Analysis & Prediction of American Graduate Admissions Process

Bhavya Ghai

Department of Computer Science, Stony Brook University
Stony Brook, NY 11794
bghai@cs.stonybrook.edu

Abstract: This project tries to understand American Graduate Admissions process by specifically analyzing MS Computer Science application over past 5 years. Edulix has been used as data source. After extensive data cleaning, I have tried to model admissions data based on patterns extracted from data and domain knowledge. The key to analyzing Graduate Admissions data is to analyze data in buckets rather than considering all in one bucket. The project aims to help students choose the right Universities by predicting whether a student will be admitted to a specific University. Similarly, this model can be used by Graduate Admission Committee to filter very low scoring or very high scoring applicants.

1 Introduction:

Graduate Admissions can be thought as a mapping problem between Students and Universities where each end strives for the best they can get. In Data Science World, this problem can be modeled as a University Recommendation problem. As per UNESCO report, US is the top destination for international students followed by UK, France and Australia [1]. The number of international students (both undergraduate and graduate) enrolling in the US crossed 1M for the first time in 2015-16 as per an IIE report backed by the State department [2]. Most of the Universities have non-refundable application fees ranging from $50 - $125 and even more for business schools. With each student applying to multiple Universities, the total application cost may lie in 100's of million USD per year. On the other hand, Admission committees have to spend a lot time evaluating tons of applications. Our objective is to help either side by providing accurate recommendations to students based on their profile. This might help students cut costs on Admission counselling and applying to a smaller set of Universities. On the other side, lesser number of application will save a lot of time to evaluate them.

2 Literature Review:

There has been various studies dealing with admission process. But there are very few papers which deal with assisting the decision making process using Machine Learning approach. In recent past, Data Science has been used to partially automate the Graduate Admission process [6]. Data Science models are being used to eliminate extremely good and bad application from the pool. This technique saves upto 74% of time for Graduate Admission Committee members. A recent study [5] published this year reveals some key factors in the decision process and, consequently, allows to propose a recommendation algorithm that provides applicants the ability to make an informed decision regarding where to apply, as well as guides the decision-makers towards a more efficient process. Some researchers have briefly reflected on the process of decision making from university's perspective, but only qualitatively [Raghunathan, 2010], [Posselt, 2016]. Raghunathan talks about his experience as part of the admission committee in Computer Science department during his graduate studies at Stanford. He documents several factors considered during the decision-making and provides his opinion regarding their

importance towards an Admit. Posselt's study provides more coverage in terms of number of universities, and the number of departments per university. Posselt focuses only on doctorate (PhD) admissions while Raghunathan had talked about only Master of Science (MS) applications.

3 Data:

Graduate Admissions process is generally carried out online via Universities own interface or third party websites like ApplyYourself. Generally, Graduate Admissions application contains lot of private information and so it is kept confidential by the respective Universities. For this project, I have used self-reported data by the students on online platforms like Edulix. This entire study is based on the assumption that majority of this data is accurate. Additionally, I have used lot of measures to eradicate noise and inconsistency in data. This data has been used in other research work too [5]. There are some other sources like theGradCafe, Yocket.in, etc. which can also be used. I used Edulix as it's been used in previous research projects and we can extract significant amount of information conveniently.

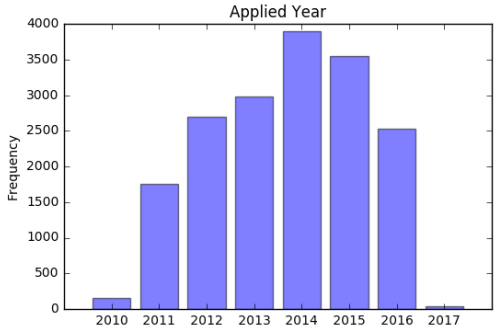


Fig. 1 Student Distribution based on Applied Year

Edulix dataset used for this project has 23,111 rows and 28 columns. Each row in the dataset represented a student profile. Each column speaks something about student profile like CGPA, GRE score, etc. As shown in Fig. 1, the dataset is fairly recent. It comprises of student profile mostly from 2011-16. Let's have a look at all the features in the dataset.

