Student Success Prediction in International Business

Mees Brinkman[1], Wout van Velzen[2], Hani Al-Ers[3], Cornelis Beyers[4], Xiao Peng[5*]

[1]Faculty Management & Organisation, The Hague University of Applied Sciences, The Netherlands

[2,3] Faculty IT & Design, The Hague University of Applied Sciences, The Netherlands

[4,5] Faculty Business, Finance and Marketing, The Hague University of Applied Sciences, The Netherlands

x.peng@hhs.nl

Abstract

Approximately half of the students of The International Business program at The Hague University of Applied Sciences (THUAS) found in this program do not successfully complete the propaedeutic phase. There has also been a significant increase in the number of enrolled students. This makes good study advice important. This research aims at improving study advice by accurately predicting if a student will successfully complete the propaedeutic phase.

This paper explores the socio-demographic factors (age, type of the previous education, ethnicity, 1st nationality), previous education grades (high-school Grade Point Average (GPA), high-school math and English grades) and intake test results (intake math and intake English) that may predict a student's success.

In total, there is data available from 3349 students who joined the International Business program in the years 2010 to 2017, each student having up to 50 different factors. This data has been used to train and test several predictive models: the Classification And Regression Tree (CART) model and the Random forest model, which were separately optimized to find the best combination of factors to predict if a student will successfully complete the propaedeutic phase.

This research concludes that average high-school grades, intake math and intake English grades are exceptionally good predictors. The most accurate model turned out to be the Random forest model with a 61.9% accuracy, similar to the accuracy of the CART model which reached an accuracy of 61.3%, but for some groups of students, a much higher accuracy might be reached.

Keywords: CART; Education; Machine learning; Modelling; Student attrition.

1 Introduction

1.1 Student success prediction

The percentage of the global population applying for tertiary education is now higher than ever (UNESCO Institute for Statistics, 2020) and many of those students pick a study they will eventually fail (Tinto & Cullen, 1973). One method to mitigate student attrition could be to provide study advice before students decide to pick a study. This assumes that students would not pick a study if they knew they would almost certainly not complete it.

There have been many studies regarding factors predicting student success already, each using their own dependent variable (for example, reaching a certain grade on a test or dropout). Student success can be predicted by using many factors according to Kovačić (2010), who identified and compared different factors that might influence students' performance. Specifically, socio-demographic features which can be difficult to objectively measure and evaluate. Kovačić found that, when combined, ethnicity, course programme and course block are the three most predictive factors. Of the models used, the Chi-square Automatic Interaction Detector (CHAID) tree provided an accuracy of 59.4% and the Classification And Regression Tree (CART) provided an accuracy of 60.5% (Kovačić, 2010). A research exploring what matters most for college completion (Chingos, 2018) discovered that children from rich parents/guardians, on average, have better results in tertiary education than children from poor parents/guardians. Though on its own, family income is hardly accurate enough to advise students about their education choice. This study also found that the American College Test (ACT) and Scholastic Assessment Test (SAT) scores (standard tests used for college admissions in the United States) are much more relevant when predicting student performance.

It was found that the Grade Point Average (GPA) in math, English, chemistry, and psychology courses are among the strongest individual predictors of attrition. Another conclusion is that there is no significant difference in accuracy between the mathematical models (Aulck, Velagapudi, Blumenstock, & West, 2016). A study about the prediction value of high-school grades (Geiser & Santelices, 2007) concluded that grades in college-preparatory subjects are good predictors for academic success and the chance of failure of a student increases when coming from lower educational backgrounds.

A case study in predicting student academic performance (Huang & Fang, 2013) found that different mathematical models offer very little to no difference in accuracy when predicting a student's success. For predicting a class of students as a whole, taking the students' cumulative high-school GPA as a sole predictive factor was found to be the best predictor, rather than using a combination of factors. For individual students however, the study found that the support vector machine model using six predictor factors was the most effective for predicting an individual student's success.

Multiple studies above found that GPA is a good predictive factor, that multiple (but not all) factors work best for predicting an individual student' success and that different models do not differ much in prediction accuracy.

