UNIVERSITY OF AMSTERDAM
Economics & Business

MSc Data Science and Business Analytics Thesis: Dutch Higher Education: Predicting Bachelor Enrolment

July 15, 2025

Roel Lust
13985736

Company: University of Amsterdam,
Faculty of Economics and Business
Contact Person: Drs. F.H.K. Pope

Supervisors:
Dr. S.T. Mol
Dr. I.M. Zwetsloot

1 Statement of originality

This document is written by Student Roel Lust who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

2 Abstract

The University of Amsterdam's Faculty of Economics and Business (FEB) receives applications from students worldwide each year across its six bachelor programs. However, it is often unclear which admitted students will ultimately enrol, as many apply to multiple programs both within the university and at other institutions. This uncertainty makes it crucial to obtain a timely and accurate estimate of expected student enrolment to support logistical planning.

This research aims to improve and augment the university's current prediction methods by employing machine learning algorithms to estimate the probability that each admission will convert into an enrolment. These individual probabilities are then aggregated to predict the total number of expected enrolments. To improve prediction accuracy, the logistic regression model incorporates additional variables such as language proficiency test results, the number of applications submitted both within and outside the university, age, and the applicant's admission group.

The central research question of this thesis is how well machine learning models can predict student enrolment compared to existing methods. The specific objectives include producing timely predictions to support operational planning, comparing model performance with the university's current manual estimation approach, and assessing the impact of recent Dutch political policy changes on Dutch student admissions, particularly due to the addition of Dutch tracks to existing programs.

Overall, the findings suggest that the models created can effectively support and potentially improve existing prediction methods. However, the added value varies across different programs. In addition to improving enrolment predictions, this research also identifies key features that influence a student's likelihood to enrol, providing valuable insights for making predictions.

Contents

1 Statement of originality . . . 1
2 Abstract . . . 1
3 Introduction . . . 3
4 Literature review . . . 4
5 Case background . . . 7
6 Method . . . 7
6.1 Business understanding . . . 8
6.2 Data understanding . . . 8
6.3 Data preparation . . . 11
6.4 Modeling . . . 14
6.5 Evaluation . . . 17
6.6 Deployment . . . 25
7 Results and discussion . . . 25
8 Product . . . 26
9 Conclusion . . . 26
10 Limitations and future research . . . 27
11 Acknowledgement . . . 28
12 References . . . 29
13 Appendix . . . 30

3 Introduction

The University of Amsterdam (UvA) receives a large number of applications each year from students around the world. Applicants apply to one or even multiple programs, including those within the Faculty of Economics and Business (FEB). However, because application deadlines are set well in advance of the formal start of the academic year, it is often unclear which applicants will ultimately enrol. While students may submit multiple applications, most will eventually choose only one program. Furthermore, some students also apply to universities outside the UvA, both within and outside the Netherlands, and the status of those external applications is unknown to the university. As a result, there is always uncertainty about how many of the formally admitted students will actually start their studies.

Accurately predicting enrolment numbers for each program is therefore essential. With the academic year beginning in early September, the university cannot afford to wait until the last moment to assess actual student numbers. Key logistical decisions, such as allocating teaching staff and reserving adequate classroom space, must be made well in advance. Although not the primary objective, predicting the number of students who will enrol can also offer insights into expected tuition return for the university.

This thesis is part of a broader forecasting project initiated by the FEB at the start of 2025. The aim of the project was to develop a one-time forecast of the total student intake for the 2025-2026 academic year, supporting both strategic decision making and educational planning across all Bachelor's and Master's programs offered by the faculty. The project was conducted in collaboration with AI4Business and involved two students: Maarten Hoogeboom and myself. While this thesis focuses specifically on forecasting the enrolment for the Bachelor's programs, Maarten Hoogeboom's thesis addresses enrolment predictions for the Master's programs (Hoogeboom, 2025).

While the University of Amsterdam already benefits from enrolment predictions based on historical conversion rates, staff experience, and known program changes, these methods take up a lot of time, whereas machine learning models might have the potential to work faster and possibly improve predictions. The goal of this project is to improve predictive accuracy by applying machine learning techniques to historical program data. By estimating the probability of each admitted student actually enrolling and aggregating these probabilities, the model aims to produce a more precise forecast of total enrolment numbers per program. This will be followed by a comparative analysis between the machine learning-based predictions and the university's current estimation approach.

This thesis examines the prediction of student conversion probabilities using binary classification methods to estimate the total number of students expected to enrol in each program within the Faculty of Economics and Business at the University of Amsterdam for the upcoming academic year. The central research question is:

"How accurately and effectively can machine learning models predict student enrolment for the University of Amsterdam's Faculty of Economics and Business bachelor programs, and how do these predictions compare to current estimation methods?"

In addition, this study investigates how recent Dutch political policy changes, specifically the introduction of Dutch language tracks, impact enrolment patterns, particularly among Dutch students. Due to recent Dutch political policy changes, universities are now required to increase their Dutch language teaching capacity. For the Faculty of Economics and Business, this has led to the creation of new Dutch language tracks within programs that were previously offered exclusively in English. As a result, each of those programs is now divided into a Dutch and an English track.

In sum, the following objectives formed the core of the research presented in this thesis:

1. Develop a machine learning model to predict the probability of individual student admissions to convert into actual enrolments.
2. Generate timely predictions by June 1st to allow sufficient time for logistical planning, including staffing and classroom allocation.
3. Analyse the impact of Dutch language related policy changes on enrolment, with a focus on student distribution between the newly introduced Dutch and English tracks.
4. Compare the performance of machine learning based predictions with the university's current method.

The remainder of this thesis is organised as follows. Section 4 discusses the importance of predicting student enrolment numbers and reviews related work. Section 5 provides background information on the importance of predicting student enrolment. Section 6 outlines the methodology in detail, following the CRISP-DM framework. Section 7 presents the results of the analysis. Section 8 describes the final deliverables developed during the project. Section 9 offers the overall conclusions drawn within this thesis. Finally, Section 10 discusses the limitations of the research and offers suggestions for future research both within the Faculty of Economics and Business and beyond.

4 Literature review

In the months leading up to the start of a new academic year, universities are typically busy preparing for the incoming first-year students. One of the main challenges is allocating classrooms and teaching staff in advance, due to uncertainty about how many students will actually enrol. While application numbers offer an indication, they are certainly not definitive, as students may be rejected by the institution or choose to withdraw their applications at any time. As a result, the actual number of students who will attend often only becomes clear at the very beginning of the academic year (Vonk, 2022).

Higher education institutions often aim to operate at or near full capacity, as doing so maximizes tuition revenue, which is a critical source of funding for many universities. To reduce financial uncertainty caused by fluctuations in student enrolment, these institutions need to accurately forecast the size of incoming student cohorts. For universities that rely heavily on tuition income, accurate enrolment predictions are essential for effective resource planning, including staffing and classroom allocation. Inaccurate forecasts can lead to misallocated resources and operational inefficiencies (Basu et al., 2019).

Dutch policy changes

Internationalisation in higher education is on the rise, not only in the Netherlands but across Europe. This trend brings economic advantages: universities benefit from the tuition fees paid by international students, and there is additional economic potential if these students remain in the country after graduation. However, the likelihood of students staying post-graduation is relatively low, particularly for those from outside the European Union. At the same time, the growing number of international students adds pressure to an already tight housing market (De Witte et al., 2020).

The Internationalisation in Balance Act (Wet Internationalisering in Balans, or WIB) is a Dutch legislative proposal intended to regulate the increasing number of international students in higher education. Its primary aim is to preserve Dutch as the main language of most bachelor's programs. To achieve this, institutions would be required to justify the use of English through a formal review process known as the Toets Anderstalig Onderwijs (TAO), a mandatory language assessment for all English-taught bachelor's programs. The proposed legislation would also allow universities to limit the intake of international students, especially those from outside the EU, and encourage them to learn Dutch during their studies. The overall objective is to protect the accessibility and quality of Dutch-language education in the higher education system while maintaining a balanced approach to internationalisation (Joubert & Fuller, 2025).

However, Dutch universities have raised concerns about the WIB proposal. They argue that the TAO introduces uncertainty and places unnecessary pressure on academic staff. In addition, the policy does not take into account the diversity between regions, academic sectors, and institutions. In response, universities have submitted a joint counterproposal to the Ministry of Education, advocating for more flexible measures. At the University of Amsterdam, this has led to a cap of 1,200 students for the English-language tracks of Business Administration and Economics and Business Economics. Other bachelor's programs are introducing Dutch-language tracks, which will allow the university to reduce enrolment in the English equivalents if necessary. At the same time, international students will receive more support to learn Dutch, increasing their chances of staying in the Netherlands after graduation (University of Amsterdam, 2025; Universiteiten van Nederland, 2025).

Earlier research and models used in previous studies

Several institutions have applied classification techniques at individual admission level to predict total student enrolment. Slim et al. (2018) demonstrate that cross-validation is an effective method for training models with limited data, allowing for reliable comparisons between different predictive approaches. In their work, logistic regression and support vector machines (SVM), both as binary classification methods, showed strong performance in forecasting enrolment, particularly when combined with cross-validation on individual student and admission data. Wanjau and Muketha (2018) found that ensemble learning, which combines the outputs of multiple trained machine learning models, is a promising approach for improving enrolment predictions. When data is limited, there is an increased risk of overfitting; ensembling helps decrease this by using the strengths of various classifiers. Their study experimented with binary classification models such as decision trees, Naive Bayes, and random forests, and concluded that ensemble techniques can enhance prediction accuracy.

