2734  1.7855  5.6852 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.83562    0.79808   4.806 1.88e-06 ***
mathLow_medium   0.55350    0.26324   2.103 0.035850 *  
mathHigh_medium  1.93001    0.28780   6.706 4.10e-11 ***
mathHigh         2.79752    0.43716   6.399 2.85e-10 ***
phyLow_medium    0.81235    0.29931   2.714 0.006809 ** 
phyHigh_medium   1.07387    0.31660   3.392 0.000733 ***
phyHigh          2.00993    0.41174   4.882 1.30e-06 ***
chemLow_medium   0.77671    0.27049   2.872 0.004208 ** 
chemHigh_medium  1.76644    0.28981   6.095 1.80e-09 ***
chemHigh         2.22708    0.41596   5.354 1.16e-07 ***
hrsMedium        2.22768    0.76567   2.909 0.003735 ** 
hrsHigh          3.11152    0.76389   4.073 5.16e-05 ***
affeLow_medium   0.07445    0.28325   0.263 0.792734    
affeHigh_medium  0.18019    0.28787   0.626 0.531545    
affeHigh         0.37524    0.34988   1.072 0.283865    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.485 on 705 degrees of freedom
Multiple R-squared:  0.3723,	Adjusted R-squared:  0.3598 
F-statistic: 29.86 on 14 and 705 DF,  p-value: < 2.2e-16
```

Figure F.1: Incremental effects of affective strategies

```
Call:
lm(formula = wavg ~ math + phy + chem + hrs + affe + goal, data = dat)

Residuals:
   Min     1Q Median     3Q    Max 
-8.144 -1.780  0.310  1.799  5.603 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.5084     0.8240   4.258 2.35e-05 ***
mathLow_medium    0.5501     0.2635   2.088  0.03716 *  
mathHigh_medium   1.9181     0.2879   6.662 5.45e-11 ***
mathHigh          2.7799     0.4380   6.346 3.95e-10 ***
phyLow_medium     0.8026     0.2997   2.678  0.00758 ** 
phyHigh_medium    1.0604     0.3168   3.347  0.00086 ***
phyHigh           2.0132     0.4130   4.875 1.35e-06 ***
chemLow_medium    0.7619     0.2709   2.813  0.00505 ** 
chemHigh_medium   1.7639     0.2900   6.081 1.96e-09 ***
chemHigh          2.2180     0.4166   5.324 1.37e-07 ***
hrsMedium         2.2507     0.7659   2.938  0.00341 ** 
hrsHigh           3.1447     0.7643   4.114 4.35e-05 ***
affeLow_medium    0.0366     0.2842   0.129  0.89756    
affeHigh_medium   0.1404     0.2893   0.485  0.62752    
affeHigh          0.2997     0.3529   0.849  0.39602    
goalLow_medium    0.3758     0.2878   1.306  0.19211    
goalHigh_medium   0.4523     0.2813   1.608  0.10836    
goalHigh          0.4674     0.3365   1.389  0.16521    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.485 on 702 degrees of freedom
Multiple R-squared:  0.3748,       Adjusted R-squared:  0.3597 
F-statistic: 24.76 on 17 and 702 DF,  p-value: < 2.2e-16
```

Figure F.2: Incremental effects of goal strategies

```
Call:
lm(formula = wavg ~ math + phy + chem + hrs + affe + goal + press,
    data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0927 -1.7092  0.2507  1.8251  5.3619 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.56699    0.84798   4.206 2.93e-05 ***
mathLow_medium    0.55751    0.26376   2.114 0.034895 *  
mathHigh_medium   1.92366    0.28777   6.685 4.73e-11 ***
mathHigh          2.73997    0.43840   6.250 7.14e-10 ***
phyLow_medium     0.81297    0.29983   2.711 0.006864 ** 
phyHigh_medium    1.08879    0.31767   3.427 0.000645 ***
phyHigh           2.05552    0.41457   4.958 8.94e-07 ***
chemLow_medium    0.78533    0.27182   2.889 0.003983 ** 
chemHigh_medium   1.76440    0.29039   6.076 2.02e-09 ***
chemHigh          2.22912    0.41792   5.334 1.30e-07 ***
hrsMedium         2.26979    0.76834   2.954 0.003241 ** 
hrsHigh           3.13900    0.76653   4.095 4.71e-05 ***
affeLow_medium   -0.02142    0.28565  -0.075 0.940247    
affeHigh_medium   0.05577    0.29213   0.191 0.848644    
affeHigh          0.21228    0.35572   0.597 0.550871    
goalLow_medium    0.42308    0.29141   1.452 0.146992    
goalHigh_medium   0.57320    0.29112   1.969 0.049356 *  
goalHigh          0.56130    0.34228   1.640 0.101477    
pressLow_medium   0.05925    0.29456   0.201 0.840631    
pressHigh_medium -0.12296    0.31496  -0.390 0.696359    
pressHigh        -0.45585    0.33299  -1.369 0.171455    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.484 on 699 degrees of freedom
Multiple R-squared:  0.3785,	Adjusted R-squared:  0.3607 
F-statistic: 21.28 on 20 and 699 DF,  p-value: < 2.2e-16
```

Figure F.3: Incremental effects of pressure preference

```
Call:
lm(formula = wavg ~ math + phy + chem + hrs + affe + goal + press +
    eff, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4803 -1.5536  0.1926  1.7364  5.7969

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.06140    0.83844   3.651 0.000280 ***
mathLow_medium    0.49281    0.25939   1.900 0.057860 .