Table 1 Features in Dataset

Features in Dataset			
Id	AppliedYear	Colleges	Conferences
Department	EdulixNickName	EdulixId	Grade
GradeScale	GRE	GreTotal	IndustrialExpDescriptor
IndustrialExpMonths	InternExpDescriptor	Journals	Major
Misc	PreviousCollege	Program	RealName
ResearchExpDescriptor	ResearchExpMonths	Specialization	TermAndYear
Toefl	TopperGrade	University	OriginalPreviousCollege

Table 1 represents all the feature names present in original data. Most of the names in the feature-set are self-explanatory. Like, GRE means Graduate Record Examination Score, Journals represents number of

accepted research papers in Journals, Program represent MS/PhD, etc. Some of these features have nested features. For Example- GRE is a complex feature consisting of GRE Quant, GRE Verbal and GRE AWA score. Similarly, TOEFL is also a complex feature consisting of TOEFL score and essay score. In the next section, we'll see how to convert raw data to processed form.

4. Exploratory Data Analysis

A big majority of the time spent on this project went to data gathering & data cleaning. It's quite easy to play with clean consistent dataset but pretty hard to build one. The data gathered from Edulix is very raw and can't be used for any significant analysis directly. I had to deal with following key issues :-

1) Missing Data
2) Inconsistent Data
3) Noise
4) Unstructured Data
5) Redundancy

After loading dataset, I used various visualization and summarization techniques to understand each feature in the dataset. Initially, all the features were in object or string form which can't be used for analysis. There were some redundant columns in the data as well. My first priority was to get rid of redundant columns and transform all features to numeric type.

Table 2 Redundant Features in Dataset

Feature	Redundant Feature
edulixId	id
researchExpDescriptor	researchExpMonths
industrialExpDescriptor	industrialExpMonths
Termandyear	appliedYear

Table 2 represents the features along with their redundant counterparts. Deleting redundant features helped to reduce dimensions of data. It enables better and faster of analysis of data. There were some unimportant features like name, edulixName which were also deleted. Furthermore, there were also features like misc and comments which might contain useful information but extracting that information might require advanced NLP techniques which is beyond the scope of this project. Next, I used regex expressions to convert different features to numeric type like edulixId, etc.
As mentioned in previous section, There were many complex features in the dataset i.e. features which represent a group of features.

Table 3 Unstructured Features

Complex Feature	Individual Features
TermAndYear	Semester, Applied Year
GRE	Quant, Verbal, AWA
TOEFL	TOEFL Score, Essay Score
Colleges	University Name, Status(Admit/Reject), Comments, Specialization

It was particularly tricky to deal with Colleges attribute. Each student can apply to multiple colleges. Unfolding this feature caused the dataset size to increase.

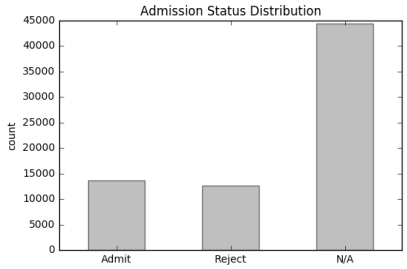


Fig. 2 Admission Status Distribution

It increased from approx. 11k to 77k. Many of the rows in newly generated data didn't have an output label Admit/Reject. So, I had to discard a lot of rows. Finally, I was left with 26k rows.

4.1 GRE

GRE stands for Graduate Record Examination. It's a prerequisite for most MS and PhD programs in US. GRE consists of 3 sections namely Quant, Verbal and AWA. GRE feature in the dataset came as a compound feature consisting of Quant, Verbal and AWA. First we split, GRE into its constituent features. Each of the 3 sections have a pre-defined range. For example- AWA score lie in between 0-6 with 0.5 point increments. Similarly in the new version of GRE, Quant and Verbal score lie in between 130-170 with 1 point increment.

