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

<table>
<thead>
<tr>
<th>Model</th>
<th>Bucketing GPA Acc</th>
<th>GPA MAPE/ MdAPE</th>
<th>Bucketing Ranking Acc</th>
<th>Ranking MAPE/ MdAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.538</td>
<td>0.145/0.0516</td>
<td>0.489</td>
<td>1.995/0.472</td>
</tr>
<tr>
<td>Bayesian Regression</td>
<td>0.541</td>
<td>0.134/0.0524</td>
<td>0.477</td>
<td>1.956/0.468</td>
</tr>
<tr>
<td>SVR</td>
<td>0.559</td>
<td>0.143/0.0483</td>
<td>0.456</td>
<td>2.126/0.478</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.550</td>
<td>0.083/0.052</td>
<td>0.496</td>
<td>2.141/0.455</td>
</tr>
<tr>
<td>K Nearest Neighbors</td>
<td>0.490</td>
<td>0.093/0.063</td>
<td>0.380</td>
<td>2.463/0.558</td>
</tr>
</tbody>
</table>
Key Findings

Feature Importance

- GPA prior inst
- inst selectivity
- gender
- international
- GRE verbal
- GRE quant
- GRE writing
- English ability
- prior education
- prior inst avg
- Penn GPA
- prior inst avg GPA
- GPA/prior inst avg GPA

Graduated: 3.0, 3.3, 3.5, 3.7
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