mathHigh_medium   1.87668    0.28296   6.632 6.63e-11 ***
mathHigh          2.66408    0.43109   6.180 1.09e-09 ***
phyLow_medium     0.84556    0.29485   2.868 0.004259 **
phyHigh_medium    1.10968    0.31221   3.554 0.000405 ***
phyHigh           2.08167    0.40725   5.112 4.13e-07 ***
chemLow_medium    0.78667    0.26711   2.945 0.003335 **
chemHigh_medium   1.59253    0.28711   5.547 4.13e-08 ***
chemHigh          2.08548    0.41147   5.068 5.15e-07 ***
hrsMedium         2.18913    0.75529   2.898 0.003869 **
hrsHigh           3.02088    0.75370   4.008 6.78e-05 ***
affeLow_medium   -0.19518    0.28255  -0.691 0.489929
affeHigh_medium  -0.29307    0.29461  -0.995 0.320182
affeHigh         -0.28108    0.36203  -0.776 0.437783
goalLow_medium    0.49899    0.28674   1.740 0.082262 .
goalHigh_medium   0.81470    0.28962   2.813 0.005045 **
goalHigh          0.99126    0.34618   2.863 0.004317 **
pressLow_medium   0.20261    0.29064   0.697 0.485967
pressHigh_medium  0.09121    0.31219   0.292 0.770242
pressHigh        -0.13244    0.33289  -0.398 0.690862
effAverage        0.78362    0.21921   3.575 0.000375 ***
effHigh           1.37504    0.27023   5.088 4.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.439 on 697 degrees of freedom
Multiple R-squared:  0.4021,	Adjusted R-squared:  0.3832
F-statistic: 21.31 on 22 and 697 DF,  p-value: < 2.2e-16
```

Figure F.4: Incremental effects of effort level

Appendix G

Results of Ordinal Regression for Model 8

```
                       OR     2.5 %    97.5 %
mathLow_medium   1.4413207 0.9532213  2.186947
mathHigh_medium  3.9337340 2.5026787  6.224628
mathHigh         8.4560868 3.9885119 18.573614
phyLow_medium    1.5111265 0.9327683  2.463806
phyHigh_medium   1.8766487 1.1304288  3.133727
phyHigh          3.6816642 1.8764644  7.303876
chemLow_medium   1.8942101 1.2336374  2.922944
chemHigh_medium  3.1695253 2.0006839  5.050754
chemHigh         6.3223030 3.1549750 12.918076
effAverage       1.8937914 1.3300412  2.703595
effHigh          3.1687356 2.0346692  4.963096
hrsMedium        3.0718592 0.8997368 11.822506
hrsHigh          5.2082088 1.5279816 20.025389
pressLow_medium  1.0787971 0.6682516  1.742177
pressHigh_medium 0.8452647 0.5050180  1.413715
pressHigh        0.8973241 0.5224856  1.539951
affeLow_medium   0.8656920 0.5501524  1.362527
affeHigh_medium  0.8259242 0.5139404  1.326357
affeHigh         0.6328787 0.3523856  1.133160
goalLow_medium   1.6056619 1.0085352  2.568226
goalHigh_medium  2.0713389 1.2918120  3.341755
goalHigh         2.3744002 1.3480511  4.206392
```

Figure G.1: R output for Odds ratios with confidence interval for model 8

```
formula: wavg_ord ~ math + phy + chem + hrs + eff + affe + goal + press
data:    ccc_log

 link  threshold nobs logLik  AIC     niter max.grad cond.H
 logit flexible  542  -483.10 1014.20 5(0)  5.71e-12 1.1e+03

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
mathLow_medium    0.40533    0.24350   1.665  0.09600 .
