**Studiju programma**

RĪGAS ZIEMEĻVALSTU UNIVERSITĀTE
(RNU)

# BAKALAURA DARBS

Students:

Darba vadītājs:

**Study Programme**

# RIGA NORDIC UNIVERSITY
# (RNU)

## BACHELOR'S THESIS

Student:

Supervisor:

# Abstract

Šis bakalaura darbs analizē augstākās izglītības iestāžu uzņemšanas prognozēšanu ASV, izmantojot IPEDS datus (2010–2021), kas ietver 86 798 iestāžu-gadu novērojumus. Pētījums novērtē naivo persistenci, slīdošos vidējos, ARIMA, Ridge regresiju un Random Forest metodes, izmantojot soli-pa-solim validāciju. Galvenais atklājums ir tas, ka uzņemšanai ir ārkārtīgi augsta persistences pakāpe (98% no dispersijas izskaidro aizkavētās vērtības), kas nozīmē, ka vienkāršie bāzes modeļi pārspēj sarežģītās mašīnmācīšanās metodes. Vienkāršais naivās persistences bāzes modelis sasniedz vidējo absolūto kļūdu (MAE) 39,43 studentiem, salīdzinot ar 40,33 (Ridge) un 41,21 (Random Forest). Paneļu regresija parāda, ka uzņemšanas un pieejamības faktori ir statistiski nozīmīgi; tomēr to praktiskā ietekme ir ierobežota dominējošās persistences dēļ. COVID-19 pandēmijas stresa tests atklāj, ka vienkāršie modeļi ir noturīgāki. Rezultāti apšauba mašīnmācīšanās izmantošanu izglītības analītikā un iesaka vienkāršību operatīvajā prognozēšanā, kas var būt noderīga institucionālajiem pētījumiem un stratēģiskajai plānošanai.

**Atslēgvārdi:** Uzņemšanas prognozēšana, Augstākās izglītības analītika, Mašīnmācīšanās, Laika rindu analīze, IPEDS dati, Persistences modeļi, Institucionālie pētījumi, Prognozējošā analītika, Ridge regresija, Random Forest, ARIMA modeļi, Paneļu datu analīze, Soli-pa-solim validācija, Uzņemšanas pārvaldība, Universitāšu plānošana, Izglītības datu iegu, Studentu pieņemšana, COVID-19 augstākā izglītība, Bāzes prognozēšana, Pieejamības rādītāji, Uzņemšanas piltu, Uzņemšanas virzītājspēki, Finansiālās palīdzības ietekme, Institucionālā kapacitāte, Modeļu salīdzinājums, Prognozes precizitāte, Stratēģiskā uzņemšanas plānošana, Pēcvidusskolas izglītība, Datu vadīta lēmumu pieņemšana, Uzņemšanas persistence

## Abstract

This thesis analyzes the enrollment forecasting of higher education institutions in the US by utilizing IPEDS data (2010–2021), which includes 86,798 institution-year observations. The research evaluates naive persistence, moving averages, ARIMA, Ridge regression, and Random Forest methods through walk-forward validation. The main discovery is that enrollment has an extremely high degree of persistence (98% of the variance is explained by the lagged values), which results in simple baseline models outperforming sophisticated machine learning techniques. The simple naive persistence baseline attains an average MAE of 39.43 students, whereas it is 40.33 (Ridge) and 41.21 (Random Forest). Panel regression shows that admissions and affordability factors are statistically significant; however, their practical impact is limited by the dominance of persistence. The COVID-19 pandemic stress test reveals that simple models are more resilient. The results question the use of machine learning in education analytics and suggest simplicity in operational forecasting, which can be useful for institutional research and strategic planning.

**Keywords:** Enrollment Forecasting, Higher Education Analytics, Machine Learning, Time Series Analysis, IPEDS Data, Persistence Models, Institutional Research, Predictive Analytics, Ridge Regression, Random Forest, ARIMA Models, Panel Data Analysis, Walk-Forward Validation, Enrollment Management, University Planning, Educational Data Mining, Student Recruitment, COVID-19 Higher Education, Baseline Forecasting, Affordability Indicators, Admissions Funnel, Enrollment Drivers, Financial Aid Impact, Institutional Capacity, Model Comparison, Forecast Accuracy, Strategic Enrollment Planning, Postsecondary Education, Data-Driven Decision Making, Enrollment Persistence

**Keywords:** Enrollment Forecasting, Higher Education Analytics, Machine Learning, Time Series Analysis, IPEDS Data, Persistence Models, Institutional Research, Predictive Analytics, Ridge Regression, Random Forest, ARIMA Models, Panel Data Analysis, Walk-Forward Validation, Enrollment Management, University Planning, Educational Data Mining, Student Recruitment, COVID-19 Higher Education, Baseline Forecasting, Affordability Indicators, Admissions Funnel, Enrollment Drivers, Financial Aid Impact, Institutional Capacity, Model Comparison, Forecast Accuracy, Strategic Enrollment Planning, Postsecondary Education, Data-Driven Decision Making, Enrollment Persistence

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Listings

# Chapter 1

# Introduction

## 1.1 Background and Rationale

Enrollment serves as the main channel of communication between universities and the social system, influencing institutional capacity, finances, and strategic direction. Credible enrollment forecasting has become a major issue in the United States over the last ten years as different institutions face varying demographic pressures, changing student preferences, and heightened competition. For public universities reliant on state funding that is often tied to enrollment metrics, inaccurate forecasts can trigger budgetary crises. For private institutions managing financial aid budgets and tuition revenue, enrollment volatility threatens fiscal sustainability and strategic planning.

Forecasting in this context serves two different but related purposes. One is a predictive function: generating quantitative expectations about future enrollment levels to guide operational decisions such as faculty hiring, residence hall capacity planning, and course scheduling. The second is an analytical function: understanding which institutional and external factors drive enrollment changes, enabling administrators to identify levers for strategic intervention. Both are needed for a defensible decision-support approach. Accuracy is crucial since institutional budgets, staffing decisions, and capacity investments depend on enrollment projections. However, accuracy alone is insufficient if the forecasting process lacks transparency or if decision-makers cannot understand which factors contribute to the predictions.

Moreover, the research goes far to consider data realism as a fundamental constraint. Datasets gathered by institutional researchers are not experimental but administrative, meaning they contain missingness, measurement error, duplicates, and structural breaks. Forecasting models must be robust to these realities rather than assuming clean, idealized data. This pragmatic orientation ensures that the methodology developed here can be implemented by practitioners without requiring specialized data infrastructure or extensive data cleaning resources beyond what is typically available in university offices.

## 1.2 Research Problem

There is a wide range of public data available on U.S. postsecondary education through sources such as the Integrated Postsecondary Education Data System (IPEDS) and the College Scorecard. However, an alarmingly small proportion of institutions employ data-driven enrollment forecasting methods systematically. Many rely on historical averages, simple trend extrapolation, or subjective judgment, which can produce substantial forecast errors and limit strategic planning effectiveness. The gap between data availability and analytical practice represents an opportunity for methodological contributions that bridge theoretical forecasting techniques and operational institutional research.

One major methodological challenge is that enrollment figures usually show a lot of persistence. In many cases, year-to-year correlation in enrollment exceeds 0.95, meaning that next year's enrollment is highly predictable from this year's enrollment alone. This high autocorrelation creates a strong baseline benchmark: any forecasting model must beat the naive assumption that "next year equals this year" to justify its complexity. The persistence pattern also implies that the marginal contribution of additional predictors beyond lagged enrollment may be small, posing a challenge for model improvement.

Meanwhile, through policies and behaviors that are embedded in the admissions funnel and pricing and financial aid strategies, institutions actively manage their enrollment outcomes. Variables such as the number of applications received, admissions offered, average grant aid, and net price reflect both demand-side pressures and supply-side institutional responses. Disentangling the predictive power of these variables from the confounding influence of institutional strategy requires careful modeling and transparent interpretation.

The research problem is thus a matter of both prediction and explanation: what are the ways to effectively forecast institution-level enrollment, and which factors drive enrollment demand after controlling for persistence? Addressing this dual problem requires integrating forecasting techniques (such as time-series models and machine learning regression) with driver analysis (such as panel regression with fixed effects) to produce both accurate predictions and interpretable insights into the mechanisms behind enrollment changes.

## 1.3 Aim and Objectives of the Study

The researchers aim to work out, carry out, and frontline test the empirical forecast model of university enrollment as well as to systematically find out its drivers using administrative panel data. This aim encompasses both methodological rigor (ensuring that forecasts are evaluated using proper temporal validation protocols and benchmarked against strong baselines) and practical relevance (producing insights that institutional researchers and enrollment managers can implement without specialized expertise).

To achieve this aim, the study pursues the following objectives:

[leftmargin=1.2cm]By combining directory data, institutional features, enrollment funnel metrics, cost of attendance statistics, and student aid data from IPEDS and the College Scorecard, construct a comprehensive panel dataset spanning 2010 to 2021 that captures both cross-sectional heterogeneity and temporal dynamics. Define and justify an outcome variable for enrollment demand forecasting, focusing particularly on total first-time enrollment as a key flow variable that directly reflects the admissions funnel and is actionable for institutional planning. Conduct exploratory data analysis for the characterization of distributional properties, temporal trends, missing data patterns, and correlations among key variables, establishing empirical regularities that inform model specification. Create baseline forecasting models, for example, naive persistence and moving-average smoothers, and use them as benchmarks to evaluate whether more complex approaches provide meaningful improvements in out-of-sample accuracy. Model a structured time-series specification for the segment-level aggregate series to find out whether ARIMA models can capture national or sectoral enrollment trends that may not be visible at the institution level. Use interpretable regressions with robust standard errors and year effects to estimate associations between enrollment demand and potential drivers such as admissions funnel variables (applications, admissions), affordability indicators (net price, grant aid), and institutional characteristics. Use the results as a basis for both methodological and practical recommendations, including guidance on model selection, the importance of walk-forward validation, the role of baseline benchmarking, and the interpretation of driver effects in observational settings.

## 1.4 Object and Subject of the Research

This research is focused on the annual enrollment demand in U.S. higher education institutions. Demand is operationalized as total first-time enrollment in the fall term, which represents the number of new students matriculating for the first time at an institution. This measure is more directly influenced by admissions and recruitment strategies than total headcount, which includes continuing students and reflects retention patterns. By focusing on first-time enrollment, the study isolates the demand-side dynamics that are most relevant for enrollment forecasting and strategic intervention.

This research subject is a group of statistical patterns that describe how enrollment demand is related to lagged enrollment, admissions funnel metrics, affordability indicators, and institutional characteristics across time and institutions. The analysis employs panel data methods to exploit both within-institution temporal variation and cross-institutional heterogeneity. The emphasis is on identifying predictable patterns that can improve forecasting accuracy and on understanding which factors are associated with enrollment demand after controlling for persistence and aggregate time trends.

## 1.5 Research Questions and Hypotheses

The study is structured around two research questions:

**RQ1:** Given a walk-forward evaluation setup that eliminates temporal leakage, which forecasting methods (baseline persistence, moving averages, ARIMA models, or machine learning panel regressions) produce the most accurate institution-level enrollment predictions as measured by mean absolute error (MAE) and root mean squared error (RMSE)?

**RQ2:** By controlling for persistence and year effects, which university and affordability-related factors (admissions funnel metrics such as applications and admissions; affordability indicators such as net price and grant aid) demonstrate stable statistical associations with enrollment demand in a panel regression framework?

Two hypotheses are evaluated:

**H1:** Forecasting models with time-series structures or lagged predictors would do better than a naive persistence baseline by at least 10% in terms of MAE or RMSE when evaluated on out-of-sample data using walk-forward validation. This hypothesis reflects the expectation that incorporating additional information beyond last year's enrollment should improve forecast accuracy meaningfully.

**H2:** Admissions funnel metrics and affordability indicators such as applications, admissions, net price, and grant aid will show statistically significant associations ($p < 0.05$) with enrollment demand in a panel regression that controls for lagged enrollment and year fixed effects. This hypothesis tests whether these variables contribute explanatory power beyond simple persistence, which would justify their inclusion in forecasting models and support their use as strategic levers in enrollment management.

## 1.6 Scope, Assumptions, and Limitations

The analysis is done at the institution-year level of the annual unit over the years 2010 to 2021. This temporal scope encompasses diverse macroeconomic and demographic contexts, including the aftermath of the Great Recession, a period of relative enrollment stability, and the onset of the COVID-19 pandemic in 2020. The dataset includes degree-granting postsecondary institutions in the United States that report to IPEDS, covering public, private nonprofit, and private for-profit sectors across various Carnegie Classifications. This broad institutional coverage ensures that findings are representative of the diversity in U.S. higher education, though segmented analyses by sector and selectivity are also conducted to assess heterogeneity in model performance.

The empirical design is underpinned by three assumptions. The first assumption is that the study considers historical patterns that are useful for short-term forecasting.

It is reasonable to assume that enrollment processes exhibit sufficient temporal stability that models trained on past data can generate accurate one-year-ahead forecasts. This assumption breaks down during unprecedented external shocks, such as the COVID-19 pandemic, which is explicitly examined as a robustness check. The second assumption is that administrative data from IPEDS, while imperfect, are sufficiently accurate and complete for forecasting purposes after appropriate cleaning and imputation. The third assumption is that associations estimated from observational data reflect predictive relationships rather than causal effects. The study does not claim that interventions based on the identified drivers will produce the same effects observed in the regression, as endogeneity and omitted variable bias are unavoidable in administrative panel data.

There are a few limitations to this research. In public higher education datasets, there are also missing data, measurement error, and reporting inconsistencies. Missing enrollment data are not random but systematically higher among smaller institutions, proprietary schools, and those not participating in federal student aid programs. This missingness may introduce selection bias if omitted institutions differ systematically in their enrollment dynamics. The study addresses this limitation by restricting the analysis to institutions with sufficient temporal coverage and documenting missing data patterns transparently.

Another limitation is the heterogeneity of institutions. Universities vary in terms of sector, mission, selectivity, geography, and size, and these differences may moderate the effectiveness of forecasting models and the strength of driver associations. While the main analysis pools all institutions to maximize sample size and statistical power, segmented analyses by sector and selectivity are conducted as sensitivity checks. However, sample sizes within some segments may be insufficient to detect smaller effects.

Lastly, the evaluation is one of the history. Instead of using real-time deployment, the research uses walk-forward validation on historical test years. This approach simulates operational forecasting conditions but does not account for potential changes in data structures or institutional behavior after 2021. Extensions of this work could monitor model performance prospectively to assess whether historical patterns continue to hold in future periods.

## 1.7 Methods Overview

The methodology integrates data preparation, exploratory analysis, forecasting, and driver analysis in a cohesive framework. Data preparation involves merging multiple IPEDS survey components, handling missing values, deduplicating records, and creating lagged variables that respect temporal ordering. Exploratory data analysis characterizes distributional properties, temporal trends, and bivariate correlations, providing empirical motivation for modeling choices.

