University of Amsterdam
Economics & Business

Predicting Enrolment Numbers for Master's Programs in Economics and Business at the University of Amsterdam

July 15, 2025

Maarten Hoogeboom
14000369

Program: MSc Data Science and Business Analytics
Company: University of Amsterdam
Contact Person: Drs. F.H.K. Pope

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

Statement of Originality

This document is written by Student Maarten Hoogeboom 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.

Abstract:

Every year, the Faculty of Economics and Business (FEB) of the University of Amsterdam faces a recurring challenge: how many students will enrol in September? The answer determines how efficiently the faculty can allocate schedules, staff, and scarce financial resources. Without accurate and timely predictions, universities are at risk of having understaffed courses, overbooked classrooms, or wasted budget. Recent work at Radboud University has shown that combining Studielink (the centralized Dutch application system) data with their internal student information system enables enrolment forecasts months in advance. However such approaches are not suitable for the context of the FEB. The Bayesian multilevel model developed in this thesis addresses key limitations of existing enrolment forecasting methods in the Netherlands by introducing a segmentation framework based on students' nationality and prior experience in Dutch higher education. The segmentation approach provided improves transparency and provides accessible insights to decision-makers who often rely on experience-based heuristics. As the dataset grows, the approach can be developed into a dynamic forecasting dashboard to support ongoing decision-making.

Contents

1 Introduction . . . 3
2 Background and Related Work . . . 4
2.1 Overview of Higher Education in the Netherlands . . . 4
2.2 Application and Enrolment Process . . . 5
2.3 Existing Forecasting methods . . . 6
2.4 Implications for the Faculty of Economics and Business . . . 8
3 Methodology . . . 9
3.1 Business understanding . . . 9
3.2 Data understanding . . . 10
3.3 Data preparation . . . 14
3.4 Modelling . . . 16
3.5 Evaluation . . . 20
3.6 Deployment . . . 20
4 Results . . . 20
4.1 Leave-one-programme-out cross-validation (LOOCV) . . . 20
4.2 Predictions 2025 . . . 22
5 Discussion . . . 23
6 Conclusion . . . 23
7 Limitations and future research . . . 24
8 Acknowledgments . . . 24
9 Appendix . . . 27

1 Introduction

Every year, the Faculty of Economics and Business (FEB) of the University of Amsterdam faces a recurring challenge: how many students will enrol in September? The answer determines how efficiently the faculty can allocate schedules, staff, and scarce financial resources. Without accurate and timely predictions, universities are at risk of having understaffed courses, overbooked classrooms, or wasted budget. Currently, a dedicated group of FEB staff already manually estimates these numbers, but this process is time-consuming. A robust forecasting model that accurately predicts student enrolments months in advance would not only enhance the faculty's operational efficiency, but also serve as a valuable complement to the existing manual efforts.

Recent work at Radboud University[1] has shown that combining Studielink (the centralized Dutch application system) data with their internal student information system enables enrolment forecasts months in advance. Their model relies on weekly historical data to train time series models such as SARIMA. Unfortunately, this approach depends on a level of data detail that the FEB currently does not have. Moreover, Radboud receives relatively fewer international applications compared to the FEB, where international students represent a significant proportion of the student population and may introduce greater uncertainty in enrolment predictions. A key challenge in this context is that, even after application deadlines have passed, it remains unclear which applicants will actually enrol. Students often submit several applications but ultimately commit to only one program. Furthermore, most students do not actively withdraw their applications, even if they decide to study elsewhere. As a result, there is a high level of noise in the application data: many entries reflect interest rather than commitment.

This thesis is part of a project initiated by the Faculty of Economics and Business (FEB) of the University of Amsterdam (UvA), aimed at predicting student enrolment across all Bachelor's and Master's programs. The objective was to produce a one-time forecast of student intake for September 2025 to support strategic and educational planning. The project was carried out in collaboration with AI4Business[2] involving two students: myself and Roel Lust. While this thesis focuses on forecasting enrolment in Master's programs, Lust (2025) addresses predictions for Bachelor's programs. Both theses were conducted as part of this broader forecasting project. In addition to the one-time forecasts, the FEB has expressed interest in developing a dashboard that could make enrolment predictions accessible and usable for policy and planning purposes.

The research question for this thesis is formulated as follows: *Can a robust and adaptable forecasting model be designed to predict student enrolment at the FEB, while addressing data noise caused by multiple applications and uncertain student commitment? Furthermore, can this model form the basis of a dynamic forecasting model that provides early and ongoing predictions to support institutional planning?*

The model developed in this thesis addresses key limitations of existing enrolment forecasting methods in the Netherlands by introducing a segmentation framework based on students' nationality and prior experience in Dutch higher education. The segmentation approach provided improves transparency

[1] Voxweb article on Radboud's model
[2] https://ai4business.uva.nl/

and provides accessible insights to decision-makers who often rely on experience-based heuristics. As the dataset grows, the approach can be developed into a dynamic forecasting dashboard to support ongoing decision-making.

The remainder of this thesis is organized as follows: Section 2, titled Background and Related Work, provides an overview of higher education in the Netherlands, the application and enrolment process, and relevant work such as that conducted at Radboud University. Section 3 outlines the methodology used in detail, following the CRISP-DM framework, followed by the results in Section 4. In Section 5, a discussion of the implications of the results is provided. Section 6 presents the conclusions of this thesis, and finally, Section 7 outlines the limitations and suggestions for future research based on this project and thesis.

2 Background and Related Work

This section provides essential context for the forecasting task addressed in this thesis. Section 2.1 describes the Dutch higher education system, Section 2.2 details the application process, Section 2.3 reviews relevant forecasting approaches, and Section 2.4 highlights their relevance to the FEB.

2.1 Overview of Higher Education in the Netherlands

The Dutch higher education system is composed of two main types of institutions: Research Universities (wetenschappelijk onderwijs, WO) and Universities of Applied Sciences (hoger beroepsonderwijs, HBO). Both types of institutions offer programmes at two levels: bachelor's and master's degrees. Bachelor's programmes typically last three to four years and provide foundational academic or professional training, while master's programmes generally span one to two years and offer more advanced, specialised education. Research Universities focus on academic and research-oriented programmes aimed at scientific or professional careers. By contrast, HBO institutions offer more practically oriented programmes designed to prepare students for the labour market (Nuffic, 2025; Rijksoverheid, 2024). This thesis focuses specifically on Research Universities.

In recent years, the Netherlands has become an increasingly attractive destination for international students, as reflected in enrolment figures. In 2024, Dutch research universities (WO) welcomed 39,402 new international students, who made up 29.9% of all new university enrolments (Nuffic, 2025). However, trends differ significantly between bachelor's and master's programmes (see Figure 1). Although international enrolment in WO master's programmes rose by 9.9% to 19,962 students, bachelor's enrolments declined by 5.2%, totalling 19,440 students and reducing their share of the intake from 27.9% to 27.0%.

The decline in bachelor-level enrolments can be attributed to several factors. According to Nuffic (2025), these include the 2022 national policy halting active recruitment of international students (with exceptions for priority sectors)[3], institutional caution due to anticipated legislation, and rising political and societal pressure to limit migration. However, much of the growth at the master's level was driven by international students who had previously studied in the Netherlands, this group made up 43% of new international master's enrolments. Excluding them, net growth was just 0.4% (Nuffic,

[3] https://nos.nl/artikel/2457440-kabinet-stop-met-werven-buitenlandse-studenten

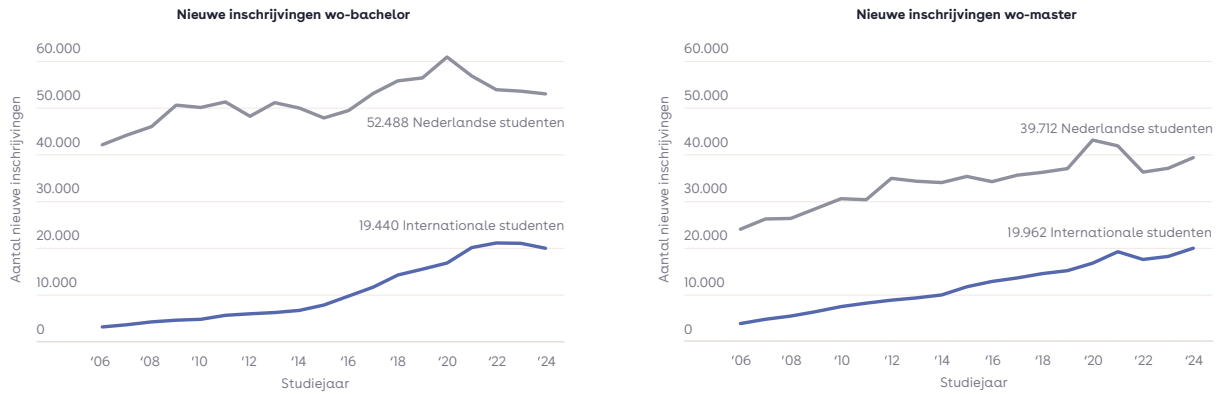

Figure 1: New international enrolments in WO-bachelor (left) and WO-master (right) programmes, 2006–2025. Blue lines represent international students.

Source: (Nuffic, 2025)

2025).

Despite national trends, institutional variation remains significant. The University of Amsterdam (UvA) hosts the largest number of international students in the country, with 15,592 international students, comprising 35.4% of its total student population. The UvA saw a 4.3% increase in new international enrolments, with a notable 15.4% rise at the master's level, suggesting that some institutions continue to attract international students despite broader restrictive trends (Nuffic, 2025).

