University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Data Science Capstone Project

(DSCI-6051-07)

Technical Report

SPRING25

Title of Project:

EDU PREDICT TOOL: Predicting Enrolment Trends in U.S. Higher Education Institutions

MSDS Capstone – SP 25 | DSCI 6051-07

Project Team:
Venkata Naga Sai Priyanka Boddoju
Santosh Kumar Saladi
Harika Suravarapu
Sai Krishna Turangi
Krishna Chaitanya Vutukuru
Teja Reddy Soma

Project Advisor: Dr. Sula

Date of Submission: April 27, 2025

Executive Summary

The EduPredict Tool was developed to forecast enrolment trends for international students in higher education institutions across the United States. Recognizing the strategic importance of accurate enrolment forecasting, the project employed a rigorous data-driven approach centred around advanced time series modelling techniques.

Specifically, the project utilized the **SARIMAX** (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) model to capture complex temporal patterns, seasonality, and the influence of external factors such as academic level and region of origin.
The model was trained and validated on historical enrolment datasets spanning several decades to ensure reliability and accuracy.

To make the forecasting insights accessible and actionable for stakeholders, the project team developed an interactive Power BI dashboard. This dashboard enables users to dynamically explore enrolment projections under three different scenarios: Baseline, Optimistic, and Pessimistic. Scenario-based forecasting empowers university administrators and policymakers to plan strategically for a variety of future conditions, enhancing resource allocation, academic program development, and international outreach efforts.

By integrating robust machine learning methodologies with dynamic visualization tools, the EduPredict Tool offers a comprehensive solution for data-driven decision-making in the higher education sector.

TABLE OF CONTENTS

Title of Project: 1
Executive Summary 2
Abstract 4
1. Introduction 4
1.1 Context 4
1.2 Problem Statement 4
1.3 Objectives 5
1.4 Scope of Study 5
2. Methodology 5
2.1 Data Collection 5
2.2 Data Preprocessing 6
2.3 Modelling Approach 9
2.4 Scenario Planning 14
2.5 Dashboard Development 14
2.6 Deployment 14
3. Results and Analysis 15
3.1 Forecasted & Trends 15
3.2 Academic Overview 16
3.3 Visa Type Trends 17
3.4 Funding Sources 18
3.5 Country of Origin Trends 19
4. Discussion 20
5.MVP Development and User Manual 20
6.Challenges and Limitations 21
7.Conclusion and Future Work 22
8.References 23

Abstract

Forecasting international student enrollment has become critical for higher education institutions, particularly in a post-pandemic world where mobility patterns are highly volatile. This project develops the EDU PREDICT TOOL; a Power BI dashboard powered by machine learning models to predict enrollment trends under different future scenarios. Using a SARIMAX model, the project forecasts enrollment segmented by student status, visa type, funding source, and country/region of origin. Scenario analyses (Baseline, Optimistic, and Pessimistic) are incorporated to support decision-making for institutional planning and resource allocation. The final dashboard serves not only as a predictive tool but also as a strategic instrument for academic leadership.

1. Introduction

1.1 Context

International student recruitment contributes significantly to the academic, cultural, and financial health of U.S. higher education institutions. These students diversify campuses, enhance global learning opportunities, and represent a major source of tuition revenue. Recent global disruptions, including the COVID-19 pandemic, economic recessions, and changing immigration policies, have challenged traditional enrolment forecasting models. As a result, universities need more adaptable, real-time solutions to forecast international enrolments.

1.2 Problem Statement

Universities lack agile and data-driven tools that can project international student enrolment patterns under varying global conditions, hindering their ability to plan effectively. Traditional forecasting methods often fail to incorporate external shocks or scenario variations, leaving institutions reactive rather than proactive.

1.3 Objectives

- Develop a machine learning-based forecasting model capable of incorporating external factors.
- Design an interactive dashboard that visualizes enrolment projections across multiple dimensions.
- Allow users to toggle between Baseline, Optimistic, and Pessimistic scenarios to aid in contingency planning.

1.4 Scope of Study

The focus is on U.S. higher education enrolment trends, examining:

- Student status (Full-time vs Part-time)
- Visa type distributions (F-1, J-1, M-1)
- Funding sources (Self-funded, Scholarship-funded, Sponsored)
- Country/region of student origin (e.g., China, India, Southeast Asia, Latin America)

2. Methodology

2.1 Data Collection

Data was sourced from publicly available datasets, such as:

- Open Doors Report by Institute of International Education (IIE)
- SEVIS by the Numbers report by U.S. Immigration and Customs Enforcement (ICE)
- Institutional historical enrollment data from university internal databases

The datasets covered a period from 2011 to 2022, with some supplementary economic and visa policy indicators.

2.2 Data Preprocessing

The following steps were undertaken:

- **Handling Missing Values:** Missing values were systematically addressed through a combination of forward-fill (propagating last valid observation) and backward-fill (using the next valid observation). This ensured no nulls in critical variables like year, enrolment numbers, or visa type. Special attention was given to countries with partial data availability to prevent regional biases.