In terms of feature importance, Slim et al. (2018) highlight that timing of admission, high school GPA, and process-related variables (e.g., whether and when a student applies for financial aid) are strong indicators of enrolment intent. For instance, early applications and timely completion of financial aid forms may reflect higher motivation, while higher grade high school students also tend to show distinct patterns in the application process. Wanjau and Muketha (2018) support that grades can be a good indicator, they focus on grades in mathematics and the final GPA as the most effective predictors in their models. However, their work does not incorporate application process related variables, which Slim et al. found to be meaningful indicators of student intent and likelihood to enrol.

Earlier research within the Netherlands

In the Netherlands, several universities have developed models to predict student enrolment, aligning with the aim of this thesis. Three notable examples come from Tilburg University, Radboud University, and Utrecht University.

At Tilburg University, a master's thesis by Hamers (2017) presented a model that diverges from student-specific data and instead utilises marketing data. The model relies on regression techniques to predict enrolment based on indicators of student motivation, such as attendance at webinars or international fairs. Using logistic regression, the study found that these motivational indicators are successful predictors of enrolment outcomes.

Radboud University has also developed enrolment prediction models, although no formal academic publication has been released at the time of writing. Their approach involves three distinct methods. The first is a basic conversion ratio model, based on the enrolment rates from the previous three years. The second method introduces individual-level prediction using a decision tree model, assigning each admission a probability of conversion. This is supported by a SARIMA time series model to forecast trends over time, resulting in a more dynamic prediction framework. The third method builds on the second but shifts the focus from individual admissions to aggregate program-level data, again using SARIMA to identify trends and patterns in enrolment on a larger scale. These models are outlined in Radboud University's publicly available dashboard tool, which can be accessed online at Radboud Dashboard.

Similarly, a master's thesis at Utrecht University also employs SARIMA time series forecasting to predict enrolment in master's programs (Vonk, 2022). Like the Radboud model, this approach demonstrates the potential of dynamic forecasting techniques for tracking enrolment over time and adjusting expectations accordingly.

A key difference among Dutch universities lies in the concentration of international students in Amsterdam, which is significantly higher than in other cities, as shown in Weber et al. (2024). International students often display different application and enrolment behaviour compared to Dutch students, a factor that must be considered when developing predictive models.

5 Case background

The Faculty of Economics and Business at the University of Amsterdam offers six bachelor's programs: Economics and Business Economics, Business Administration, Econometrics and Data Science, Business Analytics, Actuarial Science and Fiscal Economics (Fiscale Economie). All programs are originally taught in English, with the exception of Fiscal Economics, which is taught fully in Dutch.

Currently, predictions at the University of Amsterdam are made by a small group of staff members within the Faculty of Economics and Business. This group meets several times before the start of the academic year to make predictions based on historical conversion rates, the number of current admissions, and their own professional experience. They also incorporate changes in program structure, admission requirements, and other relevant factors in their predictions.

Timely and accurate predictions are important for logistical planning within the university. Insufficient or late room reservations can lead to overcrowded classrooms and last-minute schedule adjustments, while reserving rooms that are too large may waste valuable campus space that could be allocated to other programs. Similarly, understaffing increases workload and stress for existing teaching staff, potentially exacerbating the mental health crisis in academia (Pace et al., 2021), while overstaffing leads to unnecessary costs and underutilised personnel. Accurate predictions help balance these factors and support more efficient planning.

6 Method

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a widely used methodology for conducting data mining and modelling projects, making it well-suited for this project. It is chosen for its systematic, step-by-step approach that guides the data-miner through every stage of a project, from understanding business objectives and exploring data to modelling, evaluation and final deployment.

For a project such as predicting the number of students starting next year, CRISP-DM offers a clear framework that ensures systematic analysis. The methodology is divided into several phases, each consisting of a set of second-level tasks. These tasks are considered "generic" because they are designed to be broadly applicable across various data mining contexts. CRISP-DM serves both as an overview and as a detailed reference model, offering practical guidance for each phase of the process (Chapman, 2000).

CRISP-DM is structured into the following 6 phases:

Step 1: Business Understanding: Define the project objectives and success criteria.

Step 2: Data Understanding: Exploring and verifying the data relevance and quality.

Step 3: Data Preparation: Cleaning, preprocessing, and engineering features within the data.

Step 4: Modelling: Selecting, training, and evaluating predictive models.

Step 5: Evaluation: Assessing model performance and interpreting results.

Step 6: Deployment: Implementing the model into production.

6.1 Business understanding

The main objective of this project is to accurately predict the number of students who will enrol in next academic year's bachelor programs at the Faculty of Economics and Business Economics of the University of Amsterdam. Since enrolment is operationalised here as a dichotomy, this will be achieved by using binary classification methods to estimate the likelihood that each admission results in the student enrolling in the upcoming academic year. The results will be compared to the university's current prediction approach, which relies primarily on historical conversion rates and staff experience.

Another area of focus is the recent policy driven split between Dutch and English taught tracks within the bachelor programs. The project will investigate whether the number of students opting for the newly introduced Dutch tracks can be predicted, helping to assess the impact of these changes and whether a significant portion of Dutch students are expected to start these new tracks or not.

The business problem focuses on the need for timely and accurate predictions of incoming student numbers. Accurate predictions by June 1st enable the university to efficiently plan for teaching capacity, allocate classrooms and can help with estimating expected tuition revenue. When predicting individual admission conversion probabilities, it is essential to guard against overfitting, as enrolment patterns can vary from year to year. Identifying and analysing relevant predictive factors is an important aspect of this project, potentially offering valuable insights both for the university's current methods and for improving the predictive models developed during this project.

6.2 Data understanding

The University of Amsterdam has provided data from six distinct academic programs within the Faculty of Economics and Business. This includes training data covering the 2024/2025 academic year, as well as test data intended for predicting the number of students expected to enrol in the upcoming academic year 2025/2026, which is not yet known. The data is provided from two main sources: Studielink records and internal university data, both comprising multiple files. Studielink is the official online application and registration portal for higher education in the Netherlands. It is used by both Dutch and international students to apply for bachelor's and master's programmes at universities and universities of applied sciences. All datasets have been pseudonymised by removing sensitive information and assigning artificial student ID numbers to ensure the privacy of students. The table on the next page provides an overview of the datasets received from the various sources.

Name	**Source**	**Years**	**Description**
StudentSL	Studielink	24/25, 25/26	General data about students that applied (nationality, gender, etc.)
AdmissionSL	Studielink	24/25, 25/26	Data about the admission and the process (application date, application status, etc.)
Pre-educationSL	Studielink	24/25, 25/26	Data of students' prior education
SleutelSL	Studielink	24/25, 25/26	Aditional student specific data (former enrolments, new to higher education, etc.), updated weekly
TelSL	Studielink	18/19 - 25/26	Cumulative admission data, not student specific, updated weekly
SISdata	UvA	24/25, 25/26	Admission and student data as known within the university
UvAweek-to-week	UvA	24/25, 25/26	Cumulative admission data, not student specific

All of the 2024 datasets comprised static snapshots of the data, including the files that were updated weekly for the next academic year, which presented certain challenges for modelling. Specifically, all time dependent variables are already known, but the timing of changes within those variables is unknown, as no timestamps were provided apart from the date of admission. This limited the ability to reconstruct a timeline of events or to analyse the progression of the admission process over time.

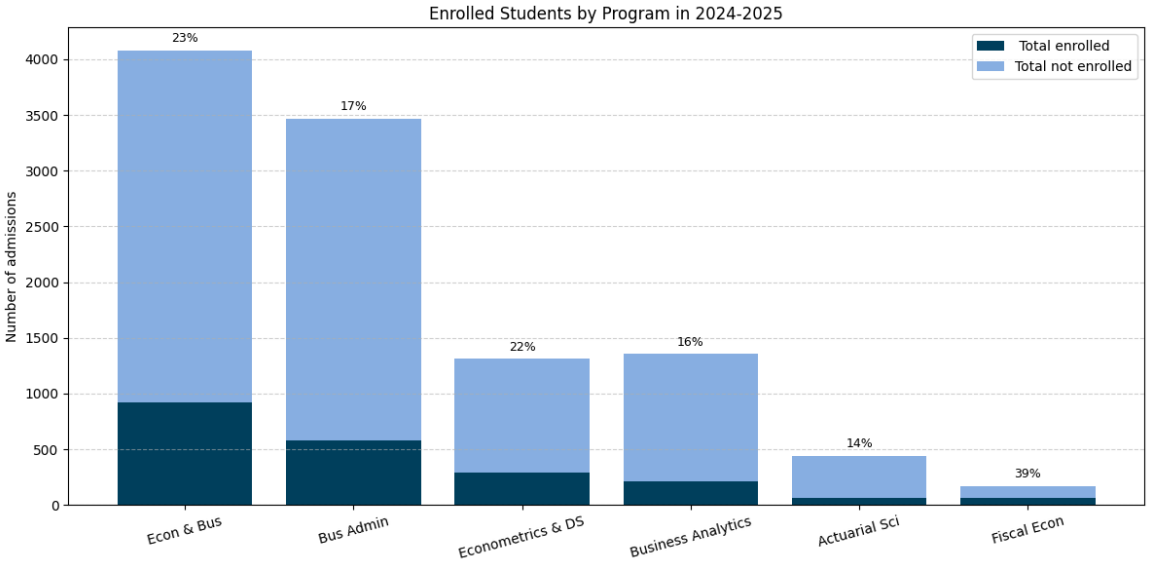
The datasets contain a wide variety of variables. These include student specific attributes such as nationality, gender, and age. Additionally, there is information about prior education, including the country and level of the student's previous studies. The data also includes variables related to the admission process, such as admission date, admission status and, results from English language proficiency tests. The table below shows a general overview of data variables available.