Several factors make the Netherlands attractive to international students: many English-taught programs, low tuition fees, and strong universities, offering a good alternative to the UK or US (Weber et al., 2024). Dutch higher education institutions actively market themselves based on academic quality, international orientation, and urban student life. While master's students often choose based on programme reputation and quality labels, bachelor's students are more influenced by affordability and the attractiveness of the city environment (Weber et al., 2024).

2.2 Application and Enrolment Process

In the Netherlands, all students, domestic and international, apply to higher education programmes through Studielink, the centralized online application portal used by all Dutch universities. While Studielink handles the core application and registration process, individual institutions may require additional documents such as motivation letters, entrance exams, or proof of language proficiency depending on the programme (StudyinNL, 2025).

Admission requirements differ by level and institution. For WO bachelor's programmes, applicants typically need a pre-university secondary education diploma (vwo) or an equivalent qualification. However, students who have completed the first year of an HBO programme (known as the propedeuse) may also apply to WO bachelor's programmes. Admission to WO master's programmes generally requires a relevant bachelor's degree from a WO institution. Alternatively, HBO graduates can enter WO master's programmes by completing a pre-master programme designed to bridge any academic gaps (Studiekeuze123, 2025).

Application deadlines vary by programme, applicant background, and institution. At the FEB, deadlines for bachelor's programmes depend on whether the programme is subject to a numerus fixus (fixed enrolment cap); selective programmes generally require applications by January 15, while non-selective programmes accept applications until May 1. For master's programmes, students requiring a visa must apply by April 1, those with a non-Dutch degree by May 1, and students with a Dutch degree by June 1. In certain cases, students already enrolled in a FEB bachelor's programme are permitted to apply after the official deadlines.

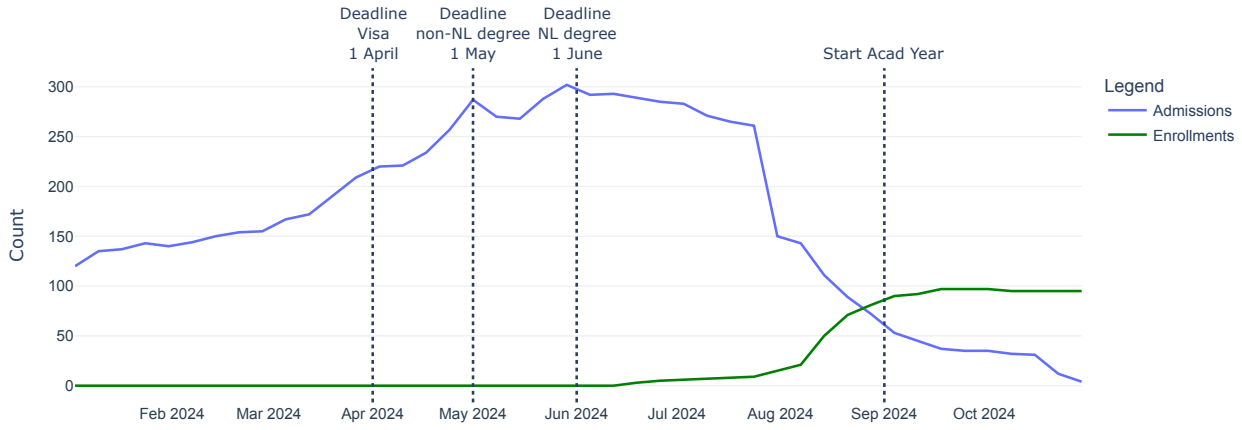

Figure 2: Weekly Monitoring Data for the Master's Program in Econometrics (Academic year 2024

Figure 2 presents week-to-week data for the Master Econometrics. The blue line shows the number of open applications, while the green line shows those who are fully enrolled. The data reveal that enrolment confirmations (green line) only begin to rise in late July to early August, just before the academic year starts. By June 1, most admissions (blue line) are already recorded, but enrolments remain low. This gap reflects delays in completing final administrative steps such as tuition payment. Notably, the blue line remains high even after the academic year begins, indicating the presence of "ghost applications". These "ghost applications" are students who neither finalise enrolment nor formally withdraw, and are only removed from the system later.

The key challenge is that accurate predictions of final enrolment numbers are needed by early June, but at that point, many students have not yet completed the enrolment process, complicating reliable forecasting.

2.3 Existing Forecasting methods

The majority of the literature in data mining for higher education is primarily focused on predicting student academic outcomes, including performance, retention, and graduation likelihood (Mao et al., 2024). Few studies have addressed the likelihood of admitted applicants actually enrolling at an institution from a predictive analysis standpoint to authors' knowledge. One notable example is Slim et al. (2018), which directly investigates admission conversion using machine learning techniques such as logistic regression and support vector machines. It found that application timing and financial aid behaviour are strong signs of enrolment, with early and timely actions suggesting higher motivation.

In the Dutch context, this topic has not yet received broad academic attention. However, a handful of Master's theses at Dutch universities have begun to explore enrolment forecasting and student influx prediction, often motivated by institutional needs for improved operational efficiency.

At Tilburg University, Hamers (2017) and van den Hurk (2017) used marketing data to model student enrolment and evaluate outreach strategies. They found that on-campus events were more effective than online campaigns, especially for bachelor applicants. Their models relied on individual factors such as travel distance, parental income, and campaign participation, though data limitations restricted the analysis to general bachelor enrolment rather than specific programmes.

At Utrecht University, recent theses have focused on forecasting Master's student influx, both in the long and short term. For long-term predictions, Leven (2022) developed a VARMA model using historical enrolment data and the number of AI bachelor diplomas awarded, though the analysis was limited to only 16 annual data points. Hsu (2022) extended this approach by applying an XGBoost regression model trained on aggregated data from eight Dutch universities, incorporating features like prior education and nationality. The study acknowledged that heterogeneity between institutions might affect the model's performance and noted that sufficient data from a single university would likely yield better results.

Short-term forecasting was addressed by Mosterd (2022), Vonk (2022), and Wirken (2022), who applied SARIMAX time-series models to predict Master's student influx for specific programmes using weekly application and enrolment counts from administrative systems. These early predictions, made by May, ahead of the June deadline, were promising but limited by the absence of individual-level data, which restricted the ability to segment applicants (e.g., by prior institution). The authors also pointed out that newer programmes, suffered from limited historical data, complicating efforts to detect stable trends or respond to sudden shifts in application patterns.

In addition to academic research, Radboud University has developed a system for enrolment forecasting, publicly available on GitHub[4]. Although this work is not accompanied by a formal publication, it represents the most closely related operational approach to enrolment prediction in the Dutch higher education context.

The Radboud model combines weekly data from Studielink with internal Student Information System (SIS) data. It uses an ensemble of time-series and machine learning models to forecast first-year student enrolment. Predictions are broken down by programme and broad nationality groups, including Dutch students, students from the European Economic Area (EEA)[5], and students from outside the EEA (Non-EEA). The results are shown in an interactive dashboard.[6].

Although a systematic evaluation of model performance is not available, an exploratory analysis of the dashboard (see Figure 3) provides some insight. The dashboard displays weekly forecasts for the 2024–2025 academic year across different faculties, programs, and nationalities. It includes performance metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

[4] https://github.com/cedanl/studentprognose
[5] See what countries are in the EEA: https://www.netherlandsworldwide.nl/eu-eea-efta-schengen-countries
[6] Radboud Dashboard

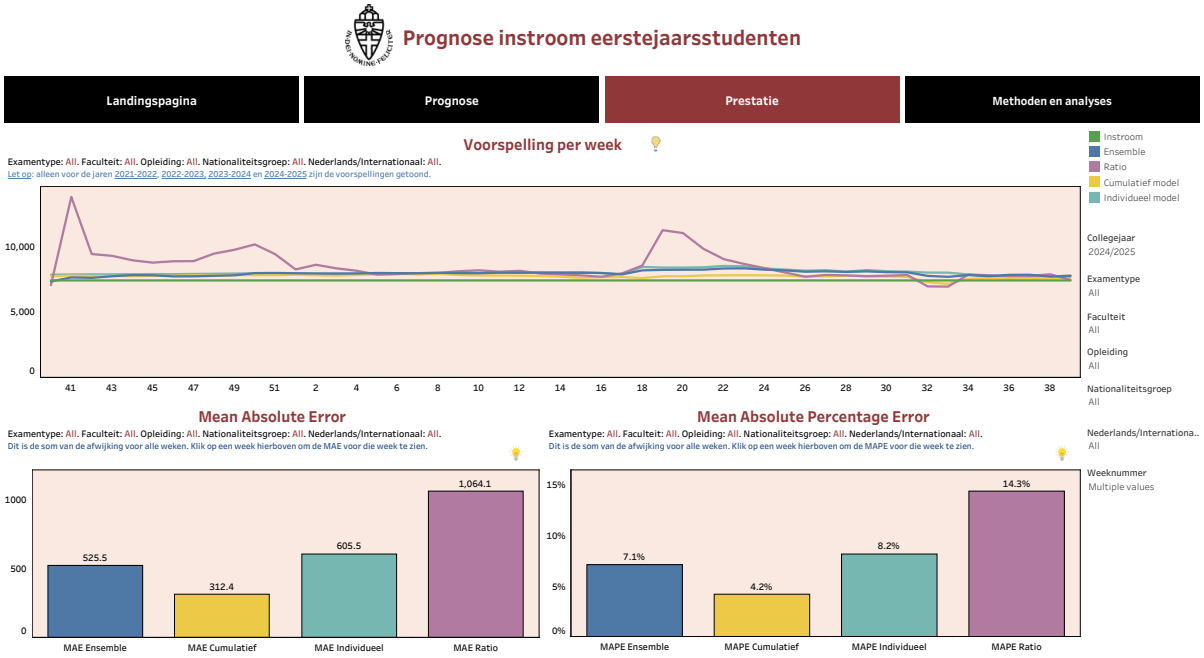

Figure 3: Screenshot of the Radboud Dashboard

For all students combined, the MAPE for the ensemble model is approximately 7.1%. When segmented by nationality, the MAPE drops to 4.9% for Dutch students, but increases significantly for EEA (23%) and non-EEA (21.8%) students.