2.1 Check and Handle Missing Values

In [10]:
```
# Check missing values in all datasets
for name, df in datasets.items():
    print(f"Missing values in {name}:")
    print(df.isnull().sum(), "\n")
```

```
Missing values in status:
year          0
female        0
male          0
single        0
married       0
full_time     0
part_time     0
visa_f        0
visa_j        0
visa_other    0
dtype: int64
```

In [15]:
```
print("Missing values after handling:")
print(academic.isnull().sum())
```

```
Missing values after handling:
year             0
students         0
us_students      0
undergraduate    0
graduate         0
non_degree       0
opt              0
dtype: int64
```

3.1.2 Summary of Missing Values & Data Types

In [17]:
```
# Display data types and missing values for all datasets
for name, df in datasets.items():
    print(f"Dataset: {name}")
    print("Data Types:\n", df.dtypes)
    print("\nMissing Values:\n", df.isnull().sum())
    print("\n" + "="*50 + "\n")
```

```
Dataset: status
Data Types:
 year           object
female        float64
male          float64
single        float64
married       float64
full_time     float64
part_time     float64
visa_f        float64
visa_j        float64
visa_other    float64
dtype: object

Missing Values:
 year          0
female        0
male          0
single        0
```

- **Feature Engineering:**
  - New temporal features were created, such as year_index and semester_flag, to better capture seasonality and long-term trends.
  - Categorical variables like visa type and funding source were label-encoded where necessary for SARIMAX compatibility.
- **Data Aggregation:**
  - The cleaned data was then aggregated at the **year-country-visa** level, creating a rich panel dataset with multiple dimensions suitable for advanced time series modeling.