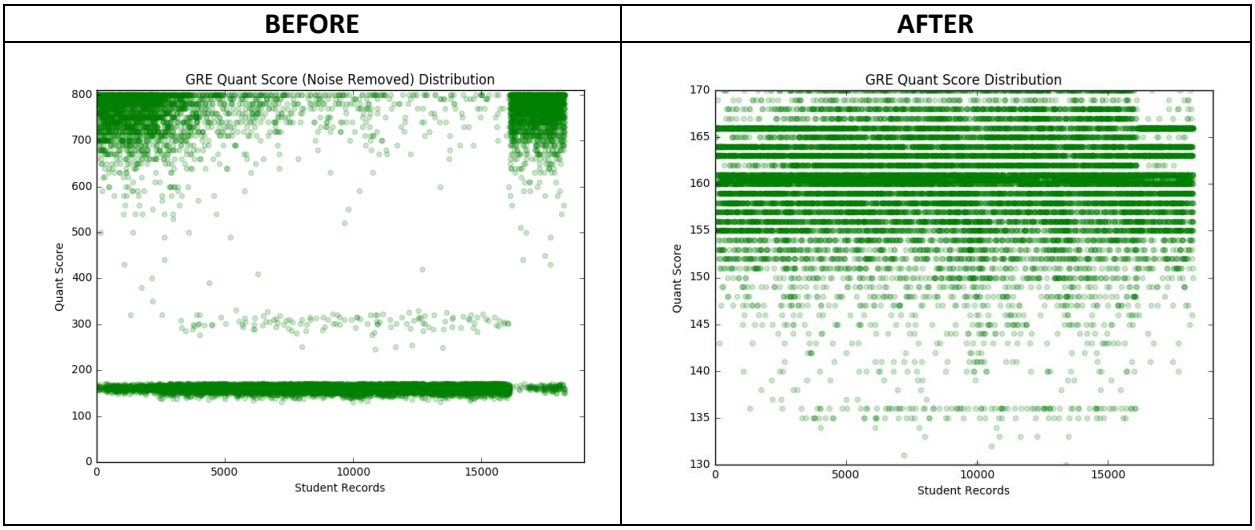


Fig. 3 GRE Quant Score before & After Processing

Fig. 2 represents GRE quant score before and after data processing. Plot on the left shows GRE Quant score representation after removing outliers and noise. The plot on right represents GRE Quant score after removing inconsistent values and scaling old and new formats. The old GRE format had Quant scores between 200-800 and new format has Quant scores between 130-170.

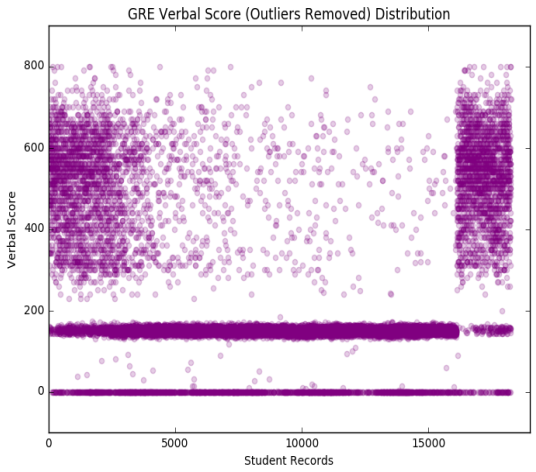
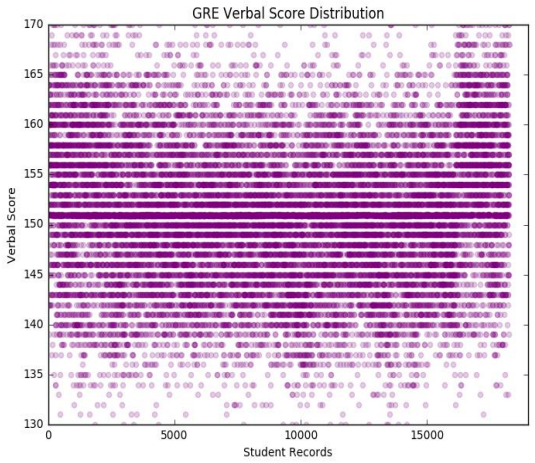
BEFORE	AFTER
	

