Predicting Student Drop Out Rates

Lia Cappellari, Branndon Marion, Elana Wadhwani

Overview

Goal: Predicting Students’ Dropout and Academic Success

Why is this important:

- Student dropout can lead to economic, social, and educational problems
- Take early interventions to improve student retention rates
- Develop specific initiatives to help students more easily and successfully access higher education

Research Question

What are the leading factors of student dropout in higher education?

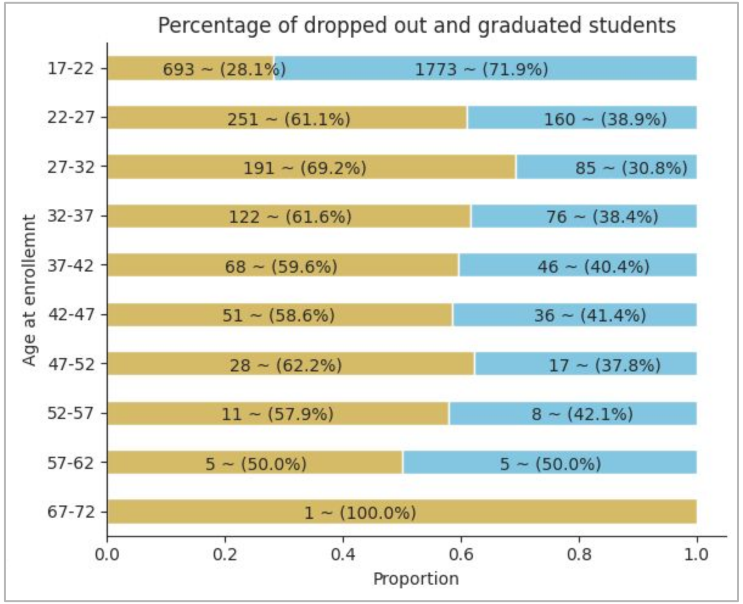
Previous research

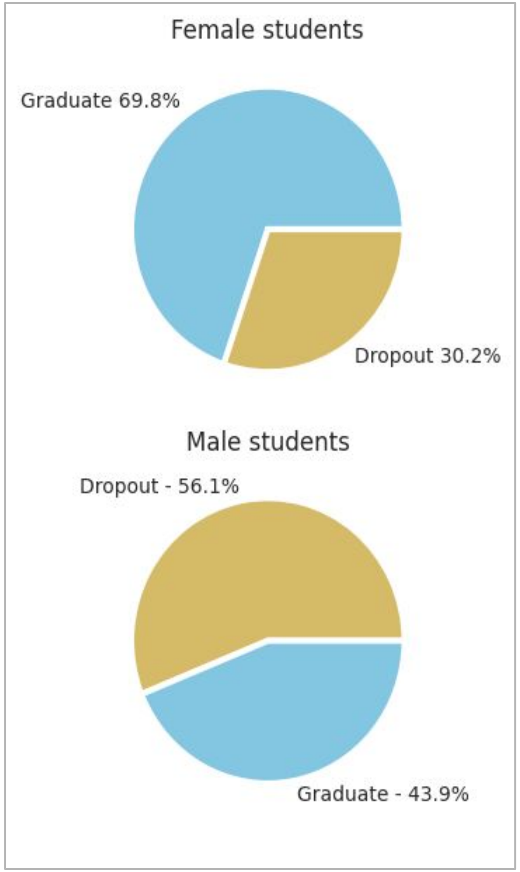
- Focuses primarily on north american universities
- Very complex models → hard to understand the importance of certain features

