

PROJECT REPORT

# PREDICTION OF ADMISSION IN US GRAD SCHOOLS

Shivaraj Nesaragi N15558346 ssn314  
Durvesh Patil N16683178 dsp372



NEW YORK UNIVERSITY



## **1) What did you propose to do? What is the motivation/background?**

### **Goal:**

The goal of our project is to build a model that will predict the list of graduate schools for MS in US. To be specific, we are predicting list of schools categorized into safe, moderate and ambitious based on the candidate profile. This helps students to decide on the universities they apply based on their profile so that they need not worry about going through troubles of contacting friends, consultancies etc. to shortlist the universities.

### **Motivation:**

To understand the background of the problem we referred to the ETS official website, where we learned numerous things such as total number of test takers from different countries and also the increasing trend in number of test takers etc. We learned how the number of Masters aspirants have doubled in the last decade and how more than 50% of the students are from the US, giving rise to TOEFL and IELTS being factored into our algorithms as main predictors. Also, one major motivation was that we had gone through the same process and we knew that one of the major issues faced by students is selection of universities based on their profile. The link we referred is: [https://www.ets.org/s/gre/pdf/snapshot\\_test\\_taker\\_data\\_2014.pdf](https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2014.pdf).

## **2) Explain the data you used and model in detail.**

### **Datasets:**

We used Facebook group posts as the main source of our data and the excel sheets provided in some of these Facebook groups. The Facebook posts we extracted using the Rfacebook package and then created a data table out of it. We collected the data from the 4 groups with each group consisting around 6000 posts out of which we eliminated few posts as they didn't contain the admit or reject posts and also gathered data from the excel datasheets and took a total of 10000 data rows. The data table contained rows with the student profile details like the year, term, GRE score, TOEFL score, IELTS score, GPA, University, Department, Admit/Reject. The excel data sheets we took contained more columns than needed such as Work Experience, Quantitative and Verbal Score, which we eliminated and considered only the required part to be in synchronization with the data collected from the Facebook group posts.

### **Model:**

The model is trained with the past Facebook data which was collected from the public groups such as MS in US Fall 2016, MS in US Spring 2016 etc. and from the excel sheets. We used the GRE, TOEFL, GPA as predictors for our model. The testing data

consists of the posts from the Fall 2017 group. The model follows the principle of supervised training.

We have divided our data into two parts in the ratio 80: 20, suggesting 80% of data is training data and the remaining 20% as testing data. From the data, we have gathered min, max and average values of all the important predictors(GRE, IELTS,TOEFL,). Using these values, a score is calculated using a formula for each university for each department. When a new input is given, a score is generated for the profile given as input. This score is compared to those generated previously and based on the comparison, the universities are classified as SAFE,MODERATE and AMBITIOUS.

### **3) What did you end up doing?**

We are predicting list of probable universities in the 3 ranges of safe, moderate and ambitious. We calculated the lowest and highest scores for an admit in a particular University for few courses for various terms and years i.e Fall 2016, Fall 2015, Spring 2016 and Spring 2015. Then we fed this to our model as training set data and the testing data was given to model in the form of manual entry as well as the posts from the Fall 2017 group. The algorithm of ours took the input from the training data and extracted the predictors out of it and computed a score for an University for a particular course by making use of weighted distribution of these predictors. The testing data was run through same algorithm and a profile score was generated for each input using the same concept of weighted distribution. Then we made use of the difference in the University and profile score to output the universities list based on the prefixed range for “Safe”, “Moderate” and “Ambitious” categories. We have predicted a list of Universities for the Fall 2017 group.

### **4) What if anything did you change about your approach and why?**

We initially wanted to have our model to be trained for 10 years of Facebook data. For this we had to collect data from 20 Facebook groups and such an approach would lead to collecting large amount of data, since each group contained around 5000 plus posts and extracting and cleaning up of data would take more time. We also intended to include “Work experience” as one of our predictors but we had to drop it eventually as many posts didn’t contain this information. Thus we changed our approach to include data from the two years 2015,2016 and two terms- “Spring” and “Fall” and dropped the “Work experience” predictor. This helped us to obtain fair amount of data with complete value for predictors. This approach of ours worked well and we could achieve the intended result.

