Using Data Science to Support the Student Success Journey: Admission - Retention - Graduation

The Decision to Enroll: Understanding Factors Shaping Student Path to Enrollment

Jack Beneigh*, Jocelyn Zhao, Rita Liu, Sohini Palit, Keegan Kang, Kevork Horissian

Office of Institutional Research and Analytics, Bucknell University, Lewisburg, PA

www.bucknell.edu/IR

Contact Info: **jjb043@bucknell.edu**

Introduction

Throughout this research, our team processed data from millions of applicants to better understand the reasons behind **student enrollment decisions**. The study aims to identify key factors that influence students' choice to attend our university, and further create a **model that will support our admissions** to support future students on their admission journey. This poster **outlines our methodology**, including decisions made during data processing, feature engineering, and model development. We employed advanced techniques to handle imbalanced data and applied hyperparameter tuning to improve model performance. As a result, our models achieved high precision and F1 scores, enabling accurate predictions of enrollment across **various stages of the admissions process**. These insights will help us proactively improve the admissions experience and foster stronger connections with prospective students.

Understanding the Data

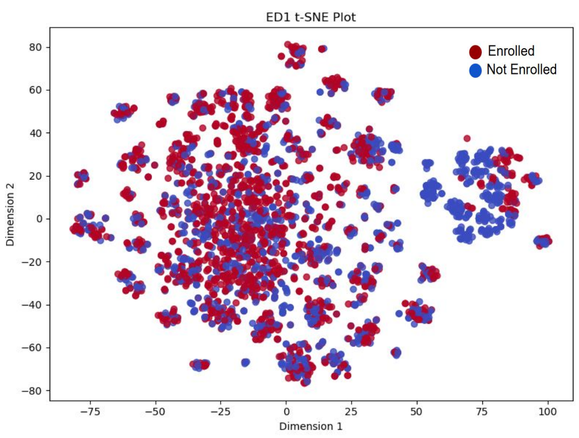
Admission Phases:

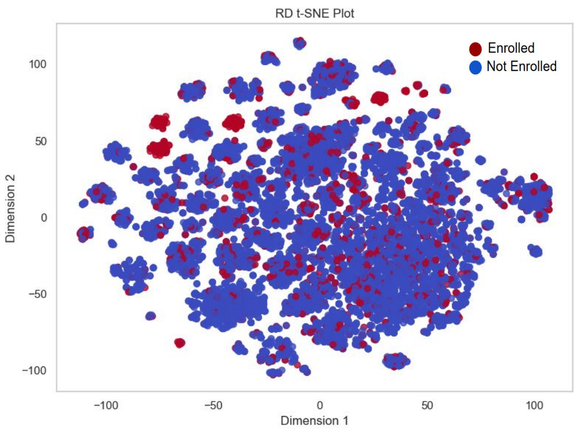
Early Decision I — Early Decision II — Regular Decision

Due to the differing timelines of each phase, certain variables were only available in specific contexts. These variations also contributed to imbalanced datasets, particularly within the Regular Decision group.

2020–2024 enrollment and student features were used to train the model, which was then applied to **predict 2025** outcomes

- **t-SNE plots** revealed large class imbalances in the **RD** phase and helped visually reduce the dataset's high dimensionality.





Exploring Statistical Significance:

During our data exploration stage, we used statistical techniques such as the **Chi-Squared Test** and **F-scores** to assess feature relevance. These methods enabled us to identify which features significantly influenced enrollment outcomes across the various admissions phases.

Feature Engineering

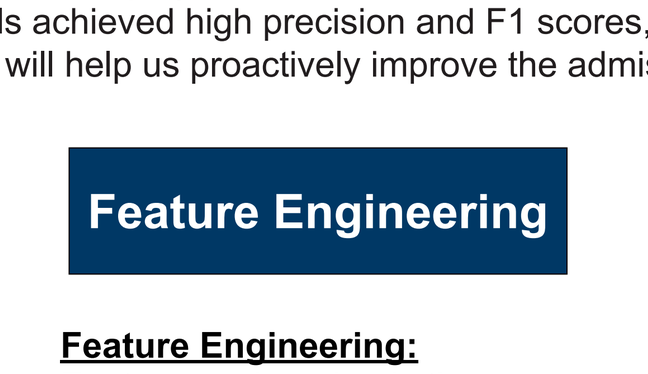
Feature Engineering:

Feature engineering is the process of converting raw data into meaningful inputs that enhance the performance of machine learning models.

