UPENN Wharton Admissions Analysis

By: Lauren Weber, Sam Bogdanovsky, Alex Beckerman, Trevor Buchanan, Maddie Bartz

Dataset

6,194 observations with 10 features:

	application_id	gender	international	gpa	major	race	gmat	work_exp	work_industry	admission
0	1	Female	False	3.30	Business	Asian	620.0	3.0	Financial Services	Admit
1	2	Male	False	3.28	Humanities	Black	680.0	5.0	Investment Management	NaN
2	3	Female	True	3.30	Business	NaN	710.0	5.0	Technology	Admit
3	4	Male	False	3.47	STEM	Black	690.0	6.0	Technology	NaN
4	5	Male	False	3.35	STEM	Hispanic	590.0	5.0	Consulting	NaN

- For race: NaN denotes international student
- For admission: NaN denotes rejected
    - There are three admission results: admit, waitlist, and reject
    - For the sake of this project, we will be predicting “admitted” versus “not admitted”, grouping together waitlist and reject

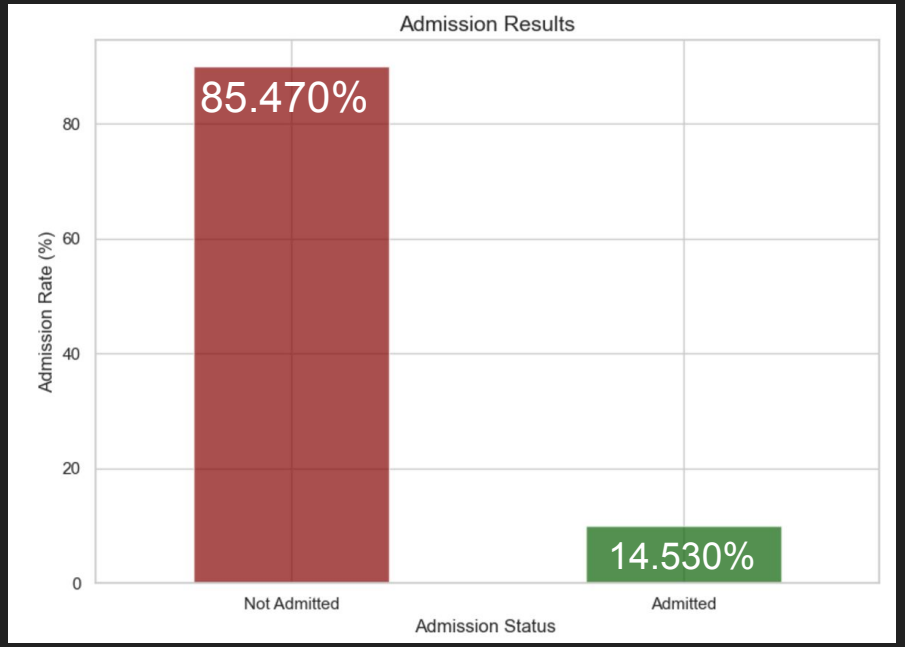
Dataset - Column Values

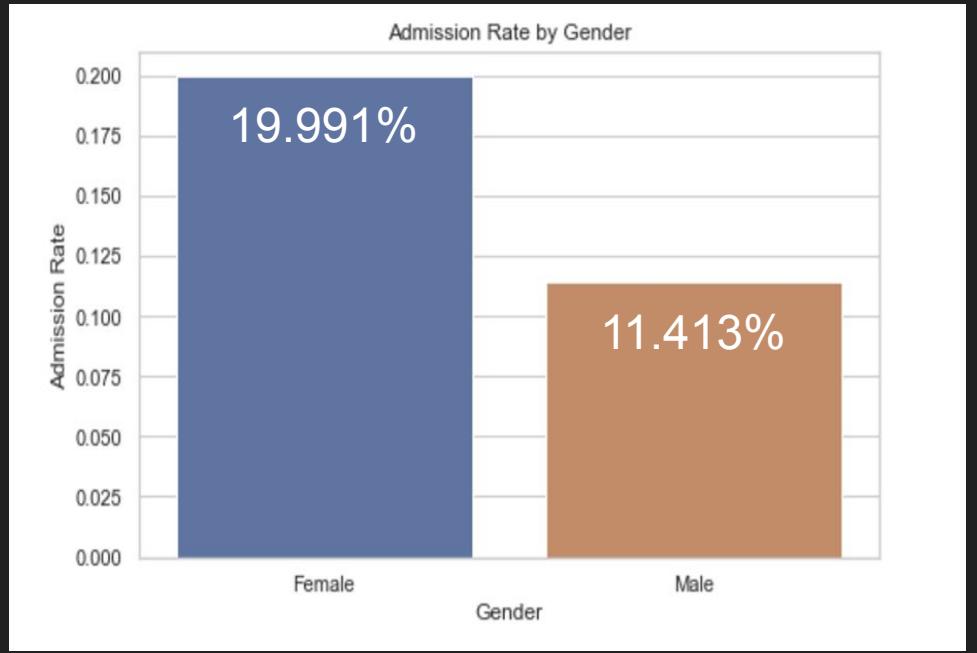
- gender
  - String - Male, Female
- major
  - String - Business, Humanities, STEM
- race
  - String - Asian, Black, Hispanic, White, Other, NaN
- work_industry
  - String - CPG, Energy, Health Care, Investment Management, Nonprofit/Gov, PE/VC, Retail
- admission
  - String - Admit, Waitlist, NaN (reject)
- gpa
  - Float
- gmat
  - Float
- work_exp
  - Float
- international
  - Boolean - True, False

Questions of Interest

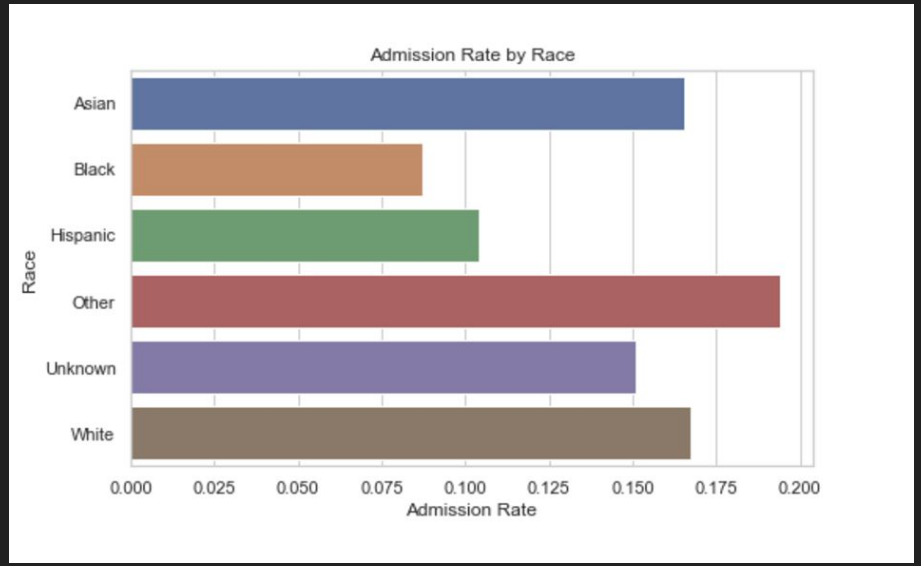
- Can we predict MBA acceptance at Wharton School of Business based on gender, GPA, GMAT, work experience, and/or undergraduate major?
- Which of these variables is most important for predicting MBA acceptance at Wharton?

