TEAM DR. DATA



Prepared By

|  |
| --- |
| Raghavendra Singh Raghav |
| Suvir Gupta  Roushan Kumar |

Contents

[Executive Summary 2](#_Toc476320588)

[Problem Statement: 2](#_Toc476320589)

[Approach: 2](#_Toc476320590)

[Findings: 2](#_Toc476320591)

[Tier Recommendation Criteria’s: 2](#_Toc476320592)

[Categorization Criteria’s: 4](#_Toc476320593)

[Sub-Categorization Weightage Criteria: 7](#_Toc476320594)

[Conclusion: 9](#_Toc476320595)

# 

# **Executive Summary**

## Problem Statement:

Teach for America (TFA) seeks to recruit remarkable leaders from a broad spectrum of universities. In

order to maximize our limited resources, we differentiate resource allocation for recruitment at

individual campuses, based on the competitive dynamics at each university. For the purpose of

this case study, we are mainly focusing on campus-based recruitment for undergraduates, though

we have separate strategies for both graduate students and professionals.

## Approach:

This report summarizes the fundamental idea behind proposed optimization of recruitment resource for TFA recruiting strategies, further categorization of school in recommended tiers to generate the most admits. The purpose of the report is to list all the idea and insights gathered while exploring, analyzing the data. The TFA data file has 13,699 and 8,919 of student records for 2016 and 2017 respectively. Our first objective is to combine the 2016 and 2017 student data with university information, so that we can analysis the data properly and can also apply modelling to gain insights about the important criteria for categorization of the universities in different tiers. We also utilized preprocessing techniques to make the data suitable for modeling.

## Findings:

1. Three-tier strategy for undergraduate is more potent than two-tier strategy.

2. Categorization of the university should be very large School centric along key parameters like GPA, US News university selectivity criteria, Recruiter, Alumni base etc.

# Tier Recommendation Criteria’s:

There are a couple of reasons or facts which suggest the effectiveness of Three Tier System.

Below are some important facts:

1. Basic Idea of Tier based categorization is to optimize recruiter efficiency as Tier 1 Universities have dedicated recruiter compared to Tier 2 and Tier 3.
2. Important factor here is to understand percentage of universities categorized in different tiers in 2016(Tier 2: approx. 75%, Tier 3: approx. 25 ) and 2017 (Tier1: 50%,Tier2: 36%, Tier 3: 14%). This shows availability of resources to each university and what effect it poses to the acceptance.
3. There are two variables in data set that corresponds to Recruiter efficiency i.e. *‘Sourced by RT vs SignUp Form’* and *‘Met?*’ will help analyze importance of recruiter in acceptance.
4. Below two distribution graph shows data from 2016 and 2017 shows only acceptance count of students who have been in interaction with Recruiter in one or the other way mentioned above.



Fig1: Sourced by RT Vs Signup & Met Acceptance count for each Tier in year 2016



Fig2: Sourced by RT Vs Signup & Met Acceptance count for each Tier in year 2017

1. It can be seen from graph that compared to 2016 in 2017 in all Tiers no of Students sourced by RT has increases. Considering these above points, we recommend three tier system for university selection as effective allocation of dedicated recruiter in case of 3 Tier system has essentially increased the admittance by **approx. 16 percent** compared to total seen in 2016.
2. Another reason for suggesting 3 tier system its due to limited number of resource available in TFA as mentioned in problem statement.
3. Further, most peculiar thing to notice is that percentage of acceptance increased by approx. 7 percentage, contrary rejection rate decrease from 2016 to 2017, exhibiting the importance or positive effect of three tier system on selection process mechanism.