2.4 Implications for the Faculty of Economics and Business

The reviewed literature shows growing interest in enrolment forecasting within Dutch higher education, but also highlights major differences in data availability across institutions. For example, Radboud University's model relies on years of historical data from Studielink and internal systems. By contrast, theses such as Mosterd (2022), Vonk (2022), and Wirken (2022) lacked access to such detailed data and were therefore unable to segment applicants meaningfully. Notably, Radboud's model predicts only broad nationality groups (Dutch, EEA, non-EEA), without further segmentation by factors such as educational background. For master's programmes in particular, it does not distinguish between applicants with a prior degree from Radboud, from another Dutch institution, or from abroad. This limits their ability to capture the diverse enrolment behavior observed in practice. This issue is particularly relevant for the University of Amsterdam, which has an high share of international students compared to other Dutch institutions (Weber et al., 2024), and therefore a greater share of applicants who either hold foreign qualifications or have already completed a Dutch degree. Moreover, many entries in the application data reflect interest rather than actual commitment, introducing significant noise. The existing literature does not fully address these issues, leaving a clear gap for models that can handle limited, noisy data while incorporating finer segmentation.

3 Methodology

This study follows the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, a widely adopted framework for structuring data mining projects. It consists of six iterative phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The process is flexible, allowing movement between phases as needed (Chapman, 2000).

The structure of this thesis aligns with these phases as follows:

- **Business Understanding** (Section 3.1): Outlines the project objectives and institutional context.
- **Data Understanding** (Section 3.2): Describes the dataset, focusing on its sources (Studielink and SIS).
- **Data Preparation** (Section 3.3): Covers data cleaning procedures and the feature engineering process, including the creation of the key variable *StudentType*.
- **Modelling** (Section 3.4): Introduces the multilevel Bayesian model used for enrolment forecasting.
- **Evaluation** (Section 3.5): Details the leave-one-programme-out validation strategy used to assess model performance.
- **Deployment** (Section 3.6): Discusses potential implementation through a forecasting dashboard and broader institutional use.

Note that results, are presented separately in Section 4.

3.1 Business understanding

The business problem addressed in this project is the ongoing challenge faced by FEB to accurately predict how many students will enrol in its master programs each September. These forecasts are essential for informed decisions about scheduling, staffing, and the allocation of limited financial resources. Inaccurate or delayed predictions can lead to understaffed courses, overbooked classrooms, or inefficient budget planning. This project focuses on enhancing the prediction of student enrolment in FEB's master programs for the upcoming academic year. This will be done by exploring the use of probabilistic modelling by assigning each admission a likelihood of converting to enrolment. The performance of this approach will be compared to the current method to determine its added value.

In addition, FEB has expressed interest in developing a dynamic forecasting model that provides early enrolment predictions throughout the admissions cycle. The business value of such a model lies in enabling staff and decision-makers to monitor enrolment trends in real time, adjust forecasts as new data becomes available, and make informed resource allocation decisions. This could be realized through a dashboard that makes forecasts easily accessible and usable for planning and policy purposes, ultimately enhancing operational efficiency and reducing risks tied to uncertain enrolment outcomes.

3.2 Data understanding

Before describing the data in detail (Section 3.2.3), it is first important to understand the data sources for this project. This project relies on two primary data sources: Studielink and the Student Information System (SIS) used by the FEB. Subsection 3.2.1 will discuss how these relate to each other and highlight their differences, followed by subsection 3.2.2, which explains the Studielink sleutelbestanden that are designed to give institutions detailed enrolment information specific to them.

3.2.1 Studielink and SIS

Studielink is the centralized national platform for managing student enrolments and applications across Dutch higher education institutions. It facilitates the submission and processing of enrolment applications between students and educational institutions. Studielink also enables data exchange with government bodies such as DUO (the Dutch Education Executive Agency) but does not directly manage educational registries. It serves as the main gateway through which students submit enrolment requests and where key enrolment-related data is collected and shared (Studielink documentation, 2023).

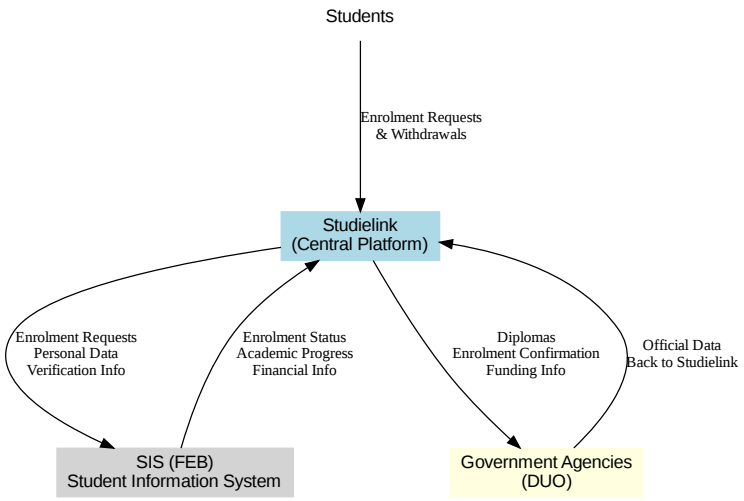

Figure 4: Overview of Data Exchange Between Studielink and SIS

SIS is the local administrative system used by the FEB to manage detailed student records. This includes personal information, academic progress, enrolment status, financial data, and prior education credentials. Each institution maintains its own SIS, tailored to its specific administrative processes. Studielink documentation (2023) shows that these two systems operate in continuous, two-way communication to ensure synchronized and up-to-date student data. When a student applies or changes their enrolment status via Studielink, this information is transferred to the institutional SIS. Conversely, updates from the university, such as confirmations of enrolment, academic progress, or financial changes, are sent back to Studielink. Most variables stored in SIS originate from Studielink data, however SIS may also contain institution-specific data not managed by Studielink. For example, FEB's admission office stores more specific enrolment details and statuses.

The overall structure of these interactions is visualized in Figure 4, which illustrates the bidirectional flow of data between students, Studielink, institutional SIS systems, and government agencies.

3.2.2 Studielink Sleutelbestanden

According to Studielink documentation (2022), the Studielink sleutelbestanden were introduced in 2013 to provide Dutch higher education institutions with detailed, institution-specific enrolment information. Unlike the previously available aggregated and anonymised data (telbestanden), the sleutelbestanden are on the student-level. The data is compiled by Studielink and enriched with information from DUO sources, specifically the 1cijferHO dataset, which indicates the student's history in Dutch Higher Education.

Studielink sleutelbestanden are generated weekly, typically on Mondays, based on a snapshot of the Studielink database that includes changes up to 1:00 AM that day. Each week, a new file is created for every institution, identified by its Brincode. These files contain detailed records for individual applications. See the table below:

Field	**Description**
Brincode	The institution's unique code.
Brin_volgnr	The branch location of the institution where the programme is offered, according to DUO
Isatcode	The programme code, also known as the education code, as defined by DUO.
Opl_vorm	The selected programme form (e.g., full-time, part-time, dual) for which an enrolment request has been submitted.
Studiejaar	The academic year to which the enrolment request pertains, indicated by the first year of the academic year (e.g., 2024 for 2024-2025).
Maand	The month in which the student intends to commence their studies, which can be adjusted by the institution.
Herkomst	The student's origin, indicating if they are a Dutch, EEA, non-EEA citizen, or unknown. Determined based on national citizenship codes from DUO.
Geslacht	The student's gender, based on data from the Studielink database.
Meercode_V	The number of enrolment requests/enrolments a student has (excluding cancelled requests). Includes all statuses except 'A'.
Meercode_A	The number of cancelled enrolment requests (status 'Cancelled') per student.
Status	The status of the enrolment request: 'V' for request, 'A' for cancellation/rejection, 'U' for withdrawal, 'I' for enrolment.
Hogerejaars	Indicates whether the student is starting in a higher academic year. Can be modified by the institution.
Fixus	Indicates whether a numerus fixus (enrolment cap) applies to the programme.

Table 1 continued from previous page

Herinschrijving	Specifies if it is a re-enrolment, based on Studielink checks and DUO's data.
1cHO_L	A code (1–6) from DUO's 1cijferHO-bestand indicating the student's history in Higher Education using a 'long sequence' (mainly for HBO).
1cHO_K	A code (1, 2, 4, 5, or 6) from the 1cijferHO-bestand indicating the student's history in Higher Education using a 'short sequence' (mainly for WO).

Table 1: Raw Studielink Sleutelbestanden

Source: (Studielink documentation, 2022)

Question	**Answer**	**Action / Value**
Does the student have a registration year ($1cijferHO$) earlier than the study year (starting in October) in higher education?	No	Set value 1cHO = 1
	Yes	Next question
Does the student have an active registration in the same type of higher education (HBO/WO)?	No	Set value 1cHO = 2
	Yes	Next question
Does the student have an active registration in the same level of higher education (bachelor/master) within the same type?	No	Set value 1cHO = 3
	Yes	Next question
Does the student have an active registration at the same institution for the same level?	No	Set value 1cHO = 4
	Yes	Next question
Does the student have an active registration in the same program at the same institution?	No	Set value 1cHO = 5
	Yes	Set value 1cHO = 6

Table 2: Decision tree for determining the 1cHO value.