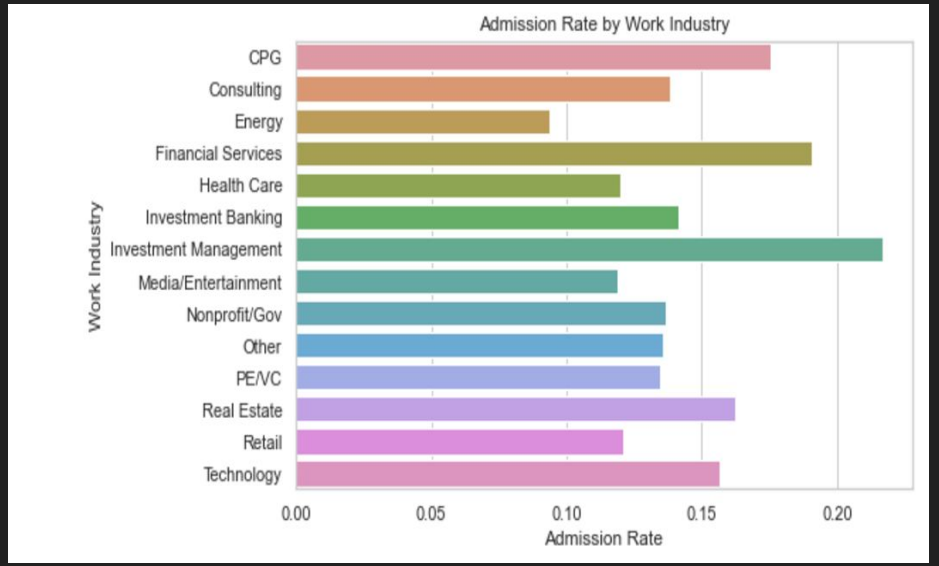
Exploratory Graphs



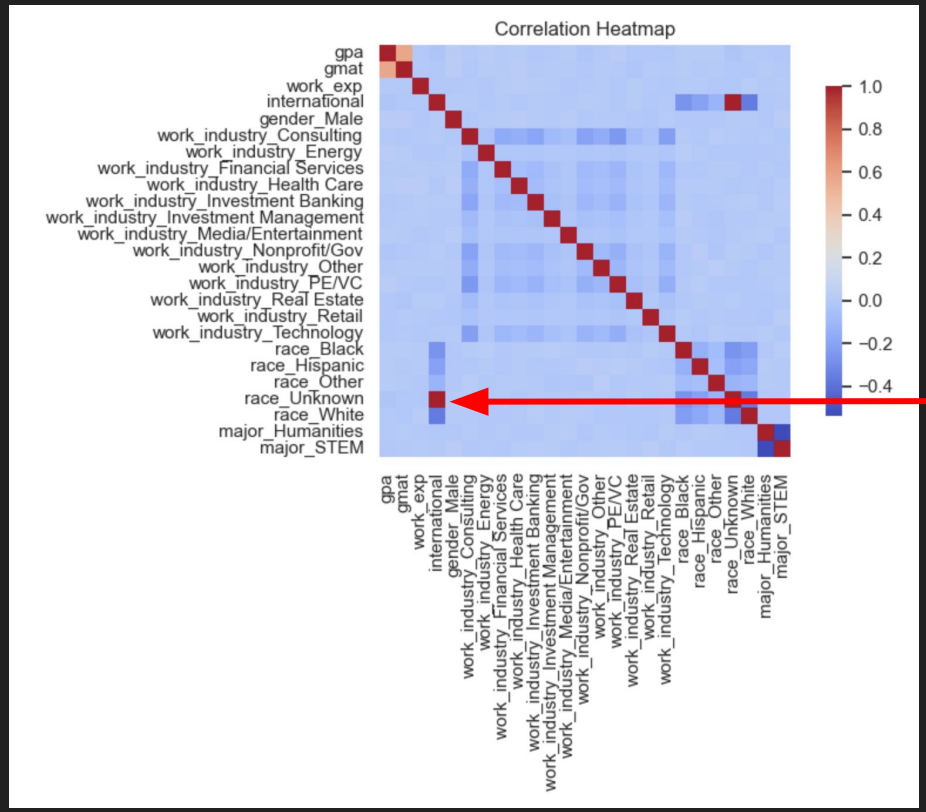


Exploratory Graphs Continued





Exploratory Graphs Continued



- race_Unknown was all the NaN values in race
  - race_Unknown = international

Feature Engineering

- Change NaN race to Unknown (International)
  - Didn’t use international as a feature since it will be represented by this
- Converted gender, work industry, race, and major to dummy variables for modeling
- Converted Admission status’ to 0 - rejected, 1 - admitted
- Used Standard Scaling for GPA, GMAT, Work Experience
  - GMAT values were significantly larger than GPA, Work Experience
- Split data into 80% Train, 10% Val, 10% Test
  - Significantly more rejects than admits -> oversampled training data

Model Selection

Grid Search Method:

- Logistic Regression with C of [.01, 1, 10, 100]
- Decision Tree with max depth of [10, 500,1000, None]
- KNN with #of neighbors [1, 2, 3, 4, 5]

Result:

- KNN (1 Neighbor) and Decision Tree (No Max Depth) both often chosen

Feature Selection

	feature	importance_mean	importance_std
1	gmat	0.171568	0.003586
0	gpa	0.169943	0.003072
3	gender_Male	0.106356	0.002651
2	work_exp	0.065682	0.003074
17	race_Black	0.046352	0.001853
20	race_Unknown (International)	0.043633	0.001826
22	major_Humanities	0.036686	0.001637
21	race_White	0.027442	0.001251
16	work_industry_Technology	0.025670	0.001673
23	major_STEM	0.025189	0.001194
13	work_industry_PE/VC	0.022454	0.000963
6	work_industry_Financial Services	0.021825	0.001189
4	work_industry_Consulting	0.020539	0.001125
18	race_Hispanic	0.019243	0.001185
8	work_industry_Investment Banking	0.015627	0.001189
11	work_industry_Nonprofit/Gov	0.012679	0.000856
12	work_industry_Other	0.011164	0.000901
9	work_industry_Investment Management	0.006646	0.000649
19	race_Other	0.005103	0.000467
7	work_industry_Health Care	0.004234	0.000447
14	work_industry_Real Estate	0.003895	0.000607
15	work_industry_Retail	0.000749	0.000183
10	work_industry_Media/Entertainment	0.000651	0.000108
5	work_industry_Energy	0.000459	0.000123

Decision Tree (No Max Depth)

	feature	importance_mean	importance_std
1	gmat	0.165846	0.003691
3	gender_Male	0.106739	0.002451
0	gpa	0.098862	0.003438
2	work_exp	0.095126	0.003629
22	major_Humanities	0.072700	0.002209
23	major_STEM	0.062252	0.001771
20	race_Unknown (International)	0.057505	0.001988
21	race_White	0.050618	0.001612
4	work_industry_Consulting	0.039372	0.001198
17	race_Black	0.028000	0.001221
13	work_industry_PE/VC	0.027683	0.000617
16	work_industry_Technology	0.022317	0.000627
18	race_Hispanic	0.020807	0.000908
8	work_industry_Investment Banking	0.020736	0.000732
11	work_industry_Nonprofit/Gov	0.020217	0.000598
6	work_industry_Financial Services	0.016218	0.000494
12	work_industry_Other	0.011881	0.000399
7	work_industry_Health Care	0.011459	0.000377
19	race_Other	0.011279	0.000428
9	work_industry_Investment Management	0.006542	0.000198
14	work_industry_Real Estate	0.004256	0.000146
10	work_industry_Media/Entertainment	0.002330	0.000089
5	work_industry_Energy	0.001482	0.000052
15	work_industry_Retail	0.001160	0.000041

KNN (1 Neighbor)

- Permutation feature selection using decision tree and KNN with 1 neighbor
- gmat, gpa, gender_male, and work_exp most important in both

Feature Selection - Continued

- Small dataset permitted this feature selection
    - Fit many models with different number of features
- Accuracy of model based off number of top features used
- After 2 features the model changes only slightly

```
{1: 0.687,
 2: 0.787,
 3: 0.789,
 4: 0.808,
 5: 0.818,
 6: 0.821,
 7: 0.83,
 8: 0.811,
 9: 0.808,
 10: 0.818,
 11: 0.821,
 12: 0.825,
 13: 0.821,
 14: 0.831,
 15: 0.825,
 16: 0.813,
 17: 0.825,
 18: 0.818,
 19: 0.816,
 20: 0.828,
 21: 0.83,
 22: 0.813,
 23: 0.816,
 24: 0.813,
 25: 0.81}
```

Decision Tree (No Max Depth)

```
{1: 0.784,
 2: 0.825,
 3: 0.815,
 4: 0.81,
 5: 0.808,
 6: 0.798,
 7: 0.821,
 8: 0.821,
 9: 0.813,
 10: 0.826,
 11: 0.807,
 12: 0.805,
 13: 0.8,
 14: 0.805,
 15: 0.8,
 16: 0.803,
 17: 0.807,
 18: 0.805,
 19: 0.802,
 20: 0.807,
 21: 0.807,
 22: 0.807,
 23: 0.808,
 24: 0.808,
 25: 0.808}
```