mathHigh_medium   1.21135    0.26349   4.597 4.28e-06 ***
mathHigh          2.41374    0.46628   5.177 2.26e-07 ***
phyLow_medium     0.70707    0.28600   2.472  0.01343 *
phyHigh_medium    0.80939    0.30279   2.673  0.00751 **
phyHigh           1.24855    0.40283   3.099  0.00194 **
chemLow_medium    0.72669    0.25139   2.891  0.00384 **
chemHigh_medium   1.09469    0.27326   4.006 6.18e-05 ***
chemHigh          1.88658    0.41025   4.599 4.25e-06 ***
hrsMedium         1.28973    0.79120   1.630  0.10308
hrsHigh           1.64359    0.78943   2.082  0.03734 *
effAverage        0.63335    0.21010   3.015  0.00257 **
effHigh           1.27154    0.26152   4.862 1.16e-06 ***
affeLow_medium    0.07156    0.26631   0.269  0.78817
affeHigh_medium  -0.09350    0.28047  -0.333  0.73886
affeHigh         -0.46347    0.34388  -1.348  0.17774
goalLow_medium    0.80032    0.28038   2.854  0.00431 **
goalHigh_medium   0.81338    0.28173   2.887  0.00389 **
goalHigh          1.07006    0.34424   3.109  0.00188 **
pressLow_medium  -0.01027    0.27510  -0.037  0.97021
pressHigh_medium -0.30282    0.30357  -0.998  0.31851
pressHigh        -0.15452    0.31471  -0.491  0.62344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
                 Estimate Std. Error z value
0-8.5|8.5-11.5     3.8094     0.9065   4.202
8.5-11.5|11.5-20   5.8826     0.9272   6.345
```

Figure G.2: Ordinal Regression output in R for model 8

Appendix H

XGBoost Parameters Associated with the Best Performance

Parameters: generalization	**XGBoost 1** **(“≤8.5” vs “>8.5”)**	**XGBoost 2** **(“≤11.5” vs “>11.5”)**
colsample_bytree	0.8	0.8
gamma	0.2	0.2
learning_rate	0.05	0.05
max_depth	3	3
min_child_weight	1	1
n_estimators	90	9
objective	‘binary:logistic’	‘binary:logistic’
subsample	0.8	0.8

Parameters: class-imbalance	**XGBoost 1** **(“≤8.5” vs “>8.5”)**	**XGBoost 2** **(“≤11.5” vs “>11.5”)**
scale_pos_weight	0.450	1.429
max_delta_step	2	10
seed	17	17

Table H.1: XGBoost parameters

Bibliography

[1] General practice to impute missing values. https://www.kaggle.com/questions-and-answers/37491. Accessed: 2018-01-05.

[2] ORDINAL LOGISTIC REGRESSION | R DATA ANALYSIS EXAMPLES. https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/. Accessed: 2018-01-05.

[3] Regression diagnostics: testing the assumptions of linear regression. http://people.duke.edu/~rnau/testing.htm. Accessed: 2018-01-05.

[4] P. L. Ackerman and E. D. Heggestad. Intelligence, personality, and interests: evidence for overlapping traits. *Psychological bulletin*, 121(2):219, 1997.

[5] 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. *Journal of Educational Psychology*, 105(3):911, 2013.

[6] A. Agresti. An Introduction to Categorical Data Analysis. 1996.

[7] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1(Dec):113–141, 2000.

[8] W. A. Arrindell and J. Van der Ende. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, 9(2):165–178, 1985.

[9] R. Barber and M. Sharkey. Course correction: using analytics to predict course success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 259–262. ACM, 2012.

[10] L. E. Bernold, J. E. Spurlin, and C. M. Anson. Understanding our students: A longitudinal-study of success and failure in engineering with implications for increased retention. *Journal of Engineering Education*, 96(3):263–274, 2007.

[11] L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[12] J. Burtner. The use of discriminant analysis to investigate the influence of non-cognitive factors on engineering school persistence. *Journal of Engineering Education*, 94(3):335–338, 2005.