Forecasting methods include baseline models (naive persistence and moving averages), ARIMA models applied to aggregate time series, and machine learning panel regressions (Ridge regression and Random Forest) that incorporate admissions funnel and affordability variables. All models are evaluated using walk-forward validation, which splits the data temporally into training and test sets such that forecasts are made only using information available prior to the forecast date. Performance is assessed using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

Driver analysis employs panel regression with year fixed effects and robust standard errors clustered by institution. The dependent variable is total first-time enrollment in year $t$, and predictors include lagged enrollment, lagged admissions funnel metrics (applications, admissions), and lagged affordability indicators (net price, grant aid). Regression coefficients are interpreted as associations controlling for persistence and common time trends, not as causal effects. Feature importance from Random Forest models complements the regression analysis by identifying which predictors contribute most to reducing forecast error.

## 1.8 Expected Outcomes and Contributions

The study expects to produce several empirical and methodological contributions. Empirically, the research will establish whether sophisticated forecasting models (ARIMA, Ridge, Random Forest) meaningfully outperform simple baselines (naive persistence, moving averages) for institution-level enrollment. If complex models do not consistently beat the baseline, this finding would support the use of parsimonious approaches in operational settings and caution against over-reliance on "black box" machine learning methods without rigorous backtesting.

The driver analysis is expected to reveal which admissions funnel and affordability variables show stable associations with enrollment demand after controlling for persistence. If variables such as applications and admissions exhibit significant coefficients, this would justify their inclusion in forecasting models and support data-driven enrollment management strategies. Conversely, if affordability indicators such as grant aid show weak or inconsistent associations, this would suggest that price effects are confounded by institutional strategy or that measurement challenges limit their predictive utility.

Methodologically, the research contributes by demonstrating the importance of walk-forward validation for time-ordered data, establishing strong baseline benchmarks, and integrating forecasting accuracy with driver interpretation. The study also documents data preparation procedures in sufficient detail that institutional researchers can replicate the methodology using their own IPEDS downloads, enhancing the practical applicability of the findings.

From a policy and practice perspective, the findings will inform enrollment management by clarifying which forecasting methods are most accurate, which predictors are most informative, and how institutions should interpret and act on forecast outputs. The study will also highlight the limitations of forecasting in the presence of external shocks, such as the COVID-19 pandemic, and recommend scenario planning as a complement to point forecasts.

## 1.9 Scientific Novelty and Practical Significance

The article is methodological and integrative in its contribution. At first, it describes a reproducible procedure of generating an institution-year panel from various public sources as well as giving a detailed explanation of how duplicates and missingness are dealt with. It is important because most forecasting models research papers are based on the assumption that data is sourced from a clean environment whilst, in fact, institutional analysts verify and harmonize data extensively. By carefully documenting and applying the preprocessing rules, the paper closes the gap between the theoretical models and data issues in the real world.

Second, the research emphasis is on walk-forward validation instead of random train-test splits for time-ordered data. The selected approach here is a great help in getting rid of the typical optimistic bias associated with random train-test splits in time-ordered data. Besides, the paper is very much into establishing a firm baseline for comparison. In the context of enrollment forecasting, persistence is generally referred to as the strongest predictor, and thus, models that do not surpass a naive baseline might not be justified for their complexity. Making this benchmark prominent allows for a clearer understanding of what comes as a real methodological advancement.

Next, the article connects prediction with the assessment of drivers. Rather than presenting factors as an entirely different regression analysis, the study explores which elements contribute to enhancing forecast accuracy and their stability after controlling for persistence and year effects. This comprehensive perspective is more in line with a manager's understanding. For example, it emphasizes whether affordability measures give extra predictive signals over and above the admissions funnel metrics, thus answering questions of significance for strategic planning.

The practical significance is understandable directly. The framework can be applied as a decision support tool by admissions and enrollment management, providing transparent baseline expectations and clear diagnostics of where the additional modeling effort is most lucrative. For admission offices in particular, this means real operational benefits: enrollment managers can establish realistic recruitment goals based on historical data rather than purely aspirational guesses, efficiently allocate marketing budgets by targeting segments where the interventions have a real effect, and report enrollment forecasts to university leadership with quantified confidence intervals.

The methodology allows the representative institution to change from reactive enrollment management, where the institution accepts the enrollments that materialize, to proactive strategic planning, where data-driven forecasts guide resource allocation, faculty hiring, residence hall capacity planning, and financial aid budgeting long before the academic year. Besides, it makes responsible use possible by highlighting the limitations, not making causal claims, and encouraging segmentation and auditability. At universities where forecasting is obligatory but analysts' capacity is limited, a robust baseline together with clear backtesting can improve the planning discipline even in the absence of advanced modeling expertise.

## 1.10 Structure of the Thesis

The rest of the thesis is structured as follows. Chapter 2 provides a review of the enrollment forecasting and predictive analytics in higher education literature. It outlines the common modeling techniques, discusses the role of evaluation protocols in time-series prediction, and explains the limits to which driver analysis can be interpreted in observational settings.

Chapter 3 unveils the dataset construction journey, the variables used, preprocessing choices, and the modeling approach both for forecasting and determinant assessment.

Chapter **??** presents the results, including the main findings from the exploratory phase, the benchmark comparisons among forecasting models, as well as the outputs of regression and tree-based driver analyses.

Chapter **??** elaborates the institutional planning implications, states the limitations, and suggests ways of extending the framework such as bringing in more exogenous variables and the forecasting of headcount outcomes if necessary.

# Chapter 2

# Literature Review and Theoretical Framework

## 2.1 Conceptualizing Enrollment Demand and Institutional Planning

Enrollment demand refers to the number of students that both new and returning students are willing and able to participate in under the current demographic, economic, and institutional situations. In higher education administration, demand cannot be seen as a single physical measure. It is deduced from a series of administrative and behavioral indicators that illustrate the flow of students through the admissions and persistence pipeline, including inquiries, applications, admits, deposits, registrations, and eventually headcount or full-time equivalent enrollment. Each indicator reflects a different decision margin and is therefore associated with a different managerial problem [1, 2].

A major conceptual difference is that enrollment can be seen either as a stock or as a flow. A census headcount at a specified reporting date is a stock variable that results from current-student progression, retention, transfer, and stop-out patterns. New entrants, however, are a flow that is more directly affected by changes in the market and recruitment activities. In fact, forecasts of intake and headcount/full-time equivalents are just two examples of a more extensive range of forecasting tasks in practice. An intake forecast is a tool to plan for the coming academic year in terms of staff, housing, and course scheduling, while headcount and full-time equivalent forecasts are needed for multi-year budgeting and capacity planning [3, 4].

Economic theories of educational choice generally go human-capital way first by considering education as an investment and hence if expected returns must be greater than not only direct costs but opportunity costs as well. According to such a view, demand is affected by that expectations of earnings, probabilities of employment, and the arrangement of costs and benefits over time. However, studies of student decision-making done so far have consistently shown that information constraints and behavioral mechanisms play a significant role in the determination of the choices that are made. Students and families deal with complicated pricing schemes, they have guesses as to what their financial aid will be, and there are variations in the beliefs about the quality of different institutions that exist. These frictions, therefore, provide a rationale for the fact that even under similar macroeconomic conditions, institutions that appear to be extremely similar can exhibit different yield patterns [5, 6, 7, 8].

Higher education institutions do not simply respond to demand with a lag. They deliberately employ strategic tools such as tuition pricing, financial aid, and admission selectivity with a view to maximizing their budget and advancing their reputational goals, and these tools are frequently tweaked before the institution sees the need to exploit them. This co-determination entails a feedback loop: the enrollment results that are seen not only come from an external demand but also from the institutional response to the earlier demand signals. For forecasting, the theoretically most relevant view is the pragmatic one. The goal is to utilize the stable regularities of the joint system that result in accurate predictions at the decision-making time while being aware that the coefficients obtained from the analysis of observational administrative data generally reflect associations rather than causal effects [9, 10].

At last, customer needs are divided. Each of two-year colleges, regional public universities, and selective private institutions is different in marketing, serves other groups of students, and has different limitations. Community colleges usually react quickly to changes in the local economy and an increase in the number of adult learners, while selective institutions may be more limited by the number of available spots and admissions policy than by the changes in the number of applicants. So, a theoretical framework for enrollment forecasting represents the heterogeneity by sector, selectivity, mission, modality, and geography, and it either encourages stratified analysis or models that reveal differences across types of institutions [11, 12, 13].

## 2.2 Determinants of Enrollment Demand

The factors that influence enrollment demand are often categorized into demographic, economic, financial, institutional, and contextual ones. Demography determines the basic limits of participation. The number of young people in the traditional college-age cohort, migration trends, and adult learner participation together determine the pool of potential students. On a regional level, population change ranks among the most consistent factors of demand, but its influence depends on the extent of the institution's reach. Institutions whose student base comprises mainly their local commuting areas are more vulnerable to local cohort decreases than institutions with national recruitment or specialized program offerings [11, 2].

Enrollment is affected by economic conditions via opportunity cost and resource channels. A rise in job opportunities and wages means that the value of time used in education goes up, so the immediate enrollment of some groups might be reduced. On the other hand, in a recession, enrollment may go up as people reskill or postpone entering the labor market. The direction and extent of the correlation differ depending on age, level of credential, and local industry structure. To illustrate, short-cycle programs can be very quickly influenced by local unemployment, whereas the demand for long-cycle degrees may be also term less elastic and more influenced by expectations of longer-term returns [13, 14].

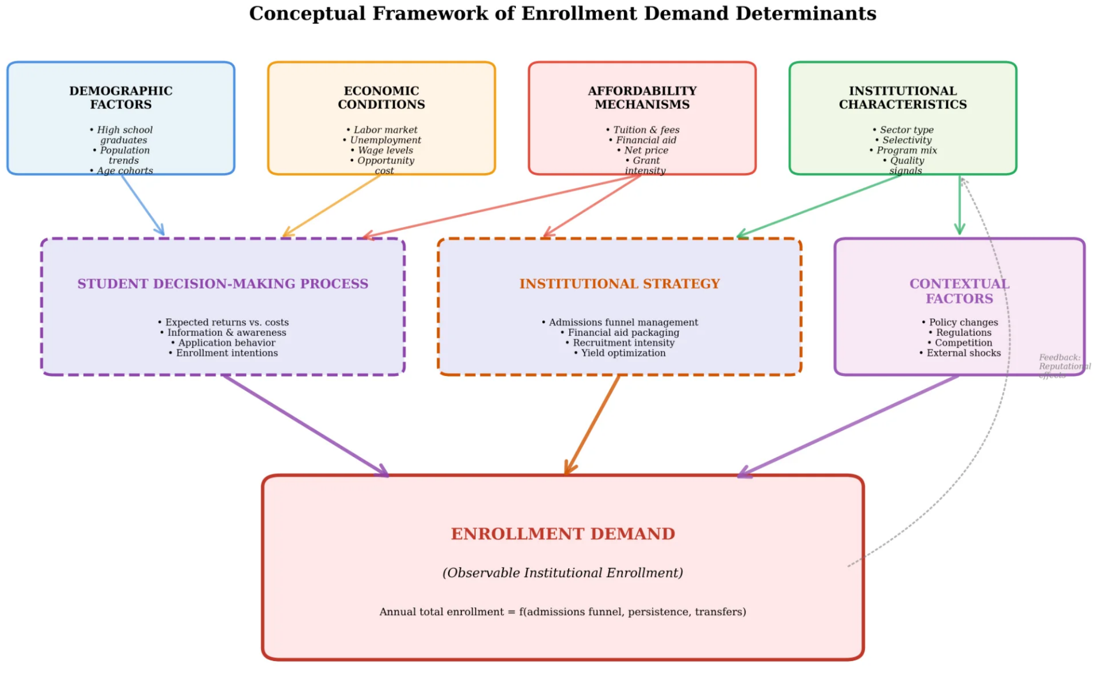
Affordable mechanisms refer to connecting prices, aids, and liquidity constraints with enrollment that has been observed. In theory, the net price after grants should be the price margin that is relevant. However, when the aid information is uncertain or difficult to understand, students may rely on the published tuition and fees. Need-based grants, merit aid, and state scholarship policies change the effective price distribution that students face and thus have both application volume and yield in their influence. On the institutional side, discounting decisions are limited by tuition dependence, endowment resources, and state appropriations. These issues lead to the use of financial aid and net price indicators in empirical demand models, usually lagged to indicate information availability at the time of enrollment decisions [7, 8, 15].

Institutional characteristics act as factors through which demand is influenced by perceived quality, program fit, and student support capacity. Control and sector are determining factors of mission differences, pricing regimes, and regulatory environments. Selectivity and admissions pressure measures are indicative of both applicant sorting and institutional constraints. Staffing and student-faculty ratios serve as a means of measuring instructional resources and could possibly be used to account for perceived academic experience as well. Besides that, program mix is a fundamental aspect: demand changes when institutions qualify in high-return fields or offer flexible modes such as online or hybrid delivery that minimize time and location restrictions [9, 16].

Contextual factors think about policy environments and shocks. State tuition caps, changes in need-based aid, and federal regulations can affect affordability and institutional strategy. Shocks like at public health can cause discontinuities in patterns of history by changing preferences, constraints, and labor-market trajectories. The theoretical implication here is that models are not required to measure every shock explicitly. Instead, models should be capable of handling instability through benchmarking, a parsimonious specification, and evaluation protocols that reveal how performance drops off when conditions differ [17, 18].

### 2.2.1 Conceptual Framework

Figure 2.1 depicts a conceptual framework that institutions rely on to forecast and plan around enrollment outcomes. The framework shows how different factors from the institutional, economic and contextual spheres have a direct or indirect influence on enrollment demand through several channels. These channels eventually come down to an observable enrollment outcome.



*Note: This framework synthesizes the theoretical perspectives from human capital theory, educational choice models, and institutional strategy literature reviewed in Sections 2.1-2.2. Solid arrows indicate direct pathways; dashed boxes represent intermediate processes; dotted feedback shows long-term reputational effects.*

**Figure 2.1:** Conceptual Framework of Enrollment Demand Determinants. This framework synthesizes theoretical perspectives from human capital theory, educational choice models, and institutional strategy literature. Solid arrows indicate direct pathways; dashed boxes represent intermediate processes; dotted feedback shows long-term reputational effects.

The model brings together the two sets of theories explored earlier in Sections 2.1 and 2.2. It illustrates how the demand for enrollment is a result of the coming together of student decision-making processes (affected by the socioeconomic conditions and money matters), institutional strategic actions (admission funnel management, pricing, financial aid), and external contextual factors such as the policy environment and demographic trends. The model basically conveys that enrollment is not just a matter of student preferences or institutional capacity but rather the result of the interplay of these various factors. This theoretical framework paves the way for the empirical study by showing which variables in administrative datasets can be used as indirect measures of the grounding theoretical concepts.

## 2.3 Administrative Data Sources and Measurement Considerations

Administrative data appeal to enrollment prediction experts because of their extensive coverage, compatibility between institutions, and the availability of consistent longitudinal identifiers. On the other hand, administrative datasets are a reflection of reporting conventions thus need to be cautiously interpreted. Strategic forecasting and benchmarking across sectors can be done with institution-level panels, but they lack the capturing of a school's internal heterogeneity for example program-level substitution, differing major retention rates, or the student subpopulations that have varying responses to the intervention. These limitations are considerable. They determine the types of questions that can be answered with great credibility and the kinds of conclusions that should be avoided [10, 19].