Fig. 4 GRE Verbal Score before & After Processing

To merge score from old and new gre format, I used official translation of old scores to new ones by ETS [4]. In Fig. 2 & Fig. 3, the left plot's y-axis goes all upto 800. Left plot represents raw data which has a mix of old and new gre score. The graph on the right represents cleaned and scaled version of GRE Verbal and Quant score.

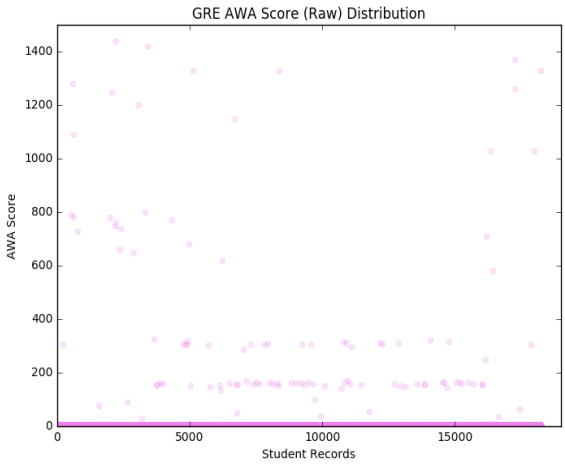
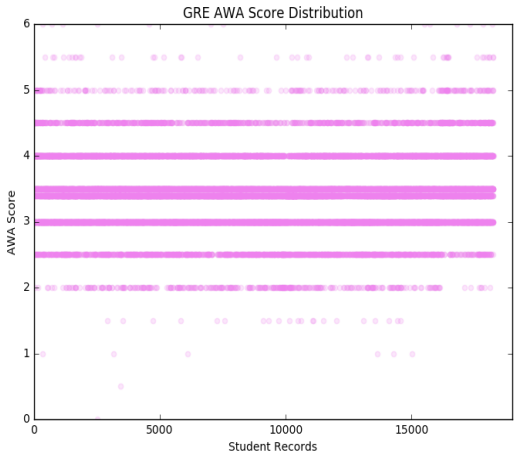
BEFORE	AFTER
	

Fig. 5 GRE AWA Score before & After Processing

Fig 4 represents left plot represents AWA score with outliers. Due to outliers, we are unable to any pattern in AWA score. The plot on the right represents cleaned AWA with outliers and inconsistent values removed.

There is one interesting observation to made from Fig 2, 3 & 4. Even after cleaning the data, there is a thick dark line in each of the right plots of these figures. They actually represent the missing values in th data which have been set to mean value.

4.2 TOEFL

There are 3 versions of TOEFL namely, iBT, cBT and pBT. In this dataset, TOEFL score belonged to iBT and cBT. Since our data is ranging from 2011-16, so we had mix of iBT and cBT scores. Many