Source: (Studielink documentation, 2022)

From the overview of fields presented in table 1, two features are particularly worth highlighting: the Meercode_V / Meercode_A fields and the 1cHO value.

The Meercode_V and Meercode_A fields are a useful set of indicators. These values represent the number of enrolment requests a student has submitted (Meercode_V) and how many of those were later cancelled (Meercode_A) during the application process. A request is counted under Meercode_V as long as it holds any status other than 'Cancelled'. If the status changes to 'Cancelled' in a later week, it moves to Meercode_A and is no longer included in Meercode_V. By comparing these counts to the number of enrolment requests directed at the institution of interest, it is possible to infer whether a student is also considering other universities. This makes the Meercode fields a valuable signal when predicting how serious a student is about actually enrolling with the FEB.

The 1cHO value is a key indicator that can be used to quickly assess a student's background in Dutch higher education. It shows how familiar a student is within the system compared to their status on October 1st of the previous academic year. Based on the decision tree that can be reviewed in table 2, the codes have the following meanings. Compared to that previous status, the student is:

1. New to Dutch higher education.
2. Known in Dutch higher education but switched to a different type of Dutch higher education (e.g., from HBO to WO).
3. Known in the same type of Dutch higher education but switched to a different level (e.g., from bachelor to master).
4. Known in the same type and level of Dutch higher education but switched institutions.
5. Known in the same type and level of Dutch higher education and institution but switched programs.
6. Known in the same type and level of Dutch higher education, institution, and program.

The 1cHO feature will serve as the essential foundation upon which the StudentType segmentation will later be based. Further details will be provided in subsection 3.3.1.

3.2.3 Data Description

After merging the Studielink data with internal SIS records, the dataset consists of application records from approximately 13,800 students across ten master's programs within the FEB. One of these programs, Business Information Technology Management, is new and will start for the first time in the upcoming 2025/2026 academic year, meaning no historical data is available for it. The dataset includes training data from the 2024/2025 academic year and test data used to forecast enrolments for 2025/2026, for which final outcomes are not yet known. Each record includes personal characteristics such as gender, nationality, and pre-education, as well as admission-related information including the application date and final admission result. The target variable is a binary indicator of whether the student ultimately enrolled. All datasets have been pseudonymised to protect student privacy. The 2024/2025 data are static snapshots, even for originally dynamic files. As a result, time-dependent changes lack exact timestamps, limiting the ability to reconstruct a exact timeline of events.

Figure 5 presents enrolment results by program and nationality for the academic year 2024, grouped into three categories: Dutch, students from the European Economic Area (EEA), and students from outside the EEA (Non-EEA). Solid bars represent enrolled students, while patterned bars indicate those who did not enrol. The Master Business Administration receives the highest number of applications, followed by the Master Finance, with the Master Data Science & Business Analytics ranking third. A key insight is that most applications come from Non-EEA students, though the balance between Dutch and EEA applicants varies by program. For example, EEA applications outnumber Dutch ones in Business Administration, while in the Master Accountancy and Control, Dutch applications dominate and EEA applicants are relatively small compared to other programs. Fiscal Economics is a notable exception, as it is taught in Dutch. The enrolment figures tell a different story. Non-EEA students are not always the largest enrolled group and, in some cases, such as the Master Business

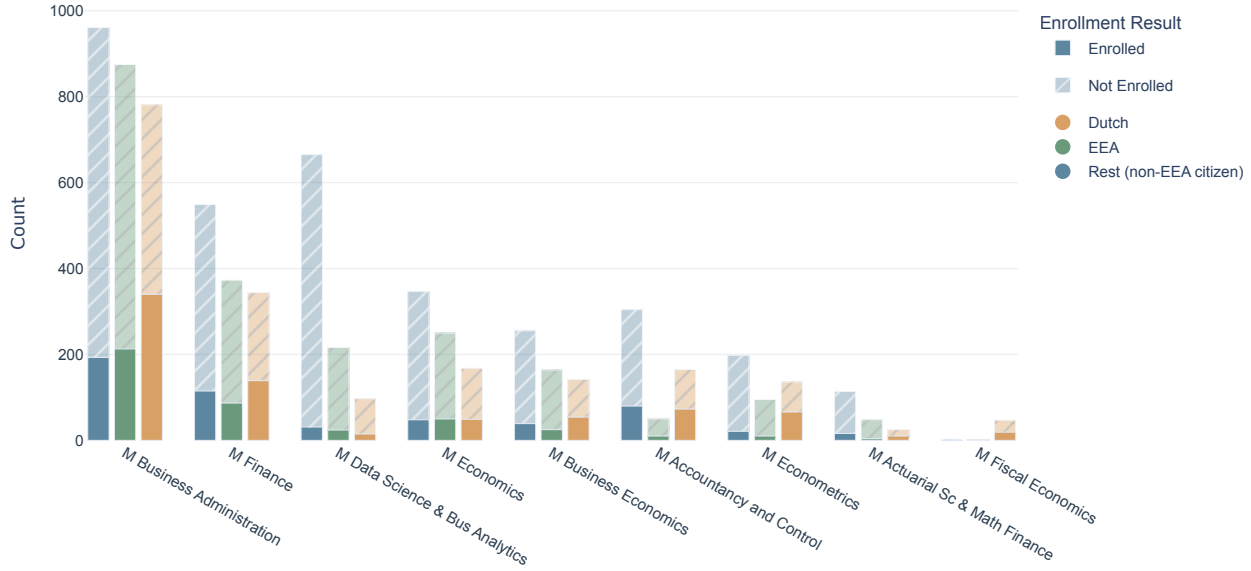

Figure 5: Enrolment results by program and nationality (Dutch, EEA, Non-EEA) for the academic year 2024. Solid bars represent enrolled students, while patterned bars indicate those who did not enrol

Administration, they represent the smallest. When comparing enrolment rates relative to applications, the Master Data Science & Business Analytics stands out: despite having the third-highest number of applications, it has relatively few enrolled students.

Figure 6 shows enrolment numbers based on the dates students submitted their applications for the academic year 2024. The chart highlights the official deadlines: April 1 for students requiring a visa, May 1 for those with a non-Dutch degree, and June 1 for students with a Dutch degree. What stands out is that some applications submitted after the June 1 deadline still result in enrolment. There are a few possible explanations for this. One is that the students are already studying at FEB, for example, those in the Bachelor's in Business Analytics who continue directly into the Master's in Data Science & Business Analytics. Because they have direct access, they may apply later than the June 1 deadline. Another explanation is that the late application is not the student's first application this academic year. For instance, a student might have applied to the Master Business Administration before the deadline but later decided to switch to the Master Finance. Since the admissions office had already reviewed their qualifications, enrolment in the new program may still be possible after the deadline.

3.3 Data preparation

During the data preparation stage, various cleaning and preprocessing actions were undertaken to enhance data quality and ensure consistency across the different sources. One of these actions was the handling of duplicate applications submitted by the same student for the same program. These duplicates arose when students cancelled and later resubmitted applications or reapplied after rejections. In such cases, only the latest application was kept. For consistency, dummy variables were created to indicate whether applications were withdrawn in each month from January through July, specifically: withdrawn_Jan, withdrawn_Feb, withdrawn_Mar, withdrawn_Apr, withdrawn_May, withdrawn_Jun, and withdrawn_Jul. The same set of dummy variables was generated for both the 2024 and 2025

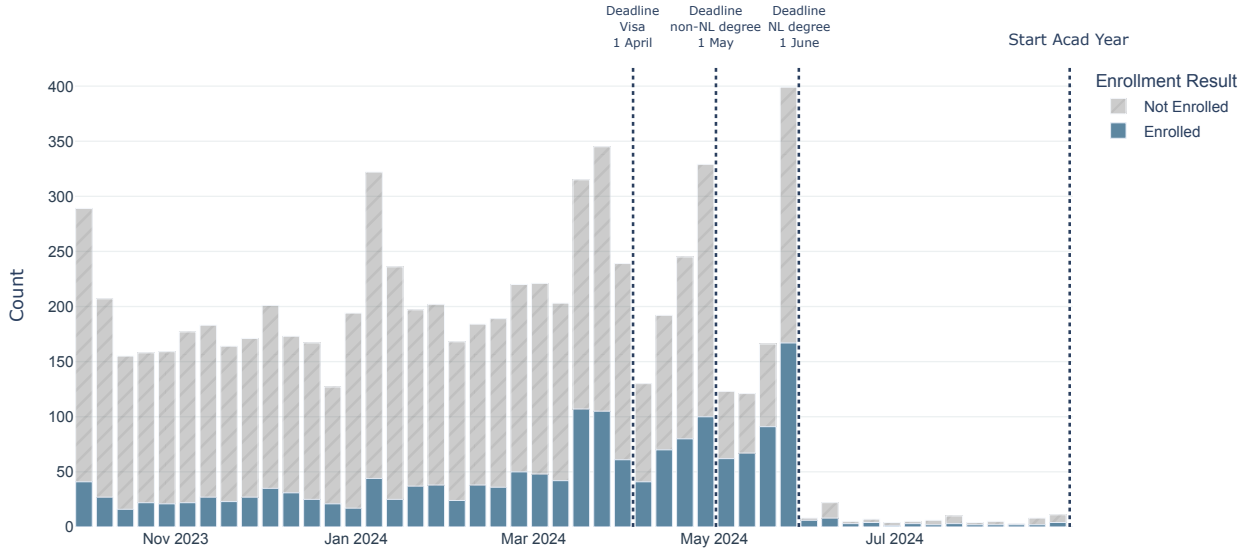

Figure 6: Enrolment results by the dates students submitted their applications for the academic year 2024. April 1 is the deadline for students requiring a visa, May 1 for those with a non-Dutch degree, and June 1 for students with a Dutch degree.

datasets to maintain consistency and comparability between training and prediction data. Depending on the objective, the dataset could be filtered accordingly to suit that specific goal.

