Report

November 30, 2019

0.1 Problem Statement

Every year approximately 10,000 students apply for graduate schools abroad (United States, Canada, United Kingdom etc.) from different Indian institutes. These articles clearly indicate an exponential rise in the number of Indian students abroad enrolled in undergraduate, graduate and doctoral programs. With such a monumental student population, one would expect an interpretable and coherent system to analyse graduate student intake, measures correlations between the different factors related to admissions, corroborate to the noise and entropy in the admission process, and predict the chances of admission into a given university. However, we could not find such a holistic system that addressed these issues.

We therefore wanted to try and disentangle the process of graduate school admissions by analysing the various factors at stake, and apply our prior inductive biases about the randomness in the process to build predictive models. Our problem statement involves exploring and comprehending the following aspects of graduate school admissions: - Studying patterns in student profiles applying to different graduate schools across geographic locations. - Comparing, differentiating and gaining insights about the admission process. - Inferring and leveraging definitive existing correlations. - Drawing conclusions about admit/reject ratios, strength of an applicant profile, probability of admission, appropriate universities for a specific profile

0.2 Data Collection

0.2.1 Source

We chose https://admits.fyi as a source to collect data from. The data (check below for description) is curated by a group of ex-grad students over the past few years. The data is checked manually to avoid any discrepancies. Key characteristics of the website are - The website has rate limting and cannot be scraped directly because of the dynamic nature of the website. - The website however allows us to set filters to search data that is relevant to us. - The results are paginated so all data cannot be collected at once

0.2.2 Collection

To overcome all this we did the following for each university: - Open up the website and set a filter to get data only for that particular university. - Deploy Javascript that would click on the next page button - Next set of results would be loaded from their server only upon this click. - We simulataneously start a network request capture mitmproxy that would allow us to snoop the

requests made by the website on the server. - We couldn't replay the same request again because of authentication that was implemented. So instead we captured all the responses sent by the server by saving the entire capture. - Later we extracted the complete data of a university by analyzing the network capture and creating jsons. - The jsons would then allow us to process data later.

Another benefit of using this type of data collection was that we would even get fields not shown on the website.

0.2.3 Challenges

The website is made by ex-grad students from India. So it is robust and they tried to prevent automation to a large extent. - We couldn't scrape all data directly because of the dynamic content. - They implemented rate limiting from a given IP. - The requests in the network capture couldn't be replayed because of server-client authentication implemented.

The above solution we proposed is a semi automated method that overcomes all these challenges and gives us clean data to work with.

0.2.4 Data Description

The data that we have is of the following format

Field	Value
University	Name of the University
Status	Boolean (Accept/Reject)q
Target Major	Major Applied For
Term	Term Applied To (Fall'19/Spring'18/etc)
GRE_Q	Quant Score in GRE (130 - 170)
GRE_V	Verbal Score in GRE (130 - 170)
GRE_AWA	Analytical Writing Score in GRE
GRE_Total	Total Score in GRE (Q + V)
TOEFL	TOEFL Score (0 - 120)
TOEFL_Reading	TOEFL Score in Reading (0 - 30)
TOEFL_Listening	TOEFL Score in Listening (0 - 30)
TOEFL_Speaking	TOEFL Score in Speaking (0 - 30)
TOEFL_Writing	TOEFL Score in Writing (0 - 30)
IELTS	IELTS Score (0 - 9)
College_Main_Form	Name of Undergraduate College
Undergrad_Major	Undergraduate Major
Grade	Undergraduate Grade (0-10/0-100/0-4)
Topper_Grade	Grade of Batch Topper
Grade_Scale	Scale of Grade (4/10/100)
Publications	Number of Research Papers Published
Work_Experience	Number of Months of Work Experience

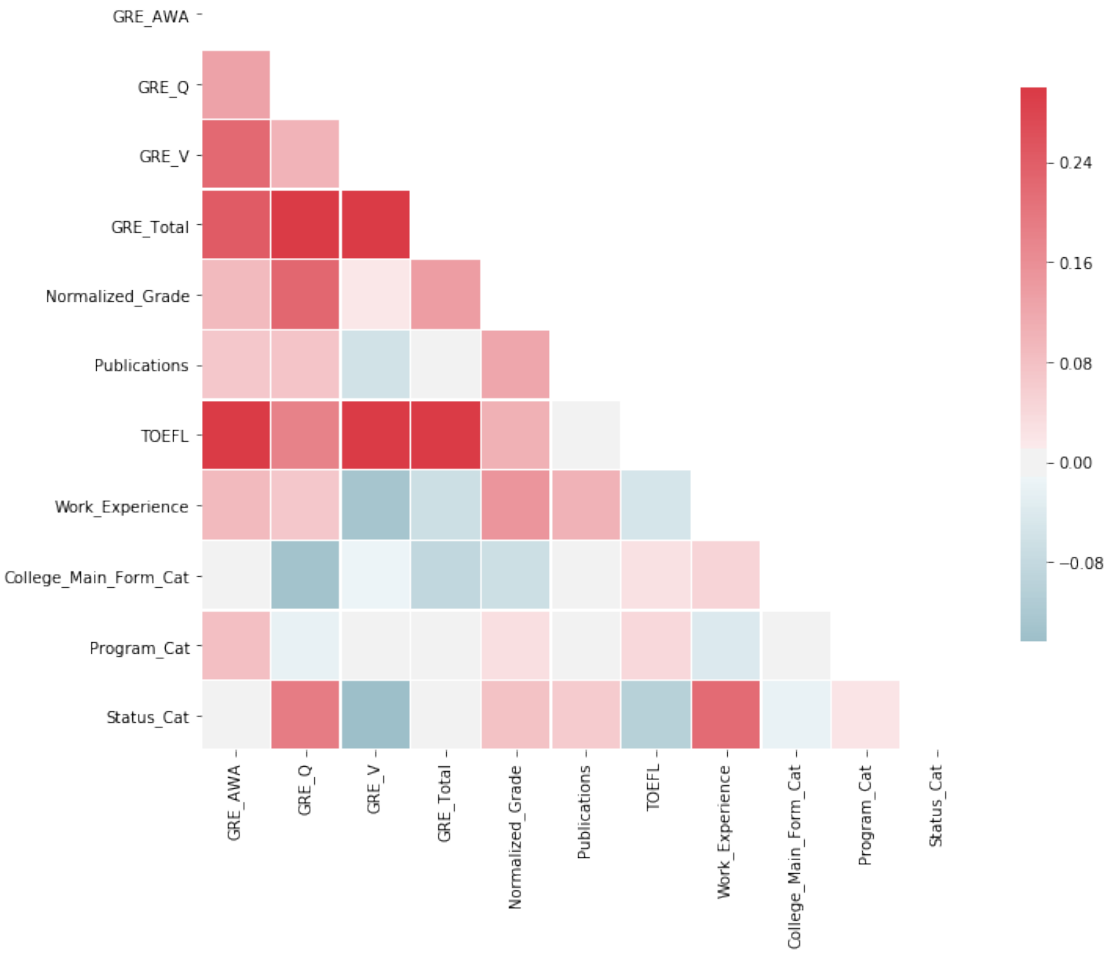
We collected data for 7 universities with each one having 3000 data points on average - Carnegie