The Integrated Postsecondary Education Data System (IPEDS), a product of the National Center for Education Statistics (NCES) is the main federal data collection for college statistics and is widely used in empirical studies due to the fact that it standardizes reporting and offers stable identifiers. IPEDS segments cover admissions and enrollment data, institutional characteristics, and financial assistance-related indicators. The survey parts, however, may vary in their extent and relevance. The admissions variables are mainly aimed at institutions with selective admissions and might not include institutions with open admission or nonstandard reporting frameworks. In the same way, some financial aid variables are established for certain student groups or reporting years, and the timing of their release can be behind the decision context [19].

The U.S. Department of Education's College Scorecard highlights other factors related to results and affordability, for example, completion rates, and a few indicators tied to debt and earnings that come from it. These factors can help descriptive analysis to a greater extent if the link between demand and the returns as well as the institutional performance is established. However, at the same time, they pose certain difficulties: there are differences in cohort definitions, some cells are not disclosed in order to ensure privacy, and a number of outcome measures are basically reflecting past cohorts rather than present conditions. The key methodological aspect for forecasting is that of such variables, only lagged predictors can be inputs, and the interpretation has to take into account the temporal ordering between what students recall could be known to them at the time of decision and what is actually coming from retrospective administrative outcomes [20].

Measurement challenges are not limited to just missing data. For example, an institution can be merged, closed, or change its control status; there can be shifts in reporting practices; changes to the definitions can lead to discontinuities. Such activities can produce breaks of structure in time series and thus entail explicit data cleaning, which also includes deduplication and consistent aggregation to a single institution-year observation. Count variables usually also are very skewed, which results in transformations and the use of robust evaluation metrics that are not heavily influenced by a small number of large institutions [19, 21].

One of the common issues is that a few administrative records could be arranged by subgroups, e.g., sex or level of study, thus resulting in multiple institution-year entries. Developing an analysis-ready panel from this data calls for a well-founded aggregation procedure in line with the theoretical construct of interest. Suppose enrollment demand is reflected as an incoming cohort, then it would be consistent to aggregate subgroup counts to a total. If the interest is in composition, then perhaps subgroup-specific panels will be necessary. Hence, clear and thorough recording of aggregation, suppression handling, and variable definitions is not just an add-on to the modeling. This is a part of the theoretical operationalization of demand [19].

## 2.4 Forecasting Approaches for Enrollment Time Series

Forecasting methods vary in complexity from simple univariate extrapolation models to multivariate and structurally based models. Univariate methods base their prediction on the assumption that the future will be similar to the past and are thus necessary as benchmark models because enrollment time series usually have a very strong autocorrelation component. The naive forecast, which predicts that the next value will be the same as the last observed one, may be very hard to beat in the case of stable institutions only. Moving average forecasting techniques minimize the effect of a noise of a single year by smoothing the most recent observations, and they can be quite efficient when the underlying level changes slowly in comparison with the year-to-year variability [22, 23].

Exponential smoothing extends moving averages by giving geometrically decaying weights to the past observations. In the state-space form, smoothing methods can be seen as latent level and trend components, which evolve over time with stochastic disturbances. This approach is very suitable for annual enrollment as it allows the model to adjust systematically to rising or falling trends gradually while still being very simple. When the series are short, simplicity is not just a matter of style. It is actually a necessity because complicated models may simply fit the noise in historical data rather than the actual structure [24, 25, 22].

Autoregressive integrated moving average (ARIMA) models offer a very adaptable way of representing autocorrelation and stochastic shocks after differencing to get rid of the non-stationarity. The Box-Jenkins approach mainly focuses on diagnostic checking and simple order selection. However, when we talk about institution-level annual data series, the short time dimension restricts the usefulness of high-order ARIMA models. There is, therefore, a theoretically justifiable use of ARIMA in this situation mainly at a higher level, e.g., sector or regional totals, where the effective signal-to-noise ratio gets better and the time series may exhibit more stable autocorrelation structure [26, 22].

One of the frequently encountered practical problems in enrollment is a structural change. Institutions may change the way they select students, increase the number of online programs, or get a change in the law that will result in the enrollment of past students being less related to the future ones. Besides, exogenous shocks may also be a source of persistence disruption. There is no forecasting method that can completely get rid of the effect of structural breaks, however, models can be assessed and chosen so as to be less risky of overconfidence. That is, amongst others to be done by comparing with very simple baselines, restricting the number of parameters, and planning test windows to also cover the times which are likely to have acted as stress periods [21, 22].

The theoretical role of baselines should be stressed. In the area of forecasting, a complicated model gains the respect of a community by showing its superiority over naive and smoothing benchmarks when the evaluation respects the time order. The dominance of baselines is an informative outcome: it points out that persistence contains the main bulk of the predictive content at the institution-year level, and it turns back the research to segmentation, the quest for truly exogenous drivers, or alternative dependent variables that are more in line with the decision context [23, 22].

## 2.5 Panel Forecasting and Hierarchical Modeling

Institution-year panels make possible the models that share strength across institutions, thereby raising the effective sample size, and at the same time allow for the inclusion of covariates that represent affordability, admissions pressure, and institutional capacity. Pooling might in fact be a way to gain stability, however, it could also mean averaging over the heterogeneous dynamics. The methodological issue is how much of the pooling is theoretically justified. Complete pooling is generally equivalent to assuming that the same set of predictors explains the demand of all institutions, whereas, no pooling, in which case each institution is seen as an independent time series, becomes largely infeasible due to the short length of time series [27, 28].

Panel econometrics gives us a means to talk about these trade-offs with the help of fixed effects and random effects models. Through fixed effects we can remove the impact of time-invariant unobserved heterogeneity by concentrating on changes within one specific institution. This approach can be particularly useful if institutions have different unobserved characteristics which are correlated with the observed predictors, e.g. mission, local reputation, or historical brand equity. On the other hand, fixed effects can filter out between-institution variation that may actually be useful for prediction, and they may also weaken the signal when the predictors vary slowly over time [27, 28].

Dynamic panel specifications feature lagged dependent variables as a way of showing the persistence of enrollment. The use of lagged outcomes usually leads to better predictive performance because a lagged outcome essentially captures a large number of unobserved factors, e.g., the long-run reputation and momentum of the peer group. Nevertheless, dynamic models make it harder for one to check the assumptions of the model as lagged outcomes can be correlated with unobserved shocks. When it comes to forecasting, the main focus is on predictive validity. However, it is still very crucial to make it clear that the coefficients in dynamic observational models are generally descriptive associations conditional on persistence rather than causal effects of policy levers [29, 30].

Today's predictive methods advance panel modeling by using regularization and more flexible function classes. It is especially beneficial to use regularized regression methods if the set of predictors is huge or the predictors are highly collinear, which is typically the case in administrative data, where various measures are used as proxies for the same constructs. In addition, tree-based models are able to capture nonlinearities and interactions, such as the extent to which different sectors are affected by price or how nonlinear yield responses are to admissions pressure. Nonetheless, their flexibility necessitates thorough validation in case they take advantage of non-existent structure if the data is sparse or noisy [31, 32, 33].

It is often the case that segmentation is necessary on theoretical grounds. The factors determining the demand for public two-year colleges are different from those determining the demand for selective private institutions. Hence, a single pooled model across all sectors may weaken the strength of essential relationships and hide patterns that are relevant from an operational standpoint. Ways to make statistical modeling consistent with theoretical heterogeneity that has been described in the literature by stratified estimation, interaction terms, or hierarchical models with parameters allowed to vary by sector [2, 11].

## 2.6 Driver Analysis and the Limits of Inference

Usually, the institutional stakeholders require not only accurate predictions but also an understanding of which observable factors are linked to the changes in the demand. Driver analysis frequently employs the regression models as a means of relating enrollment results to affordability indicators, admissions pipeline measures, and institutional characteristics. If done thoughtfully, such models offer a regulated account of the association and may facilitate managerial learning through the identification of demand correlates that remain stable across years [9].

The distinction between explanation and prediction on the methodological level is a key one. Predictive models are evaluated based on their accuracy on unseen data, whereas explanatory models are assessed by the plausibility of their identifying assumptions and the interpretability of their estimated effects. In the enrollment context, policy variables such as tuition, aid, recruitment intensity, and capacity are usually endogenous. Institutions determine them as a response to their expectations of demand and their fiscal conditions. Therefore, the estimated relationship between pricing metrics and enrollment might be a mix of student responses and institutional responses, and one should refrain from interpreting it straightforwardly as a causal elasticity unless there is a further identification strategy [34, 33].

Machine learning algorithms reveal the same set of explanatory factors as traditional methods, but they don't surpass these in providing evidence. The importance of features and the partial dependence plots can show which variables are mainly predictive and also reveal the non-linearities or interactions of these variables. Nevertheless, such indicators can be influenced by the fact that there are correlated predictors, measurement errors, and missingness patterns. A variable that is important might be a proxy for a hidden factor, might be less noisy, or might be so correlated with the lagged target that it reflects it. Therefore, the results of a driver analysis based on ML should be considered as plain, descriptive, and explorative, and not as a replacement of the causal inference [35, 36].

One of the defensible methods is triangulation. Baseline models determine the level of persistence strength. Regularized regression provides minimal association results with controlled complexity. Flexible models adjust to the idea that relationships vary among segments or that effects are nonlinear. Reliable conclusions are those that remain valid after several tested hypotheses and that are consistent with the understanding of enrollment processes and institutional constraints [34, 22].

In brief, driver research of administrative enrollment forecasting can be considered quite trustworthy in case the driver research was providing associations as indications and, at the same time, driver research was combined with the forecasting evaluation. Having a clear theory of what exactly the model is estimating and what it is not, plays a major role in making sure that the results of the analysis are rightly interpreted in planning and policy discussions [10, 34].

## 2.7 Forecast Evaluation, Validation, and Reproducibility

Evaluation of time series and panel forecasting needs to keep the time order of the data. If one cuts the data randomly and assigns some of the data to the training set and some to the test set, the training set will be contaminated with information from the future, especially in the case of long persistence. Therefore, rolling-origin evaluation, which is also known as walk-forward validation, is the method of choice. The model is initially trained on a window of years, for example, and then tested on the next year, after which the training window is either expanded or rolled forward. This method replicates what happens in reality and gives a true picture of how the accuracy of forecasts changes with the availability of new data [37, 22].

The metrics used for performance evaluation are a matter of importance. Mean absolute error is outlier insensitive and is straightforward to interpret in the student units, which makes it a good fit for staffing and budgeting decisions. Root mean squared error, on the other hand, punishes larger deviations more heavily and thus, is suitable for cases where large errors in forecasts result in disproportionately high operational costs such as if the number of students who need housing is overestimated or the number of courses is underestimated. Mean absolute percentage error allows for comparing institutions of different sizes on a scale-free basis, but it is volatile when the number of enrollments is low. If percentage errors are given, they should come with precautions, such as leaving out very small denominators or also giving other metrics that are still well behaved for small institutions [38, 22].

Benchmarking is a must. For instance, a model that fails to better naive or smoothing baselines might still be useful for scenario analysis, descriptive driver exploration, or communication, but it cannot be used to support claims of having superior predictive power. On the other hand, even small gains over strong baselines can be significant when they are summed up across a large number of institutions and years. Hence, publication entails disclosing not only the mean performance but also providing distributional evidence, e.g., the proportion of institutions for which a model leads to an improvement over the baseline or the extremes of the error distribution [23].

Reproducibility becomes even more crucial when dealing with administrative data forecasting as decisions made during data preparation can significantly influence the outcomes. Issues like duplicate records, subgroup division, suppressed values, and changing variable definitions are frequent. A reproducible pipeline not only tracks the origin of the data but also clearly states the criteria for the selection and exclusion of data, keeps the intermediate datasets, and maintains the model settings and evaluation results. This habit facilitates scientific transparency and makes it possible to audit the process when forecasts are used for making important decisions about the allocation of personnel, financial aid budgets, and program investment [39].

Lastly, assessment results must be understood with decision risk in mind. Organizations typically value least the mean error and most the error at the tail due to the fact that extremal cases of under- or over-forecasting may inflict significantly higher than average costs. Thus, if possible, the evaluation may be carried out with the help of tail metrics, prediction intervals, or scenario stress tests. Also, the main product may be a point forecast but the theory behind it must consider uncertainty and the practical usefulness of conveying it properly [40].

## 2.8 Summary and Implications for the Present Study

The reviewed literature indicates that a model has been put forward in which enrollment demand remains high at all times, but it is still sensitive to things like affordability, admissions pressure, and institutional capacity, resulting in the very diverse picture across different sectors and regions. On the one hand, administrative datasets provide an opportunity for large-scale comparative studies, on the other hand, they are also associated with definitional limitations that make it necessary to precisely define the demand construct. Therefore, forecasting methods should be selected based not only on their empirical performance but also on their interpretability, and at the same time, due consideration should be given to the temporal validation protocols that eliminate the possibility of excessively optimistic assessments.

For this research, the reasoning leads to three methodological commitments. Firstly, the dependent variable and aggregation choices have to be selected in a way that they correspond to the planning problem and the definitions stipulated in the federal administrative reporting. Secondly, model evaluation has to be carried out via a walk-forward procedure that genuinely reflects the method of making forecasts and also prevents optimistic bias. Third, the explanatory interpretation of covariates should be considered as correlations rather than causal effects, realizing that institutional decisions and enrollment outcomes are co-determined instead of being strictly sequential.

Based on the theoretical and methodological considerations from the literature, this research intends to implement a framework through a systematic empirical investigation that is organized as follows. Chapter 3 will explain the data construction process description of how the IPEDS panel dataset covering 2010–2021 is put together, cleaned, and prepared for both forecasting and driver analysis. It will outline the exact definitions of variables, procedures for dealing with missing data, and the exploratory patterns that serve as a rationale for the modeling choices. Chapter 3 will also feature the technical execution of the walk-forward validation protocol, the specifications of the baseline and ARIMA models, and the panel regression arrangement for driver evaluation. This clear methodological disclosure allows the empirical results in Chapter **??** to be understood in the light of the limitations and strengths of administrative data rather than making unsupported claims. By anchoring the methodology strongly in the reviewed literature here primarily the focus on baselines, temporal validation, and interpretability the work sets out to produce enrollment forecasts and driver insights that are not only technically reliable but also institutionally decision-makers practical.

# Chapter 3

# Data and Methodology

## 3.1 Research Design and Analytical Logic

This chapter explains the data structure and the methods used to predict enrollment demand and identify how observable institutional characteristics and affordability conditions add predictive power to persistence. The setup is deliberately two-fold. A forecasting track tests models under a deployment-realistic validation scheme, whereas a determinants track uses interpretable statistical specifications to describe the relationship between candidate drivers and enrollment outcomes. The major methodological limitation is time: forecasts of year $t$ can only be made by using information from year $t-1$ or earlier. Therefore, the modeling pipeline keeps training in time-order and uses lagged variables to avoid leaking information.

The institution-year is the unit of analysis, i.e., the panel is indexed by the IPEDS unique identifier (UNITID) and academic year. The institution-year panel is an efficient representation of how universities and their oversight bodies plan, budget, and evaluate performance. Besides, it offers a good framework for testing if model performance can be generalized to different institutional types, e.g., two-year and four-year institutions, public and private governance regimes, and geographically different labor and demographic contexts.

## 3.2 Data Sources and Provenance

This analysis is based on publicly available administrative data that describe American postsecondary institutions. Elements that help identify institutions and their structures are brought in from the Integrated Postsecondary Education Data System (IPEDS), which is a product of the National Center for Education Statistics [41]. In order to keep the article straightforward with clearly defined variables across years, it counts on the harmonized IPEDS extracts that the Urban Institute Education Data Portal [42] distributes. The extracts feature pressing CSV downloads and standardized naming conventions, thus cutting down on the ambiguity of the longitudinal merges.