[13] F. Cano. An in-depth analysis of the Learning and Study Strategies Inventory (lassi). *Educational and Psychological Measurement*, 66(6):1023–1038, 2006.

[14] G. Chen, J. E. Mathieu, and P. D. Bliese. A framework for conducting multi-level construct validation. In *Multi-level issues in organizational behavior and processes*, pages 273–303. Emerald Group Publishing Limited, 2005.

[15] 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*, pages 785–794. ACM, 2016.

[16] T. Chen and T. He. Higgs boson discovery with boosted trees. In *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, pages 69–80, 2015.

[17] J. N. Choi and S. V. Moran. Why not procrastinate? development and validation of a new active procrastination scale. *The Journal of social psychology*, 149(2): 195–212, 2009.

[18] R. H. B. Christensen. ordinal—Regression Models for Ordinal Data, 2015. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

[19] A. L. Comrey and H. B. Lee. A first course in factor analysis. 1992.

[20] A. B. Costello and J. W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7):1–9, 2005.

[21] M. W. Craven and J. W. Shavlik. *Extracting comprehensible models from trained neural networks.* PhD thesis, University of Wisconsin, Madison, 1996.

[22] L. J. Cronbach and P. E. Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[23] B. B. de Koning, S. M. Loyens, R. M. Rikers, G. Smeets, and H. T. van der Molen. Generation psy: Student characteristics and academic achievement in a three-year problem-based learning bachelor program. *Learning and Individual differences*, 22(3):313–323, 2012.

[24] K. Dembczyński, W. Kotłowski, and R. Słowiński. Ordinal classification with decision rules. In *International Workshop on Mining Complex Data*, pages 169–181. Springer, 2007.

[25] Y. Ding and J. S. Simonoff. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11(Jan):131–170, 2010.

[26] S. J. Dollinger, A. M. Matyja, and J. L. Huber. Which factors best account for academic success: Those which college students can control or those they cannot? *Journal of research in Personality*, 42(4):872–885, 2008.

[27] P. Domingos. Occam's two razors: The sharp and the blunt. In *KDD*, pages 37–43, 1998.

[28] M. J. Druzdze and C. Glymour. Application of the TETRAD ii Program to the Study of Student Retention in US Colleges. In *KDD Workshop*, pages 419–430, 1994.

[29] S. M. Elias and R. J. Loomis. Utilizing need for cognition and perceived self-efficacy to predict academic performance. *Journal of Applied Social Psychology*, 32(8):1687–1702, 2002.

[30] 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*, pages 158–161. ACM, 2012.

[31] T. Farsides and R. Woodfield. Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual differences*, 34(7):1225–1243, 2003.

[32] T. D. Fletcher. *psychometric: Applied Psychometric Theory*, 2010. URL `https://CRAN.R-project.org/package=psychometric`. R package version 2.2.

[33] B. J. Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5, 2002.

[34] L. Fonteyne, W. Duyck, and F. De Fruyt. Program-specific prediction of academic achievement on the basis of cognitive and non-cognitive factors. *Learning and Individual Differences*, 56:34–48, 2017.

[35] G. Forman and M. Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010.

[36] J. Fox. *Regression diagnostics: An introduction*, volume 79. Sage, 1991.

[37] Frank E Harrell Jr and with contributions from Charles Dupont and many others. *Hmisc: Harrell Miscellaneous*, 2017. URL `https://CRAN.R-project.org/package=Hmisc`. R package version 4.0-3.

[38] B. F. French, J. C. Immekus, and W. C. Oakes. An examination of indicators of engineering students' success and persistence. *Journal of Engineering Education*, 94(4):419–425, 2005.

[39] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[40] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2(Mar):721–747, 2002.

[41] E. Guadagnoli and W. F. Velicer. Relation to sample size to the stability of component patterns. *Psychological bulletin*, 103(2):265, 1988.

[42] B. C. Hardgrave, R. L. Wilson, and K. A. Walstrom. Predicting graduate student success: A comparison of neural networks and traditional techniques. *Computers & Operations Research*, 21(3):249–263, 1994.

[43] F. E. Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[44] D. M. Hawkins, S. C. Basak, and D. Mills. Assessing model fit by cross-validation. *Journal of chemical information and computer sciences*, 43(2):579–586, 2003.

[45] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[46] R. K. Henson, R. M. Capraro, and M. M. Capraro. Reporting Practice and Use of Exploratory Factor Analysis in Educational Research Journals. 2001.