Mellon University - University of Illinois Urbana-Champaign - University of California, Los Angeles - Columbia University - Georgia Institute of Technology - University of Maryland, College Park - University of Michigan, Ann Arbor

A sample data point would like this `json { "University": "Carnegie Mellon University", "Status": "Admit", "Program": "MS", "Target_Major": "Software Engineering", "Specialization": "-1", "Term": "Fall", "Year": "2019", "GRE_Q": 164, "GRE_V": 151, "GRE_Total": 315, "GRE_AWA": -1, "TOEFL": 111, "IELTS": -1, "College_Main_Form": "B.M.S. College of Engineering, Bangalore", "Undergrad_Major": "Computer Science", "Grade": 8.8, "Topper_Grade": -1, "Grade_Scale": 10, "Publications": 1, "Work_Experience": 18, "TOEFL_Reading": -1, "TOEFL_Listening": -1, "TOEFL_Speaking": -1, "TOEFL_Writing": -1 }`

0.2.5 Preprocessing

1. Removed features which either had more than 60% missing data and or were redundant based on correleation matrix. The features removed were
   - Topper Grade (Missing Data)
   - TOEFL R/L/S/W (Redundant)
   - Grade Scale (Irrelevant)
   - University Applied To (Same across all fields)



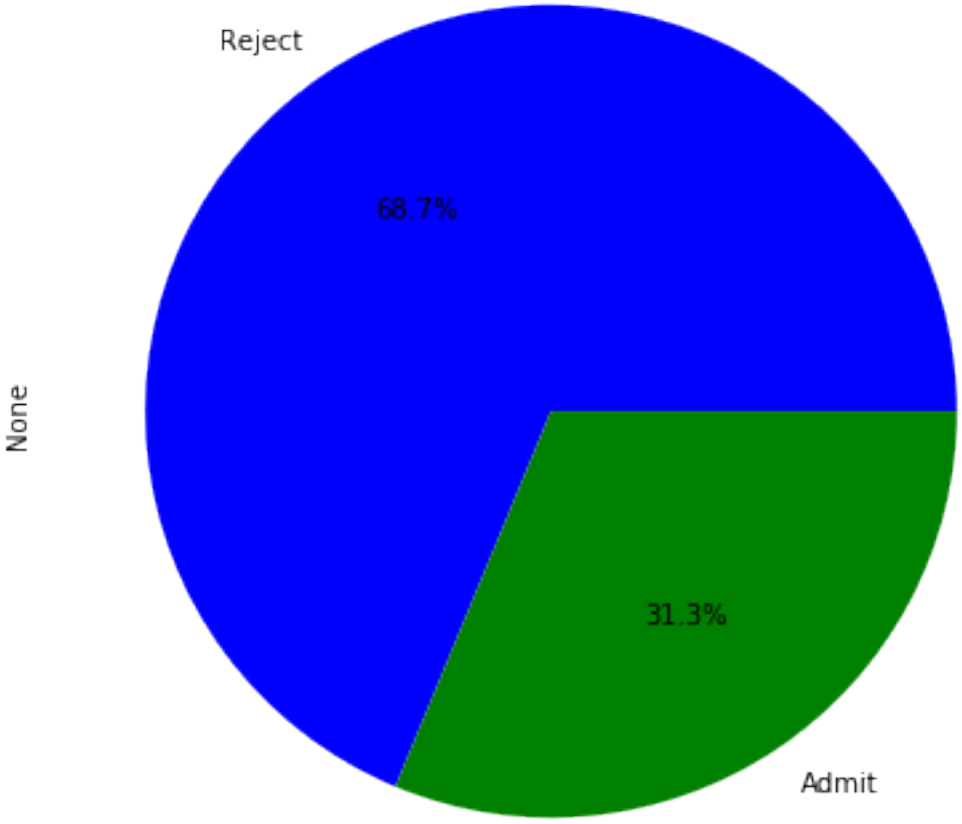
2. Feature Coalescing
   - We normalized the grades across different scales. Grade / Grade Scale
3. Columns with non numeric data were cateogorized using label encoding.