Variable Name	**Possible Values**	**Description**
Student Number	8-digit numeric code	Unique number per student
Gender	M, F, or Other	Gender of student
DEFT	DEFT or missing	Definitely enrolled
DSAC	DSAC or missing	Definitely not enrolled
Birthdate	Date in DD/MM/YYYY format	Birthdate of student
Nationality	e.g., Dutch, German, etc.	Nationality of a student
Pre-EducationLevel	e.g., VWO, BSc, etc.	The level of pre-education
Pre-EducationCountry	e.g., Italy, Netherlands, etc.	The location of pre-education
Program	e.g., Business Analytics, etc.	Academic program of admission
Year	e.g., 24/25, 25/26, etc	The academic year of the program
Taaltoets	G or missing	English language proficiency test (passed "G" or missing)

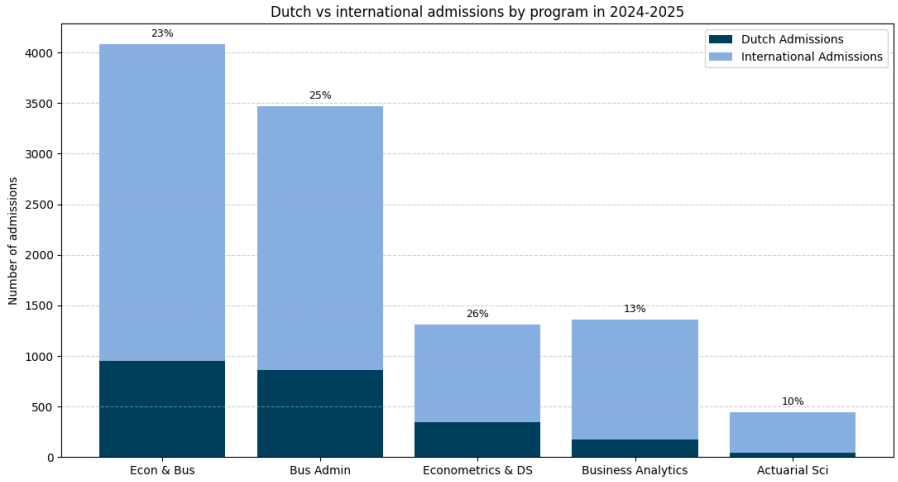
As a follow-up to the table, "DEFT" is the target variable in this project. The models developed aim to predict the chance that an admission will result in a DEFT outcome, meaning the student has officially been approved and is expected to start the program. The opposite outcome is labelled "DSAC," which indicates that the admission was either cancelled by the student or rejected by the university. DSAC outcomes can occur for various reasons, such as failure to meet admission requirements or not completing the application process on time.

One variable in this process is the "Taaltoets", which is the Dutch term for the English language proficiency test. This test is mandatory for international students to demonstrate sufficient language skills. Exceptions apply to students who have earned a diploma within the Dutch education system, students from most native English speaking countries and a few specific cases outlined by university policy. For those required to take the test, results can be either pass or fail. Failing to meet the required English level or not taking the test at all will result in the admission being rejected.

It is interesting to see the number of applications the Faculty of Economics and Business bachelor programs receive annually, alongside relatively low conversion rates, as shown in the figure below for the 2024–2025 academic year.



This highlights a significant gap between total admissions and final enrolment. It shows that the majority of admitted students ultimately did not enrol, which further illustrates why the level of uncertainty for the university is so high. With so many admissions not converting, it becomes difficult to make estimates. Also notable is that the majority of admissions come from international students, as illustrated in the chart on the next page.



There is a clear difference between the number of Dutch and international admissions within the faculty. These are two distinct groups of students with different behaviour, and this should be taken into account when developing predictive models.

6.3 Data preparation

In the data preparation phase, several cleaning and preprocessing steps were performed to ensure data quality and consistency across data sources. One key issue addressed was the presence of duplicate applications for the same student and program. These duplicates can arise from multiple scenarios. For example, a student may have cancelled an application and later reapplied, or an application might have been rejected due to incomplete documentation, after which the student reapplied with a corrected version. In such cases, multiple records for the same individual appeared in the dataset. The most recent admission was kept.

Additionally, students whose applications were cancelled or rejected (both result in admission to become DSAC) before the program's application deadline, which is May first for most bachelor's programs, were excluded from both the training and testing datasets. Since these students clearly did not intend to enrol, their inclusion would add noise and little predictive value to the model, as predictions are performed after the deadline on June first.

Due to the existence of multiple data sources, Studielink and internal UvA systems, it was necessary to merge datasets to construct a complete profile for each application. Each dataset contained different variables relevant to the admissions process, so combining them into a single dataset improved both interpretability and helped feature engineering. As Studielink and UvA data follow different data structures, adjustments were required during the merging process. By using student ID numbers, program codes, and additional linking features, all relevant data for each student admission were successfully integrated into a single dataset.

Missing values were left intentionally unfilled, as their absence often carries meaningful information. In this project, missing data typically indicates that a step in the application process had not (yet) been completed. For example, an empty field in the language proficiency test column means that the student had not taken the test. These cases were accounted for during feature engineering, as these missing values can serve as predictive indicators.

To prepare the data for modelling, non-numeric variables were transformed into dummy variables, and some related features were combined into new, more informative variables. This step aimed to reduce dimensionality and sparsity while increasing the dataset's predictive power. A summary of the most important created variables is provided with more explanations in the list below.

Variable Explanations:

- **English language proficiency test**:
  - Based on admission requirements, international students are generally required to complete an English language proficiency test. Exceptions include students who have obtained a diploma within the Dutch education system or in a native English-speaking country. Using this information, a dummy variable was created: *LanguageTest* (1 = test required but not passed or not yet completed, 0 = not required or already passed). This variable serves as a filter for international students and helps the model account for incomplete or failed language requirements.
- **Number of admissions within the UvA**:
  - By counting the number of unique applications submitted by a student to any program within the University of Amsterdam, a new variable was created to represent the total number of UvA admissions per student. The reasoning behind this feature is that students who apply to multiple programs are likely still undecided and will ultimately choose only one. Therefore, a higher number of applications within the UvA is expected to negatively correlate with the likelihood of conversion for any program.
- **Admissions outside of the UvA**:
  - While it is not straightforward to precisely count the number of admissions outside the UvA, it is possible to see whether a student has applied elsewhere within the Netherlands by using the Studielink data. Specifically, the total number of admissions listed in Studielink can be compared to the number of UvA admissions. By subtracting the UvA applications from the total, a derived binary variable was created: *OutsideUvA* (1 = student has at least one application outside the UvA, 0 = UvA only).
  - It is important to note that this method does not reveal the admission status at the other institutions, nor does it capture applications outside the Netherlands. However, the assumption remains that students with applications to other universities are still weighing their options, which may negatively impact the likelihood of enrolment at the UvA.
- **Age of student**
  - The age of each applicant is calculated using their date of birth. To ensure consistency in age comparisons, a fixed reference date is used: September 1st of the academic year for which the student applied.

  - Including age as a variable allows the model to account for potential differences in enrolment behaviour across age groups. For example, older students may be more certain about their study choice or have different motivations compared to younger applicants, which could influence their likelihood of enrolment.

- **Type of student**:
  - Using a combination of student nationality, place of prior education, and admission data from Studielink, students were categorised into distinct groups. Initially, students were grouped based on nationality into region of origin: Dutch, EEA (European Economic Area), and rest of world (not Dutch and not EEA). These groupings are largely driven by differences in tuition fees and policy regulations.
  - Additionally, students were classified based on whether they had previously obtained a diploma within the Netherlands, which can indicate familiarity with the Dutch culture and the Dutch education system.
  - By combining these characteristics with past admissions data, the following student types were identified:
    * **VWO:** Dutch high school students applying with a VWO diploma.
    * **BachelorSwap:** Students switching from a different bachelor's program within the University of Amsterdam.
    * **UniversitySwap:** Students transferring from another Dutch university, either with or without a degree.
    * **NewInNL:** Students from any country who are new to the Dutch education system this can include students with a Dutch nationality but no prior diploma from a Dutch institution, often due to dual nationality.
    * **OnCampus** OnCampus is a unique pre-education program designed for international students who wish to experience studying in the Netherlands. Upon successful completion of the program, students are offered the opportunity to progress into one of two Bachelor's programs: Economics and Business Economics, or Business Administration.
  - These classifications aim to capture differences in background and enrolment behaviour between distinct groups that could influence the likelihood of conversion.