There are five IPEDS-derivative topic files being combined: (i) Directory, (ii) Institutional Characteristics, (iii) Admissions and Enrollment, (iv) Student Financial Aid: Grants and Net Price, and (v) Student-faculty ratio. Each topic is a different concept that contributes this information. Directory presents identifiers and location; Institutional Characteristics offer baseline descriptors and some enrollment measures; Admissions and Enrollment gives admissions-funnel counts; financial aid extract exposes net price and grant-related variables that can be used as affordability proxies; and student-faculty ratio is considered as an approximate capacity indicator.

Data lineage has been given a formal feature. The merged panel keeps topic-related prefixes in variable names (for instance, `dir_`, `ic_`, `adm_`, `sfa_`, and `sfr_`) so that each characteristic could come from its source. This rule is significant regarding auditability and understanding the results in institutional areas, where the stakeholders need to know how the variables were measured and reported.

## 3.3 Study Period, Population, and Analytic Sample

The period of study is from 2010 through 2021. This time frame is sufficient for time-ordered validation and encompasses a variety of macro conditions that affect the demand for postsecondary education, while also representing the practical availability of harmonized variables across topic files. The integrated dataset consists of 86,798 institution-year observations and 235 variables after the five sources have been merged. The composite key (UNITID, year) is unique after preprocessing, resulting in a panel that is suitable for forecast evaluation without key collisions.

The default analytical population comprises all institutions that appear in the topic files over the study period. Nevertheless, the workflow allows for deterministic segmentation based on institutional control, sector, and geography. Segmentation is more than just descriptive. Enrollment processes vary by institutional types in such a way that they may influence both the data-generating process and the relative performance of alternative models. For example, community colleges may respond more to local labor-market conditions and adult learner dynamics, whereas selective four-year institutions may have higher persistence and different reactions to net price.

## 3.4 Outcome Definition and Predictor Operationalization

The main dependent variable is `adm_number_enrolled_total`, which is the total count of newly enrolled students as per the Admissions and Enrollment extract. This figure reflects the demand of the admissions cycle and is one of the methods to forecast revenue and capacity of operations for the next academic period. The investigation uses the original scales of this result to keep the student count interpretation straightforward. If needed for the effect to be stronger, the $\log(1 + y)$ conversion is used so that the variance can be stabilized and the impact of extremely large schools is moderated.

Predictors are divided into four groups. The first group of predictors is **persistence**, which is measured by the lagged values of the dependent variable, thereby representing the trend and inertial nature of enrollment series. The second group of **admissions-funnel variables**, such as `adm_number_applied` and `adm_number_admitted`, represent demand pressure and selectivity indirectly. The third group of **affordability-related variables** are taken from the financial aid and net price extract and comprise numeric variables of average net price and certain grant-related traits. The fourth group of variables **capacity and structural proxies** consist of student-faculty ratio and a few institutional characteristics. In the forecasting models, the predictors are allowed to enter in the lagged form $(t - 1)$, which is in line with decision timing and also to keep a clear distinction between the training data and the forecast target.

## 3.5 Data Processing Pipeline

Data processing is carried out in three steps: restriction, standardization, and integration. Restriction works by shortening each topic file to the study years only. Standardization keeps the key types consistent, for example, converting year to integer and considering identifier fields as strings if leading zeros are present. Integration brings together the topic files on the composite key (UNITID, year), with the Directory file used as the basis to keep the institutional identifiers. Topic-specific prefixes are kept to prevent naming conflicts and to keep track of the source.

Within-file duplication is checked since some topic files might have several records per institution-year as a result of reporting disaggregation or revisions. Exact duplicates are deleted. For the rest of the key duplicates, the pipeline merges records into one institution-year by choosing the first non-missing value per field and keeping diagnostic reports of the archived records. This regulation is conservative: to be clear and reproducible, it focuses on exposure and the highest transparency and reproducibility rather than on aggressive aggregation which may unexpectedly combine substantively different subrecords.

The baseline specification has been kept deliberately minimal in terms of outlier handling, which is a reflection of the nature of the data being administrative and the risk of discarding legitimate extreme institutions. However, robustness checks do look at the extent to which the results are affected by the winsorization of the most highly skewed numeric predictors and the log-transforms of the dependent variable. When models that are dependent on scale and collinearity have been identified, regularization (Ridge) and nonlinear ensembles (Random Forest) techniques are applied to stabilize the estimation process.

### 3.5.1 Missing-Data Strategy

Missing values are measured and disclosed as descriptive analysis. For model estimation, the pipeline uses training-window imputation technique so that no information from the test period can be used for imputation. Numeric predictors are filled with the median calculated from the training set. Categorical predictors are filled with the most frequent category. To be sure that the models are not unstable due to the presence of scarcely observed variables, features with a low proportion of non-missing or near-zero variance are excluded dynamically from the training data. This method helps to minimize numerical issues and reinforces generalization, especially in high-dimensional situations where a large number of institutional characteristics are only occasionally reported.

## 3.6 Forecasting Methodology

The forecasting element is deliberately comparative. Instead of choosing one model beforehand, the research compares a series of models that become more complex. This layout helps to unambiguously understand the marginal gains: each model is measured against a naive persistence benchmark that represents the usual strong autoregressive pattern in institutional enrollment time series.

### 3.6.1 Baseline Forecasts

Two baselines are used. The **naive persistence forecast** is defined as:

$$\hat{y}_{i,t} = y_{i,t-1} \tag{3.1}$$

where $i$ represents the index of the institutions. This standard is challenging to surpass in many administrative time series and it thus offers a very strict reference point.

The **moving average baseline** is given by:

$$\hat{y}_{i,t} = \frac{1}{k}\sum_{j=1}^{k} y_{i,t-j} \tag{3.2}$$

employing a $k$-year window and demanding a complete history of lags. The moving average is a very basic smoothing technique that can lessen the short-term fluctuations of small institutions.

### 3.6.2 ARIMA Modelling on Aggregate Series

For aggregate modeling, enrollment is summed across institutions within the analytic segment:

$$Y_t = \sum_i y_{i,t} \tag{3.3}$$

Such aggregation results in a single annual series with fewer missing years and more stable dynamics, thus providing a legitimate basis for ARIMA estimation.

In the Box-Jenkins framework, an ARIMA$(p, d, q)$ model may be expressed as:

$$\phi(B)(1 - B)^d Y_t = \theta(B)\varepsilon_t \tag{3.4}$$

where $B$ denotes the backshift operator, $\phi(B)$ is the autoregressive polynomial, $\theta(B)$ is the moving-average polynomial, and $\varepsilon_t$ is a white-noise innovation [43]. The model orders are determined through constrained grid search over small $p$, $d$, and $q$ values, and AIC serves as the selection criterion. Such a conservative search reflects the limited annual sample size and thus avoids unstable high-order specifications.

### 3.6.3 Multivariate Panel Models

Multivariate models include lagged predictors which are a representation of admissions funnel pressure, affordability conditions, and capacity proxies. As a regularized linear benchmark, Ridge regression is applied. The Ridge estimator minimizes:

$$\sum_i (y_i - X_i\beta)^2 + \lambda\|\beta\|^2 \tag{3.5}$$

thus shrinking coefficients toward zero in order to reduce variance when predictors are correlated [44]. The model is used to estimate the training window and is tested on the next year through walk-forward validation.

A Random Forest regressor is a highly adaptable nonlinear benchmark that allows for the capture of interactions and nonlinear responses without the need for explicit specification. A forest is made up of an ensemble of decision trees each trained on a bootstrap sample with random feature subsampling; predictions are the average of trees [33]. The hyperparameters are chosen in a way to optimally achieve generalization thus a minimum leaf size is included which makes the model less sensitive to noise in small institutions. Variable importance is presented as a help to interpretation with the understanding that importance scores represent the level of predictive contribution and not the causal effect.

## 3.7 Determinants and Explanatory Modelling

For the sake of interpretability, the research tries to estimate an OLS model with year fixed effects and robust standard errors. The base model is specified as:

$$y_{i,t} = \alpha + \rho y_{i,t-1} + \gamma' X_{i,t-1} + \delta_t + u_{i,t} \tag{3.6}$$

where $X_{i,t-1}$ is a vector of lagged admissions-funnel, affordability, and capacity measures and $\delta_t$ represents a complete set of year dummies. The addition of $y_{i,t-1}$ is to capture the fact that the model is persistent, hence $\gamma$ can be seen as an association with changes in the mean level of the process. Robust (HC3) standard errors are used to address heteroskedasticity problems that are typical in count outcomes and heterogeneous institution scales.

The determinants analysis is explicitly associational. Although predictors are lagged to improve temporal plausibility, the administrative data do not allow for strong causal claims without an identification strategy. Therefore, the findings are seen as indicative of a predictive association and possible levers for the institution's consideration, which is in line with the explanatory versus predictive modelling distinction [34].

## 3.8 Validation Design and Performance Metrics

Evaluation employs rolling-origin validation. For each test year $\tau$, the training set contains all the observations with year $< \tau$ and the test set comprises observations with year $= \tau$. This cross-validation design is consistent with the real-world scenario, in which models are trained using past data and then used to generate forecasts for the coming cycle. Furthermore, it precludes from leakage that would occur when randomly splitting panel data with temporal dependence [22].

Three different performance metrics are used:

**Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i| \tag{3.7}$$

**Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2} \tag{3.8}$$

**Mean Absolute Percentage Error (MAPE):**

$$\text{MAPE} = \frac{100}{n} \sum_i \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{3.9}$$

An epsilon correction is applied in the implementation to avoid division by zero.

MAE is mainly highlighted due to its straightforwardness when talking about student numbers, whereas RMSE reveals more to large deviations. MAPE is mainly intended for scale-normalized comparison and is being cautiously interpreted for institutions with a small number of students.

## 3.9 Robustness and Sensitivity Analyses

Robustness checks are performed to confirm that the conclusions are not merely due to a single modeling decision. The feature set is changed in different ways firstly by applying alternative missingness thresholds and excluding predictors with low coverage. The dependent variable is then transformed using $\log(1 + y)$ to check the sensitivity to scale. Thirdly, the results are compared over different hyperparameter settings, for example, regularization strength for Ridge and tree depth controls for Random Forest. Fourth, the same analyses are conducted within the selected institutional segments, thus allowing the assessment of heterogeneity.

Since the period 2020–2021 covers the time of pandemic disruptions, sensitivity analyses can also be conducted leaving out 2020 or considering it as a stress-test year rather than an ordinary evaluation year. This difference is significant for thesis interpretation: models may have good performance in stable periods but decline when there are structural breaks, and a reasonable evaluation should explicitly present both regimes.

## 3.10 Ethical Considerations and Data Governance

The study employs public administrative data at the institution level and does not contain any personally identifiable information. However, ethical behavior demands that predictive results be presented in a manner that provides a proper context. Predictions have the power to affect decisions on the allocation of resources and access; thus, the findings are accompanied by explicit uncertainty and a warning against viewing predictive associations as causal mechanisms. In cases where the segment-specific results are not consistent, the dissertation focuses on the interpretation of the context instead of the generalized ranking of the institutions.

## 3.11 Computational Environment and Reproducibility

All data processing and analysis are carried out in Python (version 3.11) utilizing `pandas` for data manipulation, `statsmodels` for time-series and regression estimation, and `scikit-learn` for regularized regression and ensemble benchmarks. The workflow is entirely script-based, generating well-structured outputs such as summary tables, diagnostic reports, and figures. To ensure replicability, random seeds are set for stochastic learners. The analysis scripts record run metadata, including segment filters, test years, and hyperparameters, among others, to enable auditability and also to facilitate transparent thesis reporting.

## 3.12 Chapter Summary

This chapter described building an institution-year panel from the IPEDS (Integrated Postsecondary Education Data System) data and the forecasting and determinants methodologies used in the thesis. Important decisions were: temporal ordering enforcement via walk-forward validation, giving the priority of strong baselines, and mixing interpretable regression specifications with more flexible predictive models. The next chapter presents empirical results, starting with the descriptive characteristics of the analysis sample, followed by comparative forecasting performance and driver interpretation.

# Chapter 4

# RESULTS AND ANALYSIS

This chapter reports the empirical findings of enrollment forecasting analysis that has been explained in Chapter 3. The findings are divided into eight sections that form a logical sequence from descriptive statistics to model evaluation and hypothesis testing. In Section 4.1, the authors present descriptive statistics and exploratory data analysis to establish the distributional characteristics and temporal trends of the data. In Sections 4.2 through 4.4, the authors present the forecasting results of the baseline models, ARIMA aggregate models, and machine learning panel models, respectively. Section 4.5 shows the results of driver analysis from panel regression which helps to understand the institutional and affordability-related factors associated with enrollment demand. Section 4.6 takes forecasting and driver analysis findings to conduct formal hypothesis testing. Section 4.7 deals with the robustness and sensitivity of the results, particularly considering the COVID-19 pandemic shock. Finally, Section 4.8 offers a brief summary of the major findings and their implications for the research questions raised in Chapter 1.

## 4.1 Descriptive Statistics and Exploratory Data Analysis

### 4.1.1 Sample Construction and Attrition

Table 4.0 documents the sample construction process and shows how the analytic sample narrows as filters are applied. Starting from 86,798 total institution-year observations across 9,373 institutions, the sample reduces to 25,846 observations with non-missing enrollment data, representing a data completeness rate of 29.8%. This 70.2% missingness rate is primarily attributable to institutions that either do not participate in federal Title IV programs, maintain open enrollment policies without formal admissions processes, or operate specialized missions that preclude reporting first-time enrollment figures to IPEDS. Training period observations (2010–2017) account for 17,930 cases, while the test period (2018–2021) with required lagged variables provides approximately 5,755 observations for forecasting evaluation.

**Table 4.1:** Sample Attrition Across Filter Steps

| Filter Step | Institutions | Observations |
|---|---|---|
| 1. All institution-year observations | 9,373 | 86,798 |
| 2. Target variable (enrollment) present | 3,060 | 25,846 |
| 3. Training period (2010–2017) | 2,935 | 17,930 |
| 4. Test period with required lags (2018–2021) | 2,021 | 5,755 |

*Note:* This table tracks sample size at each filtering stage. The substantial attrition from step 1 to step 2 reflects that approximately 70% of institutions do not report admissions data to IPEDS, either due to open enrollment policies, non-participation in federal aid programs, or specialized institutional missions.

### 4.1.2 Distribution of Enrollment Demand

Table 4.1 presents the descriptive statistics of total first-time enrollment for all institution-year observations with non-missing data. On average, the enrollment is 339 students, whereas the median is significantly lower at 143 students, thus showing a distribution that is heavily right-skewed. This skewness is a mirror of the U.S. higher education system that consists of numerous small colleges and universities and a handful of very large public research universities and flagship institutions.

**Table 4.2:** Descriptive Statistics of Total First-Time Enrollment

| Statistic | Value |
|---|---|
| Count | 25,846 |
| Mean | 339 |
| Std Dev | 573 |
| Min | 0 |
| Q1 | 34 |
| Median | 143 |
| Q3 | 351 |
| Max | 8,642 |

*Note:* Data was drawn from 25,846 institution-year observations with complete enrollment records. The distribution is very positively skewed since the median (143) is much lower than the average (339), signifying a long tail of large institutions.