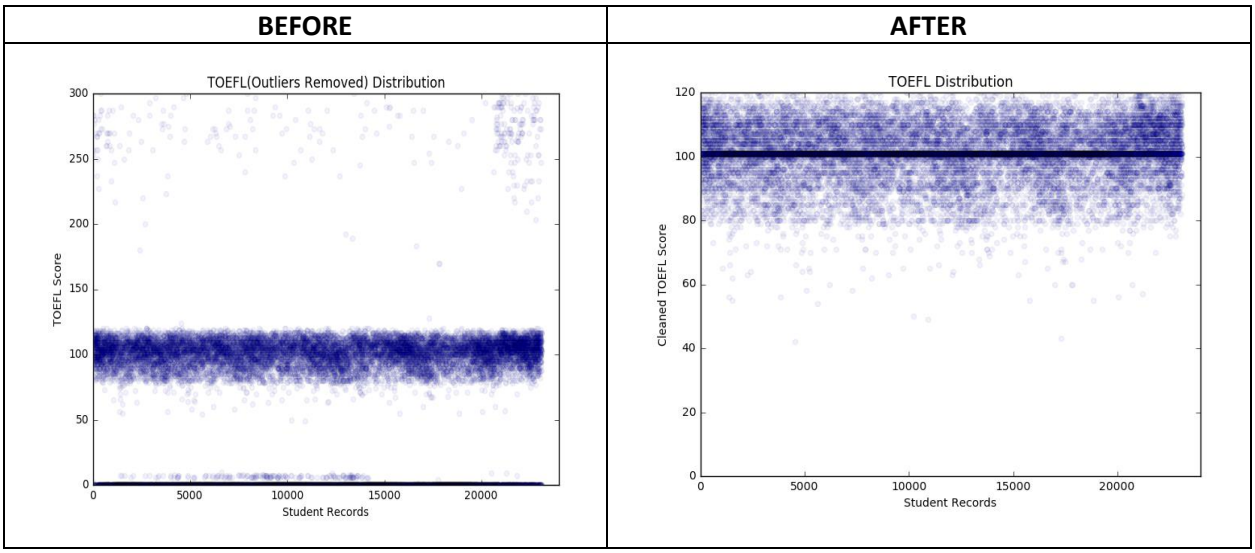


Fig. 6 TOEFL score before & after Processing

To separate them was a bit tricky because their ranges overlap a bit. iBT scores can vary from 0-120 and cBT scores can vary from 0-300. I classified the two based on the assumption that scores no greater than 120 will lie in the domain of iBT. After classification, we need to map the scores to common domain. For this, we transformed all cBT scores to iBT by following this link [3]. Finally, we got rid of missing values by assigning them mean TOEFL score(99.90).

4.3 Grade

Grade represents the academic performance of a student in his previous University. Different University have different grading criteria and different Scale. A particular University might be more stringent or lenient in awarding grades than other University. The most intuitive way is to calculate gpa as the ratio of grade and topper's grade. Given the dataset, we can easily calculate this ratio as topper grade is also available to us.

