

# GREat Expectations

---

Data-Backed Insights into  
Grad School Admissions

---

# Contents

- Problem Statement
- Data
- Case Studies
  - Case Study 1 - Carnegie Mellon University
  - Case Study 2 - University of Illinois Urbana-Champaign
  - Case Study 3 - University of California Los Angeles
- Comparative Analysis
- Next Steps

# Problem Statement

The project aims to:

- Study patterns in student profiles applying to graduate schools
- Compare, differentiate and gain insights about admissions
- Infer about definite existing correlations
- Prepare a budding applicant to build his profile in the best way
- Make conclusions about:
  - Admit/Reject status
  - Strength of an application
  - Probability of admission
  - Appropriate universities

# Data

- Changed our source to “admits.fyi”
- Rate limited for scraping
- Reverse engineered network calls to get college wise data
- Have pipeline setup
- Data has the following fields

| University | Status | Target Major | Term | GRE |   |     |       | TOEFL | UG College | UG Major | CGPA | Papers | Work Ex |
|------------|--------|--------------|------|-----|---|-----|-------|-------|------------|----------|------|--------|---------|
|            |        |              |      | Q   | V | AWA | Total |       |            |          |      |        |         |

# Case Study 1

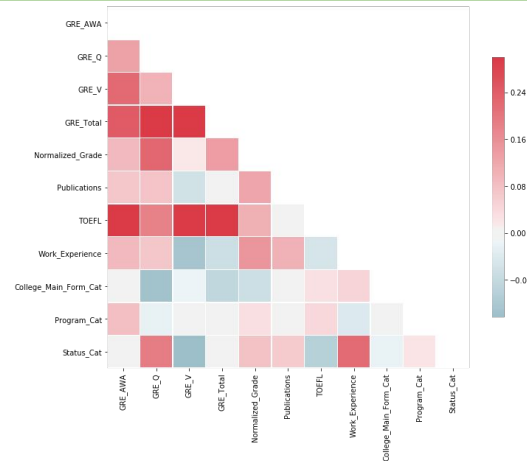
## Carnegie Mellon University

## 1. Data Statistics

- 1174 applicants, 28 features
- Sparse matrix, lot of missing data

## 2. Data Processing

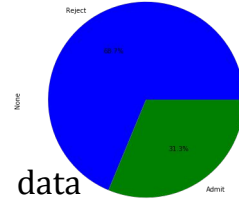
- Redundant features removed
  - Topper Grade, TOEFL R/S/L/W, Grade, Grade Scale
- Irrelevant features removed
  - Term, Year, University applied to
- Features coalesced
  - Grade normalized by dividing by upper limit & scaling:  $(\text{Grade}/\text{Grade Scale})$
- Number of features retained = 15
- Feature list - 'College\_Main\_Form', 'GRE\_AWA', 'GRE\_Q', 'GRE\_Total', 'GRE\_V', 'Normalized Grade', 'Program', 'Publications', 'Status', 'TOEFL', 'Target\_Major', 'Term', 'Undergrad\_Major', 'Work\_Experience'
- Columns with non-numeric values converted to categorical



|       | GRE_AWA     | GRE_Q       | GRE_Total   | GRE_V       | Normalized_Grade | Publications | TOEFL       | Work_Experience | College_Main_Form_Cat |
|-------|-------------|-------------|-------------|-------------|------------------|--------------|-------------|-----------------|-----------------------|
| count | 1174.000000 | 1174.000000 | 1174.000000 | 1174.000000 | 1174.000000      | 1174.000000  | 1174.000000 | 1174.000000     | 1174.000000           |
| mean  | 3.927598    | 165.217206  | 322.086031  | 156.868825  | 0.819658         | 0.248722     | 109.669506  | 10.608177       | 139.411414            |
| std   | 0.392650    | 3.680970    | 6.711718    | 5.254625    | 0.096249         | 0.531508     | 5.600646    | 15.446957       | 80.132798             |
| min   | 2.500000    | 140.000000  | 280.000000  | 134.000000  | 0.520000         | 0.000000     | 82.000000   | 0.000000        | 0.000000              |

### 3. Inferences - Data size = (1174, 18)

- Admits ratio over all students => 32%
  - On doing feature comparison across admits