[47] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2nd edition, 2000.

[48] W. IJsselsteijn, Y. de Kort, C. Midden, B. Eggen, and E. van den Hoven. Persuasive technology for human well-being: setting the scene. *Persuasive technology*, pages 1–5, 2006.

[49] M. Jamshidian and S. Jalal. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4): 649–674, 2010.

[50] M. Jamshidian, S. J. Jalal, and C. Jansen. Missmech: an r package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). *Journal of Statistical software*, 56(6), 2014.

[51] I. T. Jolliffe. Principal component analysis and factor analysis. *Principal component analysis*, pages 150–166, 2002.

[52] S. Kim. *ppcor: Partial and Semi-Partial (Part) Correlation*, 2015. URL `https://CRAN.R-project.org/package=ppcor`. R package version 1.1.

[53] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145. Stanford, CA, 1995.

[54] M. Komarraju, S. J. Karau, and R. R. Schmeck. Role of the Big Five personality traits in predicting college students' academic motivation and achievement. *Learning and individual differences*, 19(1):47–52, 2009.

[55] S. Korkmaz, D. Goksuluk, and G. Zararsiz. MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, 6(2): 151–162, 2014. URL `https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf`.

[56] W. C. Leuwerke, S. Robbins, R. Sawyer, and M. Hovland. Predicting engineering major status from mathematics achievement and interest congruence. *Journal of Career Assessment*, 12(2):135–149, 2004.

[57] P. Li, Q. Wu, and C. J. Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*, pages 897–904, 2008.

[58] J. Lin, P. Imbrie, and K. J. Reid. Student retention modelling: An evaluation of different methods and their impact on prediction results. *Research in Engineering Education Sysmposium*, pages 1–6, 2009.

[59] E. Litzler and J. Young. Understanding the risk of attrition in undergraduate engineering: Results from the project to assess climate in engineering. *Journal of Engineering Education*, 101(2):319–345, 2012.

[60] L. P. Macfadyen and S. Dawson. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & education*, 54(2): 588–599, 2010.

[61] D. P. MacKinnon, J. L. Krull, and C. M. Lockwood. Equivalence of the mediation, confounding and suppression effect. *Prevention science*, 1(4):173–181, 2000.

[62] P. A. Murtaugh, L. D. Burns, and J. Schuster. Predicting the retention of university students. *Research in higher education*, 40(3):355–371, 1999.

[63] A. A. O'Connell. *Logistic regression models for ordinal response variables.* Number 146. Sage, 2006.

[64] OECD. *Education at a Glance 2016: OECD Indicators.* OECD Publishing, Paris, 2016. doi: 10.1787/eag-2016-en.

[65] B. S. Olaussen and I. Bråten. Identifying latent variables measured by the Learning and Study Strategies Inventory (LASSI) in Norwegian college students. *The Journal of experimental education*, 67(1):82–96, 1998.

[66] S. Olejnik and S. L. Nist. Identifying latent variables measured by the Learning and Study Strategies Inventory (LASSI). *The Journal of experimental education*, 60(2):151–159, 1992.

[67] I. Olivier, M. Lacante, and V. Briers. *A re-calibration of the LASSI norm scores for a Flemish educational context.* PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2015.

[68] S. Pafka. Benchmarking Random Forest Implementations. `http://datascience.la/benchmarking-random-forest-implementations/`, 2015. Accessed: 2018-01-05.

[69] M. Pinxten, B. De Fraine, W. Van Den Noortgate, J. Van Damme, T. Boonen, and G. Vanlaar. 'I choose so I am': a logistic analysis of major selection in university and successful completion of the first year. *Studies in Higher Education*, 40(10):1919–1946, 2015.

[70] M. Pinxten, T. De Laet, C. Van Soom, and G. Langie. Fighting increasing drop-out rates in the STEM field: The European readySTEMgo Project. In *Proceedings of the 43rd Annual SEFI Conference*, pages 1–8, 2015.

[71] M. Pinxten, C. Van Soom, C. Peeters, T. De Laet, P. Pacher, P. Hockicko, and G. Langie. Learning and study strategies of incoming science and engineering students. a comparative study between three institutions in Belgium, Slovakia, and Hungary. In *Proceedings of the 44rd Annual SEFI Conference*, pages 1–8, 2016.