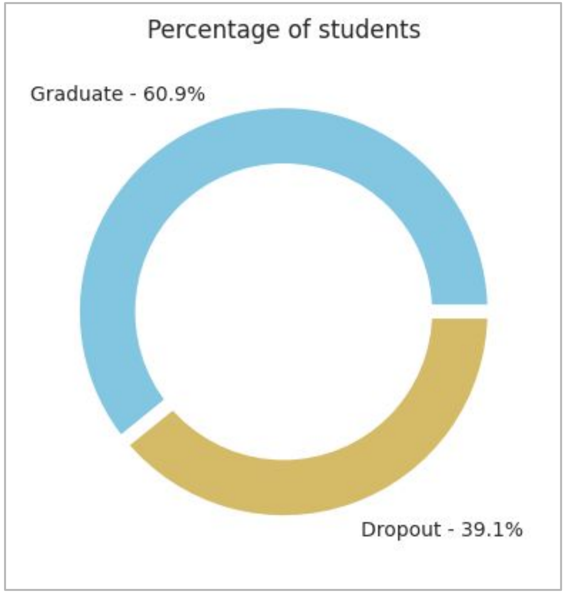
Data

Total Columns: 35

Total Rows: 4,424

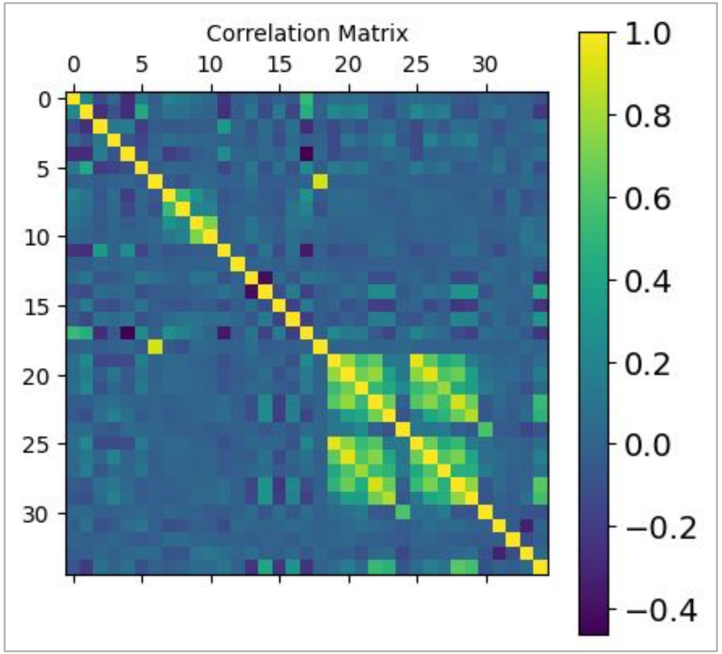


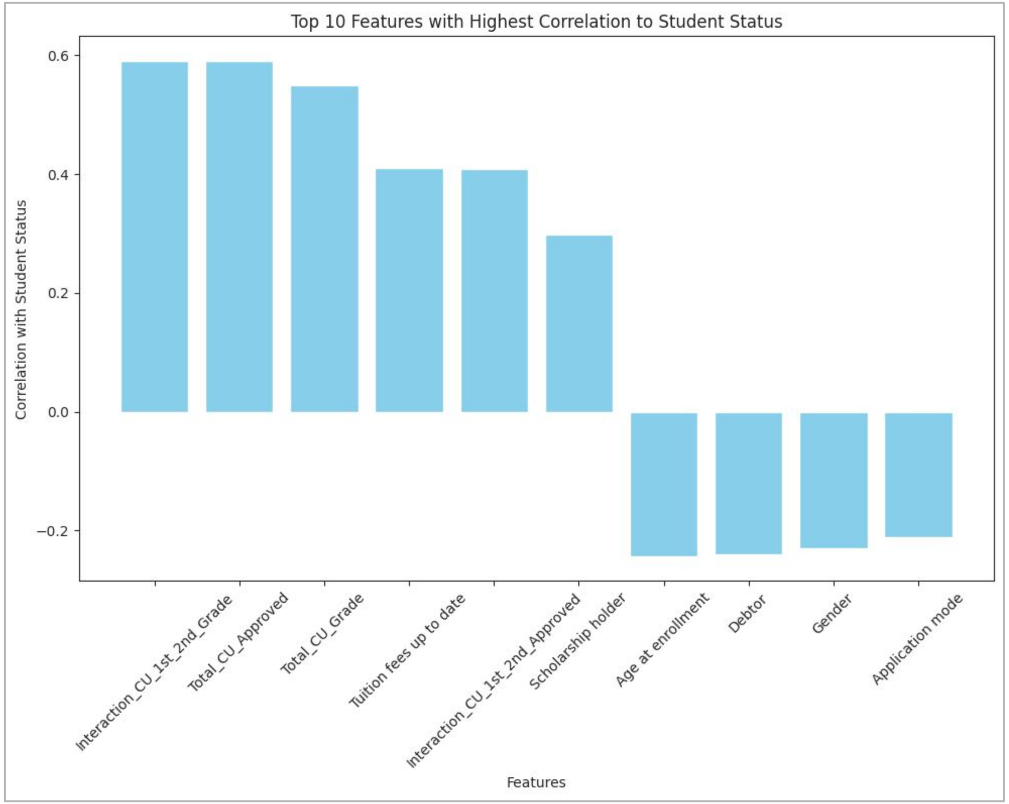




Feature Selection

Ended with 185 features



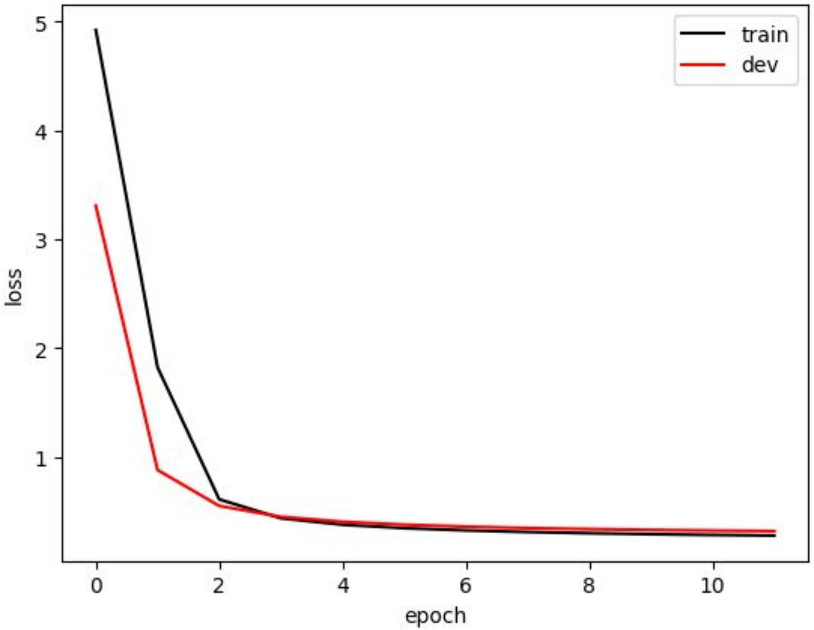


Highly correlated features were either dropped or combined

Data Preprocessing and Baseline

Preprocessing

- Parent qualifications were highly correlated - binned into 7 categories of different levels of education
- All semester features were rolled into one: enrolled - approved (failed/dropped classes)
  - Interaction calculated between first and second semesters
- Ages were binned
- Classes were imbalanced → split data proportionally for each set (60%, 20%, 20%)



Baselines

- Return the most common class (graduate)
- Logistic regression

Modeling

Majority Class Baseline

- Training
  - 61.22%
- Validation
  - 61.28%
- Test
  - 61.28%

Random Forest

- Training
  - 93.49%
- Validation
  - 91.58%
- Test
  - 90.74%

Neural Network

- Training
  - 99.33%
- Test
  - 86.36%

Logistic Regression

- Training
  - 92.31%
- Test
  - 88.38%

Overall, all three models significantly outperform the majority class baseline, with each providing notable improvements in accuracy