0.3 Data Insights

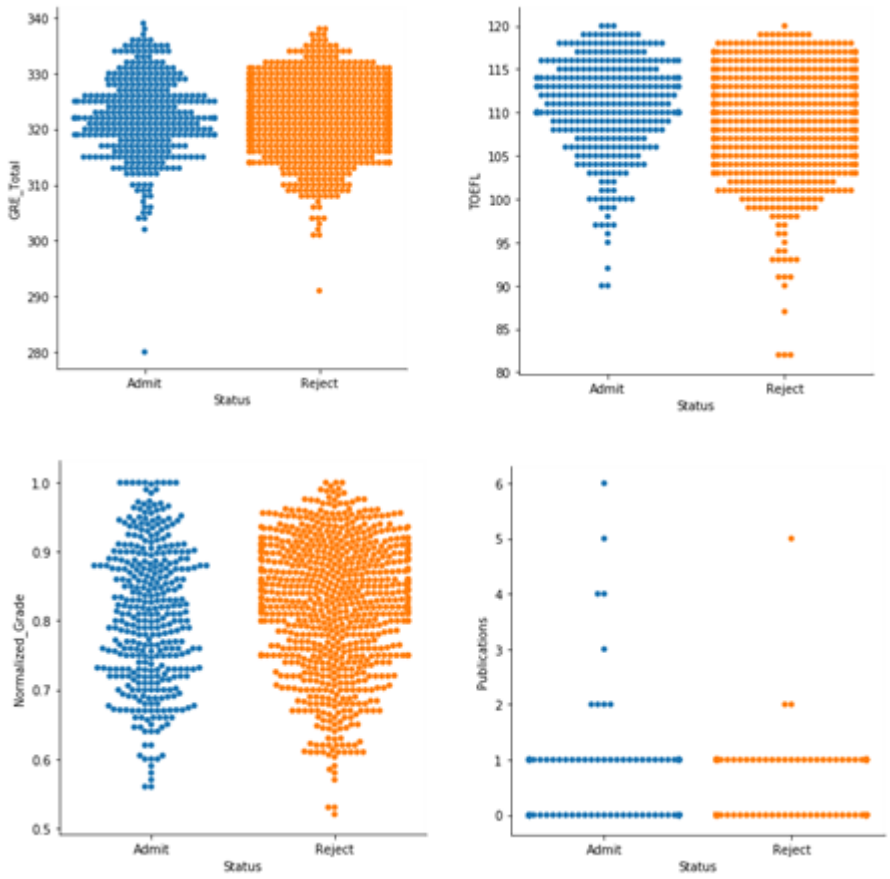
After pre-processing the data, CMU, UIUC and UCLA data were studied and the following details were derived through various plots.

0.3.1 CMU

- Data size after processing = (1174, 18)
- Ratio of students admitted = 32%

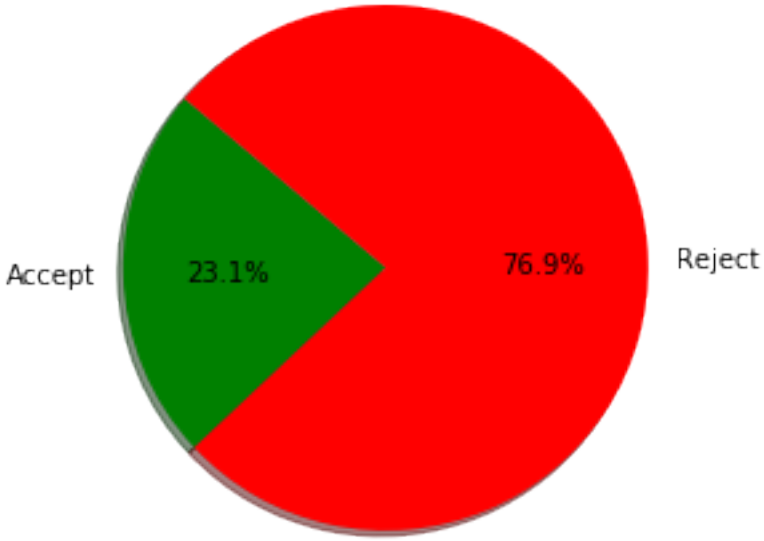


- Accept vs reject analysis showed that GRE, Normalized Grades, Publications, Work Ex most important for securing admission