[72] 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? *European Journal of Psychology of Education*, pages 1–22, 2017.

[73] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

[74] A. C. Rencher. Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician*, 46(3):217–225, 1992.

[75] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois, 2017. URL `https://CRAN.R-project.org/package=psych`. R package version 1.7.5.

[76] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[77] 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., 2004.

[78] R. M. Royall. The effect of sample size on the meaning of significance tests. *The American Statistician*, 40(4):313–315, 1986.

[79] D. W. Russell. In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and social psychology bulletin*, 28(12):1629–1646, 2002.

[80] P. H. Schonemann. Facts, fictions, and common sense about factors and components. *Multivariate Behavioral Research*, 25(1):47–51, 1990.

[81] G. Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

[82] C. B. Somers. Correlates of engineering freshman academic performance. *European journal of engineering education*, 21(3):317–326, 1996.

[83] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search.* MIT press, 2000.

[84] J.-F. Superby, J. Vandamme, and N. Meskens. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining*, volume 32, page 234, 2006.

[85] B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics.* Pearson Education, 6th edition, 2013. ISBN 9780205956227.

[86] J. P. Tangney, R. F. Baumeister, and A. L. Boone. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of personality*, 72(2):271–324, 2004.

[87] J. Vanderoost, C. Van Soom, G. Langie, J. Van den Bossche, R. Callens, J. Vandewalle, and T. De Laet. Engineering and science positioning tests in Flanders: powerful predictors for study success? In *Proceedings of the 43rd Annual SEFI Conference*, pages 1–8, 2015.

[88] T. S. Vaughan and K. E. Berry. Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. *Journal of Statistics Education*, 13(1):1–9, 2005.

[89] C. Weinstein and D. Palmer. *Learning and Study Strategies Inventory User's Manual.* H&H Publishing Company, 2nd edition.

Master's thesis filing card

Student: Ramaravind Kommiya Mothilal

Title: Statistical modeling of students' performance in an open-admission bachelor program in Flanders

Dutch title: Statistische modellering van studentenprestaties in een Vlaams bachelorprogramma met open toelating

UDC: 621.3

Abstract:

In universities with an open-admission system, students sometimes choose study programs that are inappropriate for their prior-knowledge, skills, and abilities and hence face huge challenges in successfully completing their education. The present study applies statistical modeling to extract insights about what is expected from the students to perform well in their first-semester exams. Good performance in the first semester reflects a smooth transition to higher education and acts as an early indicator for completing the study program. Although related works in this area are numerous, many have succumbed to the practice of indiscrimination between two distinct modeling approaches, namely explanatory and predictive, often leading to incorrect practical conclusions.

The present study discusses how appreciating the differences between explanatory and predictive modeling not only results in correct practical conclusions, but also reveals the advantages and disadvantages of these two approaches. This revelation, in turn, gives rise to three different modeling approaches to solve the problem at hand. The first approach employs interpretable statistical models to explain how different traits of students affect their first-semester performance. The second approach, still using interpretable models, tries to optimize both explanatory and predictive goals with an objective of generalizing the statistical inferences to "unseen" incoming students and strengthening the validity of explanations. Finally, the third approach employs complex algorithmic models to accurately predict students' performance by uncovering patterns and relationships that are difficult to hypothesize. The first and third approaches differ distinctly in their statistical goals, choice of models, and evaluation criteria, whereas the second approach combines aspects of the other two approaches. The present study discusses how each of the three approaches solves the problem at hand from different perspectives and further discusses their advantages and disadvantages.

The above three approaches result in three common insights: (1) prior knowledge and the effort exerted to acquire prior knowledge positively influences the first-semester performance strongly than affective and goal strategies, (2) affective strategies positively influence the first-semester performance only in an indirect way through prior knowledge and the effort exerted to acquire prior knowledge, (3) students with well-defined goal strategies tend to exert less effort in acquiring prior-knowledge, thereby perform poorly in the first semester. However, while the first two approaches show that students' preference for time pressure has no influence on the first-semester performance, the third approach shows that higher levels of pressure preference are associated, though not very strongly, with students who perform poorly in their first semester.

Thesis submitted for the degree of Master of Science in Artificial Intelligence, option Engineering and Computer Science

Thesis supervisor: Prof. dr. ir. Tinne De Laet

Assessors: Tom Broos
Sven Charleer
Maarten Pinxten

Mentor: Tom Broos