A short summary of the created variables can be found in the table below:

Variable Name	**Possible Values**	**Description**
LanguageTest	{0, 1}	1 if test required but not passed or not yet completed, 0 if not required or already passed
AdmissionsUvA	Positive integers	Number of admissions within UvA
OutsideUvA	{0, 1}	1 if student has at least one application outside the UvA, 0 if UvA only
Age	Positive integers	Age of student
StudentTypes	{0, 1}	1 if part of the student type group, 0 otherwise

6.4 Modeling

In this thesis, various machine learning methods were explored to predict whether an admission would result in an enrolment. These included more complex models such as Random Forest, XGBoost and Support Vector Machines, as well as more interpretable approaches like Decision Trees and Logistic Regression.

After evaluating the performance of the models with a focus on predictive accuracy, logistic regression clearly stood out. Unlike the other models, which struggled to accurately forecast the total number of expected students in the early stages, logistic regression provided reliable estimates much earlier in the process. While some of the other models improved over time, they continued to significantly under-predict enrolment.

The random forest model was the second-best performing model for making predictions. However, its performance was not strong; for example, in the Economics and Business Economics program, it still underestimated enrolment by over 80 students. Given that this program typically has an intake of around 900 students, this difference is quite substantial. In contrast, the logistic regression model came much closer, predicting within 20 students of the actual total. The remaining models underestimated enrolment even more than the random forest model. Complete results of the early models are provided in the appendix.

The combination of strong early performance and the transparency of the model, meaning that the influence of individual features on the prediction can be clearly understood, made logistic regression the best candidate for further development.

Logistic regression

Logistic Regression is a widely used statistical method for binary classification problems. In this context, it estimates the probability that a student admission will result in an enrolment, producing a value between 0 and 1. It is particularly well-suited for problems where the objective is to predict the likelihood of a specific event (outcome), such as enrolment, based on a set of input features (Harrell, 2015).

Logistic regression handles both categorical and numerical predictors effectively. Its efficiency and interpretability make it a reliable choice for classification tasks such as the one in this thesis, and it has shown potential when using student-specific data (Lust, 2024; Slim et al., 2018). The logistic regression model created is shown on the next page. Note that the variable OnCampus (X_8) applies only to students enrolled in the OnCampus program, which offers progression exclusively into the Bachelor's programs Economics and Business Economics or Business Administration. Therefore, when running the model for any of the other programs, this variable is omitted from the equation. Additionally, each student belongs to only one student type category. To avoid multicollinearity, the dummy variable for NewInNL is omitted and serves as the reference category in the model.

$$\log\left(\frac{P(\text{enrolled}=1)}{1-P(\text{enrolled}=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$

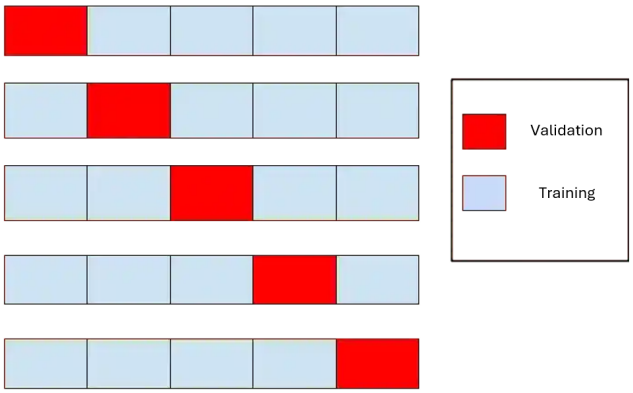
where:

$$X_1 = \text{LanguageTest}$$
$$X_2 = \text{AdmissionsUvA}$$
$$X_3 = \text{OutsideUvA}$$
$$X_4 = \text{VWO}$$
$$X_5 = \text{BachelorSwap}$$
$$X_6 = \text{UniversitySwap}$$
$$X_7 = \text{Age}$$
$$X_8 = \text{OnCampus}$$

K-Fold cross validation

To reduce the risk of overfitting, which is particularly important given that the training data is limited to a single academic year (2024 - 2025), K-Fold Cross-Validation was applied. Overfitting occurs when a model learns patterns that are specific to the training data but do not generalise well to new data. This is a concern in this context, as student behaviour may vary slightly across academic years.

In this project, 5-fold cross validation was used. The data was divided into five equal sized subsets, or folds. The model was trained five times, each time using four folds for training and the remaining fold for validation. This approach ensures that every observation is used once for validation and four times for training. As a result, the model is evaluated on unseen data in each iteration, which helps reduce variance and provides a more robust estimate of performance. The visual representation of this process is provided in the figure below.



Each of the five trained models was then used to generate predictions for the upcoming academic year (2025 - 2026), producing five predicted probabilities for each student admission to convert into an enrolment.

Ensembling

To combine the outputs from the five logistic regression models, ensembling was used. Specifically, the five predicted probabilities for each admission were averaged to produce a final probability score. This method helps smooth out variations that might be learned by individual models, resulting in more stable and reliable predictions.

Instead of applying a fixed binary threshold (e.g. 0.5) to classify admissions as enrolments or not, the predictions were aggregated using a Poisson Binomial distribution. This approach considers each admission as an independent Bernoulli trial, with its own probability of success (enrolment). Summing the individual probabilities provides an expected value for the total number of students likely to enrol. The formula for this is given below:

$$\text{Total Number of expected students} = \sum_{i=1}^{n} \bar{p_i}$$

Where:
n = Total number of admissions within the tested program
$\bar{p_i}$ = Average of the predicted probabilities for the i-th admission to convert into enrolment

This technique keeps the uncertainty present in student decisions. Some students may appear unlikely to enrol but do everything at the last minute, while others with seemingly complete applications might withdraw unexpectedly. By avoiding a strict threshold, the model maintains nuance in its predictions and reflects real-world outcomes. Every admission still has a chance to convert or withdraw, rather than being classified as a definite enrolment simply because the probability exceeds the threshold.

Confidence interval

The sum of the predicted probabilities across all admissions produces a non-integer expected value representing the total number of students likely to enrol in a given program. Rounding this value provides a straightforward estimate of expected enrolments. However, due to uncertainties such as unpredictable student decisions, this point estimate alone does not fully capture the range of possible outcomes.

Recognizing this uncertainty, the University currently reports a range of expected students rather than a single number. To align with this a confidence interval is constructed around the model's expected value. This interval accounts for the variance in predictions caused by both model limitations and the student enrolment behaviour. By providing a 80% confidence interval, the model offers a probabilistic range of expected enrolments, allowing stakeholders to consider best-case and worst-case scenarios when allocating resources such as classrooms and staff. This approach improves the utility of the predictions by quantifying uncertainty rather than ignoring it.

While a 95% confidence interval is more traditional, it is not used here due to the high variance caused by the uncertainty in predicting student decisions, such as sudden cancellations or late completion of the admission process and thus enrolment. Using a 95% confidence interval would result in a significantly wider prediction range for this model compared to the range sizes currently used by the university, which would reduce the practical usefulness and make logistical planning more challenging. To keep the prediction range comparable to current methods, the 80% confidence interval is chosen. The formula used to calculate the confidence interval is provided below.

$$\sum_{i=1}^{n} \bar{p_i} \;\pm\; 1.2816 \cdot \sqrt{\sum_{i=1}^{n} \bar{p_i}(1-\bar{p_i})}$$

Where:
n = Total number of admissions within the tested program
$\bar{p_i}$ = Average of the predicted probabilities for the i-th admission to convert into enrolment

Final model

To summarise, the final model can be divided into 5 steps:

1. Applying 5-fold cross-validation on the 2024–2025 training data to train and evaluate five logistic regression models.
2. Using each of these models to predict enrolment probabilities for the 2025–2026 admissions data.
3. Averaging the five probabilities per admission to obtain the final predicted chance of enrolment.
4. Summing all predicted probabilities using the Poisson Binomial distribution to estimate the expected number of students per program.
5. Calculating a confidence interval around the predicted number of students.

6.5 Evaluation

To evaluate the effectiveness of the prediction models, a less common evaluation method is required due to the absence of a binary classification threshold. Instead of using traditional binary metrics such as accuracy, precision, recall, or F1-score, the model's performance is assessed based on how closely the predicted range of expected students aligns with the actual number of enrolled students. If the model is effective, the actual number of students should fall within the predicted confidence interval.

In addition, the Mean Absolute Error (MAE) is chosen as the primary metric to measure the prediction accuracy, as it is both intuitive and interpretable. MAE represents the average absolute difference between the predicted and actual number of enrolled students, providing a straightforward indication of how close the model's forecasts are to the real outcomes. Unlike other error metrics, MAE is not overly sensitive to outliers, making it particularly useful in evaluating enrolment predictions where understanding the size of the deviation is important.