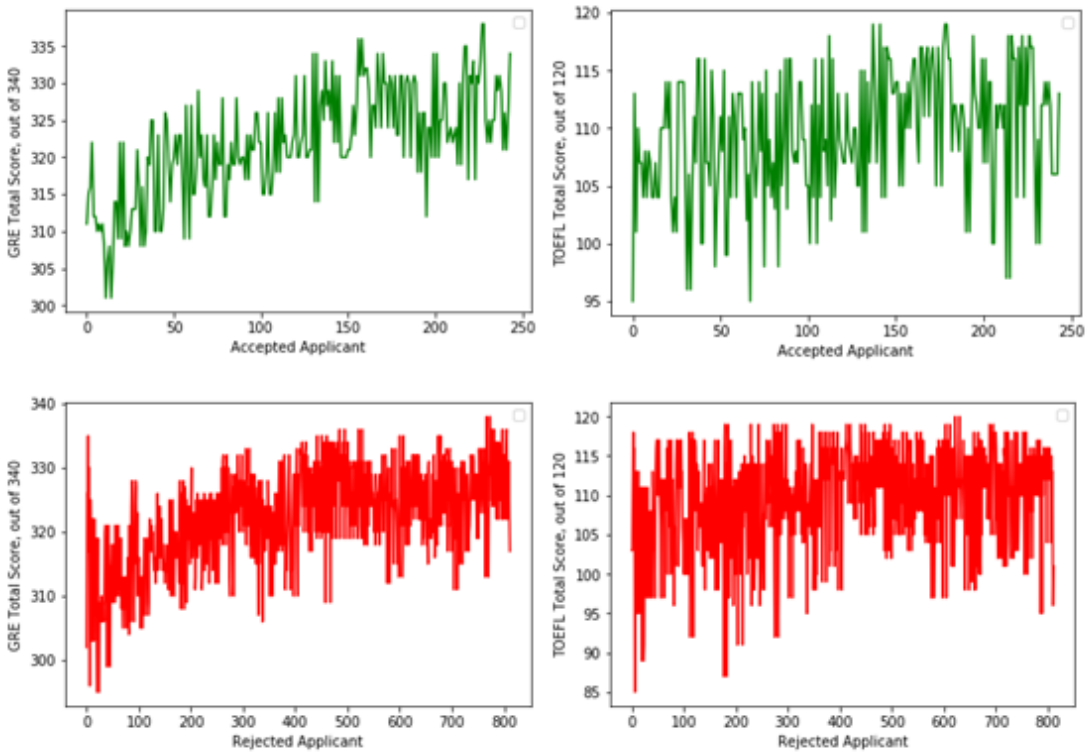


0.3.2 UIUC

- Data size after processing = (1056, 16)
- Ratio of students admitted = 23%



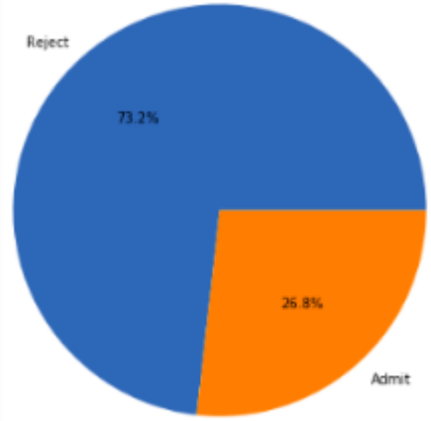
- Correlation matrix revealed correlations between GRE total & GRE verbal, GRE total & GRE quant and GRE total & GRE verbal & TOEFL. This reveals that TOEFL tests verbal skills over quantitative. In addition, the correlation between undergraduate major and target major showed that applicants mostly stick to the same area while applying
- The accept vs reject analysis showed that GRE/TOEFL, with similar mean and stdev, does not play a huge role in admission



- Grades, on the other hand show a significant difference in accepted and rejected groups which means grades play a role in admission
- Although work experience is similar in both groups, publications are slightly higher for admitted students

0.3.3 UCLA

- Data size after processing = (1583, 15)
- Ratio of students admitted = 27%



- Correlation matrix revealed that grade and GRE scores, publications and GRE scores do not correlate, whereas GRE Total, Verbal and TOEFL had high correlation and infact work experience hampers the performance in GRE and TOEFL
- Accept vs Reject performance showed that grade plays a very important role in deciding the admission whereas, TOEFL and GRE scores are not very important and an average score works

0.4 Predictive Modelling

We performed our predictive modelling experiments on three major case studies: - Carnegie Mellon University - University of Illinois, Urbana Champaign - University of California, Los Angeles

0.4.1 Classification

As a first step, we consider the prediction problem as a deterministic binary classification problem. Our data cleaning and pre-processing steps have been explained above. For the binary classification task, we considered simple models so that we can interpret the results well. We also ensured that we could retrieve the importance of every factor so that a comprehensive analysis could be performed.

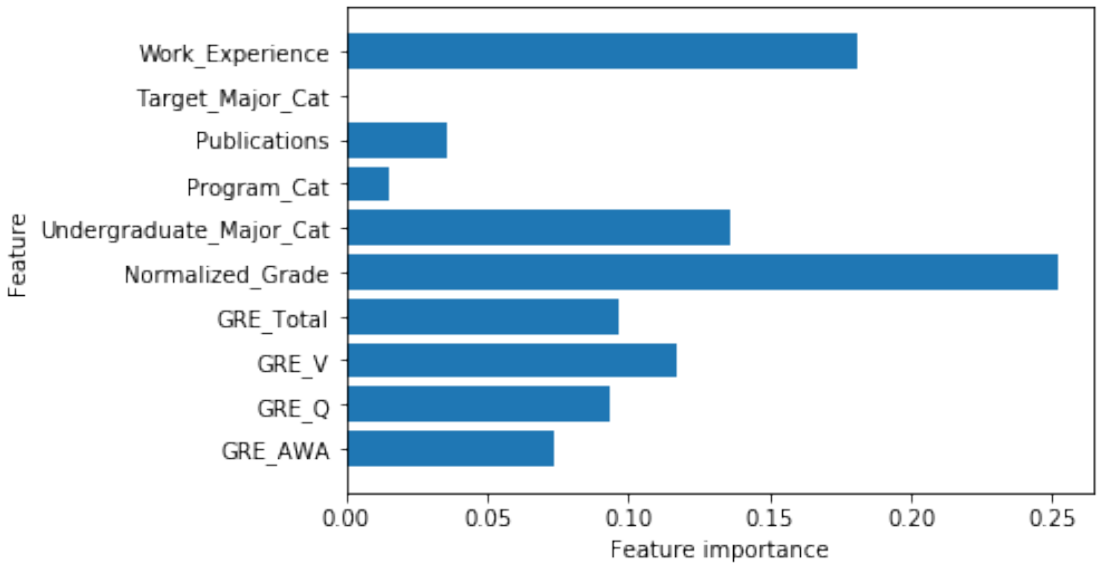
We use the following classifiers: - Logistic Regression - Decision Tree - Random Forest - Gradient Boosting - Support Vector Machine (SVM)