This study is a follow-up work on a study about how academic and personal background of the students can affect their success in the first year (Al-Ers et al., 2021) which analyzed correlations between factors (such as math grades, intake English grades and first study year)

with successfully completing the propaedeutic phase or not. This study concluded that students from MBO backgrounds and students with low high-school math GPA's have a higher dropout chance compared to other students. The study also found that parents' educational background and gender can help in predicting student success. In addition, the study also found that intake math and English tests are not a good indicator of student's success compared to high-school math and English grades. The study estimated that student success could be estimated with 61.5% accuracy. The study gathered a dataset of 3797 students of the International Business program at THUAS which will be used in the present study. The present study will perform a more extensive analysis using this dataset.

1.2 Motivation of the present study

The number of enrolled students for the International Business program at THUAS from 2015 to 2019 has been steadily increasing (Dienst Uitvoering Onderwijs, n.d.). Though there are a lot of students, only 48% of them successfully complete the propaedeutic phase, as can be concluded from the dataset. Assuming that giving students more accurate study advice will result in more suited students joining the program, improving the study advice is an important step to mitigate student attrition.

Study advice can be improved using the support of a prediction whether or not a student would successfully complete the propaedeutic phase of the international business program, or as it is called in this research: "Student success". Evaluating which of the provided factors can most effectively predict student's success might also give new insights as to what students might need from THUAS to perform better.

1.3 Research questions of the present study

In this research, it is attempted to find the most appropriate method to predict if a student will successfully complete the propaedeutic phase of the International Business program at THUAS using the current dataset. To do so, the more important factors will be selected, after which different mathematical models or techniques will be employed to predict if a student will be successful.

2 Methodology

2.1 The dataset

The dataset contains 50 factors from 3349 students. The data was collected from 2010 to 2017. All students were included. Because the course is international, high-school grades of many students could not be included in the dataset. In addition, only 6.5% of all parents presented their academic certificates. The International Business program at THUAS only started having students take intake tests on mathematics and English from 2015, hence 62.9% of students do not have this data. There are over 120 different nationalities among the students in the dataset. This has caused some trouble in training the model, which is why the nationalities have been grouped to make the model more accurate. Other factors that had this same problem and have been given the same treatment. All changes to the factors can be found in Table 1.

Some factors, like "Intake math", are converted in two steps. In the first step, the grades are rounded to whole numbers. In the second step, the whole numbers are combined into small groups. This makes the model more accurate and effective. Table 1 gives an overview with examples of what the data would look like after conversion.

Table 1, all factor changes in this study per factor

Factor	**Original factor values**	**Converted factor values**	**Description**
Intake math *(conversion step 1)*	*Example:* 54.3 / 67.4 / 94.3	*Example:* 5 / 6 / 9	Rounded grades.
Intake math *(conversion step 2)*	1, 2, 3 / 8, 9, 10	"3 or lower" / "8 or higher"	Combines small groups.
Intake English *(conversion step 1)*	*Example:* 54.3 / 67.4 / 94.3	*Example:* 5 / 6 / 9	Rounded grades.
Intake English *(conversion step 2)*	1, 2, 3 / 8, 9, 10	"3 or lower" / "8 or higher"	Combines small groups.
Age	*Example:* *(factor: cohort)* "2016" *(factor: birthdate)* "01-11-1997"	*Example:* *(=10 months, 18 years old at 01-09-2016)* 18	Calculated age on the first day of the International Business program.
High-school GPA	*Example:* *(factor: subject 1002)* 7.4 *(factor: subject 1019)* 5.4 *(factor: subject 1064)* 8.2 *(other numerical subject grades…)*	*Example:* 7	Takes all numerical subject grades from high-school, then combines and rounds them to one grade.
1st nationality	"Bulgaarse" / "Chinese" / "Nederlandse" / *(anything else)*	"Bulgaarse" / "Chinese" / "Nederlandse" / "other"	Only keeps nationalities with 200 or more students being in that category.

Gender	"Man" / "Vrouw"	"Male" / "Female"	Translation to English.
Passed propaedeutic phase (dependent variable)	0 / 1	"Failed" / "Passed"	Makes the 0s and 1s readable.
Math (high-school)	1, 2, 3, 4, 5 / 8, 9, 10	"5 or lower" / "8 or higher"	Combines small groups.
English (high-school)	1, 2, 3, 4, 5, 6 / 8, 9, 10	"6 or lower" / "8 or higher"	Combines small groups.