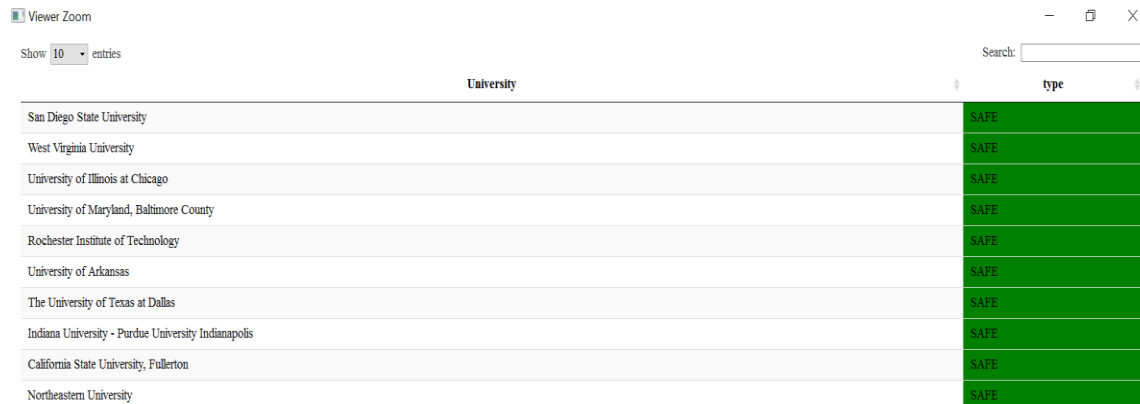
**5) What visualization(s) have you included? Explain what is conveyed in the visualization and why.**

For visualization purpose, we have given the output result as a data table and the type column in this is colored in green, orange and red colors which clearly depicts the universities in the “Safe”, “Moderate” and “Ambitious” range.

However, one of the major consideration was the seasonality since we as students knew that there is consideration difference in the profiles accepted during different terms i.e. SPRING and FALL. Also, when we analyzed the data, we found the same facts about seasonality.

We have also included couple of graphs to depict the relationship between average GRE and average GPA vs the number of profiles that received admits in Fall and Spring seasons. This can be clearly seen in the graphs which demarcates the Fall and Spring seasons in black and green colors.

The screenshots below show the visualization in our project.



The screenshot shows a web application interface for viewing data. At the top left, there is a 'Viewer Zoom' button. Below it, a 'Show 10 entries' dropdown menu is visible. On the right side, there is a search bar with the placeholder text 'Search:'. The main content area displays a table with two columns: 'University' and 'type'. The 'type' column is highlighted in green. The table contains 10 rows of data, all of which are labeled 'SAFE' in the 'type' column.

University	type
San Diego State University	SAFE
West Virginia University	SAFE
University of Illinois at Chicago	SAFE
University of Maryland, Baltimore County	SAFE
Rochester Institute of Technology	SAFE
University of Arkansas	SAFE
The University of Texas at Dallas	SAFE
Indiana University - Purdue University Indianapolis	SAFE
California State University, Fullerton	SAFE
Northeastern University	SAFE

Viewer Zoom

Show 10 entries

Search:

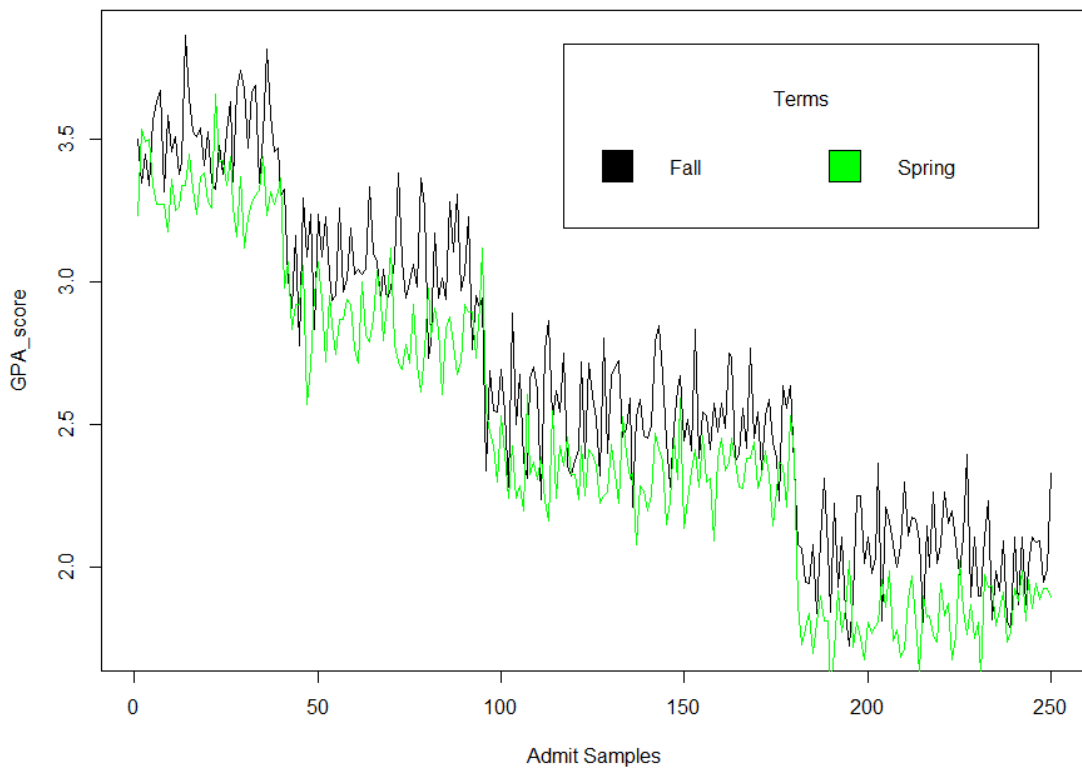
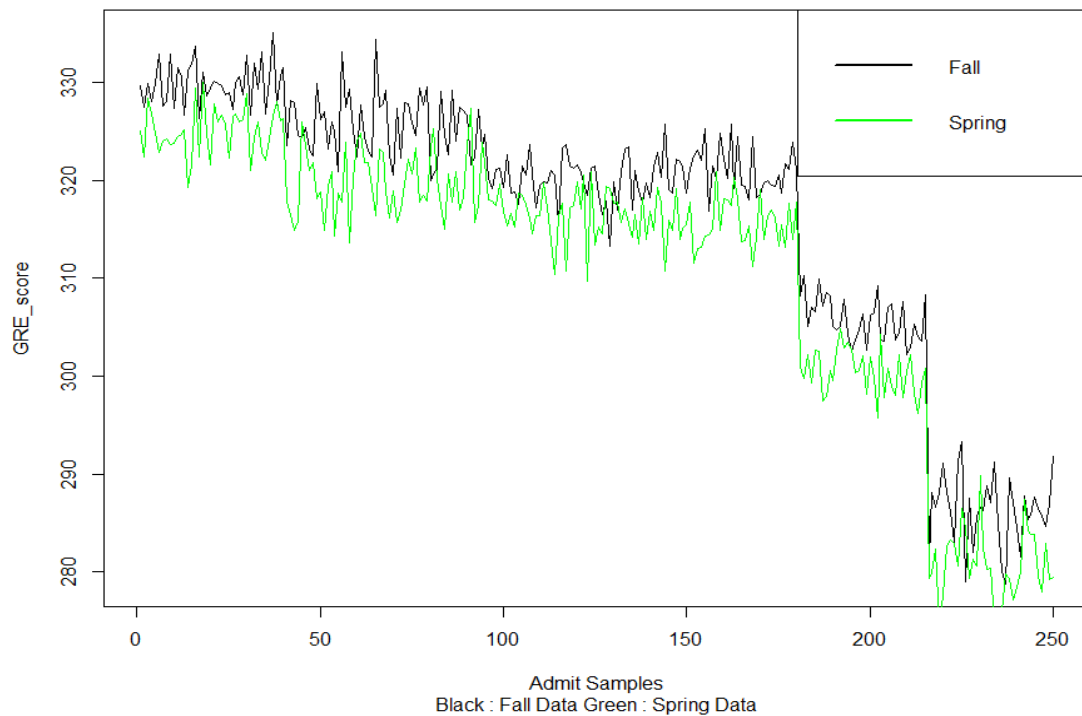
University	type
Georgia Institute of Technology	MODERATE
North Carolina State University	MODERATE
Syracuse University	MODERATE
Florida State University	MODERATE
University at Buffalo, State University of New York	MODERATE
Texas A&M University	MODERATE
Northwestern University	MODERATE
University of California, Santa Barbara	MODERATE
University of California, San Diego	AMBITIOUS
Arizona State University	AMBITIOUS