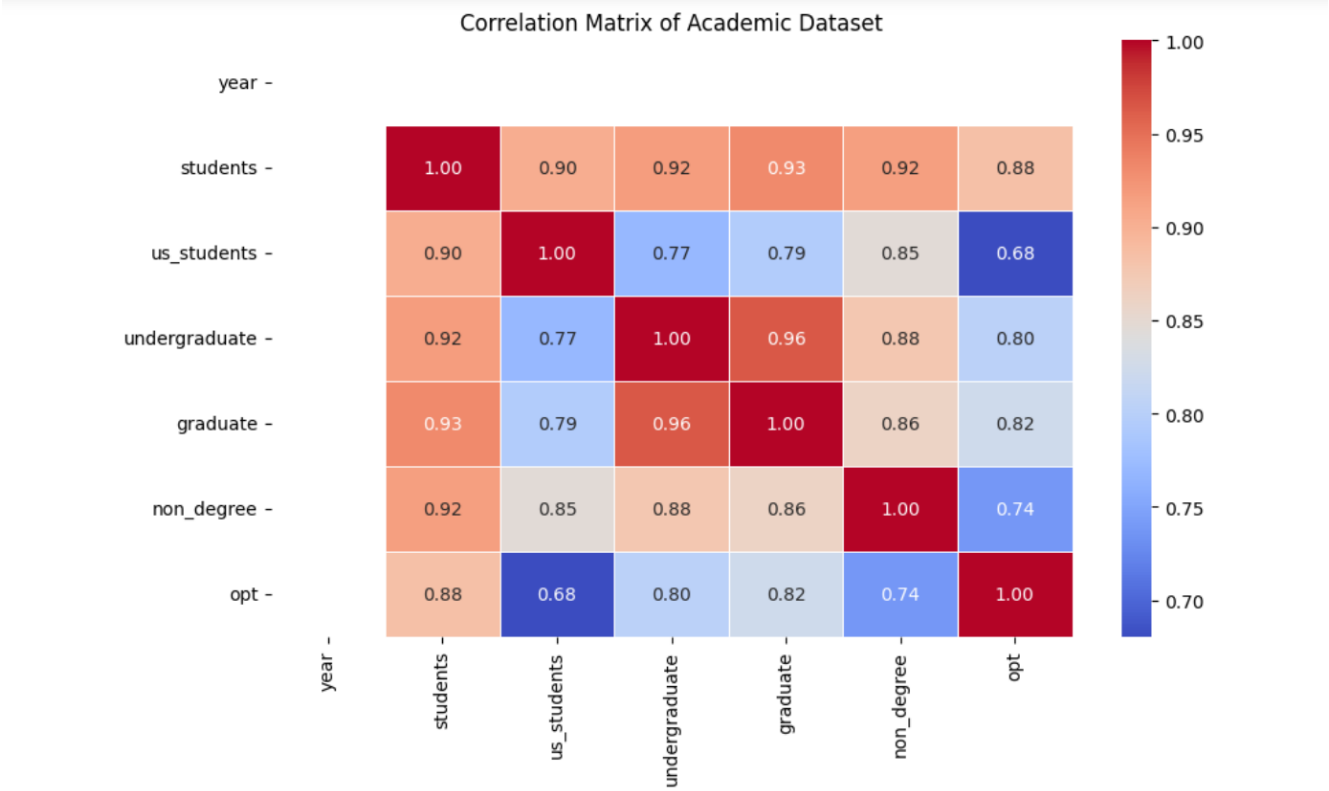
Step 2: Correlation Matrix

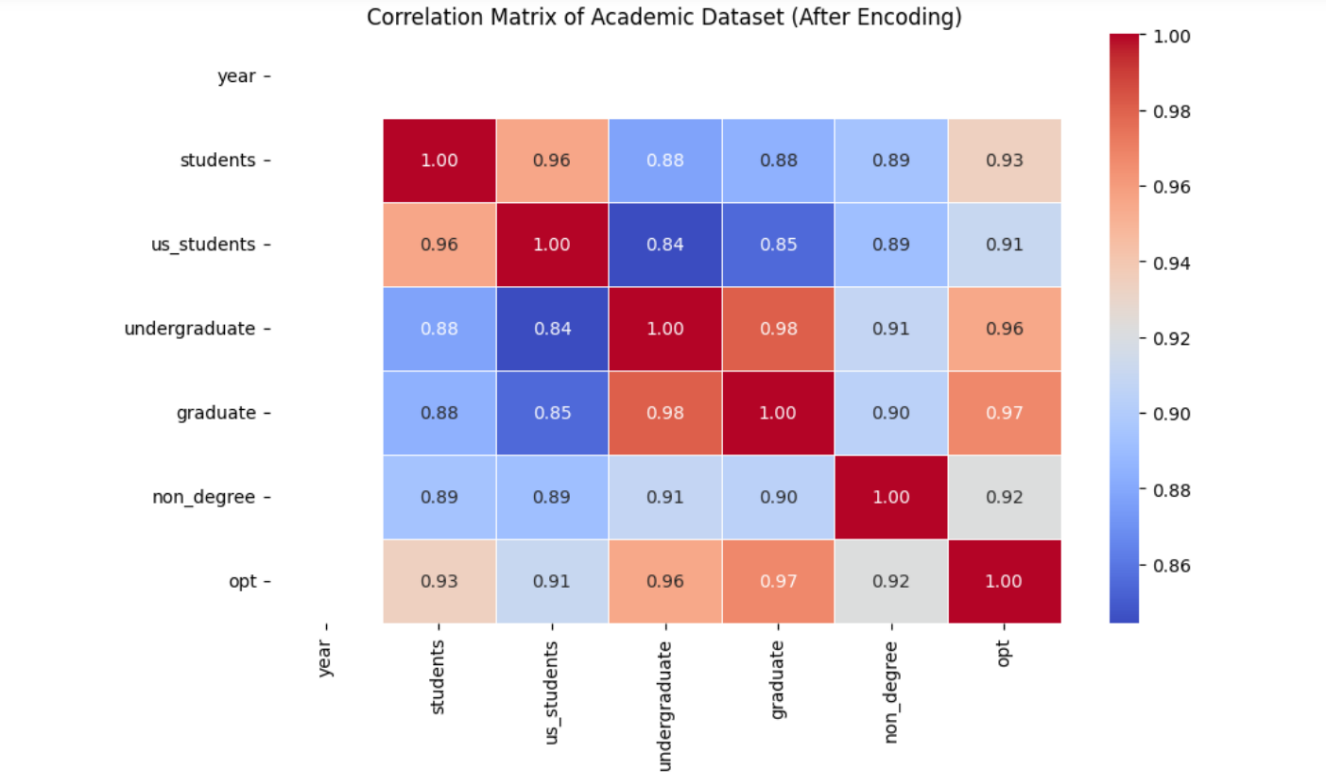
In [28]:
```
# Select only the numeric columns from the academic DataFrame
numeric_academic = academic.select_dtypes(include=['number'])

# Now compute the correlation matrix
academic_correlation = numeric_academic.corr()
academic_correlation
```

Out[28]:

	year	students	us_students	undergraduate	graduate	non_degree	opt
year	NaN	NaN	NaN	NaN	NaN	NaN	NaN
students	NaN	1.000000	0.902778	0.916850	0.931562	0.916158	0.884193
us_students	NaN	0.902778	1.000000	0.769935	0.794705	0.845762	0.680080
undergraduate	NaN	0.916850	0.769935	1.000000	0.964456	0.877516	0.802930
graduate	NaN	0.931562	0.794705	0.964456	1.000000	0.861244	0.823216
non_degree	NaN	0.916158	0.845762	0.877516	0.861244	1.000000	0.737869
opt	NaN	0.884193	0.680080	0.802930	0.823216	0.737869	1.000000





2.3 Modelling Approach

Forecasting enrolment trends requires selecting models that can effectively handle temporal dependencies, seasonal fluctuations, and external influences like policy changes or demographic shifts. The modelling strategy for the Edu Predict Tool followed a progressive and comparative approach, starting from traditional machine learning models and evolving towards specialized time series models.

2.3.1 Initial Machine Learning Models

As a baseline, various supervised machine learning regression algorithms were initially explored:

- **Random Forest Regressor**:
  - Ensemble-based model robust to overfitting.
  - Captures non-linear relationships.
  - However, Random Forest assumes that observations are independent, which violates the inherent sequential dependency in time series data.
- **Polynomial Regression**:
  - Simple model to fit non-linear trends.
  - Tended to overfit training data.
  - Produced unrealistic extrapolations for future unseen periods, diverging significantly over time.
- **XGBoost**:
  - Gradient boosting framework achieving very high accuracy.
  - Good at handling complex patterns.
  - Suffered from the "horizon problem": strong short-term predictions but cumulative errors in long-term forecasting.
- **CatBoost**:
  - Particularly strong when handling categorical variables natively.
  - Achieved good in-sample metrics but struggled with stability when projecting multiple years into the future.