A concise explanations of our results follows:

- Carnegie Mellon University

Classifier	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	SVM
Accuracy	0.72	0.65	0.7	0.67	0.75
Precision (0)	0.66	0.38	0	0.33	0.78
Recall (0)	0.27	0.29	0	0.13	0.3
Precision (1)	0.73	0.73	0.71	0.71	0.75
Recall (1)	0.93	0.81	1	0.89	0.96

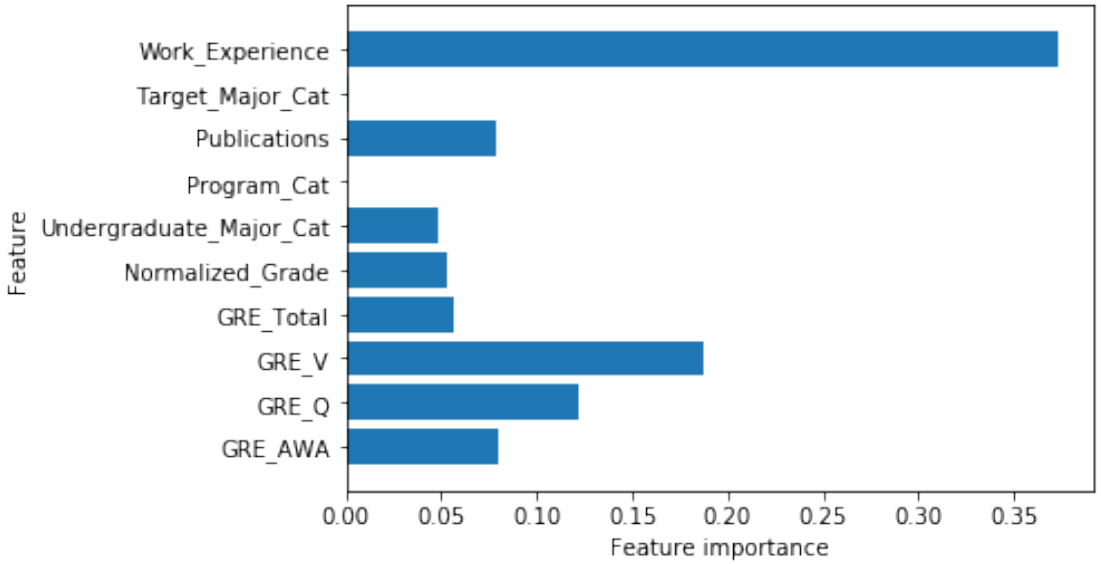
The average importance of features predicted by our classifier are depicted by the figure below.



We can infer that the GPA factor takes utmost precedence when it comes to the admit decision. Two other important factors include the undergraduate major and GRE work experience. A slightly surprising that we notice is the low importance given to the numner of publications. A common notion is that publications play a vital role in getting admitted to any university, but our analysis shows otherwise.

- University of Illinois, Urbana Champaign

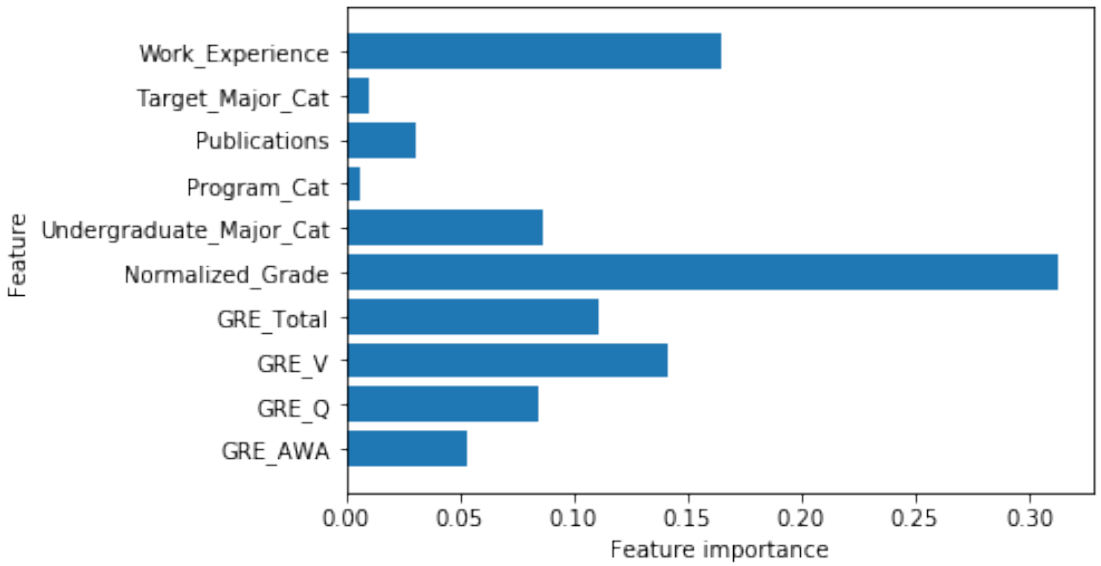
Classifier	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	SVM
Accuracy	0.74	0.844	0.726	0.87	0.75
Precision (0)	0.71	0.75	0	0.78	1
Recall (0)	0.09	0.66	0	0.74	0.1
Precision (1)	0.74	0.88	0.73	0.9	0.75
Recall (1)	0.99	0.92	1	0.92	1