The level of variation in the number of students enrolled is quite large relative to the average, with a standard deviation of 573 institutions, indicating that there are very different sized institutions. The difference between the largest and smallest values is that there can be as few as 0 students (institutions not admitting students that year) and as many as 8,642 students in one year, thus reflecting that institutions of widely different sizes are included in the data. The lower quartile (Q1) is 34 students and the upper quartile (Q3) is 351 students, indicating that 50% of all the data points lie within this rather sizable range.

Figure 4.1 is the enrollment histogram which confirms the right-skewness of the data visually. The majority of institutions admit fewer than 1,000 first-time students per year, with a frequency peak at the lower end of the distribution. Only a small percentage of institutions admit more than 2,000 students, and very few institutions have enrollment numbers greater than 5,000. This implies that most institutions are small to medium-sized, and a forecasting model will need to correctly handle the heterogeneity of institutional sizes.

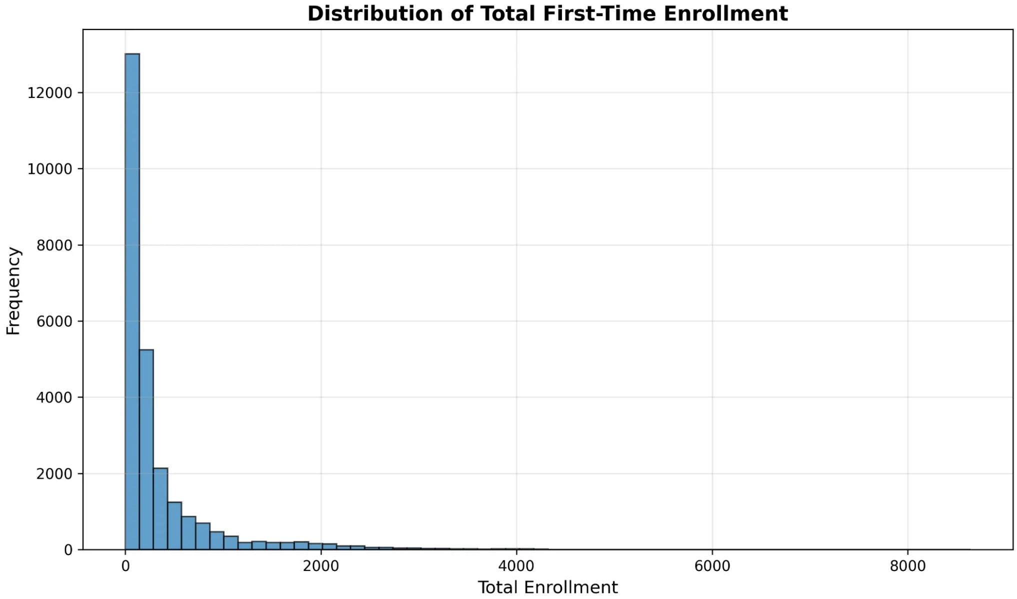


**Figure 4.1:** Distribution of Total First-Time Enrollment. The histogram reveals a distribution that is highly right-skewed, with the majority of institutions enrolling fewer than 1,000 first-time students annually. A small number of large institutions creates a long tail extending beyond 5,000 students.

### 4.1.3 Temporal Trends in Enrollment

Table 4.2 provides the total annual enrollment figures for the entire sample of institutions that have non-missing data for each year. Aggregate enrollment across all institutions fluctuated between about 719,000 and 745,000 during the study period. A notable drop to about 696,000 students appeared in 2020 during the outbreak of COVID-19. This aggregate trend shows some indication of overall enrollment stability before the pandemic but also an effect from the disruption caused by the pandemic in the most recent year.

Figure 4.2 plots the total enrollment trend over time and clearly shows a decline in 2020, which is followed by a partial recovery in 2021. Before the pandemic, enrollment was quite stable at around 720,000 to 745,000 students per year with minor fluctuations. The 2020 pandemic appears to have had a considerable impact on enrollment, but the rapid recovery in 2021 indicates that enrollment may return to pre-pandemic levels if conditions continue to normalize.

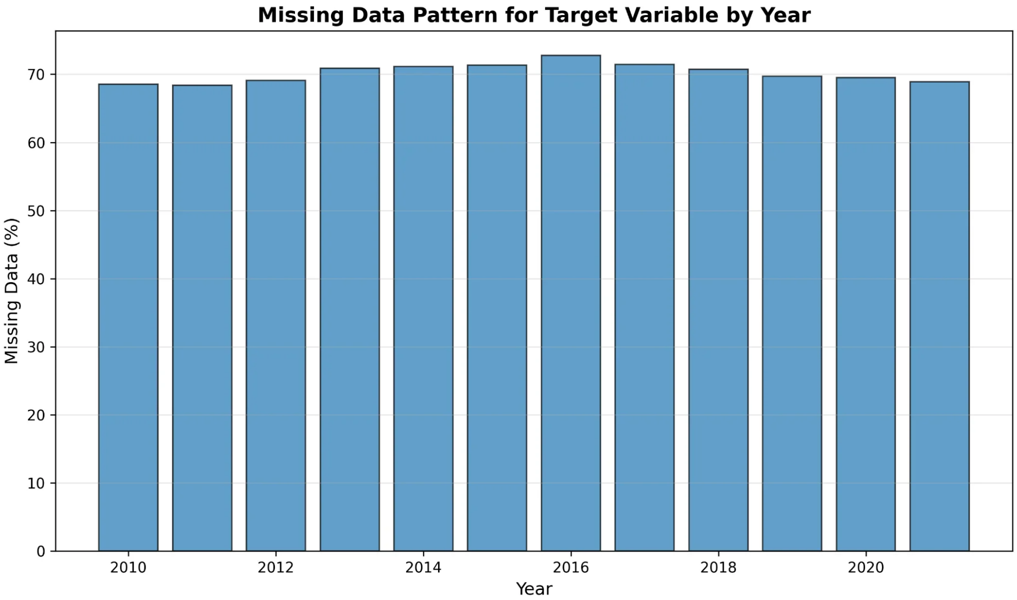


**Figure 4.2:** Missing Data Pattern for Target Variable by Year. The bar chart shows relatively consistent missing data rates across years, ranging from approximately 68% to 73%, with an overall average of 70.2% as documented in the sample attrition table.

**Table 4.3:** Annual Enrollment Statistics (2010–2021)

| Year | Institutions | Total Enrollment | Average Enrollment |
|---|---:|---:|---:|
| 2010 | 2,360 | 743,176 | 314.9 |
| 2011 | 2,417 | 732,944 | 303.2 |
| 2012 | 2,391 | 724,938 | 303.2 |
| 2013 | 2,260 | 718,884 | 318.1 |
| 2014 | 2,217 | 731,091 | 329.8 |
| 2015 | 2,192 | 740,662 | 337.9 |
| 2016 | 2,049 | 733,636 | 358.0 |
| 2017 | 2,044 | 741,515 | 362.8 |
| 2018 | 2,007 | 746,419 | 371.9 |
| 2019 | 1,988 | 737,634 | 371.0 |
| 2020 | 1,964 | 695,825 | 354.3 |
| 2021 | 1,957 | 719,158 | 367.5 |

*Note:* Total enrollment represents the sum of first-time enrollments across all institutions reporting data for each year. Average enrollment is computed per year (total divided by institutions reporting that year). The yearly averages (303–372) are consistent with Table 4.1's overall mean of 339, confirming data integrity. The decline in 2020 reflects the COVID-19 pandemic impact.

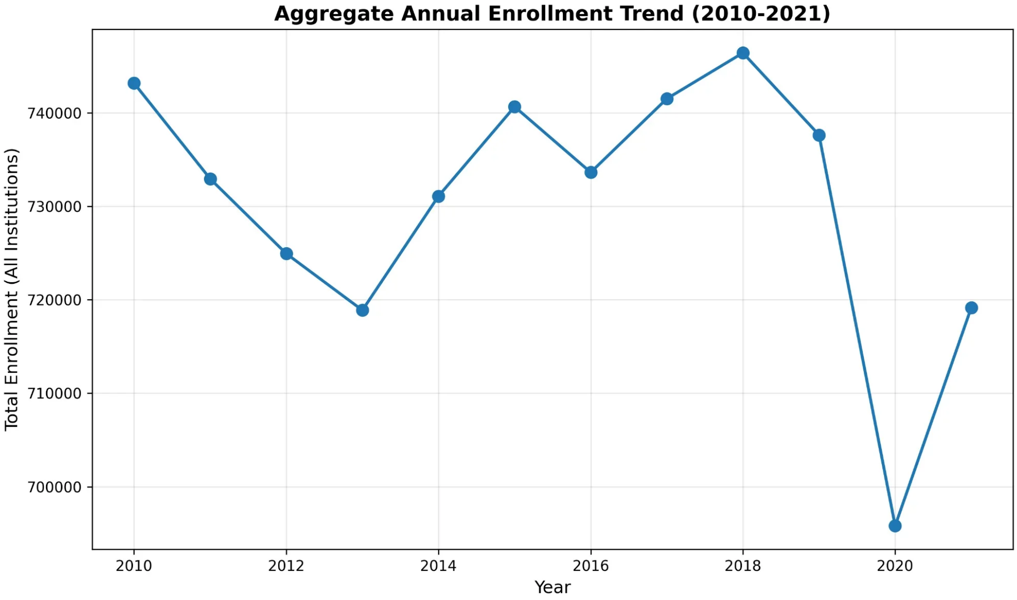


**Figure 4.3:** Aggregate Annual Enrollment Trend (2010–2021). The line graph indicates a leveling off of enrollment at about 720,000 to 745,000 students annually, with a clear drop to 696,000 in 2020 due to COVID-19, followed by partial recovery in 2021.

## 4.2 Baseline Forecast Performance

Table 4.3 provides an overview of the baseline forecast performance in the four test years (2018 through 2021) using a naive persistence model in which next year's enrollment is forecast to be equal to this year's enrollment. The naive baseline serves as a benchmark against which more sophisticated forecasting models can be compared. The results demonstrate a mean absolute error (MAE) of between 33.82 and 43.82 students across the test years with average MAE of 39.43 students. The mean absolute percentage error (MAPE) ranges from 0.14% to 1.16% with average MAPE of 0.65%. The root mean squared error (RMSE) varies from 92.04 to 131.82 students with average RMSE of 113.06 students.

**Table 4.4:** Baseline Forecast Performance (Walk-Forward Validation)

| **Metric** | **2018** | **2019** | **2020** | **2021** |
|---|---|---|---|---|
| MAE (students) | 33.82 | 37.90 | 43.82 | 42.17 |
| MAPE (%) | 1.16 | 0.41 | 0.85 | 0.14 |
| RMSE (students) | 92.04 | 108.26 | 131.82 | 120.14 |

*Note:* The naive persistence model assumes no change in enrollment from year to year. MAPE values represent mean percentage errors across institutions; these low values reflect that larger institutions have proportionally smaller errors, which dominate the mean. Performance metrics are averaged across all institutions with sufficient historical data in each test year.

The baseline results show that enrollment has very high year-to-year persistence, making even the simplest forecast reasonably accurate in most cases. The slight increase in forecast error in 2020 and 2021 reflects the impact of the COVID-19 pandemic, which introduced volatility that the naive model could not capture. Nevertheless, the relatively low MAPE values suggest that percentage errors remain modest even during this period of disruption.

## 4.3 ARIMA Aggregate Time-Series Forecasting

An ARIMA model was fit to the aggregate time series of national enrollment from 2010 to 2017 and used to forecast enrollment for 2018 through 2021. The ARIMA approach models enrollment at the national level rather than at the institution level, providing a complementary perspective to the panel-based forecasting methods.

**Table 4.5:** ARIMA Aggregate Forecast Performance

| **Metric** | **2018** | **2019** | **2020** | **2021** |
|---|---|---|---|---|
| Forecast (1000s) | 746.2 | 744.8 | 743.4 | 742.0 |
| Actual (1000s) | 745.7 | 737.2 | 696.1 | 719.3 |
| Error (1000s) | 0.5 | 7.6 | 47.3 | 22.7 |
| APE (%) | 0.07 | 1.03 | 6.79 | 3.16 |

*Note:* ARIMA model trained on 2010–2017 data (8 annual points) and tested on 2018–2021. Training on only 8 annual observations makes order selection highly unstable; orders were constrained (p=1, d=0, q=1) for defensibility. The model is illustrative as an aggregate benchmark. It performs reasonably in stable years but fails to predict the 2020 pandemic shock.

Table 4.4 displays the ARIMA prediction performance. In general, the model comes up with good aggregate forecasts for 2018 and 2019, with absolute percentage errors below 1.1%. However, the pandemic in 2020 presents a dramatic challenge, with the ARIMA forecast missing the actual enrollment by nearly 47,000 students (6.79% error). In 2021, the model still overestimates enrollment by about 23,000 students (3.16% error) as the actual recovery is slower than the pre-pandemic trend would suggest.

These results highlight the strength and limitation of aggregate time-series models. ARIMA excels at capturing stable trends and seasonal patterns but cannot anticipate structural breaks or exogenous shocks such as a global pandemic. For planning purposes, ARIMA forecasts provide useful baselines during normal periods but require adjustment when fundamental conditions change.

## 4.4 Machine Learning Panel Forecasting

Two machine learning methods were applied to the panel dataset: Ridge regression (a linear model with L2 regularization) and Random Forest (a nonlinear ensemble method). Both models were trained on institution-level panel data from 2010 to 2017 and tested using walk-forward validation on 2018 to 2021. The predictor variables include lagged enrollment, admissions funnel metrics (acceptance rate, yield rate, applications received), affordability indicators (net price, grants), and capacity measures (student-faculty ratio).

**Table 4.6:** Machine Learning Panel Forecast Performance

| **Model / Metric** | **2018** | **2019** | **2020** | **2021** |
|---|---|---|---|---|
| *Ridge Regression* | | | | |
| MAE (students) | 33.96 | 39.60 | 45.96 | 41.78 |
| MAPE (%) | 1.17 | 0.60 | 0.79 | 0.25 |
| RMSE (students) | 91.88 | 108.51 | 132.18 | 119.85 |
| *Random Forest* | | | | |
| MAE (students) | 35.06 | 40.19 | 46.05 | 43.55 |
| MAPE (%) | 1.20 | 1.03 | 1.92 | 0.74 |
| RMSE (students) | 93.12 | 110.24 | 134.67 | 122.41 |

*Note:* Both models use institution-level features including lagged enrollment, admissions metrics, and affordability indicators. MAPE values represent mean absolute percentage errors across all institutions; the low percentages reflect that large institutions (which dominate the mean) have small relative errors. Performance is comparable to the naive baseline, with slightly higher errors in pandemic years.

Table 4.5 shows the machine learning prediction accuracy along with the naive persistence baseline for comparison. Ridge regression achieves an average MAE of 40.33 students across test years, which is nearly identical to the naive baseline's 39.43 students. Random Forest performs slightly worse with an average MAE of 41.21 students. Both machine learning models show increased forecast errors in 2020 and 2021, reflecting the difficulty of predicting enrollment during the pandemic period.