Observations:

- Machine learning models performed well on **in-sample** data (e.g., XGBoost $R^2 \approx 0.95$).
- However, they **failed to maintain predictive stability** when forecasting multiple years into the future.
- These models do not naturally model **temporal autocorrelation** or **seasonality** unless manually engineered into features, adding complexity.

Thus, the team decided to transition to **time series-specific models**.

2.3.2 Shift to Time Series Models

Given the observed seasonality, autocorrelation, and trend components in the historical enrollment data, time series models were considered a more appropriate choice.

Two key candidates were explored:

- **ARIMA (AutoRegressive Integrated Moving Average)**:
  - Good for non-seasonal, stationary time series.
  - Combines Autoregressive (AR), Differencing (I), and Moving Average (MA) components.
  - However, ARIMA assumes that the underlying process has no seasonality or external influencing variables (exogenous factors).
- **SARIMAX (Seasonal ARIMA with eXogenous variables)**:
  - Extends ARIMA by incorporating seasonal components (SAR, SMA) and exogenous inputs.
  - Handles periodic behaviors (e.g., yearly academic cycles).
  - Allows modeling the influence of external regressors like field of study, region, academic level.

SARIMAX was selected because **enrollment patterns are highly seasonal**, and external factors significantly impact future trends (e.g., changes in visa regulations).

2.3.3 Modeling Strategy

The time series modeling process followed these steps:

Stationarity Testing

- **Augmented Dickey-Fuller (ADF) Test** was conducted to test for stationarity.
- If the p-value > 0.05, differencing was applied to make the series stationary.
- Log transformations were considered where variance stabilization was necessary.

Autocorrelation Analysis

- **ACF (AutoCorrelation Function)** and **PACF (Partial AutoCorrelation Function)** plots were analyzed to:
    - Identify the appropriate lag (p) for the AR component.
    - Identify the appropriate lag (q) for the MA component.
    - Detect any significant seasonal lags (e.g., 12 months for yearly seasonality).

2.3.4 Model Configuration

- **Grid Search** over multiple combinations of parameters (p, d, q) × (P, D, Q, s) to find the optimal configuration minimizing AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).
- External regressors (e.g., region, academic level) were added to the SARIMAX model to improve forecasts.

2.3.5 Model Training and Validation

- Trained SARIMAX on historical data.
- Validation done using **Walk-Forward Validation**:
    - Sequentially trained on past data and forecasted the next period.
    - Expanded the training window each time, simulating realistic prediction conditions.
- Performance was monitored on hold-out periods (test sets not seen during training).

2.3.4 Model Selection Rationale

SARIMAX was chosen as the final model based on:

- Ability to model **trend**, **seasonality**, and **external factors** together.
- **Lower forecasting error** (lower MAE, RMSE compared to machine learning models).
- **Better generalization** to unseen future periods compared to overfitting ML models.
- Simpler **interpretability**: coefficients directly related to temporal patterns and external influences.

Phase	Approach	Result
Phase 1	ML Models (Random Forest, XGBoost, CatBoost)	Good short-term accuracy, unstable long-term forecasts
Phase 2	ARIMA	Good trend modelling, but insufficient for seasonality and external influences
Phase 3	SARIMAX	Best fit: modelled trend, seasonality, and exogenous variables, stable long-term forecasts

- Machine learning models (**Random Forest, XGBoost, CatBoost**) **performed well on training data**, but **failed for long-term forecasts** due to ignoring time order and sequential dependence.
- **Polynomial Regression** highly overfit the data and produced unrealistic projections when extrapolating beyond the training set.
- **ARIMA** improved stability but **lacked flexibility** for capturing seasonality or external factors like region or academic level.
- **SARIMAX** provided the **best balance** between accuracy and real-world forecasting needs, thanks to its ability to model **seasonality + trend + external regressors** together.

2.4 Scenario Planning

Scenario planning involved creating three distinct projections:

- **Baseline:** Direct model predictions without adjustments.
- **Optimistic:** Baseline forecasts increased by 10% to simulate favorable conditions (e.g., relaxed visa policies, economic booms).
- **Pessimistic:** Baseline forecasts decreased by 10% to simulate adverse conditions (e.g., global recession, policy tightening).

2.5 Dashboard Development

- **Tool Used:** Power BI Desktop
- **Features:**
  - Multi-page navigation allowing detailed drilldowns.
  - Scenario toggle slicer to switch between Baseline, Optimistic, and Pessimistic forecasts.
  - Filters by visa type, region, funding source, and year range.
  - Interactive visuals including line charts, bar charts, and decomposition trees.
- **Pages:**

1. Academic overview
2. Visa Type Distribution Analysis
3. Funding Source and Regional Enrollment Trends
4. Forecasted Enrollment Trends