The formula for the MAE is:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |p_i - y_i|$$

Where:
n = total number of admissions within the tested program
p_i = predicted probability for the i-th admission (obtained using the logistic regression model)
y_i = actual class label for the i-th admission, 1 if enrolled, 0 otherwise ($y_i \in [0, 1]$)

This approach provides a clear understanding of the model's deviation from the real outcomes without relying on classification thresholds. However, the MAE can only be calculated for the logistic regression model, as the probabilities of each admission converting into enrolment are needed. The university's method, however, does not work with conversion probabilities, so the MAE can only be used to assess the performance of the created logistic regression models. To enable comparison with the current methods, the total absolute error (AE) will also be applied. This offers an easy comparison between the predicted total number of students and the actual results.

The formula for the total AE is:

$$\text{Total AE} = \left| \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} y_i \right|$$

Where:
$\sum_{i=1}^{n} p_i$ = total number of predicted students
$\sum_{i=1}^{n} y_i$ = total number of students that actually enrolled

It is important to note that, at the time of writing this thesis, the actual enrolment numbers for the academic year 2025–2026 were not yet available. These will only be known at the start of the academic year on September 1st 2025. Therefore, the model evaluation in this thesis is mainly based on the known results from the academic year 2024–2025.

2024-2025 results

Using the data from the last academic year (2024–2025) to validate the models, cross-validation was employed in a way similar to the approach used for predicting next year's enrolment. However, in this case, cross-validation is particularly crucial, as only one year of data is available. Since the training data also serves as the testing data, there is a significant risk of overfitting.

The exact same method described in the modelling section was applied: logistic regression with 5-fold cross-validation. In this process, instead of leaving one subset out as validation, it was used as the testing set. By iterating through all five folds, each admission was tested exactly once, generating a probability of enrolment per admission. Summing these probabilities produces the overall prediction of expected students for the academic year 2024–2025. As mentioned in the modelling section, an 80% confidence interval was applied to estimate the expected range of students, similar to the approach used in the current methods.

The model's results when applied to the 2024–2025 data are shown in the table below.

Program	**LR predictions**	**UvA predictions**	**Actual number**
Economics and Business Econ	896 – 925	900 – 950	921
Business Administration	**568 – 592**	**600 – 650**	581
Econometrics and Data Science	277 – 296	250 – 300	293
Business Analytics	200 – 215	180 – 220	215
Actuarial Science	54 – 62	50 – 60	60
Fiscal Economics	**60 – 68**	**50 – 60**	66

Two key observations stand out from the results. First, the predictions generated by the logistic regression models developed in this thesis are close to those of the current UvA prediction methods across all programs. Moreover, when using the 80% confidence interval, the prediction ranges produced by both methods are comparable in size. Second, all actual enrolment numbers fall within the predicted range of the created model, whereas the UvA's predictions are slightly off for the two programs Business Administration and Fiscal Economics. The table below shows the results of the mean absolute error and the total absolute error calculations, with the total absolute error calculated from the midpoints of the ranges (the predictions).

Program	**MAE LR**	**Total AE LR**	**Total AE UvA**
Economics and Business Econ	0.3063	11	4
Business administration	0.2548	1	44
Econometrics and data science	0.2966	7	18
Business Analytics	0.2296	8	15
Actuarial science	0.2064	2	5
Fiscal economics	0.3830	2	11

The mean absolute error can be interpreted as the average absolute difference between the predicted probabilities and the actual outcomes. For example, in the context of Economics and Business Economics, an MAE of approximately 0.31 means that, on average, the model's predicted probability differs from the actual outcome by 0.31. To illustrate: if a student actually enrolled (true outcome = 1), the model might predict a probability of 0.69, resulting in an absolute error of 0.31. This value reflects the model's predictive accuracy.

The closer the predicted probabilities are to 0.5, the less confident the model is. If the model often predicts probabilities near 0.5 while the actual outcomes are 0 or 1, the MAE will tend to be higher. Conversely, if the model makes confident predictions closer to 0 or 1 and they are correct, the MAE will be lower. A MAE of 0.31 indicates moderate prediction accuracy but also shows that predictions are less accurate on average at the individual admission level. This further highlights why no fixed threshold is used and why predicted probabilities are kept as they are when making predictions.

Comparing the total absolute error of the logistic regression model with the current method shows how far the predicted total deviates from the actual number of students, relative to the midpoint of the prediction range. As also observed from the prediction intervals, both methods estimate the total number of expected students per program quite accurately. An exception is Business Administration

in the university's method, where the total absolute error is 44 students and the actual enrolment falls outside the predicted range by about 19 students. The logistic regression model seems to outperform the university's method for this program.

Overall, these findings demonstrate the potential of the logistic regression model, at least when evaluated through cross-validation on the available training data. They also suggest that the current UvA methods are already quite effective, with most real results falling within their predicted ranges.

2025-2026 results

For next year's predictions, the logistic regression model described in the modelling section was used, applying the same 80% confidence interval as before to produce prediction ranges comparable in size to the current methods. The results are presented in the table below. As mentioned earlier, the actual enrolment numbers for the upcoming academic year are not yet available, so these predictions cannot be evaluated at this time. However, this does provide an opportunity to compare the model's forecasts with those made by the UvA.

Program	**LR predictions**	**UvA predictions**	**Actual number**
Economics and Business Econ	941 – 993	900 – 950	–
Business Administration	**607 – 650**	**500 – 550**	–
Econometrics and Data Science	288 – 317	275 – 325	–
Business Analytics	197 – 222	200 – 220	–
Actuarial Science	58 – 72	50 – 60	–
Fiscal Economics	53 – 64	50 – 70	–

Again, it can be concluded that overall the predictions from the logistic regression models are quite similar and comparable to those of the current UvA methods. However, the logistic regression model predicts a slightly higher number of students for Economics and Business Economics, as well as Actuarial Science.

The most significant difference appears for Business Administration, where the two methods differ significantly. The logistic regression model forecasts an increase in student numbers (compared to last year's 581 students), while the UvA predicts a decrease. It is particularly interesting that both models, making predictions around June 1st, produce such different outcomes, with a gap of about 50 students between the forecasts. At this stage, it is impossible to determine which prediction is more accurate. It is clear that one of the predictions will be incorrect, though there remains the possibility that actual enrolment could fall outside both predicted ranges, potentially within the gap between them.

Overall, predictions for most programs align closely with the current UvA forecasts, but these results cannot yet be fully evaluated until the actual enrolment numbers become available in September 2025. Once the real numbers are known, the model's performance will become clearer.

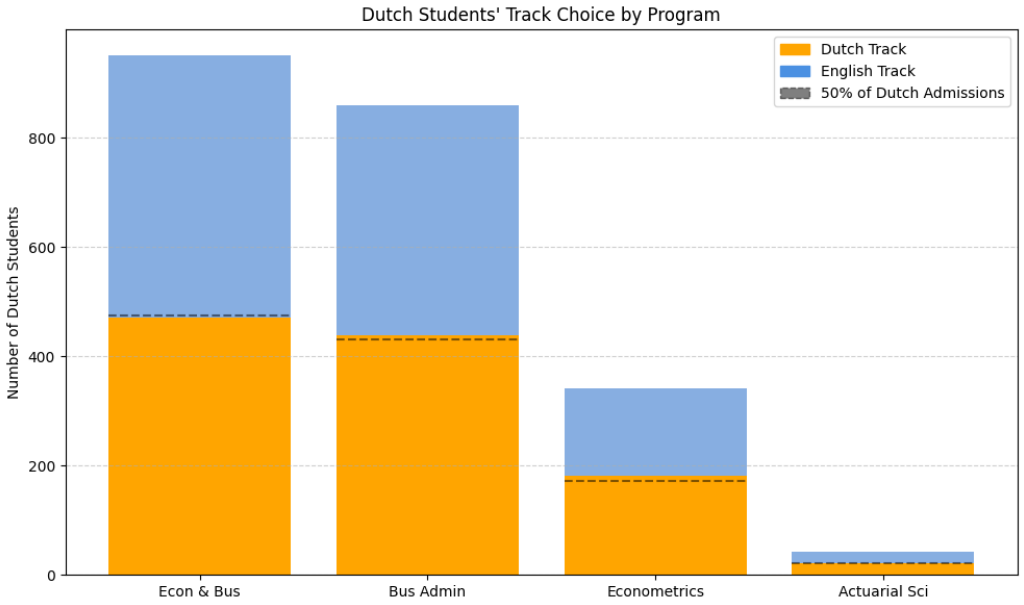
Policy impact

Due to recent policy changes introduced by the Dutch government, universities are now required to expand their capacity for Dutch language education. This initiative aims to strengthen the Dutch language education in the Netherlands. As a result, the Faculty of Economics and Business had to introduce Dutch language tracks within several Bachelor's programs that were previously offered exclusively in English.

Of the six Bachelor's programs offered by the faculty, namely Economics and Business Economics, Business Administration, Econometrics and Data Science, Business Analytics, Actuarial Science, and Fiscal Economics, five were originally taught only in English. Fiscal Economics was already taught in Dutch only. Following the new policy, four of the English-only programs (Economics and Business Economics, Business Administration, Econometrics and Data Science, and Actuarial Science) have now been split into English and Dutch tracks. Business Analytics is excluded from this change due to specific exceptions in the policy.

Dutch students now have the option to choose between the Dutch or English track when applying to these programs. It is interesting to analyse how Dutch students are responding to these newly introduced tracks and whether they show a preference for the new Dutch taught programs over the English ones.