2.2 The models

To create the best method, two models were chosen for evaluation. It was concluded from other literature that these two models were best suited for the dataset (Huang & Fang, 2013).

To be able to include the large number of different factors, the Random forest and CART model were chosen as predictive mathematical models. Using evaluation methods explained in the chapter below, these methods can also objectively be evaluated, making a suitable candidate for the purposes of this research.

The CART model is much like the description in a study about classification and regression trees (Leo Breiman; Jerome H. Friedman; Richard A. Olshen; Charles J. Stone, 1984). In short: this creates many trees and selects the best tree it can find.

The Random forest model is based on how Breiman described it in his study about random forests (Breiman, 2001). In short: this creates many trees, which for each column of data, together, attempt to predict the dependent variable.

2.3 Compiling a combination of factors

First, a baseline is created, this baseline is a compilation of all factors that could be used in the models. The baseline is used for comparison to other combinations and will be the start of the iterative process that follows.

Filtering out the factors and compiling the optimal combination of factors can be done using two methods, explained in the paragraph below. In this study, both will be employed. The importance and influence of factors on the accuracy of the model can be individually calculated. This gives an overview of which factors might be excluded from the model. The score of each factor could also indicate which factors are least influential to a student's performance. The model is iterated by removing the least important factor. If the accuracy increases, the change is definitively applied. If it does not, the second-least important factor is removed from the model. This process is repeated until the accuracy of the model does not increase anymore.

2.4 Model evaluation

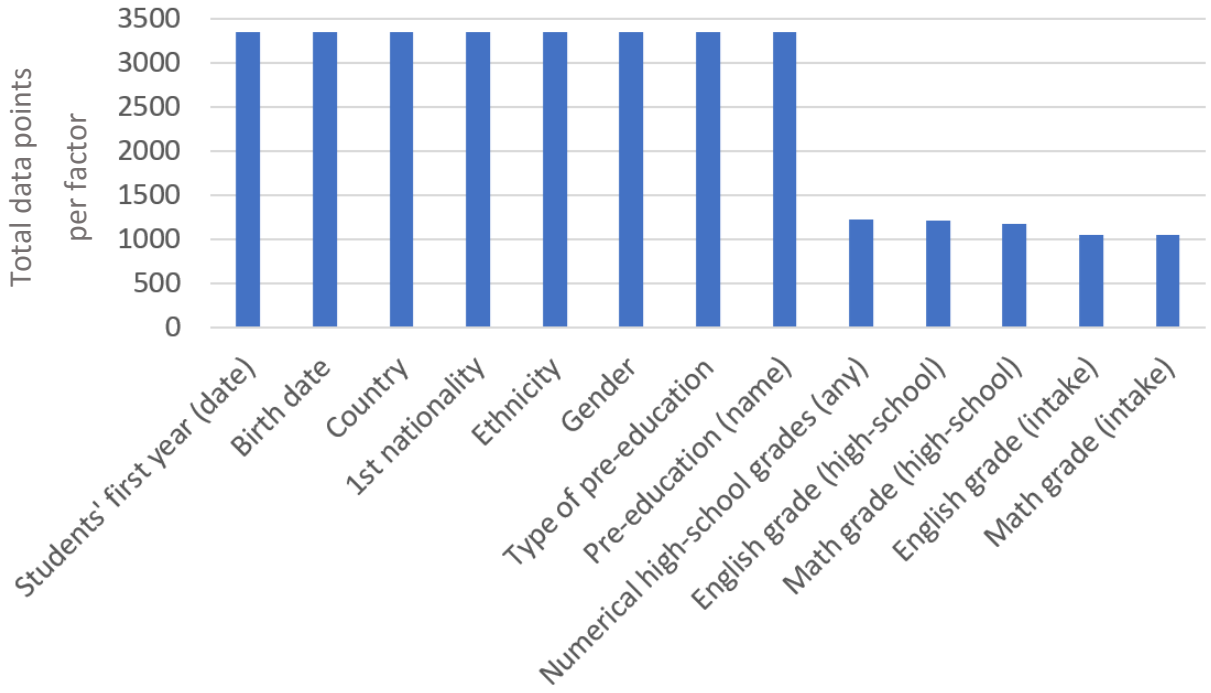
To effectively evaluate a model, 20% (670) of the dataset is randomly assigned as test data and the other 80% (2679) is used to train the model. Using the test data, the accuracy is calculated using a confusion matrix. However, if this whole process is repeated, the predicted accuracy can be very different. To increase the accuracy of the accuracy measurement, the train and test data is shuffled and a new model is made, which is again evaluated. This process is repeated at least 2000 times before an average accuracy is measured and documented.