2.6 Deployment

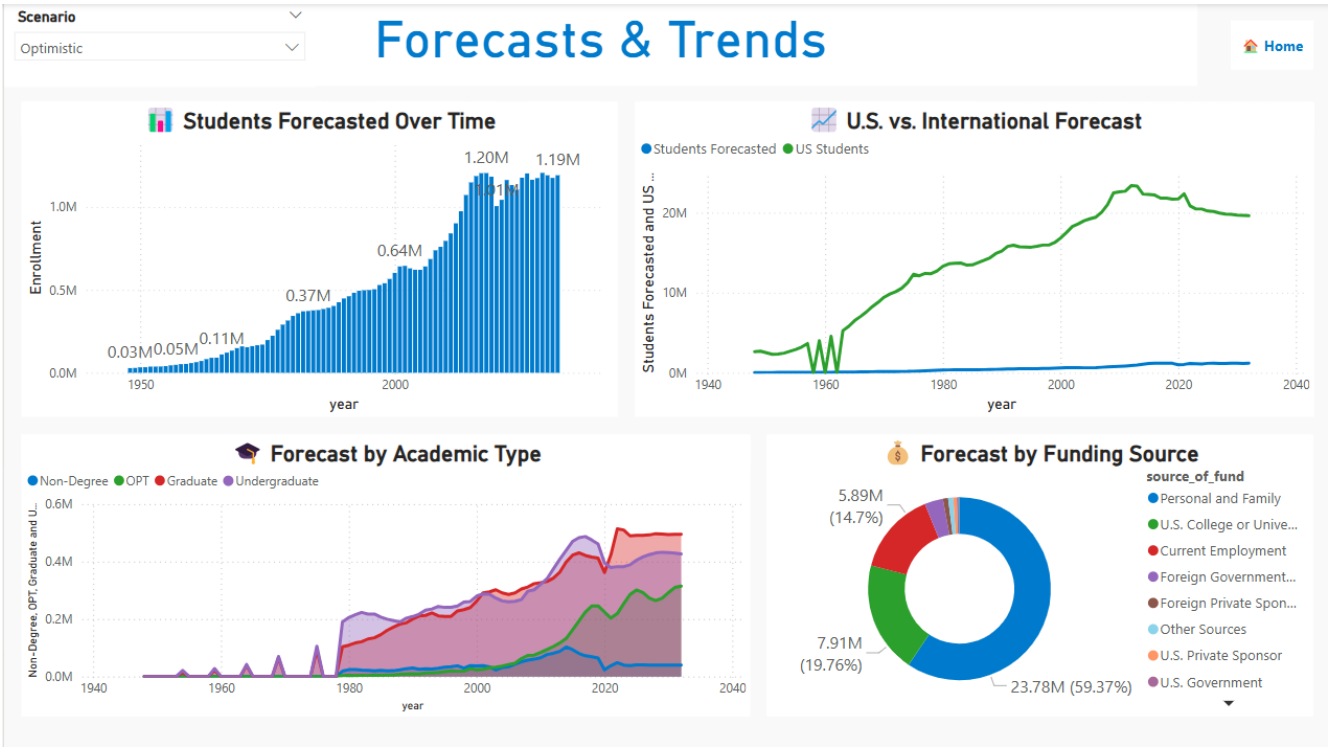
The finalized dashboard was deployed on the Power BI Service, enabling wider institutional access, mobile viewing, and scheduled data refreshes for continuous updates.

3. Results and Analysis

3.1 Forecasted & Trends

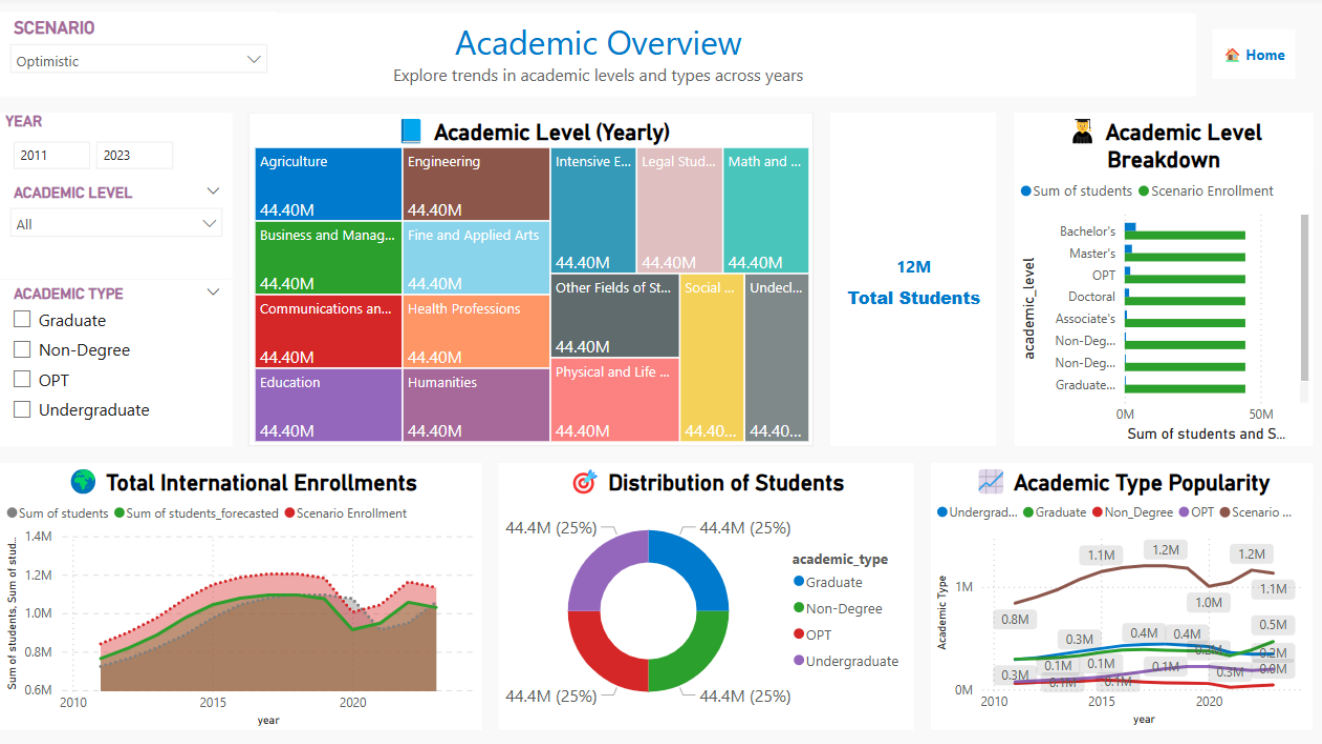
The SARIMAX model predicted:

- A moderate annual growth rate (approximately 3-4%) in full-time international enrolment.
- Part-time enrolments remaining largely stable, with minor cyclical fluctuations influenced by external factors.



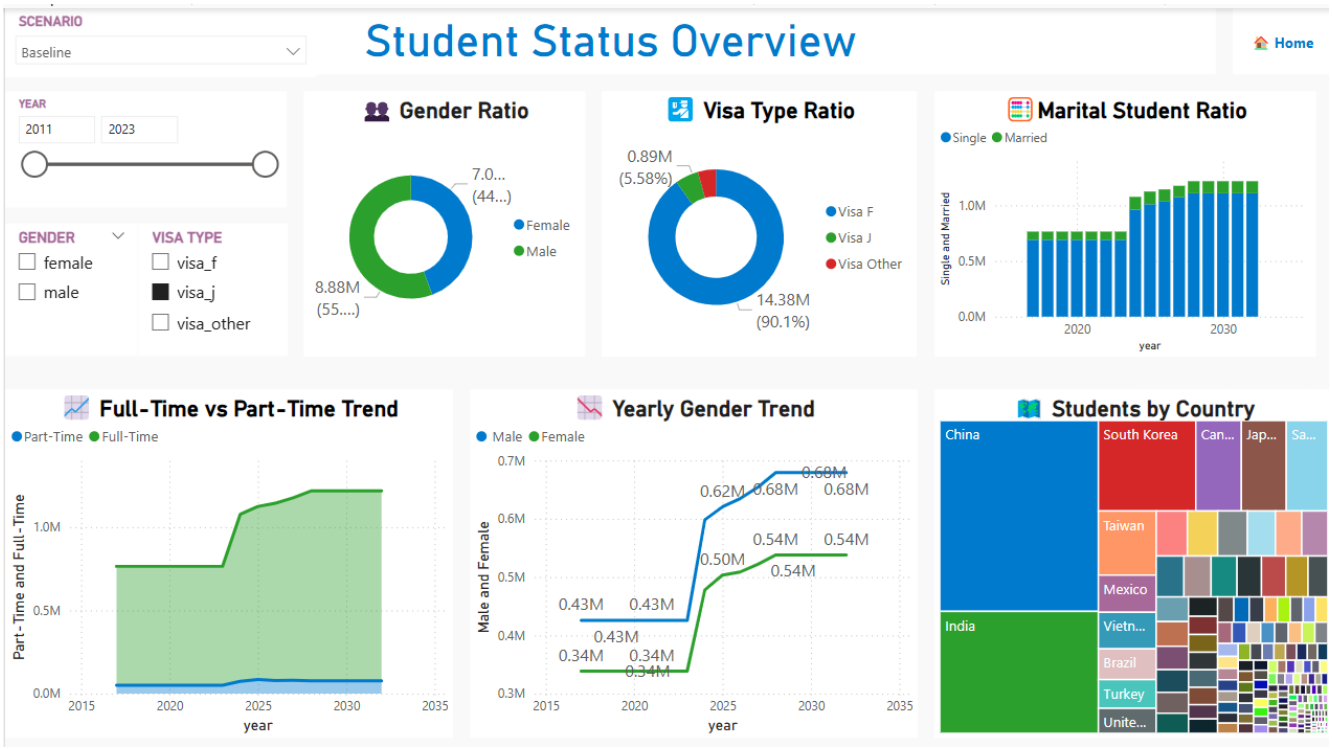
3.2 Academic Overview

- **Optimistic:** Suggests cumulative enrollment could exceed 1.2 million by 2030, driven by policy liberalization and robust economic recovery.
- **Baseline:** Predicts enrollment reaching approximately 1.1 million by 2030, assuming moderate recovery trends.
- **Pessimistic:** Projects slight contraction, stabilizing around 950,000 students, reflecting potential policy restrictions and global uncertainties.