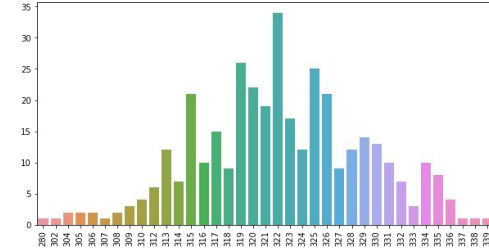
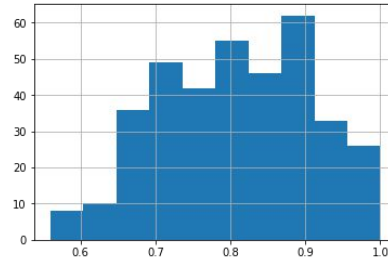
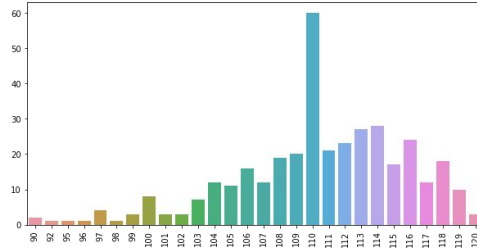


data

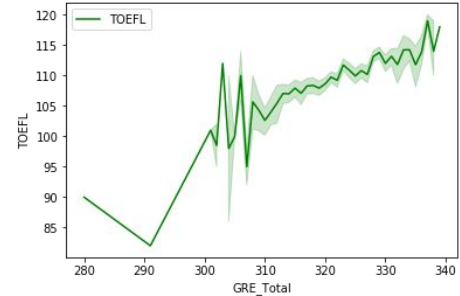
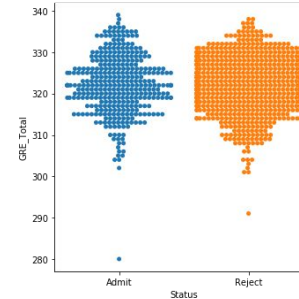
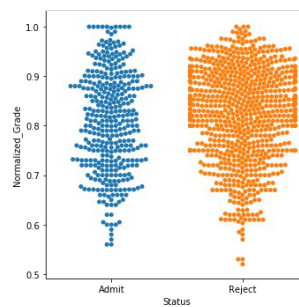
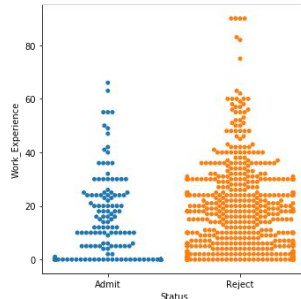
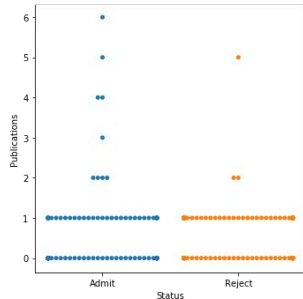
and

overall

data:

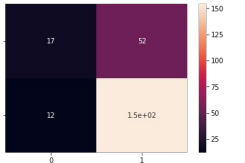
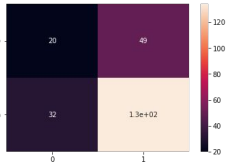
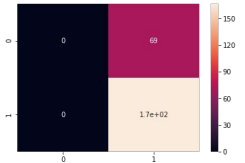
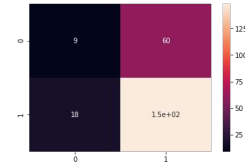
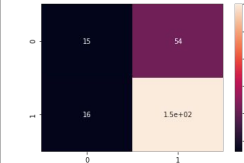


- Very high correlation between GRE and TOEFL: proves that TOEFL is just a bar to measure english
- Accept vs Reject
  - GRE, Normalized Grades, Publications, Work Ex most important



- Admission prediction

Modelled a binary classification task on an 80-20 data split

| Classifier            | Logistic Reg  | Decision Tree   | Random forest  | Gradient Boosting   | SVM   |
|-----------------------|---|---|--|---|---|
| Accuracy              | 0.72  | 0.65  | 0.70   | 0.67  | 0.75  |
| Classification Report | <pre> precision    recall 0         0.66     0.27 1         0.73     0.93 </pre>  | <pre> precision    recall 0         0.38     0.29 1         0.73     0.81 </pre>  | <pre> precision    recall 0         0.00     0.00 1         0.71     1.00 </pre>   | <pre> precision    recall 0         0.33     0.13 1         0.71     0.89 </pre>    | <pre> precision    recall 0         0.78     0.30 1         0.75     0.96 </pre>    |
| Confusion Matrix      |  |  |  |  |  |
| Important Features    |   | Normalized Grade, Undergrad major   | Work ex, GRE V, GRE Q  | Normalized Grade, Work ex, GRE  |   |