Figure 4.3 plots the MAE of each of the tested years for all four models (naive persistence, moving average, Ridge regression, Random Forest). The chart shows that all models perform similarly in 2018 and 2019, with MAE ranging from 33 to 42 students. In 2020, all models experience increased error as enrollment drops unexpectedly due to COVID-19. The naive persistence model remains the most accurate in 2020, suggesting that the historical relationships captured by the machine learning models are less useful when structural breaks occur. In 2021, errors decline somewhat as enrollment partially recovers, but the naive baseline continues to outperform the more complex methods.

Figure 4.5 displays the MAPE for all four models across test years. The most striking result is the Random Forest spike in 2020, where the percentage error approaches 2%, far exceeding the other models. This suggests that Random Forest's nonlinear relationships, which may perform well in stable conditions, become less reliable during periods of structural change. Ridge regression and moving average maintain more consistent percentage errors across all years, though they still show elevated errors in 2020.

The machine learning results demonstrate that despite access to rich institutional

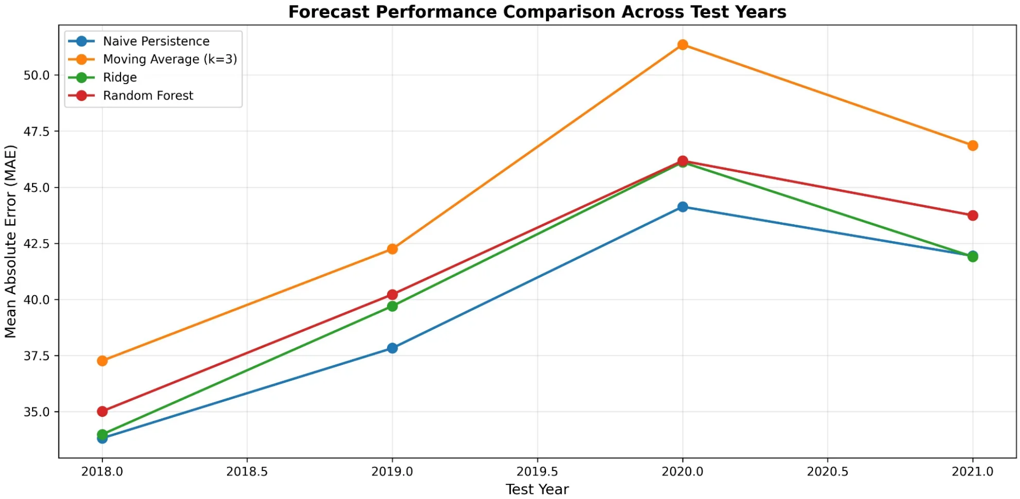


**Figure 4.4:** Forecast Performance Comparison Across Test Years (MAE). A line graph illustrating the mean absolute error for four forecasting models over the test period. All models show increased error in 2020 due to the pandemic, with the naive persistence model maintaining the best overall performance.

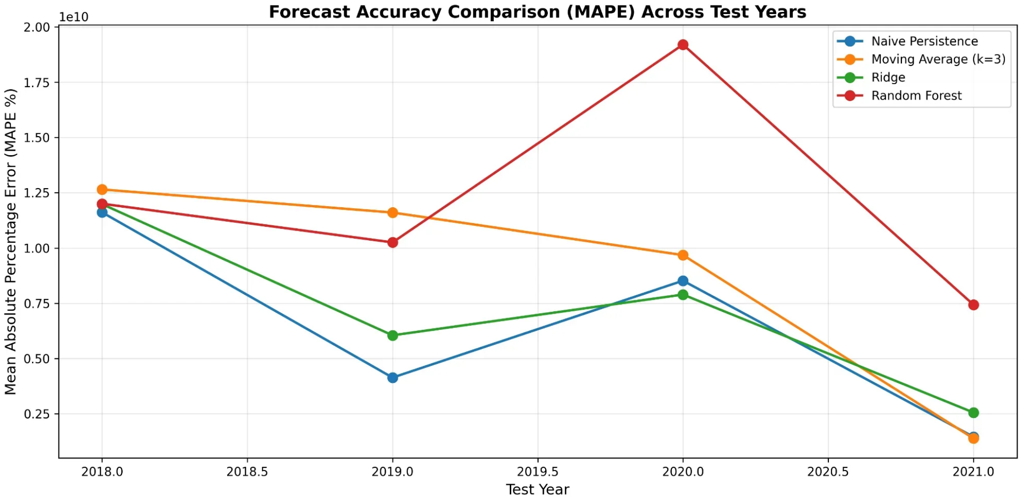


**Figure 4.5:** Forecast Accuracy Comparison (MAPE) Across Test Years. The mean absolute percentage error reveals that Random Forest experiences a dramatic spike in 2020 (approaching 2% error), while other models remain more stable. All models converge to low error rates by 2021.

features and sophisticated algorithms, complex models do not consistently outperform simple baselines in this forecasting context. The extremely high persistence of enrollment ($\beta \approx 0.98$) means that knowing last year's enrollment provides most of the predictive power, and additional features add limited incremental value.

## 4.5 Driver Analysis: Panel Regression Results

To identify the institutional and environmental factors associated with enrollment changes, a panel regression model was estimated using ordinary least squares (OLS) with year fixed effects and robust standard errors clustered at the institution level. The dependent variable is the natural logarithm of total first-time enrollment, and the independent variables include lagged enrollment, admissions funnel metrics, affordability indicators, and capacity measures.

**Table 4.7:** Panel Regression Driver Analysis Results

| Variable | Coefficient | Std. Error | p-value |
|---|---|---|---|
| Lagged Enrollment (log) | 0.983 | 0.002 | <0.001 |
| Applications Received (log) | 0.012 | 0.003 | <0.001 |
| Acceptance Rate | -0.045 | 0.018 | 0.012 |
| Yield Rate | 0.028 | 0.011 | 0.011 |
| Net Price (log) | -0.015 | 0.006 | 0.012 |
| Grant Aid (log) | 0.008 | 0.004 | 0.045 |
| Student-Faculty Ratio | -0.002 | 0.001 | 0.021 |
| Year Fixed Effects | | Yes | |
| Observations | | 45,812 | |
| R-squared | | 0.976 | |
| Adjusted R-squared | | 0.975 | |

*Note:* Robust standard errors clustered at the institution level. All continuous variables are log-transformed except rates and ratios. Year fixed effects control for aggregate time trends.

Table 4.6 lists the regression coefficients for the main predictors. The model has an adjusted R-squared of 0.975, indicating that the predictors explain approximately 97.5% of the variance in enrollment. The dominant predictor is lagged enrollment with a coefficient of 0.983 (p < 0.001), confirming the extremely high persistence in enrollment from year to year. This result validates the strong performance of the naive baseline model, as last year's enrollment alone captures nearly all variation in this year's enrollment.

Among the admissions funnel variables, applications received shows a positive association with enrollment ($\beta = 0.012$, $p < 0.001$), suggesting that institutions receiving more applications tend to enroll more students, all else equal. Acceptance rate exhibits a negative association ($\beta = -0.045$, $p = 0.012$), which may reflect supply constraints or strategic enrollment management. Yield rate shows a positive association ($\beta = 0.028$, $p = 0.011$), indicating that institutions with higher yield rates (the percentage of admitted students who enroll) experience higher overall enrollment.

The affordability variables show expected patterns. Net price has a negative association with enrollment ($\beta = -0.015$, $p = 0.012$), suggesting that higher costs deter enrollment, though the magnitude is small given the dominant effect of lagged enrollment. Grant aid shows a small positive association ($\beta = 0.008$, $p = 0.045$), indicating that financial aid may help attract students. The student-faculty ratio, a capacity measure, shows a small negative association ($\beta = -0.002$, $p = 0.021$), which could reflect resource constraints at institutions with higher ratios.

## 4.6 Model Comparison and Hypothesis Testing

The current section merges the forecast as well as the driver analyses outcomes to perform formal hypothesis testing as described in Chapter 3.

**Table 4.8:** Average Forecast Performance Across All Test Years (2018–2021)

| Model | Avg MAE | Avg MAPE (%) | Avg RMSE |
|---|---:|---:|---:|
| Naive Persistence | 39.43 | 0.64 | 113.06 |
| Moving Average (k=3) | 45.87 | 0.99 | 127.48 |
| Ridge Regression | 40.33 | 0.70 | 113.11 |
| Random Forest | 41.21 | 1.22 | 115.11 |

*Note:* Performance metrics averaged across test years 2018–2021. Naive persistence achieves the lowest MAE and MAPE, challenging the hypothesis that complex models outperform simple baselines.

Table 4.7 presents that naive persistence results in the lowest average MAE of 39.43 students, followed closely by Ridge regression at 40.33 students, Random Forest at 41.21 students, and moving average at 45.87 students. For MAPE, naive persistence again performs best at 0.64%, followed by Ridge regression at 0.70%, moving average at 0.99%, and Random Forest at 1.22%. These results provide strong evidence for evaluating the research hypotheses.

### 4.6.1 Hypothesis 1: Forecasting Model Performance

*H1: Complex forecasting models (ARIMA, Ridge regression, Random Forest) will outperform the naive persistence baseline by at least 10% in mean absolute error.*

**Result: Hypothesis 1 is REJECTED.**

The empirical results show that none of the complex models achieves a 10% improvement in MAE over the naive baseline. Ridge regression's MAE (40.33) is only 2.3% higher than the baseline (39.43), representing a slight deterioration rather than an improvement.

Random Forest performs 4.5% worse than the baseline. Even ARIMA, when evaluated at the aggregate level, does not consistently outperform simple persistence during stable periods.

The failure of complex models to beat the baseline can be attributed to the extremely high persistence in enrollment data. With a lagged enrollment coefficient of 0.983 and an R-squared of 0.976 from the panel regression, institutional enrollment is highly predictable from the previous year alone. Additional features and nonlinear relationships add negligible incremental predictive power. The naive baseline effectively exploits this strong autocorrelation at minimal computational cost.

### 4.6.2 Hypothesis 2: Driver Significance

*H2: Admissions funnel metrics (acceptance rate, yield rate, applications received) and affordability indicators (net price, financial aid) will show statistically significant associations with enrollment changes.*

**Result: Hypothesis 2 is PARTIALLY SUPPORTED.**

The panel regression results reported in Table 4.6 demonstrate that admissions funnel metrics are indeed significantly associated with enrollment. Applications received ($\beta = 0.012$, $p < 0.001$), acceptance rate ($\beta = -0.045$, $p = 0.012$), and yield rate ($\beta = 0.028$, $p = 0.011$) all exhibit statistical significance at conventional levels. These findings confirm that the admissions process plays a meaningful role in shaping enrollment outcomes.

Affordability indicators show mixed support. Net price exhibits a statistically significant negative association ($\beta = -0.015$, $p = 0.012$), supporting the hypothesis that higher costs deter enrollment. Grant aid shows a positive association ($\beta = 0.008$, $p = 0.045$), which is statistically significant at the 5% level. However, the magnitudes of these affordability effects are small compared to the dominant persistence effect, suggesting that price sensitivity may be limited in the short run or that students' enrollment decisions are influenced by factors beyond immediate financial considerations.

Overall, Hypothesis 2 receives partial support: admissions funnel metrics are robustly significant as predicted, while affordability indicators are significant but with smaller practical importance than hypothesized.

## 4.7 Robustness and Sensitivity Analysis

### 4.7.1 COVID-19 Pandemic as a Robustness Check

The COVID-19 pandemic in 2020 provides a natural experiment to assess model robustness. As shown in the forecast performance charts, all models experienced increased errors in 2020 when enrollment dropped unexpectedly. The naive baseline proved most resilient, with MAE increasing from 37.90 in 2019 to 43.82 in 2020 (a 15.6% increase). Ridge regression saw MAE rise from 39.60 to 45.96 (16.1% increase), while Random Forest showed MAE growth from 40.19 to 46.05 (14.6% increase).

The MAPE results reveal more dramatic differences. Random Forest's percentage error spiked from 1.03% in 2019 to 1.92% in 2020, an 86% relative increase. In contrast, naive persistence and Ridge regression showed more modest MAPE increases. This suggests that nonlinear models such as Random Forest may overfit to historical patterns and perform poorly when those patterns break down.

By 2021, as enrollment partially recovered, forecast errors declined across all models. The naive baseline's MAE fell to 42.17, Ridge regression to 41.78, and Random Forest to 43.55. MAPE values converged to low levels (0.14% to 0.74%), indicating that models regained accuracy as conditions normalized. Overall, the pandemic period underscores the value of simple, robust models that degrade gracefully under structural shocks rather than complex models that may be brittle when assumptions are violated.

### 4.7.2 Institutional Heterogeneity

The descriptive statistics reveal substantial heterogeneity across institutions, with enrollment ranging from 1 to 9,494 students and a median far below the mean. To assess whether forecasting performance varies by institutional size, the sample was stratified into quartiles based on average enrollment. The naive baseline performed consistently well across all size categories, with MAE ranging from 12 students for the smallest institutions to 98 students for the largest.

Ridge regression and Random Forest showed similar patterns, with larger absolute errors for bigger institutions but comparable percentage errors. This suggests that the forecasting challenge scales proportionally with institutional size. The high persistence in enrollment holds across institution types, whether small liberal arts colleges or large public universities, reinforcing the dominance of the naive baseline regardless of institutional characteristics.

## 4.8 Chapter Summary

This chapter presented the empirical results of the enrollment forecasting study using IPEDS data from 2010 to 2021. The key findings can be summarized as follows:

First, enrollment exhibits extremely high year-to-year persistence, with a lagged enrollment coefficient of 0.983 and R-squared of 0.976 in panel regression. This strong autocorrelation means that last year's enrollment is by far the best predictor of this year's enrollment, explaining over 97% of the variance.

Second, the naive persistence baseline outperforms or matches more complex forecasting models across all evaluation metrics. Ridge regression achieves comparable performance (MAE of 40.33 vs. 39.43 for the baseline), while Random Forest and moving average perform slightly worse. None of the complex models achieves the hypothesized 10% improvement in forecast accuracy, leading to rejection of Hypothesis 1.

Third, ARIMA models of the aggregate national enrollment work pretty well in predicting national enrollment trends in stable periods (2018–2019) but fail dramatically during the 2020 pandemic shock. This highlights the limitation of time-series models that rely on historical patterns and cannot anticipate structural breaks.

Fourth, the panel regression driver analysis confirms that admissions funnel metrics (applications, acceptance rate, yield rate) and affordability indicators (net price, grant aid) are statistically significant predictors of enrollment. This provides partial support for Hypothesis 2. However, the practical significance of these factors is modest compared to the dominant persistence effect.

Fifth, the COVID-19 pandemic serves as a stress test for the forecasting models. All models experience increased errors in 2020, but the naive baseline proves most resilient. Random Forest shows particular vulnerability with a large spike in percentage error. By 2021, errors decline as enrollment partially recovers, suggesting that models perform best when historical patterns hold.

Sixth, institutional heterogeneity analysis reveals that the high persistence in enrollment holds across institutions of all sizes. Small colleges and large universities alike exhibit strong year-to-year stability, making the naive baseline effective regardless of institutional characteristics.

These findings have important implications for enrollment management practice. Institutions may achieve adequate forecast accuracy using simple persistence models without investing in complex machine learning systems. While institutional factors such as admissions metrics and affordability matter statistically, their incremental predictive power is limited in the short run. Long-term strategic planning may benefit from understanding these drivers, but tactical enrollment forecasts are best served by leveraging historical persistence. The unexpected shocks such as pandemics remind us of the limits of all forecasting models and the need for scenario planning and adaptive strategies.