Lastly, extensive feature engineering was performed to improve the predictive power of the model. These features were primarily created using the Studielink sleutelbestanden to ensure reproducibility. The following subsections detail the creation of the variables: StudentType, Looking Elsewhere, and Amount of FEB Applications.

3.3.1 StudentTypes

The StudentType is built on top of the structure introduced with the studielink sleutelbestanden feature 1cHO. This value serves as an important indicator that helps quickly assess a student's background in Dutch higher education. More details about 1cHO can be found earlier in Section 3.2.2. Table 3 provides an overview of the StudentTypes created for this project. The python code used to create the StudentTypes is provided in the appendix (1).

StudentType	Description	#Applications	Avg Conversion Rate
MQP[7]	These students were preparing to qualify for their current application.	163	0.69
uva_bachelor	These students were enrolled in a bachelor program at the UvA	1380	0.54
uva_master	These students were enrolled in a master program at the UvA	149	0.32
non_uva	These students were enrolled at a different WO institution.	1686	0.25
hbo	These students were enrolled at a HBO institution.	179	0.27
new_Rest	Non-EEA students new to Dutch higher education.	2559	0.09
new_EEA	EEA students new to Dutch higher education.	1106	0.08

[7] https://www.oncampus.global/our-study-centres/oncampus-amsterdam/oncampus-amsterdam-mqp

Table 3 continued from previous page

new_NL	Dutch students new to Dutch higher education.	151	0.16

Table 3: Overview of created StudentTypes. The numbers and conversion rates are historic data from 2024 and contain all programs.

3.3.2 Individual Features

In addition to the StudentType variable, two individual features were engineered to capture specific aspects of applicant behavior:

Looking Elsewhere. This binary feature indicates whether an applicant has submitted enrolment applications to other Dutch institutions beyond the FEB. It is derived by comparing the total number of enrolment requests recorded in the Meercode_V and Meercode_A fields. More details about the Meercode fields can be found earlier in Section 3.2.2.

To adjust for students who are likely re-enrolling in their current program rather than applying elsewhere, the 1cHO field is used: if its value equals 4, indicating current bachelor enrolment at another university, the count of "other" applications is reduced by one to account for a likely re-enrolment application. This correction addresses the assumption that re-enrolments inflate Meercode counts, allowing the feature to more accurately reflect genuine applications to other institutions. Overall, this feature helps distinguish applicants focused solely on FEB from those considering multiple institutions.

Amount of FEB Applications. This feature records the number of distinct FEB programs to which a student applied.

3.4 Modelling

This section outlines the modeling approach used to forecast student enrolment at the FEB. The process is structured in two parts. Subsection 3.4.1 presents models that estimate conversion probabilities, the likelihood that an applicant ultimately enrols. These estimates then serve as inputs for the dynamic forecasting approach described in Subsection 3.4.2, which can in theory update predicted enrolment figures over time.

For this thesis, a multilevel Bayesian model was adopted. Multilevel models, also called mixed or hierarchical models, are used to analyse data that contain clusters. These models account for the similarities within each cluster and are widely used in fields like education, medicine, social sciences, and reliability (Correa-Álvarez et al., 2023). According to McElreath (2020, Chapter 13), multilevel models use varying intercepts to capture group-level differences while pooling information across clusters. This partial pooling improves estimates by balancing group-specific data with information from the overall population, leading to more accurate and stable inference, especially when some groups have limited data.

To better understand the benefits of multilevel modelling, it is useful to contrast it with two simpler approaches: complete pooling, which assumes all groups (e.g., StudentTypes) are the same and produces a single global estimate but often underfits the data, and no pooling, which estimates each group

separately without sharing information, often leading to overfitting, especially when group data are limited (McElreath, 2020, Chapter 13). The amount of pooling is governed by a group-level standard deviation parameter, typically denoted σ.

In the Bayesian framework, uncertainty in all parameters is captured through posterior distributions, which represent updated beliefs after observing the data. Because these distributions are often analytically intractable, they are approximated using Markov Chain Monte Carlo (MCMC) sampling (Gelman et al., 2013). In this thesis, models were implemented in Python using NumPyro. The code used are provided in the thesis GitHub repository[8] .

3.4.1 Conversion Model

The estimation of conversion probabilities is approached through two types of Bayesian hierarchical models: group-level models that operate on aggregated StudentTypes data, and individual-level models. These models differ in complexity and level of detail, with both leveraging partial pooling.

Group-Level Models

The group-level models (Basic and Type+) operate on aggregated data, where each observation corresponds to a unique combination of StudentType and applied program. For each group i, the number of applicants N_i and the number of enrolments Y_i are recorded, along with relevant features. An example of this grouped input data is provided in Table 4

StudentType	N_i	y_i	is_known_by_FEB	is_other_NL_institution	uva_bachelor	non_uva	new_Rest	...
uva_bachelor_BAN	90	20	1	0	1	0	0	...
uva_bachelor_EC	100	65	1	0	1	0	0	...
uva_bachelor_BA	500	285	1	0	1	0	0	...
...	...	...	...	...	...	...	...	...
non_uva_BAN	160	20	0	1	0	1	0	...
non_uva_EC	100	20	0	1	0	1	0	...
non_uva_BA	700	300	0	1	0	1	0	...
...	...	...	...	...	...	...	...	...
new_Rest_BAN	650	10	0	0	0	0	1	...
new_Rest_EC	150	4	0	0	0	0	1	...
new_Rest_BA	800	80	0	0	0	0	1	...

Table 4: Example input data for the group-level model (StudentType + Program combinations)

Basic Multilevel Model

The first and simplest model, inspired by the multilevel "Tadpole" model in McElreath (2020, p. 417), is a basic varying intercept model. It estimates the number of enrolments Y_i for a given applicant StudentType–program combination i as a binomial outcome:

[8] https://github.com/MaartenH327/forecasting-enrollment-dutch-universities

$$Y_i \sim \text{Binomial}(N_i, p_i) \tag{1a}$$
$$\text{logit}(p_i) = \alpha_i \tag{1b}$$
$$\alpha_i \sim \text{Normal}(\overline{\alpha}, \sigma) \tag{1c}$$
$$\overline{\alpha} \sim \text{Normal}(0, 1.5) \tag{1d}$$
$$\sigma \sim \text{Exponential}(1) \tag{1e}$$

This basic varying intercept model leverages the grouping structure without additional covariates, providing a robust baseline for partially pooled conversion estimates. The prior $\overline{\alpha} \sim \text{Normal}(0, 1.5)$ reflects weakly informative beliefs, centring the global log-odds around zero while allowing reasonable variation (McElreath, 2020, Chapter 5). The group-level standard deviation $\sigma \sim \text{Exponential}(1)$ places more weight on smaller values, encouraging modest variation across groups unless the data strongly suggest otherwise.

Type+ Model

The next model, referred to as *Type+*, extends the basic varying intercept model by introducing predictors at the StudentType level to better distinguish between applicant groups. These include characteristics such as prior affiliation with the FEB and enrolment at another Dutch institution. The model modifies the group-level intercepts to incorporate these predictors:

$$Y_i \sim \text{Binomial}(N_i, p_i) \tag{2a}$$
$$\text{logit}(p_i) = \alpha_i \tag{2b}$$
$$\alpha_i \sim \text{Normal}(\mu_i, \sigma) \tag{2c}$$
$$\mu_i = \beta_0 + \beta_{\text{FEB}} \cdot \text{is\_known\_by\_FEB}_i \tag{2d}$$
$$+ \beta_{\text{NL}} \cdot \text{is\_other\_NL\_institution}_i \tag{2e}$$
$$+ X_{\text{cluster}[i]} \cdot \beta_{\text{cluster}[i]} \tag{2f}$$
$$\beta_j \sim \text{Normal}(0, 1.5) \tag{2g}$$
$$\sigma \sim \text{Exponential}(1) \tag{2h}$$
$$\tag{2i}$$

Individual-Level Model

Extending beyond aggregated group-level data, the individual-level model operates on data where each row corresponds to a single application. The outcome Y_i is binary, indicating whether the applicant ultimately enrolled. This model incorporates both individual-level predictors and group-level (StudentType) effects. An example of the individual-level input data is provided in Table 5.

ApplicationID	y_i	looking elsewhere	n_applications_FEB	StudentType
1	1	1	1	uva_bachelor_BAN
2	0	1	1	uva_bachelor_EC
3	0	0	3	uva_bachelor_BA
4	1	0	1	non_uva_BAN
5	1	0	2	non_uva_EC

Table 5: Example input data for the individual-level model

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(p_i) && \text{(3a)}\\
\text{logit}(p_i) &= \alpha_{\text{StudentType}[i]} + \gamma_1 \cdot \text{looking\_elsewhere}_i && \text{(3b)}\\
&\quad + \gamma_2 \cdot \log(\text{n\_applications\_FEB}_i) && \text{(3c)}\\
\alpha_j &\sim \text{Normal}(\mu_j, \sigma) && \text{(3d)}\\
\mu_j &= \beta_0 + \beta_{\text{FEB}} \cdot \text{is\_known\_by\_FEB}_j && \text{(3e)}\\
&\quad + \beta_{\text{NL}} \cdot \text{is\_other\_NL\_institution}_j && \text{(3f)}\\
&\quad + X_{\text{cluster}[j]} \cdot \beta_{\text{cluster}[j]} && \text{(3g)}\\
\beta_0,\, \beta_{\text{FEB}},\, \beta_{\text{NL}},\, \beta_j &\sim \text{Normal}(0, 1.5) && \text{(3h)}\\
\gamma_1,\, \gamma_2 &\sim \text{Normal}(0, 1.5) && \text{(3i)}\\
\sigma &\sim \text{Exponential}(1) && \text{(3j)}
\end{aligned}
$$

This model provides the most detailed estimation of conversion probabilities, incorporating individual-level variation that might otherwise be averaged out. It leverages this detailed variation to improve prediction accuracy at the individual applicant level.