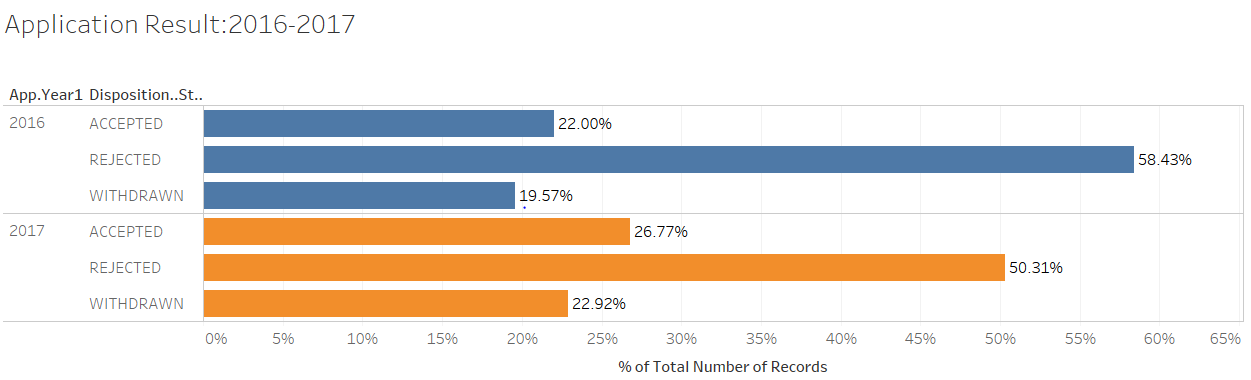


Fig3: Acceptance vs Rejection for year 2016-2017

1. Our analysis is based on the data available hence we don’t not recommend any further division of Tiers as data based on University expansion and no of dedicated HR resources available with TFA is not provided.

# Categorization Criteria’s:

To categorize universities into tiers. We follow a simple multi layer rank based approach for this analysis to rank each university based on cumulative rank calculated from weight assigned to variables and its sublevels. This helps to eliminate ambiguity of complex algorithm and put in more business sense with accurate results. In other sense we propose a linear model with calculated weights as coefficients of each predictor.

To categorize the rank into different Tiers we must bin them into different ranges based on business understanding i.e. university expansion rate and the HR resource availability in case of TFA. Based on data count available in tiers for universities, we are binning the output ranges in approximately 50%,36%,14% for Three tier system i.e. for Tier1, Tier2, Tier3.

Another advantage of this approach is if at any time, management wants to change the percentage allocation or want to move from three tier to four tier system they just have to change binning of ranks and rest complete system works the same.

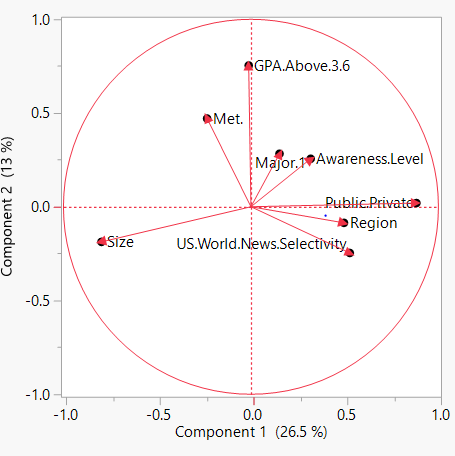
Now the question is to how to assign Weights to each Variable and its levels in case of categorical variable.

**Assigning Weights to Predictor Variables**

Firstly we union the data set from 2016 and 2017 joined with university data on university ID and clean duplicates, missing values. Etc. Now we are more concerned with variation caused by variables That leads to ACCEPTED decision by Student. So of the data cleaned we subset based on disposition step as ‘ACCEPTED’

To analyses the variation caused we are using PCA (principle component analysis) on the transformed categorical variables to standardized continuous numerical ranges. This analysis any variation in multidimension due to selected predictor variables. Variation in multidimension would cause more rotation of PCs axis causing alignment of axis to maximum variation. So Eigen vector of PCs that constitute the most variation gives the weightage of the variable.

Statistical Algorithm:

 Fig4: PCA analysis

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Major.1 | GPA.Above.3.6 | Met. | Public.Private | Size | US.World.News.Selectivity | Region | X..of.Alumni.who.attended.school.in.undergrad | Awareness.Level |
| 0.088308448 | 0.00855148 | 0.534032 | 1.148502 | 0.478513 | 0.434361 | 0.276826 | 0.042318254 | 0.282236904 |

Table 1: PC component with eigen vectors.

This just the first layer of variation explained by the predictor variable. In Second layer variation is explained by levels in each predictor variables

Further, for the generation of second layer of the recommender system, we again give weightage to sub category of each key attribute, these weightage generated based on percentage variance in each sub classification under given predictor variable for admitted candidates

For e.g. Size predictor variable have 4 sub classes Very large , Large , Medium and small , so weightage for Very Large size is percentage of total acceptance is explained by Very Large university.

Finally, we summarized all these weightage as attributes of each university to categorize them in 3 tier . **Based on total weightage our tier system divide like Tier -1: [34-26), Tier-2: [25-21), Tier-3: [20-17**).