KNN (1 Neighbor)

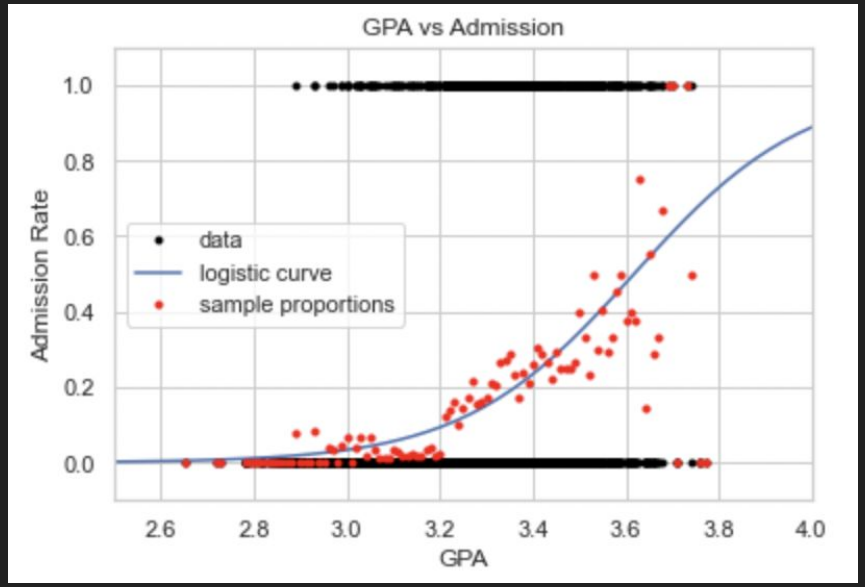
Model Evaluation

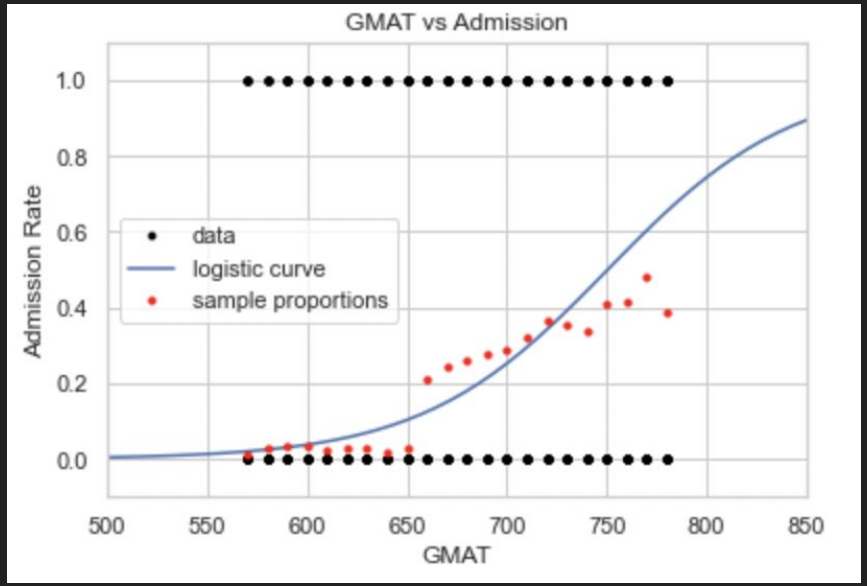
No model works well:

- Decision Tree highest accuracy value was 82.8%.
- KNN (1 Neighbor) highest accuracy value was 82.6%

For comparison, accuracy with simply guessing not admit every time is 85.47%

Logistic Regression Graphs





Conclusion

- Grid search was unable to choose a best model
- Most important variables consistently were: ‘gpa’, ‘gmat’, and “gender_male’
- Based off the given data and features, unable to explain well whether a person would be admitted or rejected
- Could be a factor outside of the dataset impacting admission rate more directly

Potential Next steps

Potential other factors that could impact admission predictions:

- Undergraduate School Ranking
- Quality of Reference Letters
- Quality of Personal Statements