3 Analysis & results

3.1 Selecting useful factors

During this research, it was decided to first see which factors were useful enough to be included in the CART and Random forest models. Excluding the factors which are unknown before the student starts with International Business and excluding factors with less than 10% of their data points filled. What remains are 13 factors, as is illustrated in Figure 1.

Figure 1: Factors sorted by available data points



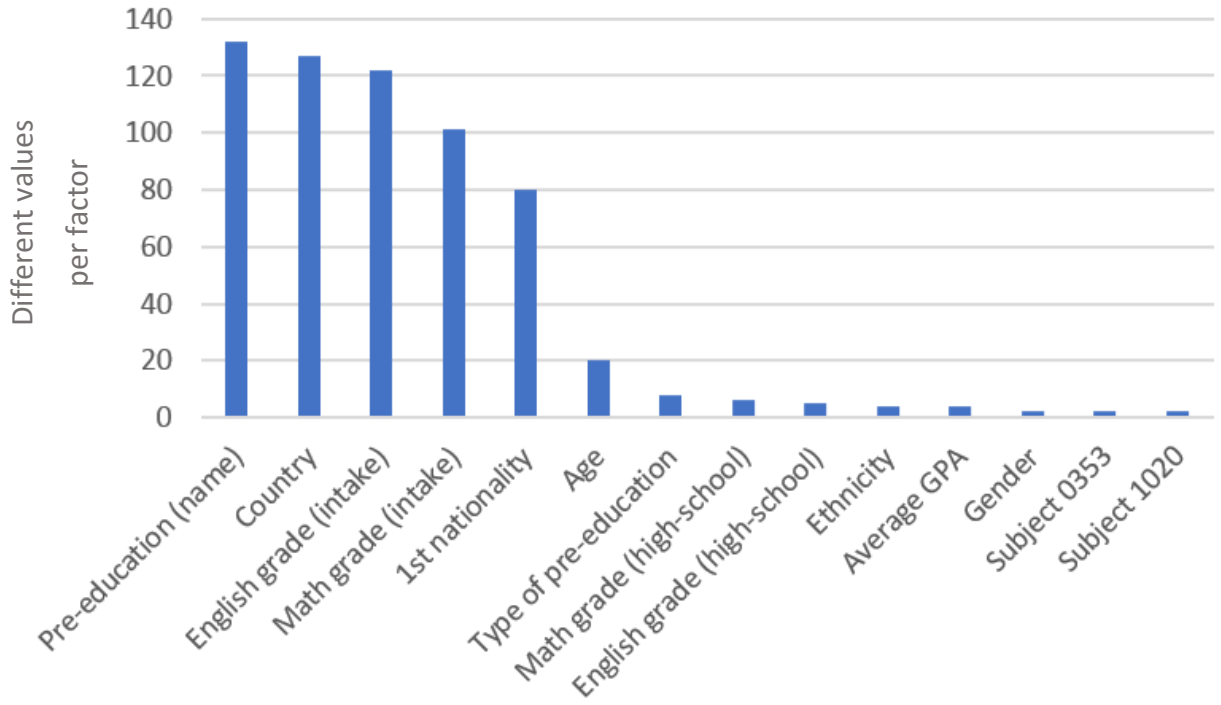
Although it is not shown in this graph, specific high-school subject grades are also available in the dataset. However, most of them only rarely have data points, so all numerical grades are combined into an average high-school grade for each student, which gives a total of 1231 data points. Literature shows that this factor can be very significant when indicating student success (Aulck, Velagapudi, Blumenstock, & West, 2016). Two subjects are mostly filled with characters instead of numbers and so could not be included in the overall GPA, but instead were saved as individual (far less important) factors. Other factors have at least 1000 entries, making overfitting much less likely.