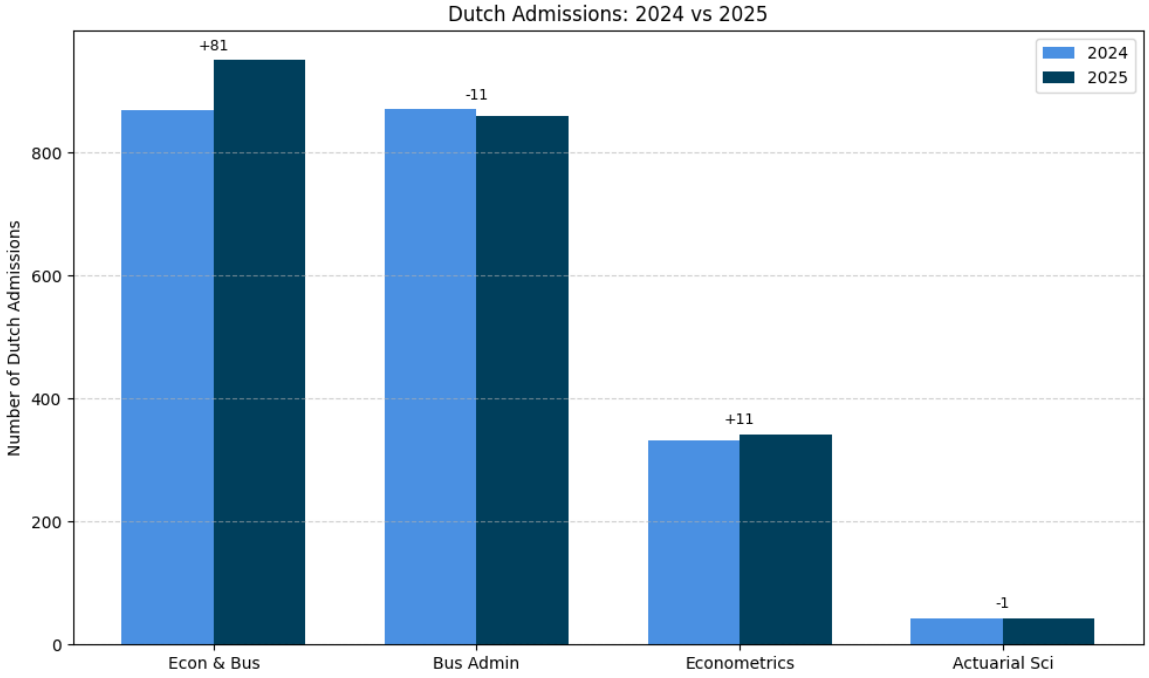
The figure below shows the distribution of Dutch student admissions for the four programs affected by the policy. Each bar is divided into two parts. The orange segment represents Dutch students who applied for the Dutch track, and the blue segment represents those who chose the English track. A dashed line at the 50 percent mark is included in each bar to make comparisons easier. A table is also included to provide a more detailed overview of Dutch admissions and their corresponding track choices.



Program	Dutch track	English track	Total Dutch Admissions
Economics and Business Econ	470	480	950
Business Administration	439	421	860
Econometrics and Data Science	182	160	342
Actuarial Science	22	20	42

There appears to be no strong preference among Dutch students for either the Dutch or English tracks. In all four programs, the distribution between the two options is roughly even. This suggests that Dutch students are nearly equally divided in their choices following the policy change.

The next question is whether the introduction of Dutch language tracks has resulted in a higher number of Dutch student admissions overall. The answer is not entirely clear. The chart below compares the number of Dutch admissions in 2024 with those in 2025. There is a noticeable increase for Economics and Business Economics, which may indicate that this program has become more attractive to Dutch students. For the other programs, however, no clear increase is visible. Dutch admissions have remained relatively stable compared to the previous year.

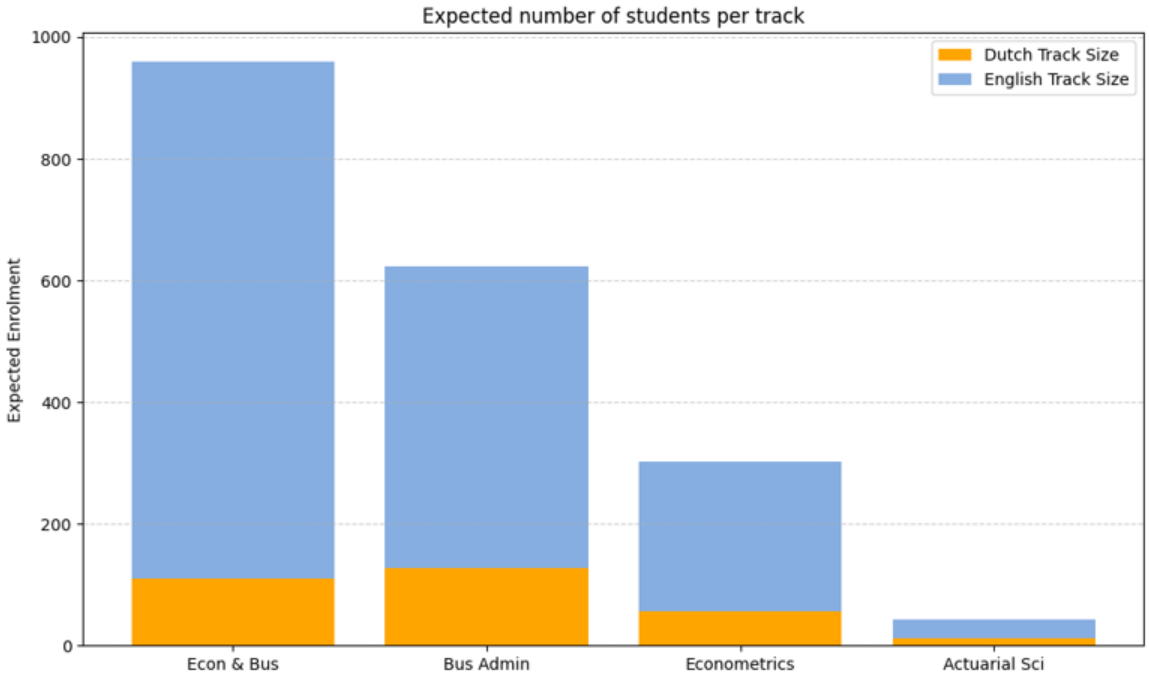


Because this analysis compares only two years, drawing firm conclusions is difficult. While there are early signs that the Dutch track may have increased interest in Economics and Business Economics, the evidence is not consistent across all programs. Additional data in coming years will be needed to determine whether this policy change has a lasting impact on Dutch student admissions for this program.

No clear conclusion can be drawn about whether the policy attracts more Dutch students overall. However, it may affect the programs in other ways. One way to explore this is by examining the expected size difference between the Dutch and English tracks within each program.

To do this, the logistic regression results are used again, but instead of summing the total probabilities for all admissions within a program, only the probabilities of Dutch students who selected the newly introduced Dutch tracks are summed. This provides an estimate of the expected number of students starting in the Dutch track. To compare this with the expected size of the English track, the same method is applied, summing the predicted probabilities of all admissions (both Dutch and international students) who selected the English track.

The figure below shows these results. Each bar represents the total number of students expected to start in one of the four Bachelor's programs. The orange segment of each bar indicates the expected number of students in the Dutch track, while the blue segment shows the expected number of students in the English track. A table is also included to provide a more detailed overview of the ranges predicted.



Program	**Dutch track size**	**English track size**
Economics and Business Econ	101 – 119	824 – 873
Business Administration	119 – 137	476 – 516
Econometrics and Data Science	51 – 63	231 – 258
Actuarial Science	5 – 9	52 – 65

A clear difference is visible: the Dutch tracks are expected to be significantly smaller than the English tracks. This is understandable, as only Dutch students can follow the Dutch tracks, while the English tracks are open to both international students and Dutch students who choose the English option. The policy essentially resulted in the creation of four new smaller programs, which the university now needs to accommodate with teaching space and Dutch speaking staff.

Feature Importance

Understanding which features influence predictions can provide insight into which input variables the model relies on most. This improves interpretability and can be important for future research. Interpreting logistic regression coefficients directly can be challenging, as they reflect changes in log-odds rather than direct changes in probability. To improve interpretability and assess the relative importance of features, Average Marginal Effects (AMEs) were calculated after each model estimation. For each model, the vector of marginal effects was stored, and a final summary matrix was constructed by averaging the effects across the five different models created through cross-validation for each of the programs. This matrix represents the average change in predicted probability of the outcome associated with a one-unit change in each predictor, holding other variables constant.

For binary variables (VWO, LanguageTest, etc.), the AME reflects the change in probability when the variable switches from 0 to 1. For continuous variables (AdmissionsUvA and Age), it reflects the impact of a one-unit increase. This approach allows for an interpretable ranking of feature importance, accounting for variation across folds while avoiding dependence on a single model. The results for each program can be found in the table below, with an average column added to highlight the most influential features across all programs.

