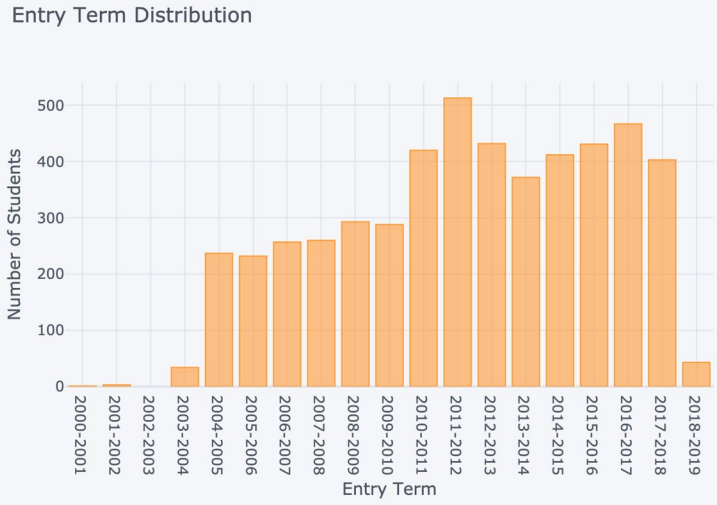
Predicting academic success of Masters students using application data

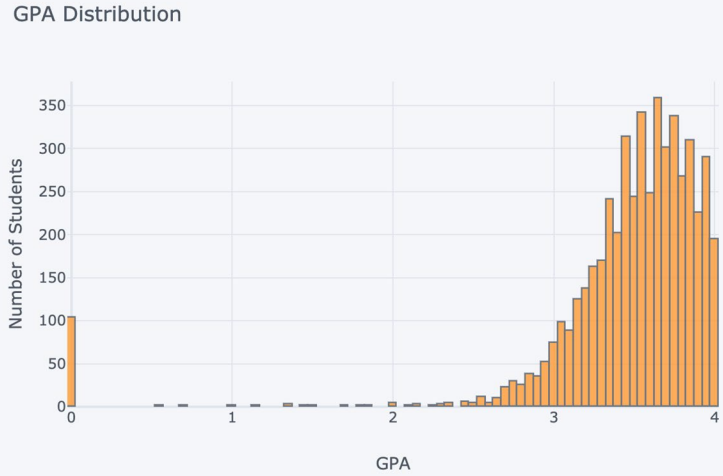
Karen Shen
Advised by: Boon Thau Loo and Ira Winston

Goals

Create a data-driven approach to help admissions staff identify which students will struggle to graduate and which students will succeed in the Penn Engineering Masters Program.

Find which factors in the application profile are most indicative of future academic performance.





Dataset

Penn Engineering Masters Program Applications

- Past education data (major, GPA, etc.), standardized test scores, gender, and international status

10,000+ application-institution records
5000+ graduation results

Tackling Data Challenges

Unexpected Missing Data

- Investigate uneven distribution of data in original dataset

Incomparable Data

- Standardize different spellings from user input, grading scales

Augmenting/Imputing Missing Information

- Institution selectivity, TOEFL/IELTS → English ability
- Match test scores to application (reverse engineer policy)

ML Approach

Classification

- Graduation status
- GPA thresholds (3.0, 3.3, 3.5, 3.7)

Evaluation

- Test set accuracy and F1 scores

Regression

- Predict GPA

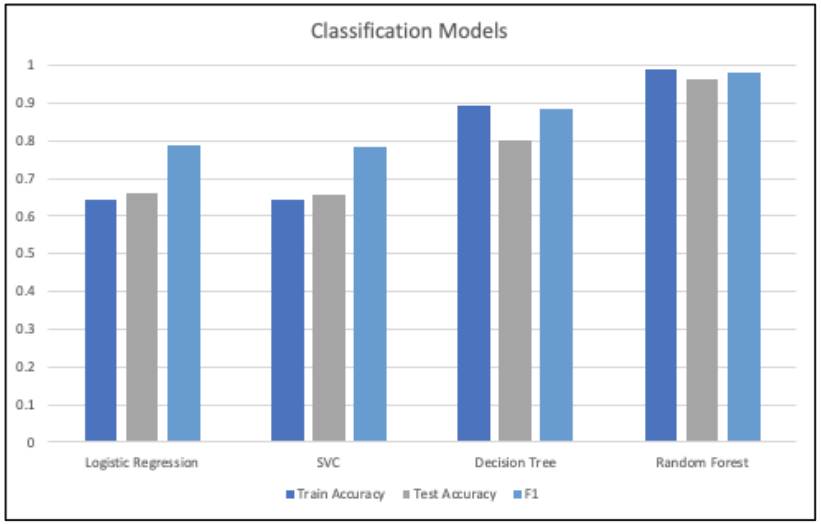
Evaluation

- GPA and percentile rank predictions
- Mean/Median Absolute Percent Error (MAPE/MdAPE)
- Bucket accuracy