The factor which indicates the year a student starts with the International Business study (factor: "Students' first year (date)") will not help with predicting a student's success for students from 2018 or later since it does not have the average success rates of those years. Therefore, this factor is unfit for practical use and is removed from the pool of factors. A student's birth date does not help when faced with students from different study years. However, it is possible to

calculate the age of each student on their first day of the International Business study. This factor will be added instead of a student's birth date.

Next comes the final problem, as stated in the Methodology chapter: factors with too many different values per factor, as seen in Figure 2.

Figure 2, Factors sorted by the number of different values per factor



To solve this problem, grouping was attempted, as stated in the Methodology chapter. In addition, the age factor was saved as a numerical value to allow for "age: > 22" expressions. "Pre-education (name)" however has only small groups or groups already defined in "Type of pre-education". Country does not have to be kept due to other factors already being in place to indicate nationality. Both factors will not be used in the next parts of this research.

The remaining factors are age, 1st nationality, gender, type of pre-education, average high-school grades, English grade (high-school and intake), math grade (high-school and intake) and ethnicity.

3.2 Individual factor importance

Using the CART model, individual factors were analyzed to see their effectiveness in predicting student success. The CART model was only given one factor at a time and has to predict every student's success using only this factor (3349 students). If there is no data available, the CART model will resort to the best guess, which brings an accuracy of 52%. The results are in Table 2.

Table 2: CART model accuracy per individual factor, sorted by highest accuracy on test

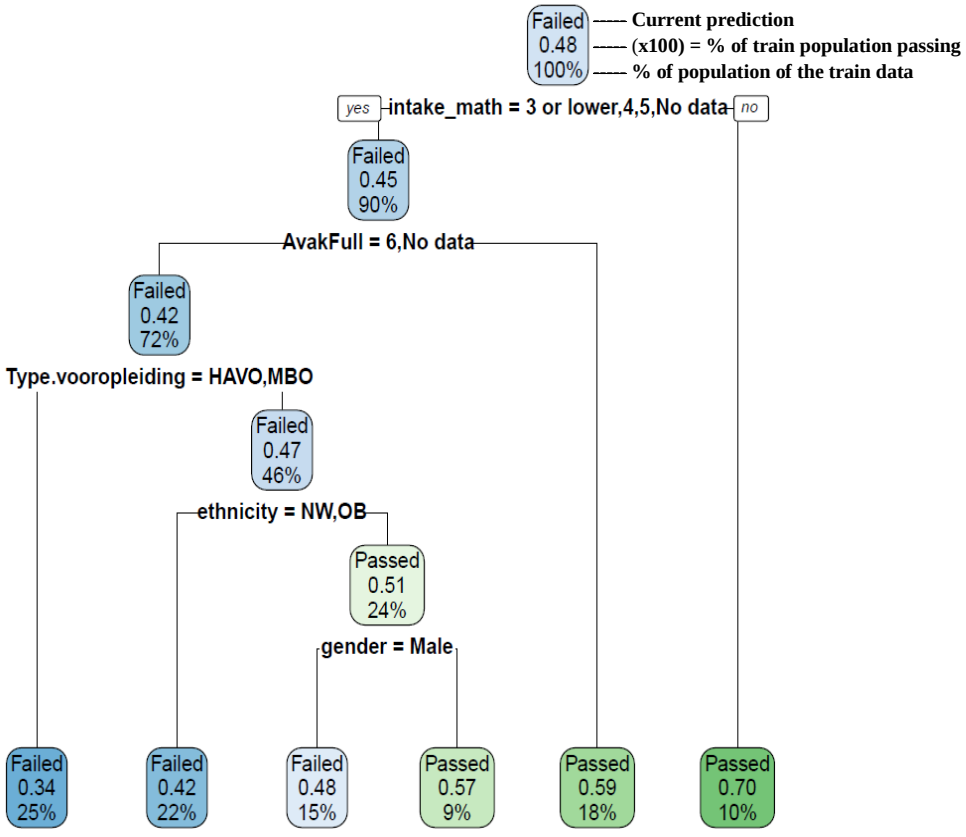
Factor	**Accuracy after 1000 runs (Test)**	**Accuracy after 1000 runs (Train)**	**Available datapoints**
High-school GPA	57.4%	57.2%	1231
Math grade (intake)	57.1%	57.5%	1048
English grade (intake)	56.8%	56.9%	1048
Math grade (high-school)	56.2%	56.0%	1182
2nd subject (character)	54.9%	54.6%	1020
1st subject (character)	54.4%	54.5%	1024
English grade (high-school)	54.3%	54.6%	1219
Ethnicity	54.3%	54.5%	3349
Gender	54.1%	54.4%	3349
Age	53.8%	54.9%	3349
Type of pre-education	53.7%	54.4%	3349
1st Nationality	52.0%	52.8%	3349
Best guess (100% fail)	52.0%	52.0%	0