# Case Study 2

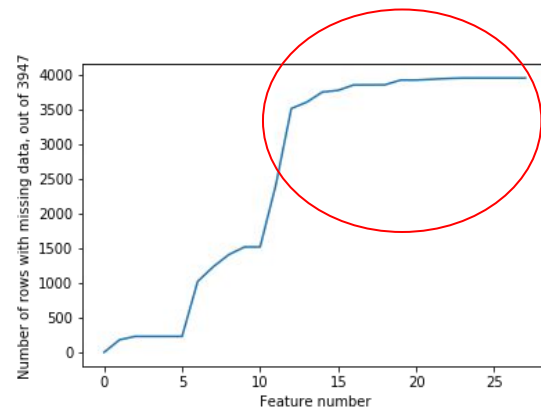
## University of Illinois Urbana-Champaign

## 1. Data Statistics

- 3947 applicants, 28 features
- Sparse matrix, lot of missing data
- 5 features had all data missing

## 2. Data Processing

- Features were removed based on number of nan value
- Number of features retained = 15
- Feature list - 'College\_Main\_Form', 'GRE\_AWA', 'GRE\_Q', 'GRE\_Total', 'GRE\_V', 'Grade', 'Grade\_Scale', 'Program', 'Publications', 'Status', 'TOEFL', 'Target\_Major', 'Term', 'Undergrad\_Major', 'Work\_Experience', 'Year'
- Rows with remaining nan values were removed
- Columns with non-numeric values converted to categorical
- Grade was normalized by dividing by the upper limit and scaling

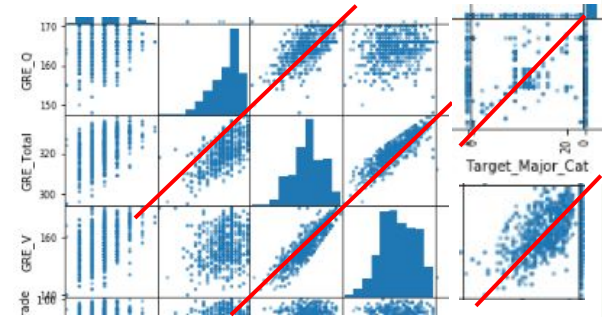


|      | GRE_AWA  | GRE_Q      | GRE_Total  | GRE_V      | Normalized_Grade | Publications | TOEFL     | Work_Experience |
|------|----------|------------|------------|------------|------------------|--------------|-----------|-----------------|
| mean | 3.861742 | 164.444129 | 322.091856 | 157.642992 | 0.819466         | 0.164773     | 109.64678 | 4.569129        |
| std  | 0.593659 | 3.475398   | 7.473307   | 5.961679   | 0.094418         | 0.662703     | 6.10937   | 11.923490       |
| min  | 2.500000 | 148.000000 | 295.000000 | 140.000000 | 0.370000         | 0.000000     | 85.00000  | 0.000000        |
| max  | 6.000000 | 170.000000 | 338.000000 | 170.000000 | 1.000000         | 7.000000     | 120.00000 | 76.000000       |

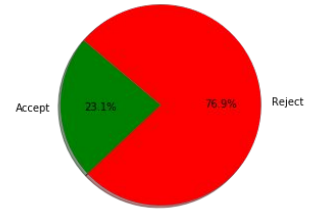
### 3. Inferences - Data size = (1056, 16)

- The scatter matrix revealed correlation between

- GRE total and GRE verbal
- GRE total and GRE quant
- GRE total, GRE verbal and TOEFL
  - This confirms that TOEFL tests verbal skills over quantitative
- Undergraduate major and target major:
  - This shows that applicants mostly stick to the same area

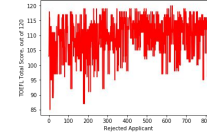
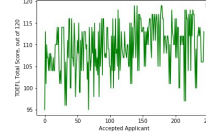
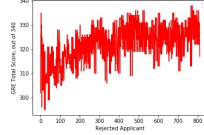
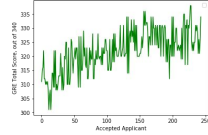


- Out of all applicants, across years, across majors, about 23% got accepted



- Accept vs Reject

- GRE, TOEFL:



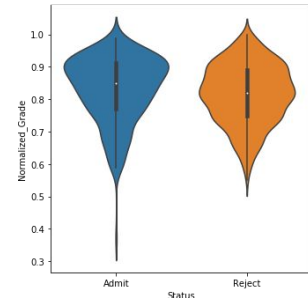
The mean, stdev were similar => GRE/TOEFL does not play a huge role in admission

- Grades:

There is a significant difference in Grades => Grades play a role in admission

- Publications/Work Experience:


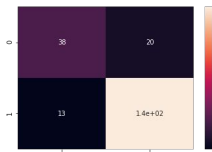

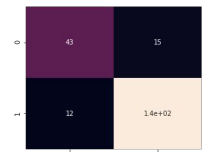

Similar work experience, publications slightly higher for admitted students



- Majority scored above 100 in TOEFL

- Classifiers

The following classifiers were used to predict admit/reject. Training - 80%, testing - 20%

| Classifier            | Logistic Reg  | Decision Tree   | Random forest  | Gradient Boosting   | SVM   |
|-----------------------|---|---|--|---|---|
| Accuracy              | 0.74  | 0.844   | 0.726  | 0.87  | 0.75  |
| Classification Report | <pre> precision    recall 0      0.71    0.09 1      0.74    0.99 </pre>          | <pre> precision    recall 0      0.75    0.66 1      0.88    0.92 </pre>          | <pre> precision    recall 0      0.00    0.00 1      0.73    1.00 </pre>           | <pre> precision    recall 0      0.78    0.74 1      0.90    0.92 </pre>            | <pre> precision    recall 0      1.00    0.10 1      0.75    1.00 </pre>            |
| Confusion Matrix      |  |  |  |  |  |
| Important Features    |   | Target major, Grade, College  | Undergrad major/Target major, Grade  | Grade, College, GRE/TOEFL   |   |

# Case Study 3

## University of California Los Angeles

## 1. Data Statistics

- 1583 applicants, 28 features
- Sparse matrix, lot of missing data
- Some features like Grade had to be handled explicitly because of differences

|                         |  |
|-------------------------|--|
| Status                  | Admit                                      |
| Program                 | MS   |
| Target_Major            | Computer Science                           |
| GRE_Q                   | 166  |
| GRE_V                   | 158  |
| Normalized_Grade        | 0.938                                      |
| GRE_Total               | 324  |
| GRE_AWA                 | 4  |
| TOEFL                   | 120  |
| College_Main_Form       | Chaitanya Bharathi Institute of Technology |
| Undergrad_Major         | Computer Science                           |
| Publications            | 0  |
| Work_Experience         | 2  |
| College_Main_Form_Cat   | 38   |
| Program_Cat             | 2  |
| Status_Cat              | 0  |
| Target_Major_Cat        | 9  |
| Undergraduate_Major_Cat | 8  |

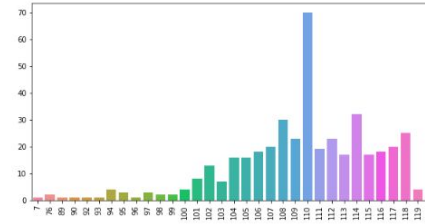
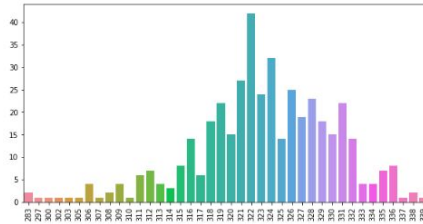
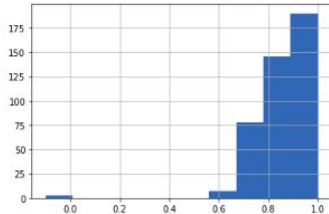
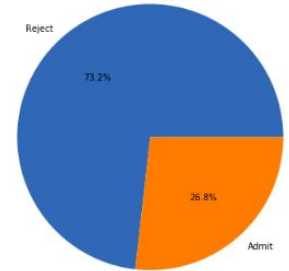
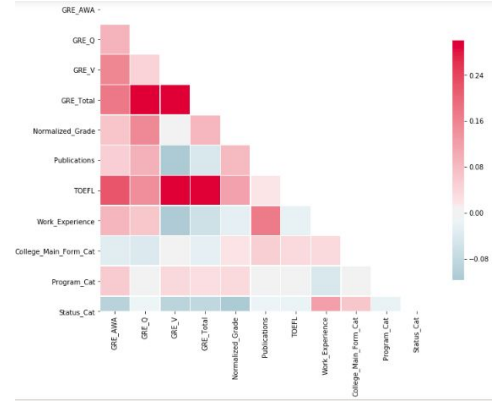
## 2. Data Processing