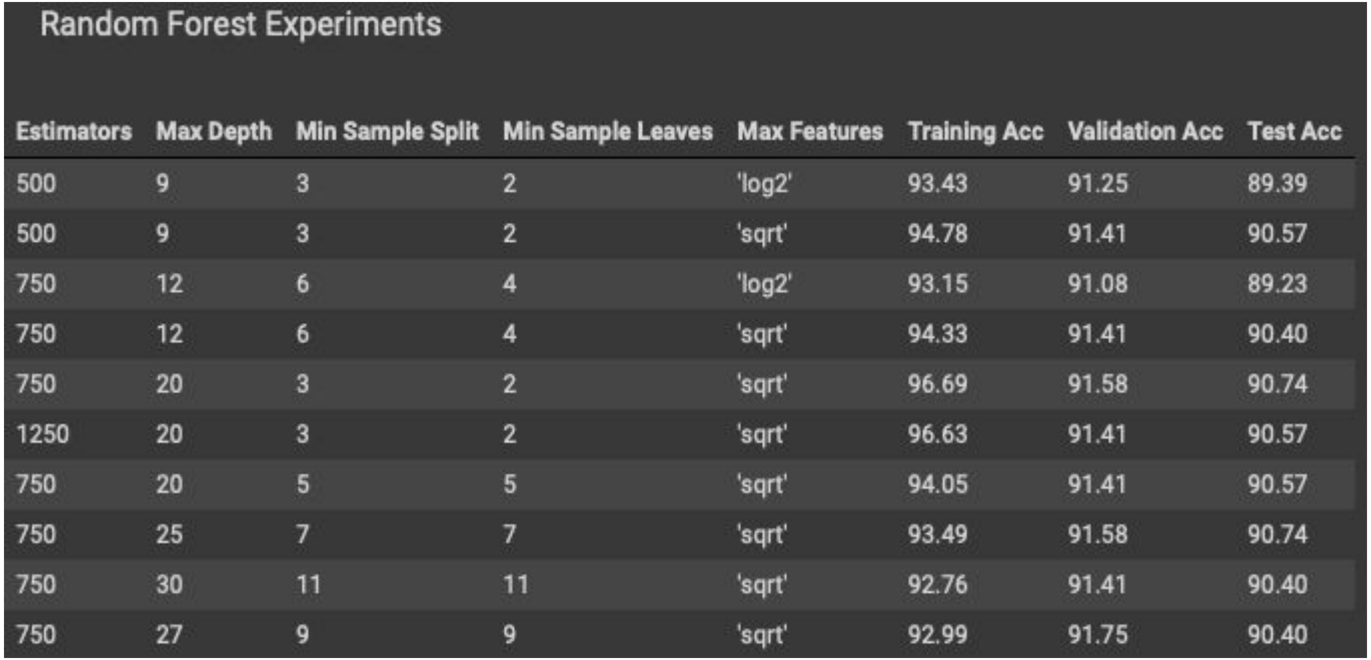
Experiments

Hyperparameters

1. Estimators
2. Max Depth
3. Min Samples Split
4. Min Samples Leaves
5. Max Feat

Best Models

1. Model 8
2. Model 5
3. Model 2

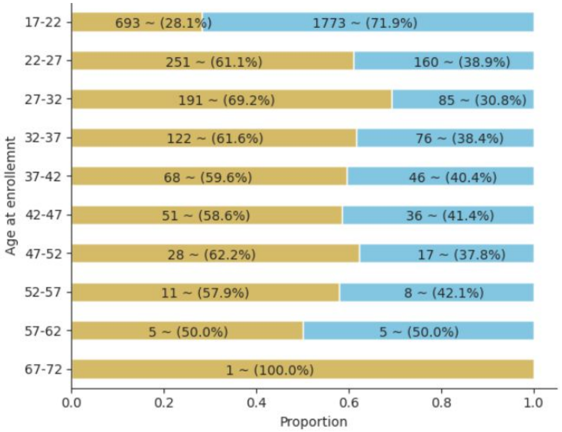


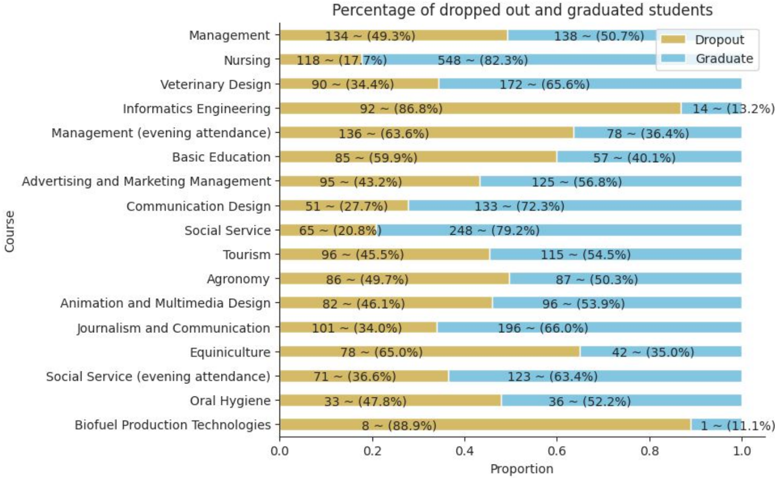
Random Forest Experiments

	Estimators	Max Depth	Min Sample Split	Min Sample Leaves	Max Features	Training Acc	Validation Acc	Test Acc
	500	9	3	2	'log2'	93.43	91.25	89.39
★	500	9	3	2	'sqrt'	94.78	91.41	90.57
	750	12	6	4	'log2'	93.15	91.08	89.23
	750	12	6	4	'sqrt'	94.33	91.41	90.40
★	750	20	3	2	'sqrt'	96.69	91.58	90.74
	1250	20	3	2	'sqrt'	96.63	91.41	90.57
	750	20	5	5	'sqrt'	94.05	91.41	90.57
★	750	25	7	7	'sqrt'	93.49	91.58	90.74
	750	30	11	11	'sqrt'	92.76	91.41	90.40
	750	27	9	9	'sqrt'	92.99	91.75	90.40

Conclusion - Important Features (Logarithmic Model)

- **More likely to graduate**
    - Mother’s and Father’s occupations had a large impact on whether a student graduated
    - Students who enroll at ages 18-20
    - Students who have one or more parent complete some/all of highschool
- **More likely to drop out:**
    - Students who have not paid all of their tuition
    - Students with a large number of failed/dropped classes
    - Students taking particular classes → management (evening), equinculture, social service (evening attendance)
- **Future research:**
    - Effects of tuition waivers on dropout rates
    - additional resources for working students





Contributions

Lia: EDA and Data Cleaning

Elana: Feature Engineering, Logarithmic baseline, common class baseline

Branndon: Random Forest and Neural Network models, Experiments