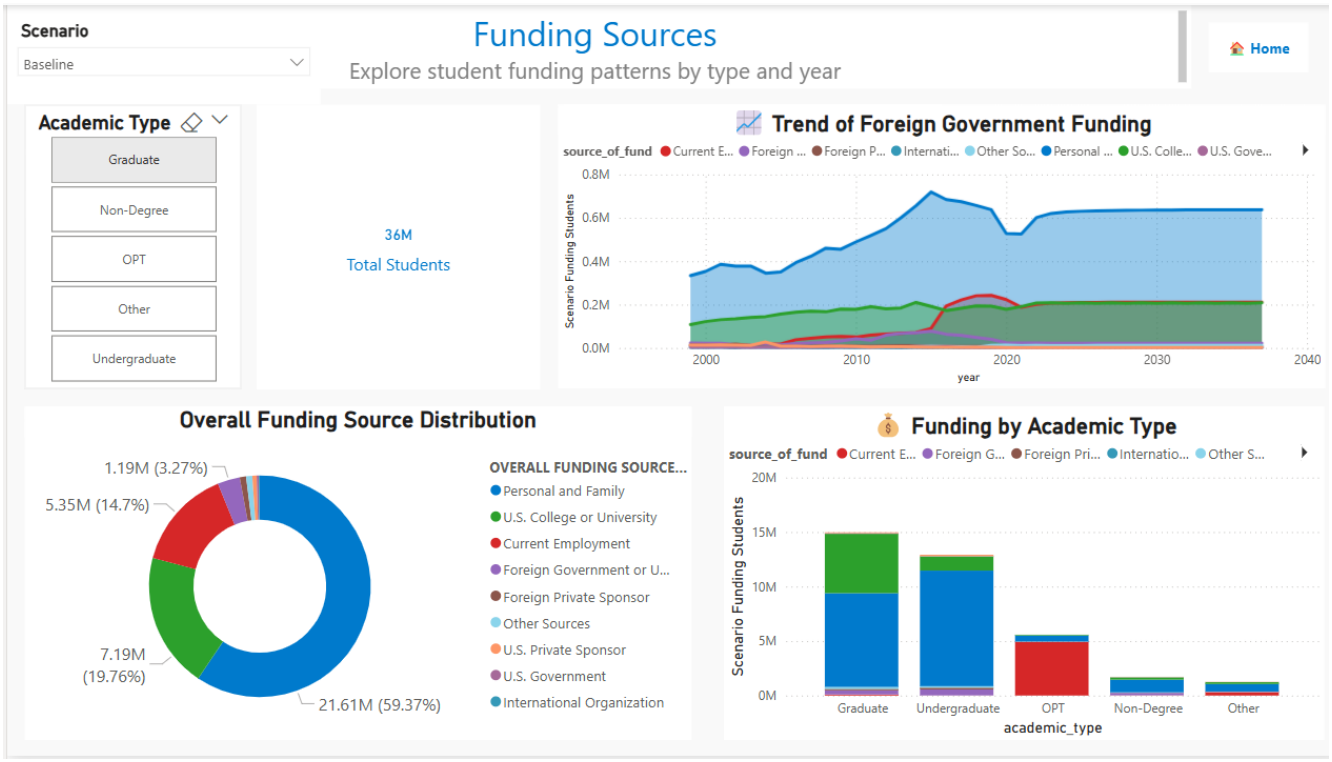
3.3 Visa Type Trends

- F-1 visa students constitute the largest proportion and are projected to continue dominating international enrollments.
- Declines were observed in J-1 visa holders, suggesting reduced non-degree exchange activities and research scholar participation.