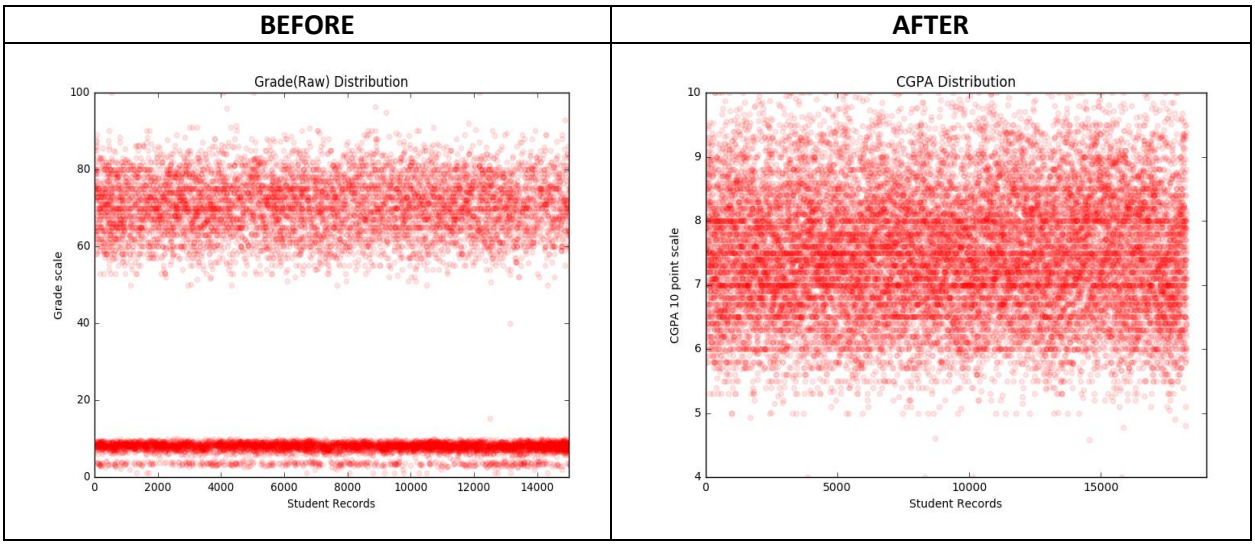


Fig. 7 CGPA before & After Processing

Initially, I thought of scaling w.r.t. topper's grade. But if we consider real scenario, Admission committees don't always have topper's grade with them. So, I decided to simply calculate CGPA based on grade scale. Some Universities use 100 points scale while others use 10 point scale. We can easily see two groups of values in Fig 6 left plot. The group at bottom uses 10 point grade scale while other uses 100 point scale. After removing inconsistent values and scaling to 10 point scale, I constructed right plot which represents the processed data.

4.4 Other Features

As a part of Exploration Analysis, I explored other features like number of Conference submissions, number of Journal submission, Industrial Experience, Research Experience, etc. Many of the student profiles didn't have any research paper submissions or significant research experience. But they might prove to be important later on, So I cleaned all this data until it looked satisfactory. Following are the plots for each of them :-

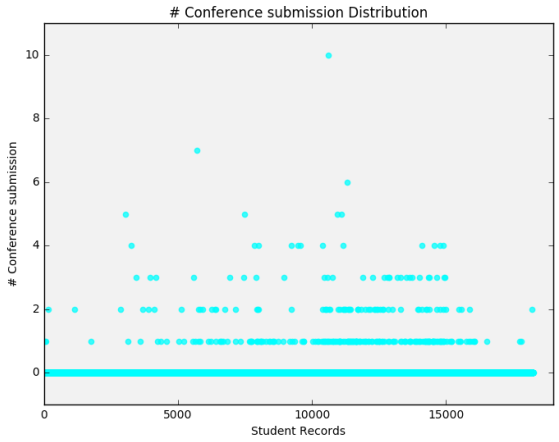


Fig. 8 Conference Submission Distribution

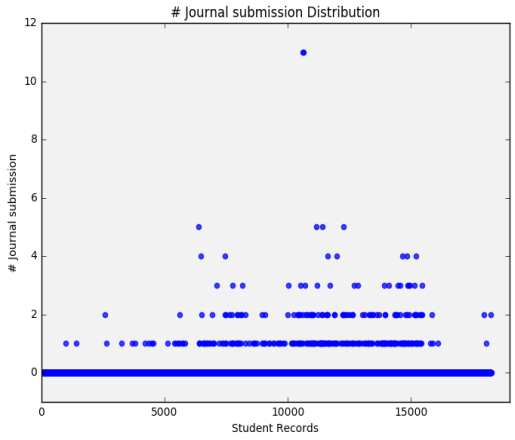


Fig. 9 Journal Submission Distribution

Let's see the distribution for Intern Experience, Industrial Experience & Research Experience.

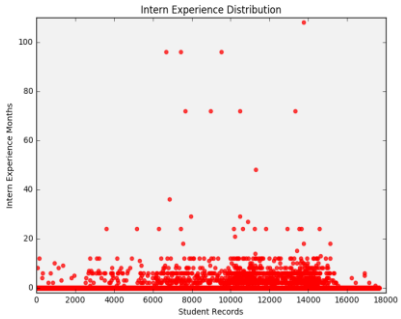


Fig. 10 Intern Experience Distribution

Fig. 11 Research Experience Distribution

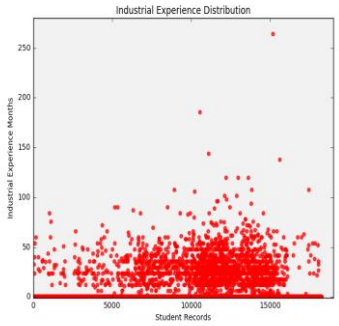


Fig. 12 Industrial Experience Distribution

As clear from Fig 9,10 & 11, we can see many of the students have more Industrial experience than research or intern experience. Intern and Research experience is generally under 20 months but industrial experience can be more than 50 months in significant amount of cases.