We immediately see the huge spike in the importance of the work experience factor. This could be slightly counter-intuitive to our understanding of the admissions process, but we believe that these sorts of biases differ from college to college, and hence a uniform criterion cannot be applied. We also see that publications are an important factor for an admit to UIUC, clearly orthogonal to our results for CMU.

- University of California, Los Angeles

Classifier	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	SVM
Accuracy	0.73	0.76	0.77	0.75	0.74
ROC Score	0.51	0.66	0.58	0.65	0.53
F1 Score	0.63	0.75	0.71	0.74	0.66



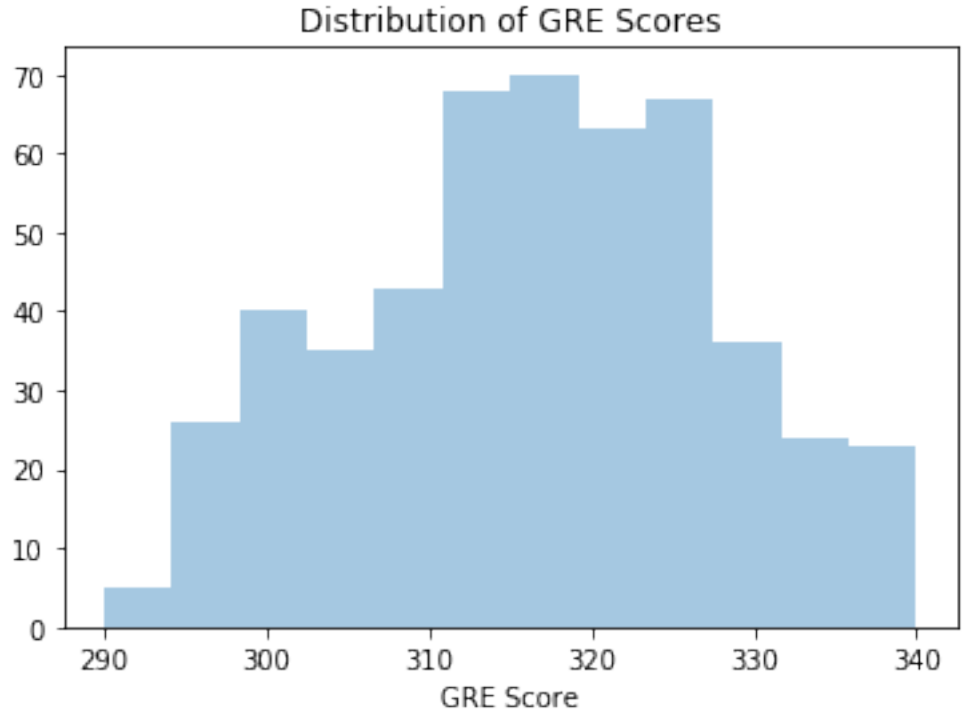
Here too, GPA takes the driver's seat when it comes to importance of factors. It outweighs every factor by almost a double margin, and this result might be slightly concerning to students with a low GPA but a strong profile otherwise. We also see that work experience is the second most important factor for getting an admit. Similar to the case of CMU, publications are not as important a factor as they are made out to be.

On the basis of the above classification experiments, we can state with some confidence that the different factors weigh distinctly for the admission criteria for different universities. Hence, students looking to target particular universities, cannot blindly follow a uniform 'gold standard' trajectory for getting an admit. Each university needs to be researched independently and a student must portray his/her profile to meet the specific requirements of each university.