Program	LanguageTest	AdmissionsUvA	OutsideUvA	VWO	BachelorSwap	UniversitySwap	Age	Foundation
Economics and Business Econ	-0.264	-0.029	-0.002	0.348	0.279	0.137	0.003	0.145
Business Administration	-0.244	-0.002	-0.057	0.363	0.224	0.278	-0.009	0.099
Econometrics and Data Science	-0.186	-0.178	-0.001	0.349	0.268	0.157	-0.001	–
Business Analytics	-0.261	-0.020	-0.072	0.390	0.090	0.910	-0.006	–
Actuarial Science	-0.167	-0.029	-0.066	0.319	0.241	0.162	-0.014	–
Fiscal Economics	–	-0.188	-0.225	0.189	0.234	0.154	-0.015	–
Average	-0.224	-0.074	-0.071	0.326	0.222	0.300	-0.006	0.122

For Fiscal Economics, the LanguageTest variable is omitted since it is a Dutch-only program.

It can be seen that, on average, Dutch students with a VWO diploma are an influential factor for predicting enrolment, as are students applying after previously studying at another university within the Netherlands. On the other hand, not passing the language proficiency test appears to be a strong indicator of non-enrolment, as expected. For the continuous variables, a higher number of admissions within the UvA has a notable effect. As the count of admissions increases, the predicted probability of enrolment decreases steadily, indicating that students with many applications elsewhere are less likely to enrol.

6.6 Deployment

The university's current prediction methods rely mainly on historical conversion data and personal experience. Insights about feature importance found using the logistic regression models could guide adjustments, such as increasing the weight for Dutch students with a VWO diploma or reducing it for those with multiple admissions. Language test results could also help filter international applicants. Incorporating these elements may improve prediction accuracy and can be considered by current employees.

The logistic regression models show potential to effectively predict student enrolment and could be deployed as a tool for current employees to use and compare against their predictions. However, the models have not yet been evaluated on new academic year data. If the model proves effective again for the 2025–2026 year, it could be considered not only as a comparison tool but also as a replacement for current methods. The combination of effectiveness and interpretability makes the logistic regression models a strong candidate to replace the current time-consuming manual process.

7 Results and discussion

The CRISP-DM methodology effectively structured the research by providing a clear, step-by-step framework to guide the project. Interestingly, the current prediction methods used by the UvA already perform quite well, despite relying only on conversion rates and experience rather than a machine learning model. This suggests that institutional experience and an understanding of yearly trends can result in accurate forecasts.

Data preparation proved effective for the logistic regression models. The inclusion of newly created variables allowed the models to generate predictions that showed promising performance on the 2024-2025 data. In particular, the combination of the language test as a admission process variable and the total number of admissions stood out as especially useful in identifying admissions with a lower likelihood of enrolling.

On the 2024-2025 data, the logistic regression model appears to outperform the existing faculty method, indicating potential for improved forecasting. However, since this evaluation is based on cross-validation using same year data, the model's future performance is not guaranteed. Student behaviour may vary in upcoming years, which introduces uncertainty. Nonetheless, the results suggest that logistic regression shows potential as a more data-driven alternative to the current approach.

Regarding the main objective, predicting the number of students for the upcoming academic year, no clear conclusions can yet be drawn. While the model's predictions align closely with those of the university for most programs, there is a significant difference for Business Administration. The logistic regression model predicts an increase in enrolment, whereas the UvA expects a decrease. At least one of these predictions will be incorrect, but this wont be known until final enrolment numbers become available in September 2025.

When inspecting the Dutch policy impact of the introduction of Dutch tracks in four previously English-only bachelor programs, analysis shows that Dutch students are evenly divided in their track choices, with approximately 50% picking the Dutch version and the other half choosing the English version. This suggests no strong overall preference between tracks for Dutch students.

Furthermore, there is no clear evidence that the policy change has led to a faculty wide increase in Dutch student admissions. While Economics and Business Economics shows a rise in Dutch admissions, the other programs do not show a similar trend. This may suggest growing interest in that specific program, but with only a single year of data, no definitive conclusions can be drawn. It is possible that the observed increase would have occurred regardless of the track changes.

Finally, the newly established Dutch tracks are expected to be significantly smaller in size than the English tracks. From the university's perspective, this may be less beneficial. The UvA needs to accommodate each of the smaller new Dutch tracks by providing additional teaching spaces and Dutch speaking staff who can teach and grade in Dutch. Previously, English speaking staff would have sufficed. These adjustments are required for a relatively small group of students, meaning the return is lower. Additionally, the pool of qualified Dutch speaking teachers is smaller compared to the broader international teaching staff pool, making it more difficult to find the necessary personnel.

8 Product

The deliverable of this project is divided into 2 different components:

1. **Final predictions** - The table containing the predictions for the next academic year 2025/2026, which can be used by the university to compare to their current methods.
2. **GitHub for code** - The code created during the project in user-friendly code files for STATA, all in one GitHub repository. The repository for this thesis can be accessed through the following URL: https://github.com/Roellust/Predicting-Bachelor-Enrolment

9 Conclusion

The primary objective of this thesis was to accurately predict bachelor student enrolment at the University of Amsterdam's Faculty of Economics and Business using logistic regression, as an alternative or complement to the university's current manual forecasting methods. By leveraging individual application-level features, the model aimed to estimate enrolment conversion probabilities and aggregate these into program-level forecasts.

The results indicate that the logistic regression model performs comparably to the university's existing estimation approach on historical (training) data, and in some instances, even outperforms it. For enrolment predictions for the 2025–2026 academic year, the predictions from both methods were closely aligned for most programs. However, there was a notable difference in the Business Administration program: the logistic regression model predicted higher enrolment numbers and an increase compared to the previous year, whereas staff predict a decline. At present, it is not possible to determine which

method is more accurate; this can only be evaluated once actual enrolment numbers become available after the academic year begins in September 2025.

In addition to forecasting, the thesis also explored the early effects of recent Dutch language education policy changes, specifically the introduction of Dutch-language tracks in several bachelor programs. The findings suggest that Dutch students do not show a clear preference between Dutch and English tracks, and the new Dutch-language options have not led to a significant increase in total Dutch enrolments for the faculty. These tracks, however, are expected to be significantly smaller in size compared to the English tracks. From a practical standpoint, this raises concerns: creating Dutch-language tracks requires additional resources for a comparatively small student population, resulting in limited return and posing new challenges for university resource planning that, without the policy change, would not have been there.

In conclusion, the logistic regression models developed in this study demonstrate real potential to enhance the university's enrolment forecasting process by delivering timely and accurate predictions. It could significantly reduce the time and effort required from staff currently relying on manual methods, offering them a data-driven tool to support or validate their forecasts. However, further validation across multiple academic years is needed to ensure the model's robustness and accuracy on unseen academic year data. While it cannot fully replace manual judgement before further validation, the model provides a valuable comparison that staff can use to check and possibly refine their own predictions.

10 Limitations and future research

One of the main limitations of this research is the availability of only a single year (2024–2025) of training data. Relying on just one academic year increases the risk of overfitting, as the model may learn patterns specific to that year and incorrectly assume that student behaviour will remain the same in future years. Incorporating multiple years of historical data would likely improve the model's stability, reduce the risk of overfitting, and improve its predictive power by exposing it to a wider range of student behaviours and program specific trends.

Another challenge relates to the training data being a snapshot in time. While the training dataset provides a detailed overview of the academic year, it lacks time stamps for the various process steps students complete. In practice, this means that process variables are either completed or missing in the data, but the exact timing of these actions is unknown. This creates a mismatch when using the model on the upcoming academic year's data, which is a live dataset extracted at a specific moment in time (shortly before June 1st). Without knowing when a step was completed, it becomes difficult to properly incorporate process variables into predictive models.

Incorporating time stamps would significantly enhance the usability of process variables by allowing the model to consider not only if a step was completed but also when. This would enable dynamic modelling over time, allowing for rolling predictions as the admission process progresses. For example, early completion of key steps (like submitting documents or passing the language proficiency test) may indicate motivated students who are more likely to convert to enrolment. Conversely, delays or missing steps may be early indicators of non-enrolment. Currently, process variables like the language test

already provide predictive value, but only for international students. Introducing more time sensitive process features could substantially improve the model's predictive performance.

Additional promising process related variables that were not yet available during this project include whether a student has paid the required pre-enrolment fee (particularly relevant for international students), or whether they have submitted personal identification to verify their identity. Including such variables in future models could further increase prediction accuracy.

Beyond predictions, the FEB has expressed interest in developing a dashboard or interface to make predictions more accessible and actionable for planning and decision making. Currently, enrolment forecasts are made through several manual meetings per year, which is both time-intensive and costly. A model that can generate accurate forecasts quickly, with minimal manual input, could reduce the workload on staff significantly. If integrated into a dashboard, such a model could serve as a valuable decision support tool, especially around June 1st, but potentially even earlier in the year if a dynamic, time sensitive model is implemented.

In summary, expanding the training dataset, integrating process timing through time stamps or live data, and incorporating additional features could greatly improve the model. Additionally, operationalising the model through a user-friendly dashboard would support more efficient and data driven predictions at the faculty.