It should be noted that these accuracy measurements have some inaccuracy (as seen by some factors scoring 0.3% higher on the test than the training data) and that these measurements are for the dataset as a whole. However, some factors may be much more significant if given a selection of the dataset. For example, if given only students from a HAVO or VWO background, high-school GPA will turn out to be much more significant because it has more data points available for these groups.

3.3 Regression tree model

The best combination of factors and settings found in this study for the CART model resulted in an average accuracy of 61.3% after 10,000 trees. The factors used are intake math grade, type of pre-education (translations in the appendix), ethnicity, gender and high-school GPA. All trees were also required to have at least a sample size of 500 before attempting a split and each split having at least a third of that to prevent overfitting. Figure 3 is a visualization of the CART model with the highest accuracy on the test data. If a student is in one of the mentioned groups they go left and if they do not belong to any of the mentioned groups, they go right. For example, if a student doesn't have an intake math grade, they go left at the first question. It should be noted that the accuracy may strongly increase or decrease depending on how many of those factors have data available. Most importantly, intake math and high-school GPA help in analyzing students. Almost all variations of CART models use these two factors as their first two predictors, including the one below.

Figure 3, Best CART model out of 10,000.



Referring to Figure 3, it is worth noting the big difference in the amount of "Passed" and "Failed" predictions. Only 37% of the students are predicted to pass, while 63% of the students are predicted to fail. This is mostly since there is a category of 15% with about the same prediction accuracy as at the start. The model was configured not to further split this group due to potential overfitting, as overfitting would lower the average accuracy of the model.

The CART model in this visualization first attempts to exclude two groups of students with good success chances using the two most important factors but does not use the #3 and #4 best predictors. It does this because, after the intake math predictor and high-school GPA (which includes math and English high-school grades), those are much less significant. It should be noted that students from HAVO and WVO generally have high-school grades, so basically, only those students with a 6 (lower is not possible) go to the third step. This means we have a selection of low-performing HAVO and WVO students and most of the international and MBO students (since most of them did not do an intake math test). MBO students already have a low success chance in general, so do low-performing HAVO students. WVO, however, is a very strong group when it comes to success chances, so the model decides to only put MBO and HAVO students into the "failed" category, despite their high-school GPA being low or, in rare cases, being unknown.

Due to lack of grades (if someone did not do the intake math test they rarely did intake English either), the selection is continued using less predictive but always available demographic data.