3.4.2 Dynamic Forecasting (Concept)

This section outlines a conceptual approach to dynamic forecasting that was initially planned for this thesis. Although simulations based on this framework were not implemented due to time constraints, the concept is included here because it directly relates to the research question and offers a foundation for future work.

To develop a dynamic forecasting model, the objective is to predict the number of applications N_j for each StudentType j, while allowing the estimate $\hat{p}_j$ of the conversion probability to improve over time. The estimate $\hat{p}_j$ can be derived from the individual-level model by averaging the predicted probabilities for all applications belonging to a given StudentType. The number of applications N_j could be forecast using a time series model that captures temporal application patterns. In theory, this would allow enrolment to be predicted early in the application cycle, well before all applications have been submitted.

$$y_j \sim \text{Binomial}(N_j, \hat{p}_j)$$

where y_j denotes the number of enrolled StudentType j, N_j is the total number of applications from StudentType j, and $\hat{p}_j$ is the estimated probability that an applicant of StudentType j will enrol.

Even though this dynamic forecasting approach was not implemented in this thesis, Appendix Figures 7 and 8 illustrate the distribution of application times across StudentTypes, revealing notable differences in application behaviour.

3.5 Evaluation

Due to the absence of fully observed outcome data for the 2025/2026 academic year, a traditional train-test evaluation is not feasible. To address this, an evaluation strategy inspired by leave-one-out cross-validation (Hastie et al., 2009) was developed. Specifically, a leave-one-programme-out cross-validation approach was implemented, using the complete 2024/2025 dataset. This approach was applied across three multilevel models and four different values of the hyperprior σ (ranging from 0.1 to 1.5), enabling comparison of generalisation performance when one programme is held out as a pseudo-test set.

The evaluation metric is the mean absolute error (MAE), calculated at the programme level. For each programme k left out during cross-validation, the model predicts the conversion probability for every application within that programme. The predicted conversion rate $\hat{p}_k$ is then obtained by averaging these predicted probabilities. The absolute difference between the observed conversion rate p_k and the predicted rate $\hat{p}_k$ is computed for each programme. The overall MAE is the mean of these absolute differences across all programmes:

$$\text{MAE} = \frac{1}{K} \sum_{k=1}^{K} |p_k - \hat{p}_k|$$

where K is the total number of programmes, p_k is the observed conversion rate for programme k, and $\hat{p}_k$ is the predicted conversion rate for programme k.

3.6 Deployment

The modelling results could be integrated into an interactive dashboard, similar to the one developed at Radboud University, allowing FEB stakeholders to access updated enrolment forecasts in real time. Such a dashboard can support ongoing monitoring of predicted student inflow and conversion rates throughout the application cycle, offering valuable insights for strategic planning and resource allocation. As of writing this thesis, the FEB is planning to further develop and implement such a dashboard in the future.

4 Results

This section reports the main findings of the analysis. Section 4.1 evaluates model performance using leave-one-programme-out cross-validation. Section 4.2 presents enrolment forecasts for the 2025/2026 academic year.

4.1 Leave-one-programme-out cross-validation (LOOCV)

			AE Score								
Model name	σ	Combined (MAE)	BA	FIN	BAN	ECON	ACC	BUS ECON	EC	ACT	FISC
Basic	0,1	0,0894	0,0948	0,0418	0,2209	0,0597	0,0821	0,0180	0,0021	0,0812	0,2036
Basic	0,5	0,0880	0,0939	0,0474	0,2109	0,0459	0,0933	0,0069	0,0024	0,0793	0,2115
Basic	1	0,0901	0,1044	0,0636	0,1962	0,0291	0,1018	0,0071	0,0140	0,0726	0,2223
Basic	1,5	0,0946	0,1192	0,0807	0,1825	0,0128	0,1132	0,0197	0,0222	0,0624	0,2388
Type+	0,1	0,0458	0,0766	0,0139	0,1377	0,0177	0,0863	0,0106	0,0106	0,0517	0,0072
Type+	0,5	0,0472	0,0845	0,0310	0,1223	0,0016	0,1059	0,0085	0,0054	0,0386	0,0275
Type+	1	0,0519	0,0909	0,0366	0,1140	0,0114	0,1163	0,0170	0,0137	0,0322	0,0353
Type+	1,5	0,0569	0,1028	0,0406	0,1116	0,0144	0,1240	0,0262	0,0142	0,0293	0,0493
Individual	0,1	0,0459	0,0613	0,0107	0,1294	0,0078	0,0706	0,0242	0,0015	0,0525	0,0551
Individual	0,5	0,0492	0,0709	0,0289	0,1153	0,0100	0,0892	0,0392	0,0135	0,0416	0,0338
Individual	1	0,0529	0,0799	0,0334	0,1113	0,0178	0,0993	0,0519	0,0183	0,0344	0,0301
Individual	1,5	0,0571	0,0885	0,0396	0,1055	0,0201	0,1113	0,0617	0,0224	0,0356	0,0289

Table 6: Summarises leave-one-programme-out cross-validation results for three multilevel models and four values of the hyperprior σ (ranging from 0.1 to 1.5). The evaluation metric is absolute error (AE), where lower values indicate better predictive performance. Each programme-specific column shows the AE when that entire programme was excluded from training and used solely as the test set. This setup assesses how well the model generalises to previously unseen programmes

The LOOCV results that can be reviewed in table 6 reveal clear improvements in predictive accuracy as the models increase in complexity. The Basic model, which includes only varying intercepts for StudentTypes, performs the worst overall. Introducing group-level predictors in the Type+ model results in a substantial drop in MAE across most held-out programs, indicating that background information at the StudentType level carries meaningful predictive power. Interestingly, the addition of individual-level predictors in the Individual model does not lead to a significant further improvement over the Type+ model. This suggests that, in this context, most of the predictive gains can be achieved by accounting for StudentType characteristics, while individual features offer marginal benefit.

The parameter σ, which governs the degree of pooling in the hierarchical structure, plays a critical role in shaping predictions. A lower σ implies stronger shrinkage toward the global average, whereas a higher σ allows for greater variation between groups. The LOOCV results reveal that the relationship between σ and predictive performance is not uniform across programs. For some programs, increasing σ leads to worse performance, whereas for others, performance improves.

To interpret this pattern, it is useful to consider σ as controlling how strongly group predictions are pulled toward the global mean. Some programs, such as Econometrics (EC), appear to be close to this global center: even the Basic model performs well when EC is excluded from training. This suggests that EC is well-explained by the overall pattern in the data and does not behave as an outlier. In contrast, a program such as Data Science & Business Analytics (BAN) appears to be more atypical in this particular year. For BAN, higher σ, meaning less shrinkage, results in better predictions, indicating that it benefits from being treated more independently of the rest.

This behaviour implies that there is no single "best" level of pooling that works equally well for all programs. From a practical standpoint, it may be advisable to generate predictions using multiple

values of σ and then aggregate them to form a more robust final prediction. This approach acknowledges the uncertainty about whether a given program in a given year will behave like an outlier or align more closely with the global average. Embracing this uncertainty, rather than committing to a fixed degree of pooling, may yield more stable and generalizable predictions over time.

4.2 Predictions 2025

Model name	σ	BA	FIN	BAN	ECON	ACC	BUS ECON	EC	ACT	FISC	BITM
ESC[9]	-	630-670	320-350	75-100	80-90	130-150	110-130	80-90	40-45	45-55	60-80
Basic	0,1	649-733	342-404	142-186	161-206	111-148	116-155	70-101	35-58	13-27	46-71
Basic	0,5	710-790	356-414	68-101	139-182	136-173	109-146	82-110	32-53	23-38	45-70
Basic	1	714-793	356-413	58-87	138-180	143-180	108-144	82-109	31-50	28-42	42-67
Basic	1,5	715-794	356-413	55-84	138-179	145-181	107-143	82-108	30-48	29-43	40-65
Type+	0,1	659-737	357-415	122-161	148-190	119-153	114-149	82-110	37-57	27-42	67-92
Type+	0,5	708-787	354-410	67-99	137-178	140-176	106-142	83-110	33-52	30-44	63-88
Type+	1	713-793	355-412	57-86	137-179	144-181	107-143	82-109	31-50	31-45	62-86
Type+	1,5	715-794	356-413	55-83	138-179	145-181	107-143	82-108	30-48	31-44	60-84
Individual	0,1	672-749	366-423	125-163	152-193	127-161	108-142	82-110	38-58	32-47	59-83
Individual	0,5	709-787	362-417	71-102	144-185	145-181	111-145	85-112	34-53	32-46	56-80
Individual	1	713-791	362-417	60-90	145-186	148-183	112-147	85-111	32-51	31-45	53-77
Individual	1,5	714-792	363-418	58-87	146-187	149-184	113-148	84-111	31-49	31-45	53-77

Table 7: The first model is a varying-intercepts model, allowing partial pooling across StudentTypes. The second model adds StudentType-level predictors to explain differences between groups. The final model incorporates individual-level variables to capture student-specific behavior.

Table 7 presents enrolment forecasts for the 2025/2026 academic year across all three models: Basic, Type+, and Individual. These predictions are based solely on data from the 2024/2025 cycle and include 95% credible intervals. The table also shows results for different values of the hyperparameter σ, which controls the degree of pooling in the hierarchical model. The first row (ESC) displays the manual estimates provided by FEB staff. Compared to the models, these estimates have much narrower ranges.