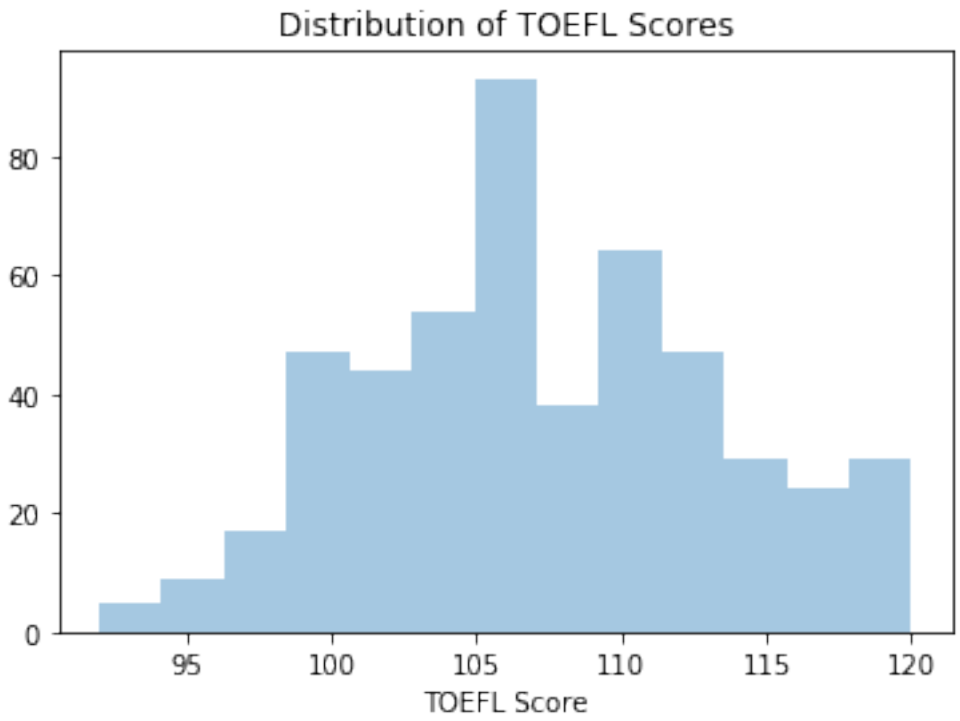
0.4.2 Regression

As a follow-up experiment, we wanted to explore the exact chance of admit as a function of the student profile. For this, we use the public UCLA admission dataset to perform a regression analysis.

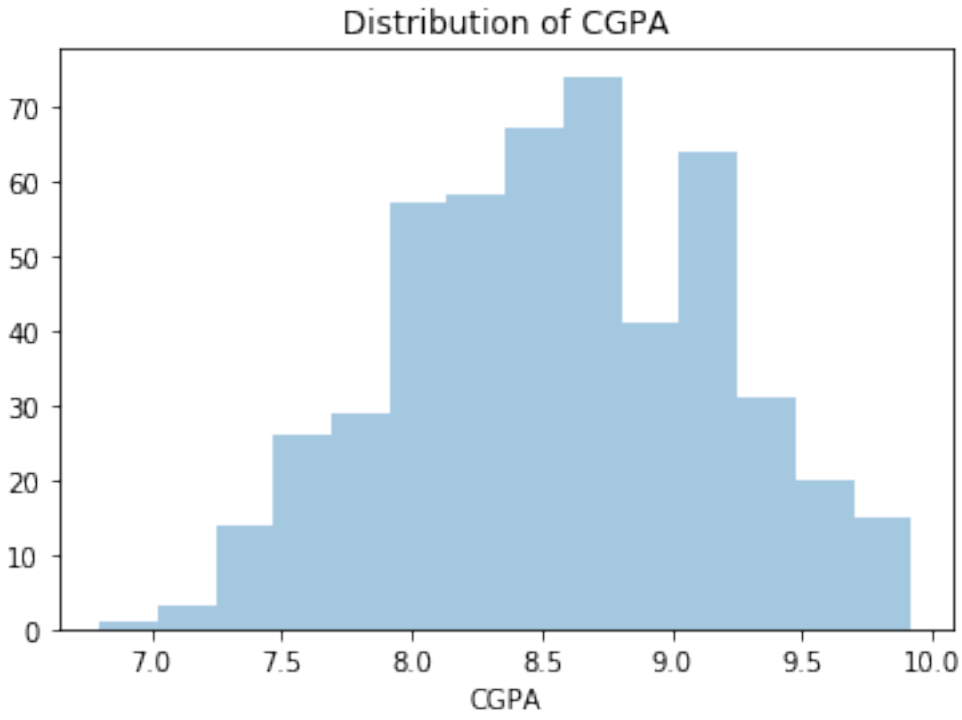
The data contains the following fields: - GRE scores: The GRE scores range from 260 to 340. The GRE scores were distributed as:



- TOEFL scores: The TOEFL scores have a range of 0 to 120, however our dataset has aleast TOEFL score of 93. The TOEFL scores were distributed as:



- CGPA: The CGPA distribution of the candidates in the dataset is:



- University rating: This field gives the rating of the undergraduate university of the student on a relative scale of 1 to 5.
- SoP and LoR: These two fields give relative indicators about the strength of the candidate's

statement of purpose and recommendations.

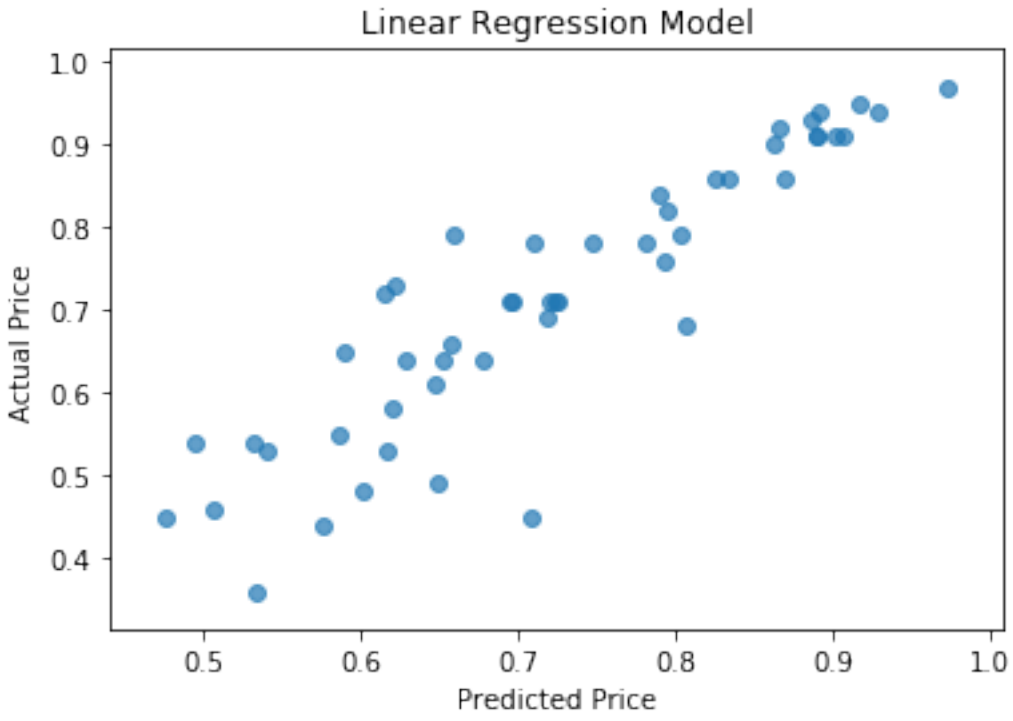
- Research: This is a binary attribute about the candidate's research experience. If the candidate has done research before it take a value of 1, else 0.