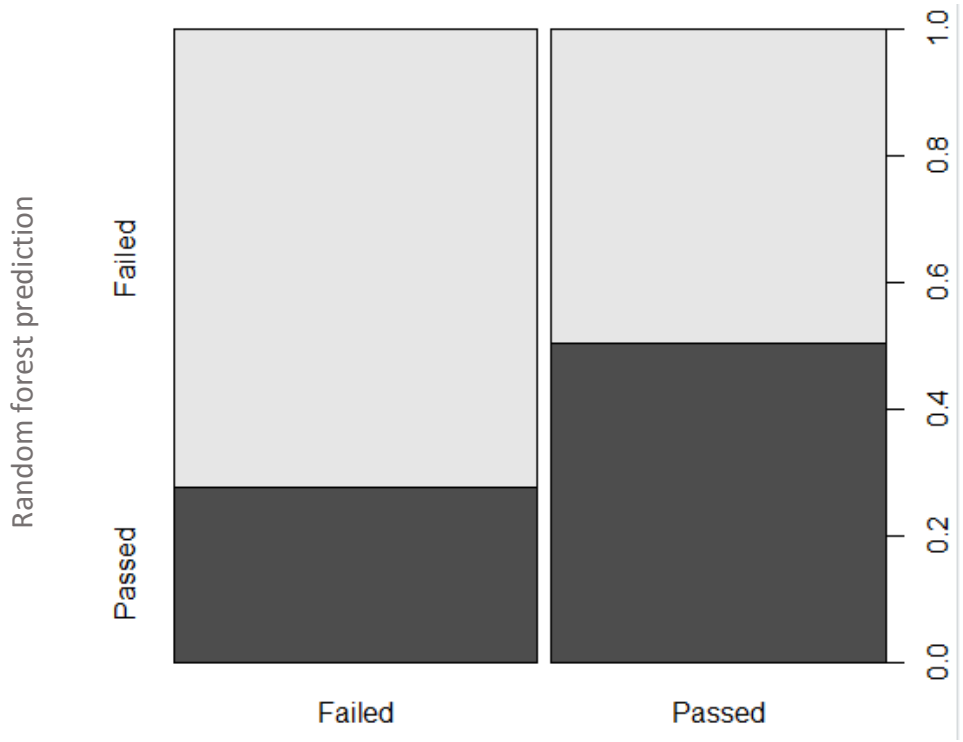
Firstly, students are selected based on ethnicity, with NW (non-western foreign) and OB (unknown) being put in the failed category and WE (western foreign) and AU (autochthonous, ethnic Dutch) being put to the last test. This is an ethically doubtable factor, but it could be seen as an alternative Dutch/English test. The exact reason for this however could be worth a study on its own.

After having used the most important demographic factor, the CART model calculates that the 2nd most important factor (gender) is good predictor, as females have a higher passing chance than males. However, this results in another problem: the students in the group of males end up with the exact same accuracy as they started with. This group should be treated as not predictable for the model. This group could be further selected by putting everyone with an intake math or math grade of 4 or lower or an intake English or English grade of 3 or lower in the "failed" category and the others in the "passed" category, but this is simply an educated guess.

3.4 Random forest model

The best Random forest model included the same factors as the CART model. In total, the Random forest used the factors: intake math grade, type of pre-education, ethnicity, gender and high-school GPA. This model has an average accuracy of 61.9% after 2,000 forests, being only 0.6% better than the CART model. Visualized in Figure 4, the estimate of the Random forest can be seen on the left side (dark grey = Random forest predicts Passed, light grey = Random forest predicts Failed). More students belong to the "Failed" category (52%) than to the "Passed" category (48%), which is the reason for the unequal distribution of predictions.

Figure 4, predicted student result compared to actual result



3.5 Comparing the models

Every model type attempted in this study and its results, including the best guess without using any factors, can be found in Table 3.

Table 3: Comparison of models used in this study

Model type	Average accuracy (test)	Average accuracy (train)	Total models created	Factors used	Simple visualization
Random forest	61.9%	65,3%	2,000	5	No
CART	61.3%	61.6%	10,000	5	Yes
Best guess (100% fail)	52.0%	52.0%	1	0	Yes

Given that both the Random forest and the CART model are almost identical regarding the accuracy, the question now is which model is the easiest to use. The CART model is by far the easiest to use since it's a simple scheme, easy to understand and substantiate.

4 Conclusions and recommendations

In this paper, many different factors and combinations of factors were investigated to predict if a student will successfully complete the propaedeutic phase during International Business at THUAS. The main research goal was to find the most effective method to predict if a student will successfully complete the propaedeutic phase in this environment.

Many factors in the dataset were not included in the analysis, due to the reasons stated in the Analysis chapter. After analyzing the remaining individual factors, almost all factors turned out to be predictive to some degree. However, only a few factors could predict student success with 57% or higher accuracy, which is 5% above the best guess accuracy. Those factors were high-school GPA (being widely supported by other research on this topic) and the intake math grade. The CART model almost exclusively uses these two factors as its first two selectors to predict student success.