In the previous academic year, Data Science & Business Analytics (BAN) behaved as an outlier, and this characteristic strongly influences how the forecast changes with σ. A higher σ value (e.g., 1.5) assumes BAN will continue to deviate from the average behaviour, resulting in less shrinkage toward the global mean and thus more independent predictions. Conversely, a lower σ (e.g., 0.1) implies stronger pooling, treating BAN more like the typical program in the dataset.

The sensitivity of BAN's predictions to different σ values underscores the uncertainty surrounding whether the programme will continue to behave atypically in the upcoming year. In the absence of actual enrolment outcomes for 2025/2026, it is not possible to determine definitively which level of pooling will yield the most accurate forecast. This uncertainty supports the consideration of multiple σ values and the combination of their predictions to produce more robust and flexible estimates. However, if a single value must be selected, a σ of 0.5 is recommended to limit the risk of overfitting by encouraging partial pooling toward the overall average. Although not definitively optimal, the LOOCV results indicate that this value performs reasonably well across programmes.

[9] dedicated group of FEB staff manual estimates

5 Discussion

This study addresses key limitations in existing enrolment forecasting methods in the Netherlands by introducing a segmentation framework based on nationality and Dutch higher education history. The resulting StudentType-based model offers two major advantages: it reduces noise in the data by grouping applicants with similar enrolment behaviour, and it improves interpretability, making the model more accessible to decision-makers who often rely on experience-based heuristics. Categories such as *uva_bachelor* and *EEA_new_in_Dutch_Higher_Education* resonate with institutional knowledge, increasing transparency and trust in the forecasts.

These improvements directly respond to challenges identified in the literature. The Radboud model, for example, leverages extensive historical data from Studielink and internal systems, but lacks segmentation beyond broad nationality groups. Meanwhile, the models proposed in Mosterd (2022), Vonk (2022), and Wirken (2022) lacked access to such data and were limited in their ability to capture applicant heterogeneity.

The Bayesian framework further strengthened the model's applicability by introducing partial pooling through hierarchical structures, allowing the model to generalize across StudentType while still capturing StudentType-specific patterns. This was particularly valuable in settings with sparse data, such as smaller programs or rare StudentTypes. Moreover, the inclusion of hyperparameters such as σ, which governs the extent of pooling between programs, allowed the model to flexibly adapt to program-specific deviation from the overall trend. The results showed that different programs benefited from different degrees of pooling, emphasising the need for adaptable rather than one-size-fits-all solutions.

Despite the Individual model included applicant-level features such as "Looking Elsewhere", the additional predictive benefit over the simpler Type+ model was modest. This indicates that much of the signal relevant for enrolment forecasting is captured at the StudentType level, highlighting the utility of well-designed group-level predictors in noisy application environments.

6 Conclusion

This research aimed to determine whether a robust and adaptable forecasting model could be developed to predict enrolment numbers at the FEB, while accounting for the data noise introduced by multiple applications and uncertain commitment. Additionally, it examined whether such a model could serve as a foundation for a dynamic, real-time forecasting tool that supports institutional planning.

The results suggest that a Bayesian multilevel model combined with StudentType segmentation provides a promising approach for generating interpretable and nuanced forecasts without relying on extensive historical data. The hierarchical structure allows flexible pooling across diverse master's programs, meanwhile StudentTypes improve transparency and reduces noise. Additionally, using StudentTypes enables early forecasting before the admission cycle completes by estimating enrolment probabilities and projecting applicant numbers based on partial data. Although this approach fa-

cilitates timely decision-making, it remains conceptual and would require further development and validation.

While further validation is needed to assess robustness across academic years, the model already offers practical value. It provides FEB staff with a structured, data-driven comparison to complement and possibly refine experience-based forecasts. As the dataset grows, the approach can be developed into a dynamic forecasting dashboard to support ongoing decision-making.

7 Limitations and future research

The main limitation of the research was the lack of multiple years of historical data. The model was trained and evaluated using data from only a single academic cycle (2024/2025), and the enrolment outcomes for the forecasted cycle (2025/2026) are not yet available. As a result, external validation of the model's predictive accuracy will only be possible once future outcomes are known. This constrains the current evaluation to retrospective cross-validation and internal performance metrics.

Future research can take many directions. This includes refining the StudentType segmentation, improving the model itself, evaluating its dynamic performance over time, and conducting a more thorough comparison between this model and the Radboud model once more data becomes available. Additionally, further research could explore the predictive value of more specific enrolment-related variables, such as application timing and financial aid behaviour

This thesis acknowledges that other universities in the Netherlands may be interested in developing similar enrolment forecasting models. Therefore, a key focus of this thesis was to thoroughly understand and prepare the Studielink sleutelbestanden data, a task that proved essential. To best support ongoing and future research in this area, the code developed during this project has been shared publicly on GitHub[10]. It is hoped that this resource will provide a solid foundation for other Dutch universities and researchers working with sleutelbestanden data to build upon and further improve enrolment forecasting models.

8 Acknowledgments

I would like to thank the individuals and institutions who supported the completion of this thesis:

- Drs. F.H.K. Pope and the University of Amsterdam for providing the data and resources used in this research, and for their commitment to supporting the project throughout.
- Dr. I.M. Zwetsloot and Dr. S.T. Mol for their supervision and helpful feedback throughout the thesis process.
- Roel Lust for sharing ideas and insights during the development of this work. His related research can be found in (Lust, 2025).

[10] https://github.com/MaartenH327/forecasting-enrollment-dutch-universities

References

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

Correa-Álvarez, C. D., Salazar-Uribe, J. C., & Pericchi-Guerra, L. R. (2023). Bayesian multilevel logistic regression models: A case study applied to the results of two questionnaires administered to university students. *Computational Statistics*, *38*(4), 1791–1810.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Hamers, Y. (2017). *Predicting student enrollment logistic regression on attended marketing events* [Master's Thesis]. Tilburg University.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd). Springer.

Hsu, S. (2022). *Long-term prediction of master student influx in computing science* [Master's Thesis]. Utrecht University.

Leven, N. v. (2022). *Long-term prediction of master student influx in programs artificial intelligence and applied data science* [Master's Thesis]. Utrecht University.

Lust, R. (2025). *Dutch higher education: Predicting bachelor enrolment* [Master's Thesis]. University of Amsterdam.

Mao, S., Zhang, C., Song, Y., Wang, J., Zeng, X.-J., Xu, Z., & Wen, Q. (2024). Time series analysis for education: Methods, applications, and future directions. *arXiv preprint arXiv:2408.13960*.

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Chapman; Hall/CRC.

Mosterd, T. (2022). *Prediction of master student influx in the faculty of science*.

Nuffic. (2025). *Inkomende diplomamobiliteit in het hbo en wo 2024–25* (Rapport No. NUF2025/04) (Auteurs: Jonatan Weenink, Anneloes Slappendel-Henschen, Ece Arat, Saoradh Favier. Redactie: Acolad. Opmaak: Osage. Fotografie: Marit Hazebroek.). Nuffic.

Rijksoverheid. (2024). Hogeronderwijs vraag & antwoord [Accessed: 2025-07-11]. https://www.rijksoverheid.nl/onderwerpen/hoger-onderwijs/vraag-en-antwoord/met-welke-diploma-s-kan-ik-naar-de-universiteit-of-hogeschool

Slim, A., Hush, D., Ojah, T., & Babbitt, T. (2018). Predicting student enrollment based on student and college characteristics. *International Educational Data Mining Society*.

Studiekeuze123. (2025). Hogeronderwijs vraag & antwoord [Accessed: 2025-07-11]. https://www.studiekeuze123.nl/wat-ga-jij-kiezen/hbo-naar-wo

Studielink documentation. (2022). Programma van levering telbestand studielink, versie: Definitief 2.8 [Unpublished internal document].

Studielink documentation. (2023). Koppelvlak a versie 5, versie: Definitief 4.3 [Unpublished internal document].

StudyinNL. (2025). How to apply [Accessed: 2025-07-11]. https://www.studyinnl.org/plan-your-stay/how-to-apply

van den Hurk, T. (2017). *Predicting student conversion using machine learning techniques* [Master's Thesis]. Tilburg University.

Vonk, S. (2022). *Prediction of master student influx in the faculty of science.*

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.