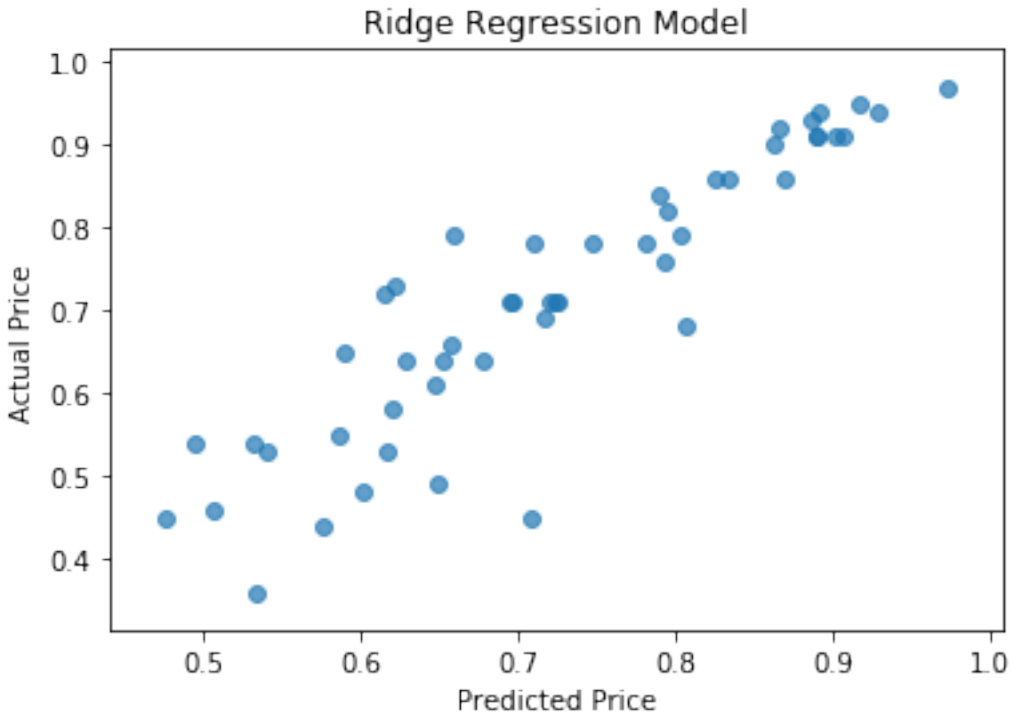
We use the following regressors: - Linear Regression - Ridge Regression - Lasso Regression - Bayesian Ridge Regression - AdaBoost Regression - Gradient Boosting Regression

The results we obtain are:

Regressor	Linear	Ridge	Lasso	Bayesian Ridge	Ada Boost	Gradient Boost
R2 score	0.80	0.8	0.21	0.8	0.77	0.78
RMSE	0.005	0.005	0.02	0.005	0.006	0.006
MAE	0.05	0.05	0.12	0.05	0.06	0.05

We analyse the correlations between the predicted and actual dependent variables to check the goodness of fit. The correlation plots obtained were:





Therefore, we purport that such a regression based predictive model can be used to extend the previously reported classification technique. We believe that such a combined representation of predictive models could be a viable option for mitigating the randomness in the grad school admissions process.

0.5 Comparative Analysis

Comparing the data and analysis from across different colleges, we can infer the following comparisons, 1. Importance of GRE and TOEFL * TOEFL is not a deciding factor and is just a bar/baseline/threshold * Most colleges have a specific GRE threshold where the freq of applicants is max 2. Difficulty levels * CMU is relatively harder to get admitted to as compared to other universities * UCLA pays lesser attention to Work Experience and Publication compared to other universities 3. Model Performance * F1 & ROC score needs to be analysed rather than accuracy to incorporate for data imbalance 4. Important features * Grade is one of the most important factor across all colleges * Work Experience matters in most of the colleges but not all

0.6 Conclusions

We perform detailed analysis of grad school admit results for Indian Students. We started off by collecting high quality data from https://admits.fyi. After preprocessing the data, we selected 3 of the most competitive schools and analyzed how various factors such as Undergraduate GPA, GRE, Publications affect an admission result. We found that undergraduate GPA is one of the most important factor and seems to be consistent across all the three universities we selected. GRE score though important doesn't seem to be a deciding factor for admission. Students with similar GRE scores have got different results when applying to the same program. TOEFL score does not

affect your admission result in these top places. This result is consistent with the fact that it is only used as a language requirement for Indian Students. We also find that UCLA plays lesser attention to Work Experience and Publications. However, one shortcoming of our analysis is that we cannot take into account subjective factors like the strength of a student's LoR's or how well written an SoP is. These also play a crucial role in admission but cannot be modelled directly.