Techniques our team used included **One-Hot Encoding**, **Binning**, and **Data Imputation.**

Feature Transformation:

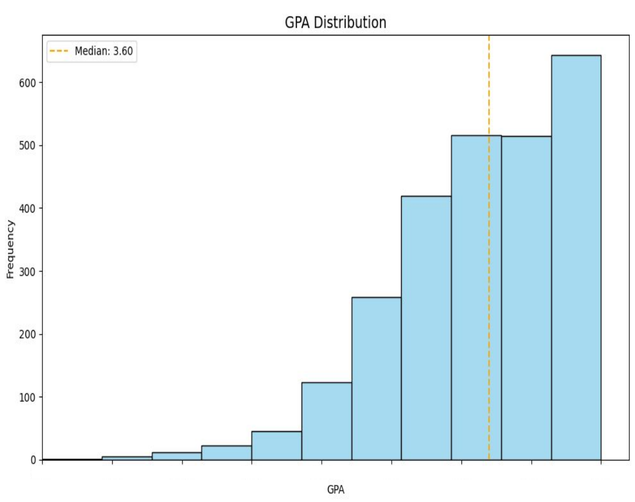
To help increase model's performance, we modified original features such as **ACT** and **SAT** scores to Categorical Bins using **Quartiling**.



Data Imputation for Missing Values:

Data imputation is a technique that preserves most of a dataset's integrity by replacing missing values with substitute estimates. Understanding the trends of certain features is important to decide which imputation technique to use, whether **KNN**, **Mean/Median**, or **Regression**.

Because of the **left-skewed** distribution in the graph below, our team used a Median Imputation to help restructure the data.



Modeling

Picking a Model:

Our team used multiple models throughout the research timeline, comparing their strengths and weaknesses using **Decision Trees, Random Forest, Support Vector Classifier, Logistic Regression, and Adaptive Boosting Models**.

Final models:

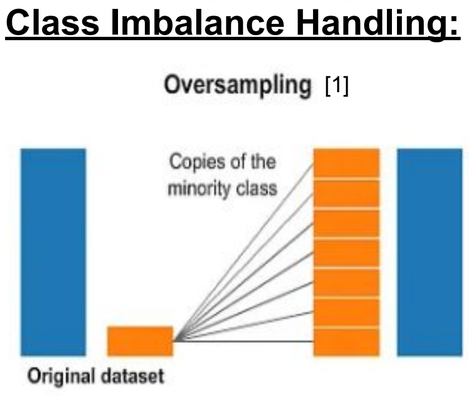
Early Decision I Model:

- **Logistic Regression**

Regular Decision Model:

- **Adaptive Boosting Classifier**

Re-Sampling Strategies for Class Imbalance Handling:



Our group used two oversampling techniques: **SMOTE** and **ADASYN** for our Regular Decision data.

Hyperparameter Tuning:

Hyperparameter tuning is the experimental process of identifying the most effective hyperparameter values to optimize the training of a predictive model. Parameters we tuned include:

Early Decision I Model:

- **solver:** Algorithm to use in the optimization problem
- **penalty:** Model configuration parameter

Regular Decision Model:

- **estimator:** Base model from which the boosted ensemble is built
- **criterion:** The function to measure the quality of a split
- **n-estimators:** The number of trees (or models) in an ensemble method

Results

Model Metrics and Results:

With large, imbalanced data in the RD phase, our metrics were lower than those of the previous ED1 model. Using a classification boosting model, the goal was to achieve high recall and F1 scores. Our model struggled with false positives on the minority class, but resulted in high precision for the majority class.

ED1 Metrics

	precision	recall	f1-score
Not Enrolled	0.91	**0.82**	**0.86**
Enrolled	0.94	**0.97**	**0.96**

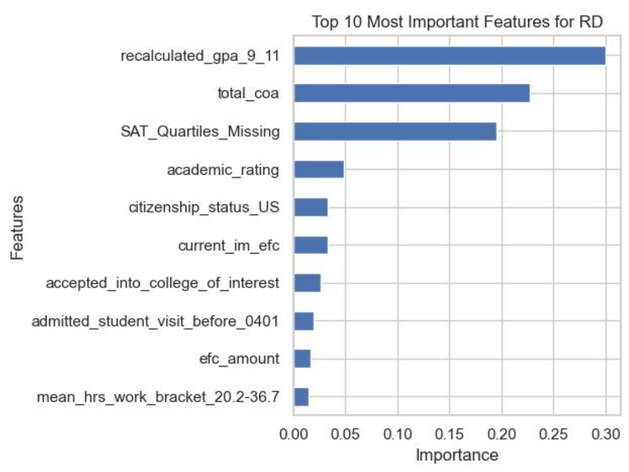
RD Metrics

	precision	recall	f1-score
Not Enrolled	0.99	**0.89**	**0.94**
Enrolled	0.38	**0.90**	**0.53**

Conclusions

Feature Importance:

The model's most influential predictors of RD enrollment reflect a combination of academic strength and financial context. Academic variables like recalculated GPA and institutional ratings are strong predictors for our model. Likewise, financial factors such as total cost and institutional EFC also carry weight, conveying how affordability impacts decisions.



References:

1. Alencar, R. (2017, November 15). Resampling strategies for imbalanced datasets. Kaggle. https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets
2. Goenner, C. F., & Pauls, K. (2006). A predictive model of inquiry to enrollment. Research in Higher Education, 47(8), 935–956. https://doi.org/10.1007/s11162-006-9021-8
3. DesJardins, S. L. (2002). An analytic strategy to assist institutional recruitment and marketing efforts. Research in Higher Education, 43(5), 531–553.