Wirken, T. (2022). *Prediction of master student influx: Faculty of science (adsm-hcim-gmte).*

9 Appendix

```
typestudentmap= {
    1:"nieuw_in_hoger_onderwijs",
    2:"hbo_instroom",
    4:"universiteit_switcher",
    5:"studie_switcher",
    6:"herinschrijver",
}
df_SL["StudentType"] = df_SL["1cHO_K"].map(typestudentmap)

def StudentTypeExtra(row):
    # if isinstance(row["MSc_MQP"],float) == False: # of het nan is
    #     return "MQP"

    if row["AcademicLevel"] == "M" and row["StudentType"] =="studie_switcher" and row["1cHO_L"] == 3:
        return "uva_bachelor_student"

    if row["AcademicLevel"] == "M" and row["StudentType"] =="studie_switcher" and row["1cHO_L"] == 4:
        return "uva_master_student"

    if row["AcademicLevel"] == "M" and row["StudentType"] =="studie_switcher" and row["1cHO_L"] == 5:
        return "uva_master_student"

    if row["AcademicLevel"] == "B" and row["StudentType"] =="studie_switcher" and row["1cHO_L"] == 3:
        return "uva_master_student"

    if row["AcademicLevel"] == "B" and row["StudentType"] =="studie_switcher" and row["1cHO_L"] == 4:
        return "uva_bachelor_student"

    if row["AcademicLevel"] == "B" and row["StudentType"] =="studie_switcher" and row["1cHO_L"] == 5:
        return "uva_bachelor_student"

    if row["AcademicLevel"] == "M" and row["StudentType"] =="universiteit_switcher" and row["1cHO_L"] == 3:
        return "nietuva_bachelor_student"

    if row["AcademicLevel"] == "M" and row["StudentType"] =="universiteit_switcher" and row["1cHO_L"] == 4:
        return "nietuva_master_student"

    if row["AcademicLevel"] == "B" and row["StudentType"] =="universiteit_switcher" and row["1cHO_L"] == 3:
        #Some students apply for a bachelor's program after completing a master's degree.
        return "nietuva_master_student"

    if row["AcademicLevel"] == "B" and row["StudentType"] =="universiteit_switcher" and row["1cHO_L"] == 4:
        return "nietuva_bachelor_student"

    #Below are only students who are new to higher education.
    if row["StudentType"] !="nieuw_in_hoger_onderwijs":
        return row["StudentType"]

    if row["Herkomst"] == "N":
        return "NL_nieuw_in_hoger_onderwijs"

    if row["Herkomst"] == "E":
        return "EER_nieuw_in_hoger_onderwijs"

    if row["Herkomst"] == "R":
        return "rest_nieuw_in_hoger_onderwijs"

    if row["Herkomst"] == "0":
        return "rest_nieuw_in_hoger_onderwijs"

    return row["StudentType"]

#If a student has applied for both a master's and a bachelor's program
def ChooseStudentType(studentTypes):
    _studentTypes = set(studentTypes)
    _studentTypes.discard("herinschrijver")

    if len(_studentTypes)==0:
        return "herinschrijver"

    if len(_studentTypes)==1:
        return list(_studentTypes)[0]

    if('nietuva_master_student' in _studentTypes):
        return "nietuva_master_student"

    if('uva_master_student' in _studentTypes):
        return "uva_master_student"

    return _studentTypes
```

Listing 1: StudentType Feature Engineering based on Studielink Sleutelbestaden

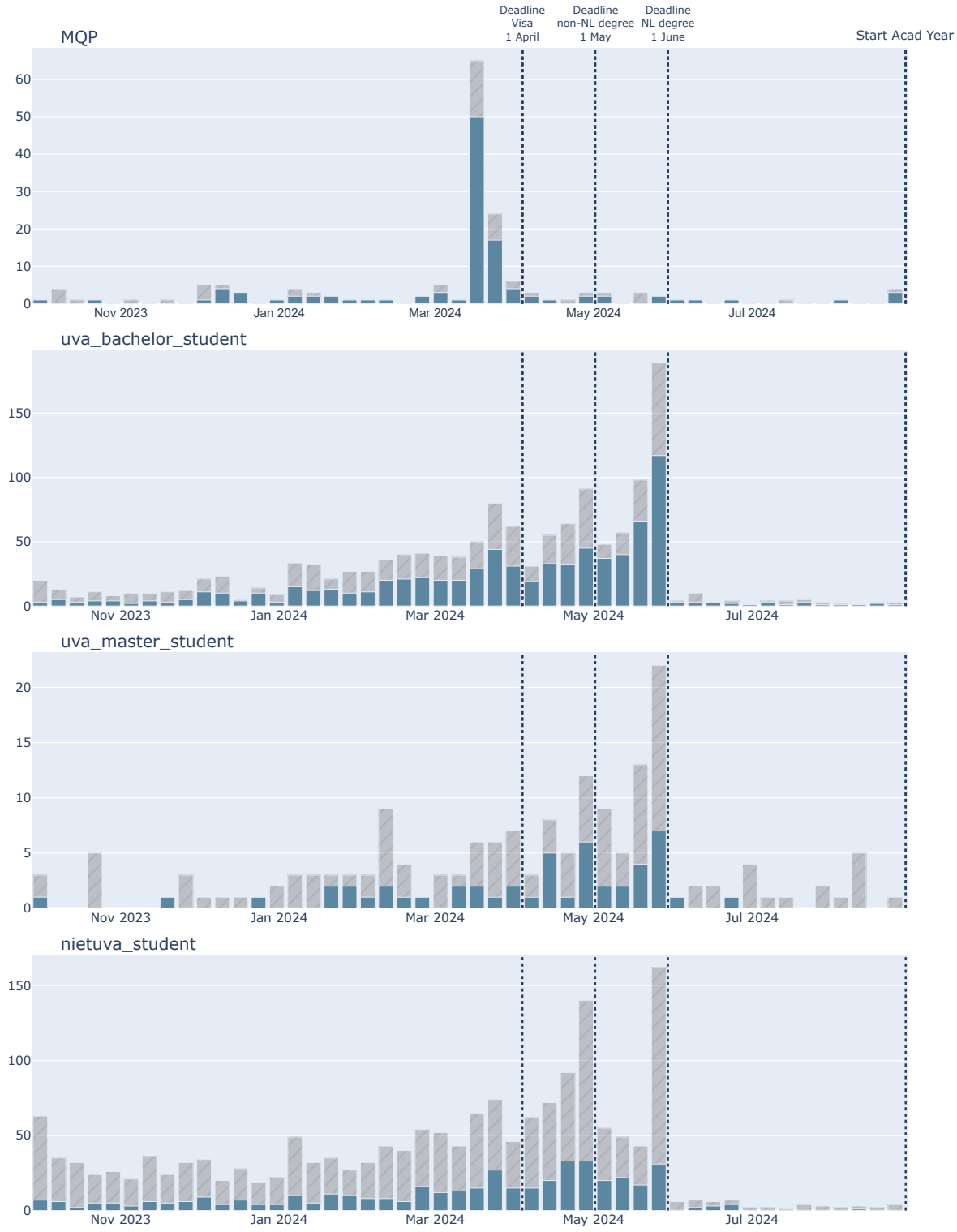

Figure 7: Application dates for different StudentTypes.

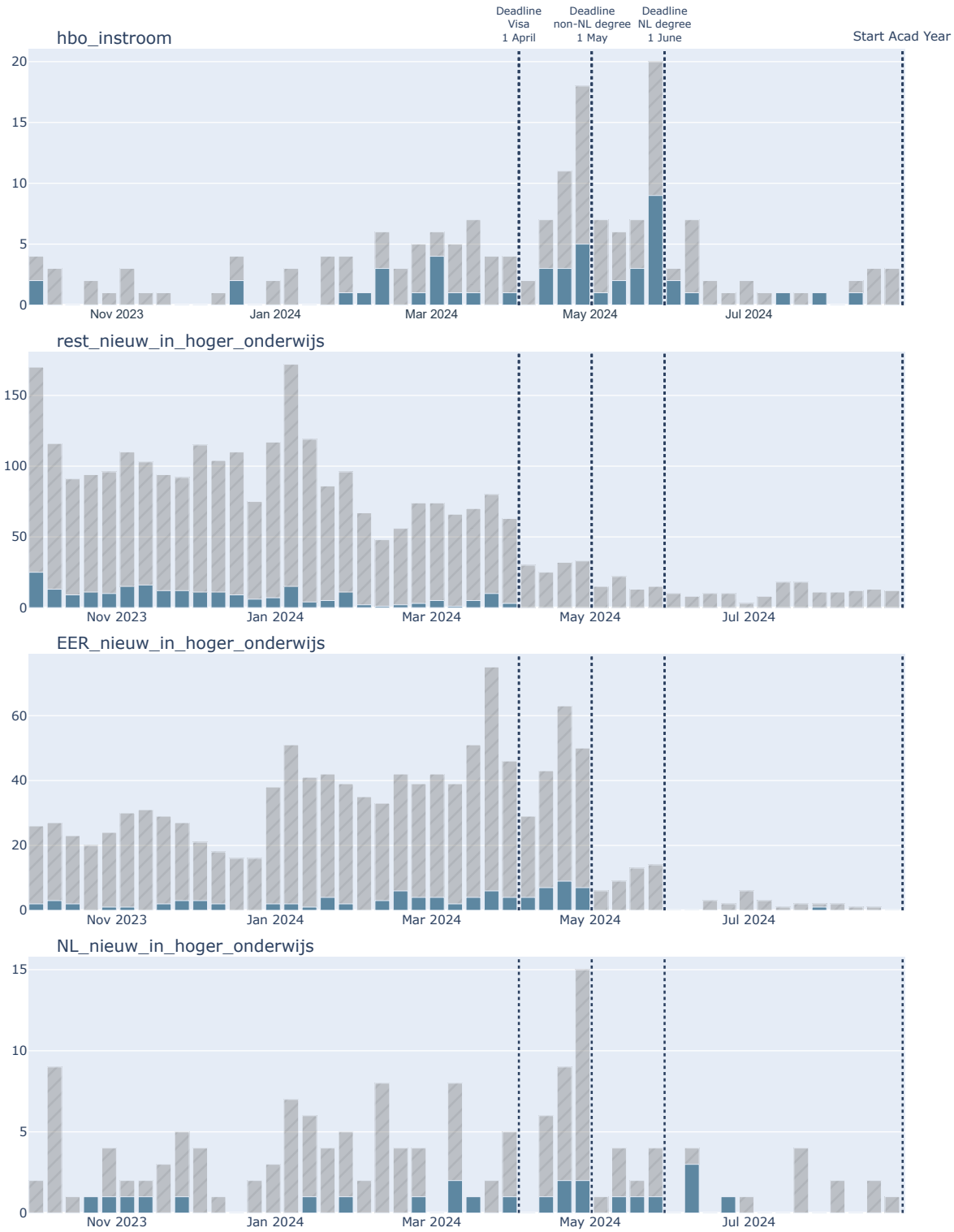

Figure 8: Application dates for different StudentTypes.