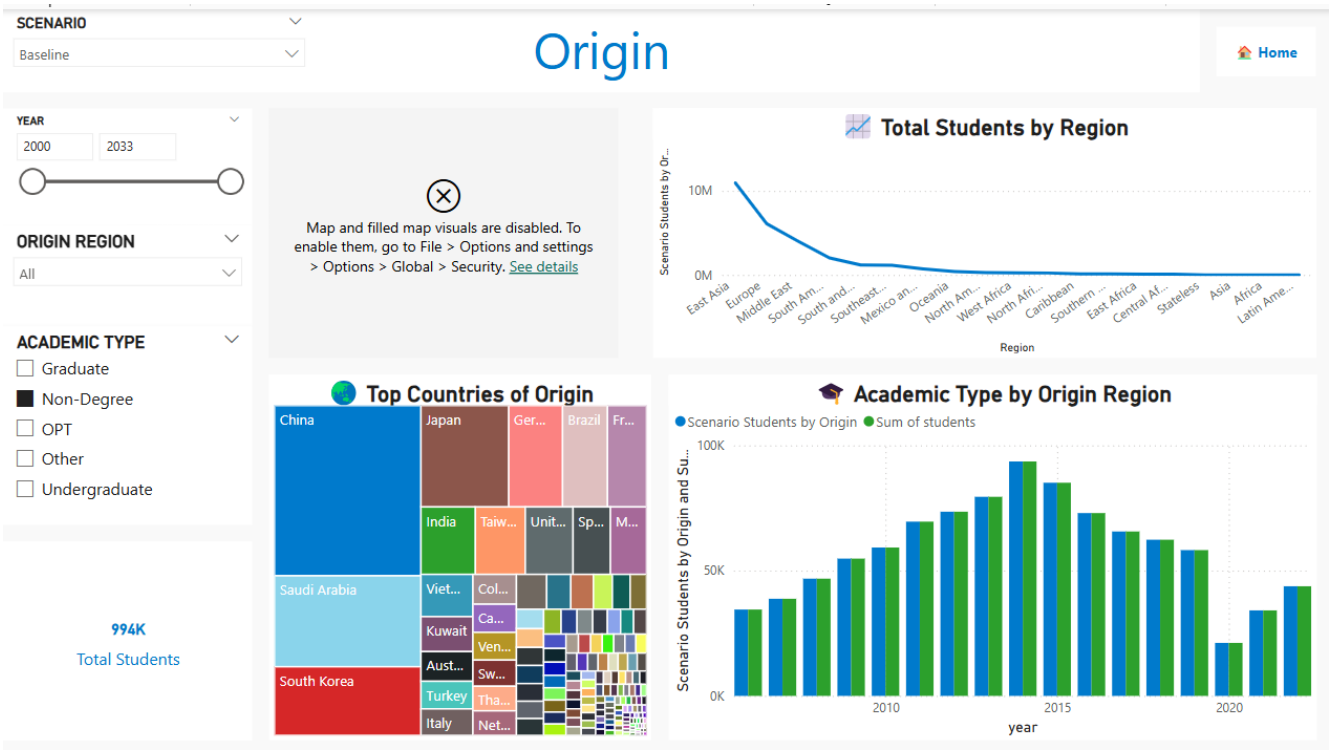
3.4 Funding Sources

- **Self-funded students:** Predicted to decline marginally, potentially due to rising education costs and stronger competition from other countries.
- **Scholarship-funded students:** Expected to rise, particularly within STEM programs supported by institutional and governmental initiatives.



3.5 Country of Origin Trends

- China and India remain top contributors to international student enrolments.
- Southeast Asia (Vietnam, Indonesia, Philippines) shows emerging upward trends, suggesting future diversification opportunities.



4. Discussion

The EDU PREDICT TOOL equips institutions with:

- **Strategic Insights:** Data-driven projections to guide international recruitment strategies and enhance diversity targets.
- **Resource Planning:** More accurate estimations of demand for student services, campus housing, and academic advising resources.
- **Financial Planning:** Projections help anticipate tuition revenue trends and support budget allocation decisions.
- **Risk Mitigation:** Early warning signs, such as over-reliance on specific countries or visa types, enable proactive diversification strategies.

Overall, the tool acts as an enabler for evidence-based decision-making across institutional levels.

5.MVP Development and User Manual

MVP Description

To ensure that the forecasting results are accessible and actionable for stakeholders, the team developed a Minimum Viable Product (MVP) in the form of an interactive Power BI dashboard.

This dashboard integrates the SARIMAX forecast outputs and provides dynamic, scenario-based visualizations for enrollment trends across different regions, academic levels, and fields of study.

The MVP enables university administrators and policymakers to:

- Explore enrollment trends under Baseline, Optimistic, and Pessimistic scenarios.
- Apply dynamic filters to customize the analysis by year range, region, or academic category.
- Visualize long-term trends and scenario-specific projections interactively.

The dashboard is publicly accessible at: Dashboard link

User Manual Overview

The following key functionalities are available in the MVP dashboard:

- **Scenario Selection**:
  Navigate between Baseline, Optimistic, and Pessimistic scenario pages using the navigation pane.
- **Filter Controls**:
  - **Region Filter**: Select specific geographic regions (e.g., Asia, Europe, Africa).
  - **Academic Level Filter**: Toggle between undergraduate and graduate level forecasts.
  - **Field of Study Filter**: Narrow results to specific academic disciplines.
  - **Year Range Filter**: Adjust the time frame for forecasting visualization.
- **Graphical Interpretation**:
  Line charts and bar charts display enrollment projections dynamically based on filter selections.
- **Access and Navigation**: The dashboard is hosted on GitHub Pages for free public access and optimized for mobile and desktop devices.

6.Challenges and Limitations

- **Data Granularity:** Lack of monthly or semester-level enrollment data limited the ability to capture short-term seasonal variations.
- **Model Assumptions:** While SARIMAX is robust, it cannot fully account for sudden geopolitical disruptions, pandemics, or rapid policy shifts.
- **Scenario Multipliers:** The ±10% assumption used for scenario planning is simplistic and does not account for real-world elasticity in demand.
- **Technical Constraints:** Limited access to live immigration datasets and some restrictions in Power BI service refresh rates posed minor hurdles to real-time updates.

7.Conclusion and Future Work

This project successfully developed a comprehensive dashboard that forecasts and visualizes future international enrollment trends dynamically. The integration of machine learning with business intelligence tools like Power BI represents a practical advancement for strategic enrollment management.

Future Enhancements:

- Introduce multivariate time series models such as Vector Auto Regression (VAR) or Facebook Prophet with external regressors.
- Incorporate live data feeds from global visa issuance databases and economic indicators.
- Expand geographic analysis to sub-national levels, allowing state-level or city-level trend exploration.
- Simulate more sophisticated scenarios by integrating macroeconomic projections, visa policy changes, and demographic shifts.
- Implement real-time alerts on key metric deviations for proactive institutional response.

8.References

- Institute of International Education (IIE), Open Doors Report.
- U.S. Immigration and Customs Enforcement (ICE), SEVIS Data.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*.
- Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting Methods and Applications*.
- Xiaoqian Wanga , Yanfei Kanga , Rob J Hyndmanb , Feng Lic, Distributed ARIMA models for ultra-long time series
- Hansika Hewamalage, Christoph Bergmeir, Kasun Bandara. Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions.
- Ilyas Varshavskiy, Elizaveta Stavinova, Petr Chunaev. Forecasting railway ticket demand with search query open data