5. Modeling & Results

After extensive Data cleaning & Exploration, Let's model data. On observing the program feature, we see that students are interested in applying for Master, PhD or both. Fig 12 tells us that there are a lot more MS applicants than PhD applicants. To give a perspective, there are about 17k MS samples and just 500 samples for PhD. The admission criteria for PhD might be a bit strict than MS admissions as it is generally fully funded. Moreover, PhD admissions have more focus on research projects and research experience. Due to these differences, we can't model MS and PhD in the same bucket. Since, we don't have significant data for PhD admissions, so we'll just focus on MS admissions.

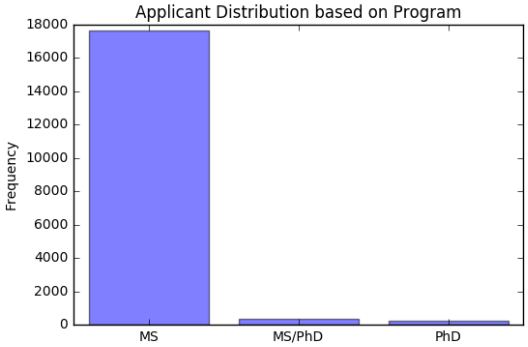


Fig. 13 Applicant Distribution based on Program

Some interesting observations :-

i) CGPA has the most discriminative power i.e. CGPA is most important feature to decide admissions decision of a student.

ii) Most Applied Universities are not the most famous ones. The most popular Universities as per Edulix Data are :-

a) University of Southern California
b) Arizona State University
c) University of Texas Dallas
d) North Carolina State University
e) SUNY Stony Brook

Let's model all Universities data at once considering that they all are in one bucket. We'll use a baseline line model which returns admit for all input profiles. We'll divide the dataset into train and test in the ratio 80:20

	Precision	Recall	F-Score
Baseline	0.5088	0.9998	0.6743
Decision Tree	0.5673	0.4746	0.5168
AdaBoost	0.5487	0.6779	0.6065
Random Forest	0.5546	0.5507	0.5527
Naïve Baye's	0.5277	0.8630	0.6549
SVC	0.5562	0.5547	0.5523

As we can see from above table, No model performs better than Baseline model. Let's try to visualize this using PCA.

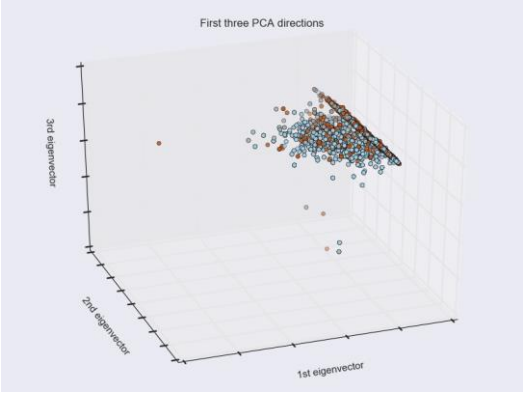


Fig. 14 Visualizing data using 3 Principal Vectors

We can observe from Fig.14 that data points are not clearly separable. They seem highly intermingled so let's reduce the variability a bit by trying to model each University at a time.

Let's take the case of USC. We'll take only those student profiles who have applied to SUNY Stony Brook. Now, we'll use the same modeling techniques as before and hope to see better results.

	Precision	Recall	F-Score
Baseline	0.5088	0.9998	0.6743
Decision Tree	0.5684	0.4757	0.5179
Random Forest	0.5588	0.5526	0.5557

AdaBoost	0.5597	0.6779	0.6065
Naïve Baye's	0.5427	0.8930	0.6979
SVC	0.5562	0.5547	0.5523

6. Conclusion

We have used Edulix to perform extensive analysis of American Graduate Process. We have used many real features to model admission process. First we tried to model all universities at once, but this seems to be quite difficult since it involves lot of variability. Each University has slightly different admission criteria so predicting results all universities at once doesn't seem a good idea. Next, I tried to approach each university at a time. For Example- I used the case of Stony Brook University. This time it seems to work better and give good results. To generalize to all Universities, we should model each University individually or probably cluster similar Universities together.

