Predicting Student Decisions

Abdul Wasay, Alina Dang, Edward Song, Malia Oguna

Duke University Data+ Program 2025

Project Motivation and Background

This project aims to develop a more accurate, data-driven model to predict whether admitted students will choose to enroll. While Duke's extensive admissions data provides a strong foundation, current yield models rely on limited variables and may overlook key student factors.

By analyzing 3 years of admissions records, we seek to uncover hidden patterns and improve yield prediction. Stronger models can support better institutional planning, including:

- Housing
- Financial aid
- Class sizing

Our approach was shaped by internal data analysis and a review of literature on yield modeling in higher education.

Year	ED/RD	Yield
2023	ED	0.97
2023	RD	0.36
2024	ED	0.96
2024	RD	0.37
2025	ED	0.97
2025	RD	0.38
2023	OVERALL	0.55
2024	OVERALL	0.59
2025	OVERALL	0.6

Figure 1: Yield rates by decision round, 2023–2025

Data Collection and Exploratory Analysis

Extensive data wrangling was performed to prepare the dataset for analysis, including handling missing values, converting categorical fields to standardized formats, and ensuring all modeling variables were properly encoded as numerical features. This preprocessing allowed for consistent analysis across all three years of admissions data.

Key factors influencing yield included:

- **Demographics** (income, gender, location)
- **Financial aid** (award amount, aid status)
- **Institutional and social influences** (reputation, programming, mentorship)

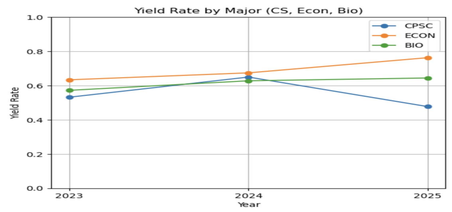

Figure 2: Yield Rates by Major Over time

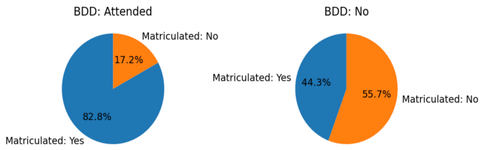

Figure 3: Yield Rates by Blue Devil Days Attendance Status

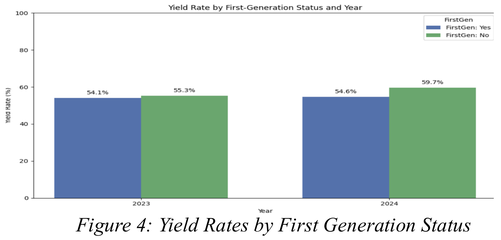

Figure 4: Yield Rates by First Generation Status

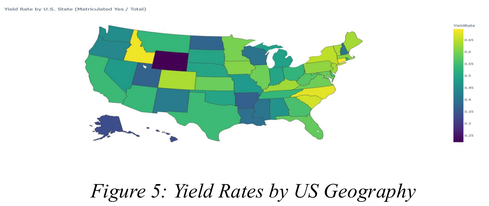

Figure 5: Yield Rates by US Geography

Prediction Models

Logistic Regression

- Two logistic regression models were trained: one using both ED and RD applicants, and another using only RD data to reflect undecided applicants.
- Features included financial aid, first-gen status, test scores, school, and one-hot encoded fields for country, major, state, and BDD attendance.
- The full model performed well (72.9% accuracy, 84.9% precision) but had moderate recall (63.7%), missing some matriculants. The RD-only model struggled, with low recall (16.9%) and precision (45.5%).

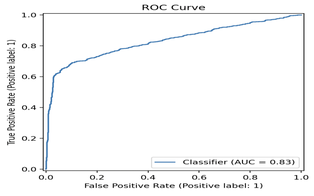

Figure 6: Logistic Regression ROC Curve

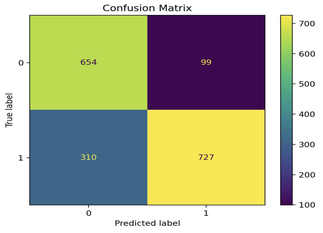

Figure 7: Logistic Regression Confusion Matrix

Random Forests

- Two Random Forest models were trained: one using both ED and RD applicants, and another using only RD data to focus on undecided applicants.
- Features included financial aid, first-gen status, standardized scores, school, and one-hot encoded variables for major, state, country, BDD attendance, and outreach category.
- The full model performed strongly (79.6% accuracy, 96.4% precision), with balanced recall (67.3%) and strong specificity (96.5%). The RD-only model showed lower recall (62.2%) and precision (71.1%), indicating greater difficulty predicting decisions in that group.