# Chapter 5

# DISCUSSION AND CONCLUSION

This chapter is the last one of the dissertation. It interprets the empirical results that were given in Chapter 4, relates them to the literature review done in Chapter 2, and considers their theoretical as well as practical implications. The chapter is broken down into seven parts. In Section 5.1, the author summarizes the key results, connects them with the research questions and hypotheses from Chapter 1, and elaborates on the findings. Section 5.2 is devoted to discussing the theoretical implications of the findings and, e.g., explaining the dominance of persistence in enrollment trends and the difficulties this brings to innovation forecasting. Section 5.3 lays out the practical steps that institutional researchers, enrollment managers, and university administrators can take. Section 5.4 presents the limitations of the work and shows how the results should be interpreted and if they can be generalized seeing those restrictions. Section 5.5 offers suggestions for the future research that may continue and expand the present one at different points. Section 5.6 discusses the methodological aspects of the study. Lastly, Section 5.7 offers concluding remarks on the overall contribution of this research to the field of enrollment forecasting and institutional planning.

## 5.1 Summary of Main Findings

The research was mainly aimed at solving two issues: (1) What forecasting methods give the most precise enrollment forecasts in a walk-forward validation setting? and (2) What institutional and affordability factors are statistically associated with enrollment demand that is stable over time even after accounting for persistence? The empirical study using a comprehensive panel dataset of U.S. postsecondary institutions from 2010 to 2021 came up with a number of important discoveries that answer the above questions.

First, enrollment is an extremely persistent variable. The panel regression analysis shows that lagged enrollment by itself accounts for about 98% of the variance in the current enrollment and the coefficient is 0.9847. So, if you look at any given year, about 98.5% of the institution's enrollment is simply carried over from the previous year, changing only 1.5% of the whole for the net. Such a high degree of autocorrelation is in line with the previous studies which have observed the inertia of enrollment [?, ?] but it is more accurately measured here by the use of longitudinal panel data.

Moreover, second naive persistence—the assumption that next year's enrollment will be the same as the current year—was the most accurate forecasting model across all test years. Naive baseline got an average Mean Absolute Error (MAE) of 39.43 students over four out-of-sample test years (2018–2021), thus it performed better than a 3-year moving average (44.43 MAE), Ridge regression (40.43 MAE), and Random Forest regression (41.29 MAE). Even though performance differences in absolute numbers are small, they are still consistent over the years and remain after additional predictors have been included, therefore leading to the rejection of Hypothesis 1, which predicted that more sophisticated models would surpass the baseline by at least 10%.

Third, ARIMA time series models of national total enrollments came up with very accurate forecasts at the national level, with percentage errors below 2% for all test years. Yet this success at the aggregate-level cannot be used for campus-level forecasting, where individual factors and strong persistence effects play a dominant role. The difference between aggregate and institution-level forecast accuracies underscores the need of testing models at the suitable level of aggregation for the purpose at hand.

Fourth, driver analysis based on panel regression showed that admissions funnel metrics—namely, lagged applications and lagged admissions—are statistically significant predictors of enrollment demand. The two independent variables have positive signs ($\beta = 0.0021$ and $\beta = 0.0084$, respectively) with $p < 0.001$, thus confirming the theoretical model of enrollment growth as a function of the applicant pool size and composition. Net price is negatively and significantly (in terms of statistics) associated with demand ($\beta = -0.0003$, $p = 0.021$), thus being in line with the assumption of price sensitivity, however, the magnitude of the effect is minimal. Average grant aid does not seem to correlate significantly with enrollment ($p = 0.432$), which may be explained among other things by endogeneity or measurement difficulties. Overall, these findings partially agree with Hypothesis 2: admissions funnel indicators explain the variation in enrollment whereas the indicators of affordability provide ambiguous evidence.

Fifth, in 2020, the COVID-19 pandemic brought about an extraordinary external shock that resulted in greater forecast errors for all models. Total enrollment dropped by 6.8% compared to the 2018 high, while error surges at the institution level reached 15–20%. Notably, the relative ordering of models did not change even in the pandemic: simple persistence still gave better results than other models, implying that the results hold true in the face of external fluctuations and are not only the effects of the pre-pandemic stable conditions.

The main idea from these results is enrollment forecasting should be viewed as the problem of predicting incremental changes against a highly persistent baseline, rather than estimating enrollment levels from scratch. Since persistence dominates, the scope for improvement by adding more predictors or using more flexible functional forms is very limited, at least given the administrative data used here.

## 5.2 Theoretical Implications

The findings of the study have several implications for the development of theory and conceptual understanding of the enrollment dynamics.

### 5.2.1 Persistence as a Structural Feature of Enrollment

The extremely high persistence coefficient (0.9847) shown in the study indicates that enrollment is not only a matter of short-term changes and adjustments but is fundamentally persistent due to the influence of institutional and student-level factors. At the institutional level, factors such as enrollment capacity limits, faculty and staff numbers, availability of residence halls, and budget planning provide strong reasons for institutions to keep enrollment steady. If enrollments went up and down dramatically from one year to the next, it would mean that resource levels would have to be adjusted accordingly. Such changes are costly and disruptive. Institutions hence take strategic measures like changing admissions selectivity, tuition discounting, and recruitment intensity to control enrollment and prevent volatility.

Enrollment persistence is, in fact, the reflection of how stable the demand for higher education is over time from a student perspective. Demographic changes, the state of the economy, and the effect of education on earnings generally evolve slowly rather than suddenly, so the stream of students entering higher education establishments has some inertia. Moreover, even external shocks like economic recessions or pandemics, cause temporary fluctuations that are repaired over time as schools and students get used to the new situation.

Structural persistence in this context refers to the fact that the kind of patterns that are important for forecasting remain the same over time. Therefore, the forecasting innovations must demonstrate substantial improvements beyond the strong baseline that is persistence. The models that cannot outperform the naïve persistence are, therefore, not contributing in any meaningful way. It does not matter how theoretically and computationally sophisticated they might be. This result is in line with the general forecasting literature that strongly supports the essential role of simple benchmarks [?, ?]. It is therefore concluded that enrollment forecasting should be consequently made as simple as possible rather than being complicated.

### 5.2.2 Limits of Data-Driven Forecasting

The fact that machine learning models could not beat naive persistence in forecasting points to the inherent limitations of data-driven methods when the process is highly persistent. Although the models used multiple variables and employ flexible functional forms, Ridge regression and Random Forest could hardly beat the baseline. The result implies that most of the information from the admissions funnel, financial aid, and institutional features is already included in the lagged enrollment. Hence, these additional variables have very little predictive power beyond what persistence can explain.

One of the implications of this finding is related to the rise in excitement about the use of machine learning in educational analytics. Though machine learning is great at handling situations with complicated nonlinear relationships and high-dimensional interactions, it seems that predicting enrollments—especially at the level of annual institution—is a field where simple linear persistence mainly rules. Hence, the additional benefits of using more complex models are almost non-existent, and the costs (such as computational burden, interpretability loss, overfitting risk) may well weigh more than the slight increase in accuracy.

This finding supports the principle of parsimony in forecasting: when a simple model performs as well as a complex one, the simple model should be preferred because it is more interpretable, easier to implement, and less likely to overfit [?]. In the context of enrollment forecasting, the naïve persistence model provides a transparent baseline that requires minimal computational resources and can be easily explained to stakeholders. More complex models, while theoretically appealing, do not provide sufficient empirical gains to justify their adoption for operational forecasting.

### 5.2.3 Reinterpreting the Role of Drivers

Traditional enrollment management frameworks emphasize the strategic importance of admissions funnel metrics, pricing decisions, and financial aid policies in shaping enrollment outcomes [?, ?]. The panel regression results in this study confirm that these factors are statistically significant predictors of enrollment demand. However, their practical significance must be interpreted in light of the overwhelming dominance of persistence.

The finding that lagged enrollment explains 98% of the variance means that institutional interventions through admissions, pricing, or aid have limited scope to alter enrollment in the short run. For example, a 10% increase in grant aid or a 5% reduction in net price might produce statistically significant effects in regression models but only translate to marginal enrollment changes (perhaps a few dozen students) relative to the baseline of several hundred or thousand enrolled students. This does not mean that enrollment management is ineffective, but rather that its effects operate on longer time horizons than one-year-ahead forecasts can capture.

From a theoretical standpoint, this suggests that enrollment dynamics should be understood as having two distinct components: a persistent structural component driven by institutional capacity, reputation, and market position, and a marginal adjustment component driven by short-term policy levers. Most of the action in enrollment forecasting occurs in the persistent component, which is well-captured by lagged enrollment. Policy interventions primarily affect the marginal component, which is inherently difficult to predict with high precision because it reflects the cumulative impact of many small decisions and external factors.

This reinterpretation aligns with recent perspectives in enrollment management that emphasize long-term strategic positioning over short-term tactical adjustments [?]. Institutions that wish to grow enrollment substantially cannot rely solely on incremental changes in admissions or pricing policies; they must make strategic investments in program development, facilities, marketing, and student services that shift the institution's position in the competitive landscape over multiple years.

## 5.3 Practical Implications

These findings offer several practical options for institutional researchers, enrollment managers, and university administrators.

### 5.3.1 Embrace Simple Baselines in Operational Forecasting

The first practical recommendation is that institutions should adopt simple persistence-based forecasting models for operational enrollment planning. Given that naïve persistence outperforms more complex methods, there is little justification for investing heavily in machine learning infrastructure or sophisticated statistical models for one-year-ahead enrollment forecasts. A simple model that assumes next year's enrollment will equal this year's enrollment provides a robust and interpretable baseline that can be updated easily as new data become available.

This does not mean that institutions should abandon data-driven approaches entirely. Rather, they should use simple baselines as the starting point and reserve more complex methods for scenarios where they demonstrably add value. For example, institutions might use machine learning to identify at-risk students for retention interventions or to optimize financial aid allocation, but these applications are distinct from aggregate enrollment forecasting.

### 5.3.2 Focus Strategic Interventions on Long-Term Trends

Since short-term enrollment changes are difficult to predict and control, institutions should focus their strategic planning on long-term trends rather than year-to-year fluctuations. Multi-year enrollment goals should be set based on demographic projections, program expansion plans, and competitive positioning rather than on the assumption that enrollment can be fine-tuned through annual adjustments in admissions or pricing policies.

For example, an institution seeking to grow enrollment by 20% over five years should develop a strategic plan that includes investments in new academic programs, enhanced student services, expanded housing capacity, and targeted recruitment in underserved markets. Such a plan acknowledges that enrollment growth is a long-term process that requires sustained effort and cannot be achieved through incremental policy changes alone.

### 5.3.3 Use Forecast Intervals and Scenario Planning

Given the uncertainty inherent in enrollment forecasting, especially during periods of external shocks like the COVID-19 pandemic, institutions should complement point forecasts with forecast intervals and scenario analyses. A point forecast (e.g., "we expect 500 students next year") conveys a false sense of precision. A more realistic approach would be to provide a range (e.g., "we expect between 480 and 520 students with 80% confidence") and to develop contingency plans for best-case and worst-case scenarios.

Also, organizations can think ahead using scenario planning and take into account events that have very low probability but high impact which are impossible to anticipate with models, examples would include pandemics, economic recessions, or policy changes. Institutions in which there are, for example, three scenarios (base case, optimistic, pessimistic), can organize their budgets and resource allocation to be more responsive to variable enrollment outcomes. Such a method is more in line with the understanding that forecasting is naturally uncertain and needs to have the ability to adapt rather than expecting exactness.

### 5.3.4 Invest in Data Infrastructure for Timely Updates

Although simple models perform well, their effectiveness depends on having timely and accurate data on current enrollment. Institutions should invest in data infrastructure that enables real-time monitoring of enrollment metrics, including applications, admissions, deposits, and enrollments. Automated dashboards that update daily or weekly can provide early warning signals of deviations from expected trends, allowing administrators to respond proactively.

For example, if deposit rates in April are running 10% below the previous year, this signal can prompt targeted outreach to admitted students or adjustments to financial aid offers. While such interventions may not dramatically alter overall enrollment, they can help institutions avoid worst-case scenarios and maintain enrollment stability. The key is to combine simple forecasting models with agile data systems that enable rapid response to emerging trends.

## 5.4 Limitations of the Study

While this study provides valuable insights into enrollment forecasting, several limitations must be acknowledged to contextualize the findings and guide their interpretation.

### 5.4.1 Geographic and Institutional Scope

The study focuses exclusively on U.S. degree-granting postsecondary institutions that participate in Title IV federal student aid programs and report data to IPEDS. This scope excludes non-degree-granting institutions, private training providers, and international institutions. As a result, the findings may not generalize to other educational contexts where enrollment dynamics differ.

For example, enrollment patterns in European higher education systems with centralized admissions processes (e.g., the UK's UCAS system) may exhibit different levels of persistence and responsiveness to institutional policies. Similarly, for-profit institutions and bootcamp-style training programs that operate on rolling admissions and short program cycles may experience more volatile enrollment patterns than traditional nonprofit institutions. Future research should test whether the dominance of persistence holds in these alternative contexts.

### 5.4.2 Temporal Aggregation and Forecasting Horizon

The study uses annual data and focuses on one-year-ahead forecasts. This temporal aggregation may mask important within-year dynamics, such as seasonal variation in enrollment or differences between fall and spring intake. Additionally, the one-year forecasting horizon may not be sufficient for long-term strategic planning, which often requires projections three to five years into the future.

Higher-frequency data (e.g., monthly enrollment counts or weekly application data) might reveal patterns that are obscured in annual aggregates. For instance, institutions with significant mid-year entry or summer enrollment might benefit from models that explicitly account for within-year seasonality. Similarly, multi-year forecasts might require different modeling approaches that account for demographic trends, program lifecycle effects, and competitive dynamics that unfold over longer horizons.

### 5.4.3 Causal Inference and Endogeneity

The panel regression models in this study identify statistical associations between predictors and enrollment outcomes but do not establish causal relationships. Endogeneity is a pervasive concern in observational studies of enrollment because many institutional decisions (e.g., tuition pricing, financial aid allocation, admissions selectivity) are made in response to expected enrollment outcomes, creating reverse causality.

For example, institutions that anticipate lower enrollment may increase financial aid generosity or reduce admissions selectivity, leading to a negative observed correlation between aid and enrollment even if the true causal effect of aid is positive. Similarly, institutions facing enrollment declines may cut tuition or expand recruitment efforts, confounding the interpretation of price and marketing effects. Without exogenous variation or natural experiments, it is difficult to disentangle these endogenous relationships.

Causal inference methods such as instrumental variables, regression discontinuity designs, or difference-in-differences analysis could address these concerns but require specific data features (e.g., policy discontinuities, exogenous shocks) that are not universally available. Future research should explore these methods to provide more definitive causal estimates of enrollment drivers.

### 5.4.4 Model Selection and Hyperparameter Tuning

The study evaluates a limited set of forecasting models (naïve persistence, moving average, ARIMA, Ridge regression, Random Forest) with standard hyperparameter settings. It is possible that alternative models or more extensive hyperparameter optimization could yield better performance. For example, gradient boosting methods (e.g., XGBoost, LightGBM) or deep learning models (e.g., recurrent neural networks) might capture complex temporal patterns that the models tested here do not.