- Redundant features removed
  - Topper Grade, TOEFL R/S/L/W, Grade Scale
- Irrelevant features removed
  - Term, Year, University applied to
- Number of features retained = 15
- Feature list - 'College\_Main\_Form', 'GRE\_AWA', 'GRE\_Q', 'GRE\_Total', 'GRE\_V', 'Grade', 'Program', 'Publications', 'Status', 'TOEFL', 'Target\_Major', 'Term', 'Undergrad\_Major', 'Work\_Experience'
- Columns with non-numeric values converted to categorical
- Data was made uniform by replacing either with nan values or by removing the row

|       | GRE_Q       | GRE_V       | Normalized_Grade | GRE_Total   | GRE_AWA     | TOEFL       | Publications | Work_Experience | College_Main_Form_Cat | Progra      |
|-------|-------------|-------------|------------------|-------------|-------------|-------------|--------------|-----------------|-----------------------|-------------|
| count | 1583.000000 | 1583.000000 | 1583.000000      | 1583.000000 | 1583.000000 | 1583.000000 | 1583.000000  | 1583.000000     | 1583.000000           | 1583.000000 |
| mean  | 165.585597  | 156.685407  | 0.836864         | 322.271004  | 3.941883    | 109.145294  | 0.284270     | 8.758054        | 164.527479            | 2           |
| std   | 3.462473    | 5.393570    | 0.100465         | 6.540390    | 0.357910    | 6.339957    | 0.547421     | 12.926524       | 85.709667             | C           |
| min   | 146.000000  | 137.000000  | -0.100000        | 283.000000  | 2.500000    | 7.000000    | 0.000000     | 0.000000        | 0.000000              | C           |

### 3. Inferences - Data size = (1583, 15)

- The correlation plot revealed
  - Grade and GRE scores do not correlate.
  - Publications and GRE scores also do not correlate
  - GRE total, GRE verbal and TOEFL
    - High Correlation between GRE Total, Verbal and TOEFL
  - Work Experience and Exams:
    - This shows that work experience hampers the performance in exams like GRE and TOEFL
- Out of all applicants, across years, across majors, about 27% got accepted
- Accept vs Reject
  - Grade plays a very important role in deciding the admission
  - TOEFL and GRE scores aren't super important, Average works



- **Classifiers**

The following classifiers were used to predict admit/reject. Training - 80%, testing - 20% along with 5 fold cross validation.

| Classifier         | Logistic Reg | Decision Tree            | Random forest                        | Gradient Boosting        | SVM   |
|--------------------|--------------|--------------------------|--------------------------------------|--------------------------|-------|
| Accuracy           | 72.87        | 76.34                    | 76.97                                | 75.07                    | 74.13 |
| F1_Score           | 0.63         | 0.75                     | 0.71                                 | 0.74                     | 0.66  |
| RoC - Score        | 0.51         | 0.66                     | 0.58                                 | 0.65                     | 0.53  |
| Important Features |              | Target major, Grade, GRE | Undergrad major, Target major, Grade | Grade, Target, GRE/TOEFL |       |



# Comparative Analysis

- Importance of GRE and TOEFL
  - TOEFL isn't a deciding factor but just a bar.
  - All of the colleges have a specific GRE threshold where the freq is max
- Difficulty levels
  - CMU is relatively harder to get compared to other universities
  - UCLA does not pay attention to Work Ex and Publication comparatively
- Model Performance
  - F1 & ROC score needs to be analysed rather than accuracy
- Important features
  - Grade and GRE score were the most important across all colleges
  - Work Ex matters in most of the colleges but not all

# Next Steps

- Can an admit ever be guaranteed?
  - Study this question on a wider scale by using cumulative data
- Analyse a regression based task
  - Predict chance of admission rather than a binary decision
- Generate model profile for each institute
  - Ideal GRE, TOEFL scores, number of publications, GPA etc

# Group Members

Group Number - 4

- Vishaal Udandaraao (2016119)
- Suryatej Reddy (2016102)
- Surabhi S Nath (2016271)
- Suril Mehta (2015104)



Thank You