Features and Models

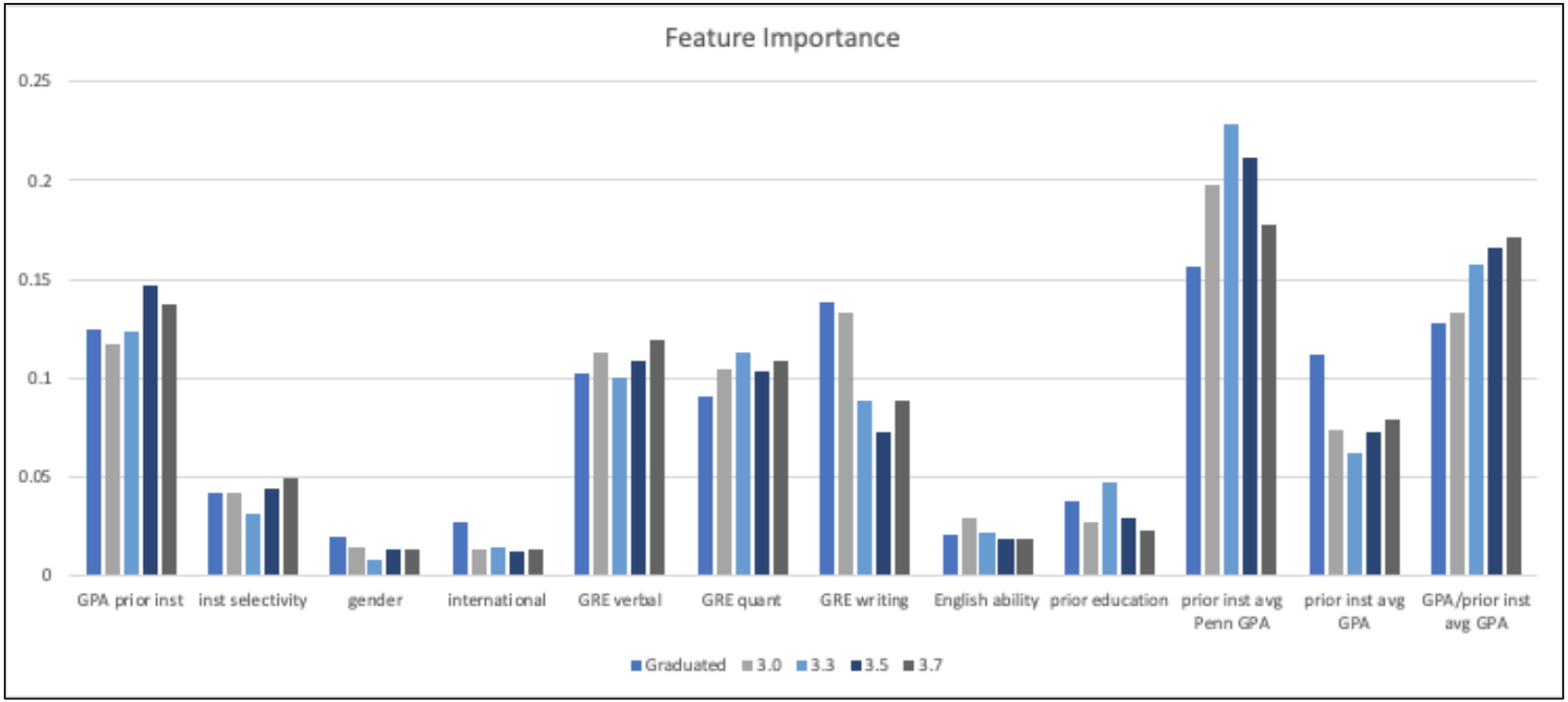
Class Imbalance

GPA (at prior institution), institution selectivity, gender, international status, GRE percentiles (verbal, quant, writing), English ability, number of prior experiences, prior institute's average GPA at Penn and prior institute's average GPA



Model	Bucketing GPA Acc	GPA MAPE/ MdAPE	Bucketing Ranking Acc	Ranking MAPE/ MdAPE
Linear Regression	0.538	0.145/.0516	0.489	1.995/0.472
Bayesian Regression	0.541	0.134/.0524	0.477	1.956/0.468
SVR	0.559	0.143/.0483	0.456	2.126/0.478
Random Forest	0.550	0.083/0.052	0.496	2.141/0.455
K Nearest Neighbors	0.490	0.093/0.063	0.380	2.463/0.558

Key Findings



Limitations

- Failure to graduate due to extenuating circumstances vs. lack of academic preparedness indistinguishable
- Curriculum difficulty and course load not reflected in GPA
- Limited school ranking data

Recommendations/Next Steps

- Prioritize applicants GPA over school ranking
- Use as many predefined options in application as possible
- Archive recommendation data (including ratings and text)
- Obtain more institution data on average GPAs and rankings
- Build a tool based on feature importance weights