This project was a great learning opportunity. I started right from website scrapping to data cleaning and finally modeling. It helped me to experience the real life challenges faced by a Data Scientist.

7. Limitations & Future Work:

1) **More Diverse Data**
   Data is the fuel which drives Data Science. We can improve our analysis and predictions significantly by having more data. In this regard, I have been trying rigorously to find more data sources. I came across some new data sources like Yocket.in, Gradcafe.com, Facebook Groups, etc. The Data gathered from Facebook Groups is in structured form and can be used after some cleaning. Yocket.in offers wide range of parameters for significant number of users. We can also get updated dataset from Edulix.com.

2) **Authenticity**
   The data that I have used is self reported data which might be wrong. If input data is Grabage then we'll get only Garbage out. If we can get authentic data, we can come up with more accurate predictions.

3) **Diversity: Demographic-Bias**
   We might look for data from more diverse sources like Chinese, German or Korean websites. This will help us see the big picture. It will also enable to compare the strengths, aspirations and priorities of International Students based on demography. We can compare the strength from GRE scores, TOEFL score, CGPA, etc. We can study the top priorities to shortlist a University like location, reputation, ranking, etc. by observing the popular Universities in each country. We can gauge the aspirations by comparing the Universities applied and the universities admitted to.

4) **Bias (Admit/Reject)**
   The data is biased in the respect that it has lot more accepts than in real case scenario. In real case scenario, the acceptance rate may lie between 5%-15%. In the Edulix data, People have reported more accepts than rejects. The rate of acceptance is 65% which is far beyond the accepted range. Probably, we should hunt for other sources which provides genuine data.

5) **Advanced Modeling (NLP)**

We should advanced NLP techniques to capture SOP, LOR, etc. SOP & LOR play a major role in Graduate student admissions. We can't achieve high accuracy if we don't included them in our analysis. We should try to rank the Organizations where Internship or Full-time work is carried out by the applicants. Furthermore, We should also categorize conferences and journals where applicant has submitted his work.

References:

[1] UNESCO Stats http://www.uis.unesco.org/Education/Pages/international-student-flow-viz.aspx

[2] IIE Report http://www.iie.org/Services/Project-Atlas/United-States/International-Students-In-US

[3] TOEFL Score Conversion http://transint.boun.edu.tr/toefl/belgeler/puanlar.pdf

[4] GRE Score Conversion https://www.ets.org/s/gre/pdf/concordance_information.pdf

[5] American graduate admissions: both sides of the table http://hdl.handle.net/2142/92866

[6] Waters, Austin, and Risto Miikkulainen. "GRADE: machine learning support for graduate admissions." *AI Magazine* 35.1 (2014): 64.

[7] [Agrawal et al., 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, VLDB, 1215:487-499.

[8] Bruggink, T. H., and Gambhir, V. 1996. Statistical models for college admission and enrollment: A case study for a selective liberal arts college. Research in Higher Education 37(2):221–240.

[9] Lux, Thomas, et al. "Applications of supervised learning techniques on undergraduate admissions data." *Proceedings of the ACM International Conference on Computing Frontiers*. ACM, 2016.

[10] Ali, Alnur, et al. "Preferences in college applications-a nonparametric Bayesian analysis of top-10 rankings." *NIPS Workshop on Computational Social Science and the Wisdom of Crowds, December 10th 2010, Whistler, Canada*. 2010.

[11] [Raghunathan, 2010] Raghunathan, K. (2010). Demystifying the American Graduate Admissions Process Web.

[12] [Posselt, 2016] Posselt, J. (2016). Inside graduate admissions: Merit, diversity, and faculty gatekeeping. Harvard University Press.