The CART model and Random forest model were both used to predict if a student will pass the propaedeutic phase. Both models were separately given different combinations of factors, assuming that the models would have different optimal factors. However, both models achieved the highest accuracy when using 5 factors: intake math grade, type of pre-education, ethnicity, gender and high-school GPA. This resulted in the CART model achieving an average accuracy of 61.3% and the Random forest model achieving a slightly higher average accuracy of 61.9%.

The random forest is the most accurate. However, the most preferable method would be the CART model. Unlike the Random forest model, the CART model uses a simple scheme, is easy to understand and substantiate, making it much fitter for practical evaluations. In addition, the CART model can give up to 70% certainty for some groups as shown in Figure 3, allowing the model to only be partially used to give study advice, but with much higher accuracy.

For further research, it would be advisable to have all students participate in the intake math test. Currently, not even a third of the students in the dataset has an intake math test result and yet it turned out to be the 1st or 2nd best predictor. The intake test seemed to be more effective than the average high-school math grade, being 4th place in the list for predictive factors. In addition, the English intake grade is currently the 3rd best predictor with less than a third of the students in the dataset having this grade available. If more students would take this test, this factor could be much more influential. In addition, further research could explore the reasons as to why certain ethnicities have a higher success chance than others.

By exploring different predictive models and many combinations of factors, this research has shown how a student's success can be effectively predicted using different approaches and how to possibly make future predictions even better for International Business.

5 References

Al-Ers, H., Yilmaz, V., Lengkeek, J., Syoufi, M., & Peng, X. (2021). *Students' success.* Holland, The Netherlands.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). *Predicting Student Dropout in Higher Education.* Seattle: DataLab, The Information School, University of Washington.

Breiman, L. (2001). *RANDOM FORESTS.* Berkeley: University of California.

Chingos, M. M. (2018). *What Matters Most for College Completion?* American Enterprise Institute and Third Way Institute.

Dienst Uitvoering Onderwijs. (n.d.). *Aantal hbo-ingeschrevenen (binnen domein ho).* Retrieved from duo.nl: https://duo.nl/open_onderwijsdata/databestanden/ho/ingeschreven/ingeschrevenen-hbo1.jsp

Gardner, J., & Brooks, C. (2018). *Student Success Prediction in MOOCs.* Michigan: School of Information; University of Michigan.

Geiser, S., & Santelices, M. V. (2007). *VALIDITY OF HIGH-SCHOOL GRADES IN PREDICTING STUDENT SUCCESS BEYOND THE FRESHMAN YEAR: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes.* California, Berkeley: Center for Studies in Higher Education.

Huang, S., & Fang, N. (2013). *Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models.* Elsevier.

KDnuggets. (2019). *Top Analytics/Data Science/ML Software in 2019 KDnuggets Poll*. Retrieved from kdnuggets.com: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

Kovačić, Z. J. (2010). *Early Prediction of Student Success: Mining Students Enrolment Data.* Wellington, New Zealand: Open Polytechnic. Retrieved from http://proceedings.informingscience.org/InSITE2010/InSITE10p647-665Kovacic873.pdf

Leo Breiman; Jerome H. Friedman; Richard A. Olshen; Charles J. Stone. (1984). *Classification and regression trees.* Monterey: Brooks/Cole.

Tinto, V., & Cullen, J. (1973). *Dropout in Higher Education: A Review and Theoretical Synthesis of Recent Research.* Columbia Univ., New York, NY. Teachers College.

UNESCO Institute for Statistics. (2020, 09). *School enrollment, tertiary (% gross)*. Retrieved from data.worldbank.org: https://data.worldbank.org/indicator/SE.TER.ENRR

6 Appendix: Translations for Type.vooropleiding

Type.vooropleiding has 8 values:

BD	Buitenlands diploma
	Foreign certificate
BUITENL_SL	Buitenlandse vooropleiding via SL – 01
	Foreign preliminary training
CD	CD Hogeschool Examen HHS
	Admission test (Hague University of Applied Sciences)
HAVO	Hoger algemeen voortgezet onderwijs
	General secondary education
HO	Propedeuse Bachelor HBO / Bachelor HBO
	Higher education
MBO	Middelbaar beroepsonderwijs
	Secondary vocational education
VWO	Voorbereidend wetenschappelijk onderwijs
	Pre-university education
OVERIG	Anders: Europese School
	Other: European education