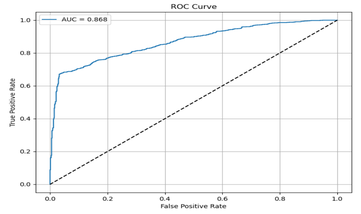

Figure 8: Random Forests ROC Curve

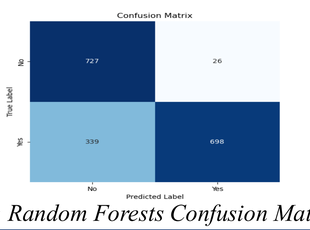

Figure 9: Random Forests Confusion Matrix

Predictive App Development

In order to visualize these effects, a Streamlit app was created that allows users to input hypothetical student profiles and receive predicted matriculation probabilities in real time. This interactive tool enables admissions officers to explore "what-if" scenarios—for example, how the probability of matriculation changes based on whether a student receives financial aid, or how different academic majors and geographic locations influence the outcome. The app generates predictions using the trained logistic regression and random forest models, all within a user-friendly interface.

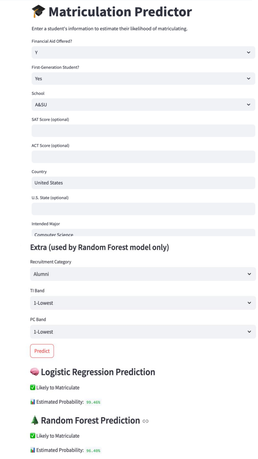

Figure 10: User interface of the app

Interactive Dashboard and Insights

To help visualize and interpret yield-related patterns, we developed an interactive Tableau dashboard for Duke Admissions. Users can explore how factors such as **major**, **state**, **financial aid**, **first-gen status**, and **admission round** influence the likelihood of matriculation.

The dashboard enables filtering by multiple attributes and includes yield rates across:

- Geographic regions (U.S. state map)
- Application cohorts
- Financial aid groups (e.g., with/without aid)
- Demographics (e.g., first-gen, gender)

In addition to exploratory analysis, the dashboard includes a **model performance panel** showing how **accuracy**, **precision**, **recall**, and **F1 score** vary across **conditional populations** (e.g., by ED/RD group, aid status, and more). This helps highlight where models perform well — and where predictive uncertainty remains.

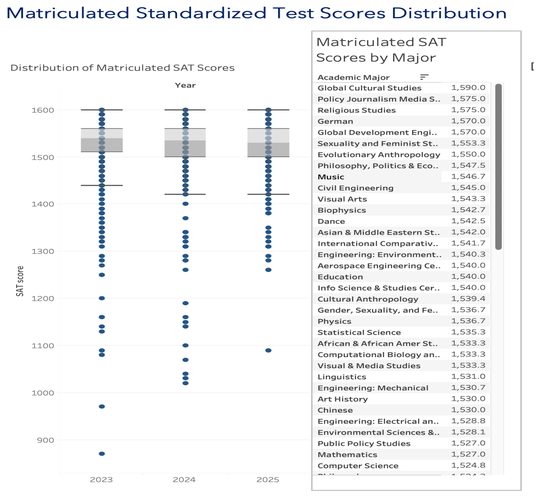

Figure 11: Standardized test score distribution for matriculants, segmented by academic year and major

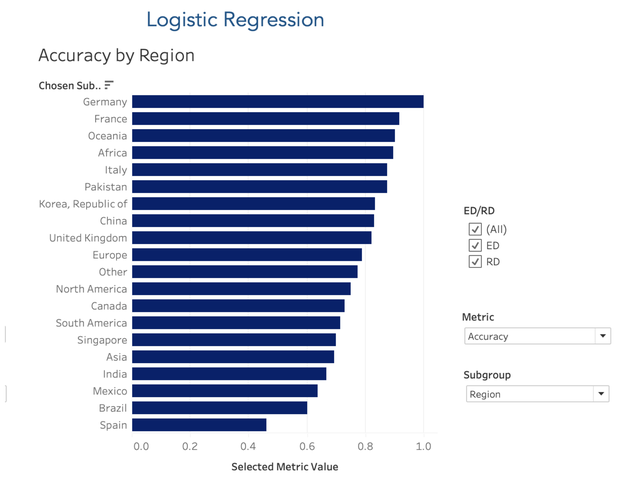

Figure 12: Logistic regression model accuracy by applicant region

Conclusion

This project demonstrates the effectiveness of integrating data preprocessing, statistical modeling, and interactive visualization to improve the prediction and interpretation of student enrollment decisions at Duke University. The Random Forest model exhibited strong predictive performance, while the Tableasu dashboard offers an accessible platform for exploring scenario-based insights across applicant subgroups. Collectively, these tools contribute to more data-driven decision-making in admissions strategy and institutional outreach.

Acknowledgement

We would like to thank the following mentors and collaborators for their guidance, support, and contributions to the success of this project:

Project Lead: Heather Mechler
Project Manager: Diego Rodriguez
Data + Program Manager: Gregory Herschlag