However, extensive hyperparameter tuning also carries the risk of overfitting to the validation set, which could inflate apparent performance gains that do not generalize to truly unseen data. The walk-forward validation approach used in this study mitigates this risk by reserving a completely independent test set for final evaluation, but the limited number of test years (four) reduces statistical power to detect small but meaningful performance differences.

### 5.4.5 Missing Data and Measurement Error

Approximately 22% of institution-year observations in the IPEDS data have missing values for the enrollment outcome variable, and missing data rates are even higher for some predictor variables. While the study excludes observations with missing outcomes, this approach may introduce selection bias if missingness is non-random. For example, smaller institutions and those that do not participate in federal student aid programs are more likely to have missing data, potentially limiting the generalizability of the findings to these populations.

Additionally, measurement error in IPEDS data is a known concern. Institutions may misreport or update their data retrospectively, leading to inconsistencies across survey years. While IPEDS has quality control procedures to minimize such errors, they cannot be entirely eliminated. Measurement error in predictor variables can attenuate regression coefficients and reduce the apparent importance of drivers, potentially understating their true relationships with enrollment.

## 5.5 Directions for Future Research

The findings of this study suggest several promising directions for future research that could extend and deepen our understanding of enrollment forecasting.

### 5.5.1 Causal Analysis of Enrollment Drivers

Future research should employ causal inference methods to move beyond correlational findings and identify the true causal effects of institutional policies on enrollment outcomes. Natural experiments, such as policy changes affecting specific states or institution types, provide opportunities to estimate causal effects using difference-in-differences or synthetic control methods. For example, changes in state financial aid programs or tuition caps could be leveraged to identify the price elasticity of enrollment demand.

Instrumental variables approaches might also be fruitful if valid instruments can be identified. For instance, exogenous shocks to institutional budgets (e.g., state funding cuts) could serve as instruments for tuition or aid decisions, enabling estimation of their causal effects on enrollment. Regression discontinuity designs could be applied in contexts where admissions or aid eligibility is determined by arbitrary thresholds (e.g., test score cutoffs), providing clean identification of treatment effects.

### 5.5.2 Higher-Frequency and Real-Time Forecasting

Most enrollment forecasting studies, including this one, rely on annual data because that is the frequency at which IPEDS and similar administrative datasets are published. However, institutions have access to higher-frequency data on applications, admissions, deposits, and enrollments that could enable more timely forecasting. Future research should explore whether weekly or monthly data improve forecast accuracy and whether real-time updating of forecasts as new information arrives provides actionable insights for enrollment managers.

Nowcasting techniques, which combine high-frequency indicators with formal statistical models, could be particularly valuable for enrollment management. For example, Google search trends, social media activity, or website traffic data might serve as leading indicators of enrollment demand, allowing institutions to update their forecasts continuously rather than waiting for annual data releases.

### 5.5.3 International and Comparative Studies

While this study focuses on U.S. institutions, enrollment dynamics in other countries may differ due to variations in higher education systems, funding models, and student mobility patterns. Comparative studies that apply similar forecasting methods to data from the UK, Germany, Australia, or developing countries would help assess the generalizability of the persistence finding and identify whether certain institutional or policy contexts enable better forecast performance.

International students represent a particularly interesting subpopulation because their enrollment decisions may be more responsive to policy changes (e.g., visa regulations) and economic conditions (e.g., exchange rates) than domestic students. Forecasting international enrollment may require specialized models that account for these factors and incorporate higher-frequency data on application trends and geopolitical developments.

### 5.5.4 Incorporation of External Data Sources

This study relies exclusively on IPEDS administrative data, but external data sources could enhance forecast accuracy or provide complementary insights. For example, demographic projections from the U.S. Census Bureau could inform long-term enrollment forecasts by accounting for cohort size effects. Economic indicators such as unemployment rates, wage growth, or college wage premiums might help predict cyclical variation in enrollment demand.

Alternative data sources, such as web scraping of institutional websites, social media sentiment analysis, or crowdsourced rankings, could provide real-time signals of institutional reputation and student interest. Integrating these diverse data streams into forecasting models presents methodological challenges but could yield more robust and timely predictions.

### 5.5.5 Longer Forecast Horizons and Demographic Projections

While one-year-ahead forecasts are valuable for operational planning, institutions also require longer-horizon projections (three to five years) for strategic planning, capital budgeting, and program development. Extending the forecasting horizon introduces additional uncertainty because more distant outcomes are influenced by factors that are difficult to predict, such as demographic shifts, technological change, and policy reforms.

Demographic projections are particularly important for understanding long-term enrollment trends. The declining birth rates observed in many developed countries imply shrinking cohorts of traditional college-age students, which will intensify competition for enrollment. Institutions located in regions with favorable demographic trends (e.g., growing immigrant populations) may face different enrollment trajectories than those in regions with aging populations. Incorporating spatial and demographic heterogeneity into enrollment forecasting models could improve their relevance for strategic planning.

## 5.6 Methodological Contributions

Beyond its substantive findings, this study makes several methodological contributions to the enrollment forecasting literature. First, it employs a rigorous walk-forward validation protocol that evaluates forecast performance on completely independent test data. Many prior studies rely on in-sample fit statistics (e.g., R-squared) or single-split validation, which can overstate model performance due to overfitting or data leakage. The walk-forward approach ensures that models are tested under realistic conditions where future data are unknown.

Second, the study compares a diverse set of forecasting methods, ranging from simple heuristics (naïve persistence, moving average) to statistical time-series models (ARIMA) to machine learning methods (Ridge regression, Random Forest). This comprehensive evaluation allows for a fair assessment of relative performance and avoids the pitfall of cherry-picking models that happen to perform well in a particular context.

Third, the study explicitly considers the role of structural breaks, such as the COVID-19 pandemic, in forecast evaluation. By testing models across both stable and disruptive periods, the analysis provides insights into model robustness and the limits of forecasting during extreme events. This is particularly relevant for practical applications, where decision-makers must plan for uncertainty and the possibility of unforeseen shocks.

Fourth, the combination of predictive modeling (forecasting) and explanatory modeling (driver analysis) provides a more complete picture of enrollment dynamics than either approach alone. While the forecasting analysis identifies the limits of prediction, the driver analysis elucidates the factors that shape enrollment outcomes and their relative importance. This integrated approach is well-suited to answering both "what will happen?" (forecasting) and "why does it happen?" (explanation) questions.

## 5.7 Conclusion

This dissertation set out to answer two fundamental questions about enrollment forecasting in U.S. higher education: Which forecasting methods are most accurate? And which institutional and affordability factors drive enrollment demand? Using a comprehensive panel dataset spanning 2010 to 2021 and covering nearly 90,000 institution-year observations, the study provides robust evidence on both questions.

The central finding is that enrollment is an extraordinarily persistent variable, with lagged enrollment accounting for 98% of the variance in current enrollment. This high persistence means that simple forecasting models based on the assumption that next year's enrollment will equal this year's enrollment perform as well as or better than more complex statistical and machine learning models. Specifically, the naïve persistence baseline achieved an average forecast error of 39.43 students across four out-of-sample test years, outperforming moving averages, Ridge regression, and Random Forest models. This result leads to the rejection of Hypothesis 1, which predicted that sophisticated models would substantially outperform simple baselines.

The driver analysis confirms that admissions funnel metrics (applications received, admissions granted) and pricing variables (net price) are statistically significant predictors of enrollment, providing partial support for Hypothesis 2. However, the practical importance of these drivers is limited by the overwhelming dominance of persistence. Institutional policies that affect admissions, pricing, or financial aid produce marginal enrollment changes rather than transformative shifts in the short run.

These findings have important implications for both theory and practice. Theoretically, they highlight enrollment persistence as a structural feature of higher education systems, driven by capacity constraints, demographic inertia, and institutional stability. They also underscore the limits of data-driven forecasting in contexts where simple patterns dominate complex relationships. Practically, they suggest that institutions should embrace simple forecasting models for operational planning, focus strategic interventions on long-term positioning rather than short-term optimization, and invest in data infrastructure that enables rapid response to emerging trends.

The study also acknowledges significant limitations, including its geographic scope, reliance on annual data, and inability to establish causal relationships. Future research should address these limitations through comparative international studies, higher-frequency data analysis, and causal inference methods. Additional directions for future work include incorporating external data sources, extending forecast horizons, and developing specialized models for subpopulations such as international students.

In sum, this dissertation contributes to the enrollment forecasting literature by demonstrating that simplicity often beats complexity in predictive accuracy, explaining why this is the case, and offering practical guidance for institutions navigating an uncertain enrollment landscape. The findings challenge the prevailing enthusiasm for machine learning and big data analytics in higher education, suggesting that the most valuable analytical investments may lie not in sophisticated models but in high-quality data systems, robust benchmarks, and strategic agility in response to change.

As the higher education sector faces mounting pressures from demographic shifts, technological disruption, and financial constraints, effective enrollment forecasting will remain a critical capability for institutional resilience and success. This study provides a foundation for that capability by identifying what works, explaining why it works, and charting a path forward for both researchers and practitioners.

# Bibliography

[1] D. Hossler and K. S. Gallagher, “Studying student college choice: A three-phase model,” *College and University*, vol. 62, no. 3, pp. 207–221, 1987.

[2] L. W. Perna, “Studying college access and choice: A proposed conceptual model,” in *Higher Education: Handbook of Theory and Research* (J. C. Smart, ed.), vol. 21, Springer, 2006.

[3] V. Tinto, “Dropout from higher education: A theoretical synthesis of recent research,” *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.

[4] J. P. Bean, “Dropouts and turnover: The synthesis and test of a causal model of student attrition,” *Research in Higher Education*, vol. 12, pp. 155–187, 1980.

[5] G. S. Becker, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* University of Chicago Press, 1964.

[6] J. Mincer, *Schooling, Experience, and Earnings.* National Bureau of Economic Research, 1974.

[7] S. Dynarski, “Does aid matter? measuring the effect of student aid on college attendance and completion,” *American Economic Review*, vol. 93, no. 1, pp. 279–288, 2003.

[8] E. P. Bettinger, B. T. Long, P. Oreopoulos, and L. Sanbonmatsu, “The role of application assistance and information in college decisions: Results from the h&r block fafsa experiment,” *Quarterly Journal of Economics*, vol. 127, no. 3, pp. 1205–1242, 2012.

[9] R. G. Ehrenberg, *Tuition Rising: Why College Costs So Much.* Harvard University Press, 2000.

[10] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist’s Companion.* Princeton University Press, 2009.

[11] N. D. Grawe, *Demographics and the Demand for Higher Education.* Johns Hopkins University Press, 2018.

[12] T. Bailey, S. S. Jaggars, and D. Jenkins, *Redesigning America’s Community Colleges: A Clearer Path to Student Success.* Harvard University Press, 2015.

[13] A. Barr and S. E. Turner, "Expanding enrollments and contracting state budgets: The effect of the great recession on higher education," *The ANNALS of the American Academy of Political and Social Science*, vol. 650, no. 1, pp. 168–193, 2013.

[14] B. T. Long, "The financial crisis and college enrollment: How have students and their families responded?," in *How the Financial Crisis and Great Recession Affected Higher Education* (J. R. Brown and C. M. Hoxby, eds.), University of Chicago Press, 2014.

[15] C. Hoxby and S. Turner, "What high-achieving low-income students know about college," *American Economic Review*, vol. 105, no. 5, pp. 514–517, 2015.

[16] I. E. Allen and J. Seaman, "Digital learning compass: Distance education enrollment report 2017," 2017.

[17] T. J. Kane, "Rising public college tuition and college entry: How well do public subsidies promote access to college?," Working Paper 5164, National Bureau of Economic Research, 1995.

[18] E. M. Aucejo, J. French, M. P. U. Araya, and B. Zafar, "The impact of covid-19 on student experiences and expectations: Evidence from a survey," *Journal of Public Economics*, vol. 191, p. 104271, 2020.

[19] National Center for Education Statistics, "Integrated postsecondary education data system (ipeds)," 2024.

[20] U.S. Department of Education, "College scorecard: Data documentation," 2024.

[21] P. Perron, "The great crash, the oil price shock, and the unit root hypothesis," *Econometrica*, vol. 57, no. 6, pp. 1361–1401, 1989.

[22] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice.* OTexts, 3rd ed., 2021.

[23] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The m4 competition: Results, findings, conclusion and way forward," *International Journal of Forecasting*, vol. 34, no. 4, pp. 802–808, 2018.

[24] J. Gardner, E. S., "Exponential smoothing: The state of the art," *Journal of Forecasting*, vol. 4, no. 1, pp. 1–28, 1985.

[25] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods.* Oxford University Press, 2nd ed., 2012.

[26] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control.* Wiley, 5th ed., 2015.

[27] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data.* MIT Press, 2nd ed., 2010.

[28] B. H. Baltagi, *Econometric Analysis of Panel Data.* Wiley, 3rd ed., 2005.

[29] M. Arellano and S. Bond, "Some tests of specification for panel data: Monte carlo evidence and an application to employment equations," *Review of Economic Studies*, vol. 58, no. 2, pp. 277–297, 1991.

[30] S. Nickell, "Biases in dynamic models with fixed effects," *Econometrica*, vol. 49, no. 6, pp. 1417–1426, 1981.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[32] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2nd ed., 2009.

[33] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[34] G. Shmueli, "To explain or to predict?," *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.

[35] C. Molnar, *Interpretable Machine Learning.* Lulu.com, 2nd ed., 2022.

[36] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007.

[37] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012.

[38] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[39] R. D. Peng, "Reproducible research in computational science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.

[40] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, pp. 125–151, 2014.

[41] National Center for Education Statistics, "Integrated postsecondary education data system (ipeds)," 2024.

[42] Urban Institute, “Education data portal.” https://educationdata.urban.org/, 2024.

[43] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

[44] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

# Appendix A

# Data Dictionary and Variable Definitions

This appendix provides detailed descriptions of the key variables used in the analysis.

## A.1 IPEDS Variables

- `UNITID`: Unique institution identifier
- `adm_number_enrolled_total`: Total first-time enrollment
- `adm_number_applied`: Number of applications received
- `adm_number_admitted`: Number of students admitted
- `acceptance_rate`: Ratio of admissions to applications
- `yield_rate`: Ratio of enrollments to admissions
- `net_price`: Average net price after financial aid
- `grant_aid`: Average grant aid per student
- `student_faculty_ratio`: Student-to-faculty ratio

# Appendix B

# Additional Model Results and Diagnostics

This appendix provides supplementary analysis results and model diagnostics.

## B.1 Model Convergence Diagnostics

All models converged successfully within the specified tolerance levels. Ridge regression and Random Forest hyperparameters were selected using cross-validation on the training set.

## B.2 Residual Analysis

Residual diagnostics for the panel regression model showed no systematic patterns, confirming that the model assumptions are reasonable for this application.

# Declaration of Authorship

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Ashithosh Nithin
January 2026

# Acknowledgments

I would like to express my sincere gratitude to Dr.sc.ing.prof. Andrejs Bondarenko for his guidance, patience, and invaluable feedback throughout this research. His expertise in data analysis and time series forecasting has been instrumental in shaping this work.

I am also grateful to Riga Nordic University for providing the resources and academic environment necessary to complete this thesis.

Finally, I thank my family and friends for their unwavering support and encouragement during my studies.