11 Acknowledgement

I would like to extend my gratitude to several individuals and institutions who have contributed to the completion of this thesis:

- I am grateful to Drs. F.H.K. Pope and the University of Amsterdam for providing the data and resources necessary for this project.
- I would also like to express my appreciation to Dr. S.T. Mol and Dr. I.M. Zwetsloot for their supervision and advice throughout the course of this thesis.
- A special thanks to Maarten Hoogeboom for exchanging ideas during the project.

12 References

1. Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive Models of Student College Commitment Decisions Using Machine Learning. Data, 4(2), 65. https://doi.org/10.3390/data4020065

2. Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide. https://api.semanticscholar.org/CorpusID:59777418

3. De Witte, K., Soncin, M., Vansteenkiste, S., & Sels, L. S. (2020). De economische effecten van internationalisering in het hoger onderwijs. Studie in opdracht van de 'Vlaamse Universiteiten en Hogescholen Raad'(VLUHR) en de Vlaamse Adviesraad voor Innoveren en Ondernemen (VARIO).

4. Hamers, Y. (2017). Predicting student enrollment. Tilburg University.

5. Harrell, F. E. (2015). Binary logistic regression. In Regression modeling strategies (pp. 219–274). Springer, Cham. https://doi.org/10.1007/978-3-319-19425-7_10

6. Hoogeboom, Maarten (2025). "Predicting Enrollment Numbers for Master's Programs in Economics and Business at the University of Amsterdam ". Master's Thesis. University of Amsterdam.

7. Joubert, A., & Fuller, J. (2025). Internationalization in balance. Linguistic Diversity in Professional Settings, 7, 19.

8. Lust, Roel (2024). "Predicting Binary BSA". Bachelor's Thesis. University of Amsterdam.https://scripties.uba.uva.nl/search?id=c11245756

9. Pace, F., D'Urso, G., Zappulla, C., & Pace, U. (2021). The relation between workload and personal well-being among university professors. Current Psychology, 40, 3417–3424. https://doi.org/10.1007/s12144-019-00294-x

10. Slim, A., Hush, D., Ojah, T., & Babbitt, T. (2018). Predicting Student Enrollment Based on Student and College Characteristics. International Educational Data Mining Society.

11. Universiteiten van Nederland. (2025, 15 april). Universiteiten gaan zelf internationalisering verder in balans brengen. https://www.universiteitenvannederland.nl/actueel/nieuws/universiteiten-gaan-zelf-internationalisering-verder-in-balans-brengen

12. University of Amsterdam. (2025, April 16). Universities make joint proposal on internationalisation. https://student.uva.nl/en/articles/2025-universities-make-joint-proposal-on-internationalisation

13. Vonk, S. (2022). Prediction of master student influx in the faculty of science (Master's thesis).

14. Wanjau, S. K., & Muketha, G. M. (2018). Improving student enrollment prediction using ensemble classifiers.

15. Weber, T., Van Mol, C., & Wolbers, M. H. (2024). Destination choices of international students in the Netherlands: A meso-level analysis of higher education institutions and cities. Population, Space and Place, 30(4), e2744.

13 Appendix

Program	Actual	LR	RF	SVM	DT	XG
Economics and Business Econ	921	902	839	812	778	1025
Business administration	581	567	499	471	446	782
Econometrics and Data Science	293	278	237	223	198	355
Business Analytics	215	201	160	158	139	316
Actuarial Science	60	51	39	33	28	112
Fiscal Economics	66	52	39	34	27	237

Logistic regression (in the early stage) outperforms other machine learning models.

Below are six examples of logistic regression output, showing the first cross-validation model for each program (foundation is another term for the OnCampus variable)

Economics and Business Economics:

```
Logistic regression                                     Number of obs =  3,253
                                                        LR chi2(8)    = 313.73
                                                        Prob > chi2   = 0.0000
Log likelihood = -1539.712                              Pseudo R2     = 0.0925
```

deft	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
study_switch	1.786606	.1946089	9.18	0.000	1.40518	2.168033
vwo	2.418995	.1907462	12.68	0.000	2.04514	2.792851
university_switcher	.9138702	.2369198	3.86	0.000	.4495159	1.378225
languagetest	-1.821378	.1498508	-12.15	0.000	-2.11508	-1.527676
AdmissionsUvA	-.1679309	.0553851	-3.03	0.002	-.2764837	-.0593781
OutsideUvA	.0619459	.114449	0.54	0.588	-.16237	.2862619
age	.0326992	.0248532	1.32	0.188	-.0160122	.0814106
foundation	1.03256	.1994259	5.18	0.000	.6416924	1.423427
_cons	-1.57349	.474894	-3.31	0.001	-2.504265	-.6427145

Business Administration:

```
Logistic regression                                     Number of obs =  2,796
                                                        LR chi2(8)    = 277.64
                                                        Prob > chi2   = 0.0000
Log likelihood = -1154.1877                             Pseudo R2     = 0.1074
```

deft	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
study_switch	1.892218	.2702959	7.00	0.000	1.362448	2.421988
vwo	2.8605	.2233067	12.81	0.000	2.422827	3.298174
university_switcher	2.279925	.2944486	7.74	0.000	1.702817	2.857034
languagetest	-1.956761	.20035	-9.77	0.000	-2.34944	-1.564082
AdmissionsUvA	-.0077154	.0664836	-0.12	0.908	-.1380209	.12259
OutsideUvA	.4209609	.141744	2.97	0.003	.1431478	.6987739
age	-.0835796	.036259	-2.31	0.021	-.154646	-.0125132
foundation	.780057	.2211399	3.53	0.000	.3466308	1.213483
_cons	-.1000561	.6864371	-0.15	0.884	-1.445448	1.245336

Econometrics and Data Science:

Logistic regression — Number of obs = 1,034; LR chi2(7) = 83.82; Prob > chi2 = 0.0000; Log likelihood = -481.27739; Pseudo R2 = 0.0801

deft	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
study_switch	1.717492	.3916228	4.39	0.000	.9499258	2.485059
vwo	2.267933	.2861446	7.93	0.000	1.7071	2.828766
university_switcher	1.053281	.3584514	2.94	0.003	.3507296	1.755833
languagetest	-1.163825	.241618	-4.82	0.000	-1.637388	-.6902629
AdmissionsUvA	-.1388004	.0862463	-1.61	0.108	-.3078401	.0302392
OutsideUvA	.0189247	.2165291	0.09	0.930	-.4054645	.443314
age	.0201341	.0342953	0.59	0.557	-.0470834	.0873515
_cons	-1.593873	.6695148	-2.38	0.017	-2.906098	-.2816481

Business Analytics:

Logistic regression — Number of obs = 1,077; LR chi2(7) = 117.11; Prob > chi2 = 0.0000; Log likelihood = -399.21071; Pseudo R2 = 0.1279

deft	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
study_switch	.5945997	.5886223	1.01	0.312	-.5590787	1.748278
vwo	3.496891	.4941674	7.08	0.000	2.528341	4.465442
university_switcher	.9556383	.5655239	1.69	0.091	-.1527682	2.064045
languagetest	-2.528709	.4199483	-6.02	0.000	-3.351792	-1.705625
AdmissionsUvA	-.1488018	.0949252	-1.57	0.117	-.3348518	.0372482
OutsideUvA	.9407888	.2541074	3.70	0.000	.4427474	1.43883
age	-.0954015	.0497052	-1.92	0.055	-.1928219	.002019
_cons	.4587181	.9501855	0.48	0.629	-1.403611	2.321047

Actuarial Science:

Logistic regression — Number of obs = 349; LR chi2(7) = 30.74; Prob > chi2 = 0.0001; Log likelihood = -114.89999; Pseudo R2 = 0.1180

deft	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
study_switch	2.512949	.8976603	2.80	0.005	.753567	4.272331
vwo	3.051577	.9195779	3.32	0.001	1.249238	4.853917
university_switcher	2.057308	.8721086	2.36	0.018	.3480061	3.766609
languagetest	-1.618659	.5905843	-2.74	0.006	-2.776183	-.4611353
AdmissionsUvA	-.2359474	.175768	-1.34	0.179	-.5804464	.1085515
OutsideUvA	.4594848	.5100387	0.90	0.368	-.5401726	1.459142
age	-.2052609	.1079184	-1.90	0.057	-.416777	.0062552
_cons	2.434311	2.062567	1.18	0.238	-1.608246	6.476868

Fiscal Economics, LanguageTest is omitted as it is a Dutch only program.

```
Logistic regression                                     Number of obs =    125
                                                        LR chi2(6)    =  35.08
                                                        Prob > chi2   = 0.0000
Log likelihood = -65.219094                             Pseudo R2     = 0.2119
```

deft	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
study_switch	1.743267	.9100037	1.92	0.055	-.0403075	3.526841
vwo	.991489	.7325369	1.35	0.176	-.444257	2.427235
university_switcher	.9551099	.7879852	1.21	0.225	-.5893127	2.499533
languagetest	0	(omitted)				
AdmissionsUvA	-1.331747	.3450412	-3.86	0.000	-2.008015	-.6554783
OutsideUvA	-1.268893	.4710394	-2.69	0.007	-2.192113	-.3456726
age	-.0273476	.0851294	-0.32	0.748	-.1941981	.1395029
_cons	1.603915	1.968621	0.81	0.415	-2.254512	5.462342