Viewer Zoom

Show 10 entries

Search:

University	type
New York University	AMBITIOUS
University of North Carolina at Chapel Hill	AMBITIOUS
Penn State University	AMBITIOUS
Cornell University	AMBITIOUS
The University of Texas at Austin	AMBITIOUS
University of Pennsylvania	AMBITIOUS
Stanford University	AMBITIOUS
University of California, Los Angeles	AMBITIOUS
Massachusetts Institute of Technology	AMBITIOUS
Purdue University	AMBITIOUS



This graph depicts the average GRE and GPA for the admit profiles in Fall and Spring term. We can see that the average GRE and GPA for successful admits is higher in Fall than Spring.

## **6) What evaluation method did you propose?**

Since our output consists of list of universities we cannot use any of the existing evaluation method such as F-1, K-fold, CV etc. So we created our own evaluation method. The method takes input as the user profile from the test data and the list of predicted universities from our model for this input. The method checks for a term such as Spring or Fall and then takes the list of predicted universities in “Safe” and “Moderate” range and checks it with the list of admits a profile has actually received for that profile. As mentioned above, a score is generated for each university and for each department. While going through test cases, the score is generated using the same formula for that test case and a list of safe, moderate and ambitious universities is created. If the test case university is present in the safe or moderate university list, then the accuracy is 100. Otherwise it is set to 0. The accuracy varies from 0 to 100% depending on the total number of universities present in the predicted list, for example: If there are 4 universities present in predicted list out of the 5 admits the profile has got in actual, then the accuracy is 80%.

## **7) How did your model perform according to this evaluation?**

So, what we did was we looped through the testing data and generated a list of Safe, Moderate and Universities. If the actual University was present in the safe or the moderate list, accuracy was set to 100. Otherwise it is set to 0. After looping through 2000 test cases, and averaging accuracy, we got an accuracy of 79%(78.9 %to be precise.) Considering, our previous evaluation strategies that we had devised, this one had a very high accuracy.

## **8) Based on your results what conclusions do you draw?**

Based on our above-mentioned evaluation model we got an average accuracy of 79% for our model and we can say the model is feasible.

Using the results:

- The students can make use of the list of the universities and can consider where to apply.
- Also, they can equally divide the universities in to 3 categories as per their liking.

**9) Based on your results what further studies would you do or are warranted?**

For future studies, we can incorporate large number of data by making use of the numerous Facebook group posts. Currently we are considering 50 universities and 5 departments. However, this could be expanded to include more universities and departments. Also, we only considered 5 admits and 5 rejects for each University-Department Combination. This amount of admits and rejects can also be increased. This will help us to predict the list of universities for all courses.

We can also include the scope of listing profiles for a particular university, where user can select a university and course so that they will be able to see the list of admitted profiles.