**Weightage of Parameters and Sub-Classification:**

We have excluded some of the data in tables based on business and binned some as shown in the table below with weightages we use multiplier of 20 for eigen vector to get a wider range:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.no | **Parameter** | **Weightage1** | **Sub Classification** | **Weightage2** |
| 1 | Size | 9.57 | Very Large | 0.47 |
| Large | 0.28 |
| Medium | 0.12 |
| Small | 0.13 |
|  | | | | |
| 2 | Met | 10.68 | Yes | 0.37 |
| No | 0.63 |
|  | | | | |
| 3 | US News Selectivity | 8.69 | Most | 0.47 |
| More | 0.37 |
| Selective | 0.12 |
| Less | 0.04 |
| Least |  |
|  | | | | |
| 4 | Major1 | 1.77 | > 5 % Acceptance in most Common Major | 0.58 |
| 3-5 % Acceptance in most Common Major | 0.29 |
| 1-3 % Acceptance in most Common Major | 0.11 |
| < 1 % Acceptance in most Common Major | 0.02 |
|  | | | | |
| 5 | Awareness Level | 5.64 | High | 0.47 |
| Medium | 0.30 |
| Low | 0.03 |
|  | | | | |
| 6 | GPA | 0.17 | > 3.6 | 0.56 |
| < 3.6 | 0.44 |
|  | | | | |
| 7 | Public / Private | 22.97 | Public | 0.53 |
| Private | 0.47 |
|  | | | | |
| 8 | Alumni | 0.85 | > 100 | 0.55 |
| < 100 | 0.45 |
|  | | | | |
| 9 | Region | 5.54 | East | 0.43 |
| West | 0.16 |
| North | 0.20 |
| South | 0.21 |

Table 2: Variable weightage table

For an example based on the weightage table, we can calculate the total weightage for a university having below attributes:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Size | Met | US News Selectivity | Major1 | Awareness | GPA | Public / Private | Alumni | Region |
| Very Large | No | More Selective | Psychology | Medium | <3.6 | Public | 257 | North |

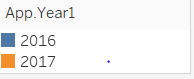
Total Weightage: 9.57\*0.47+10.68\*0.63+8.69\*0.37+1.77\*0.11+5.64\*0.30+0.17\*0.44+22.97\*0.53+0.85\*0.55+5.54\*0.2 = 30.15

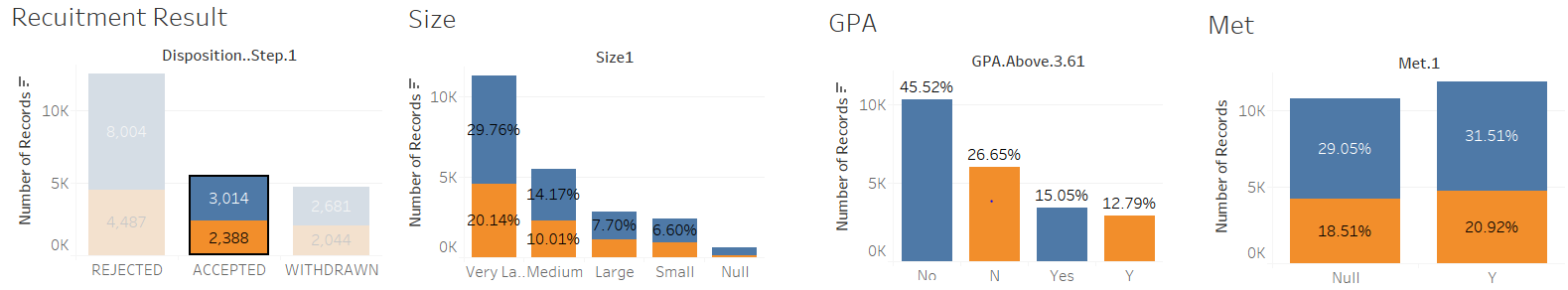
Recommended Tier: Campus 1

This weightage has been calculated by averaging predictor variable contribution from each model, which are specified above.

# Sub-Categorization Weightage Criteria:

Below is the screen shot from dash board that shows the variation caused by sublevels for Recruitment result accepted. This constitutes the bases for calculation of weightage for each level. That is the percentage of total acceptance explained by predictor levels.





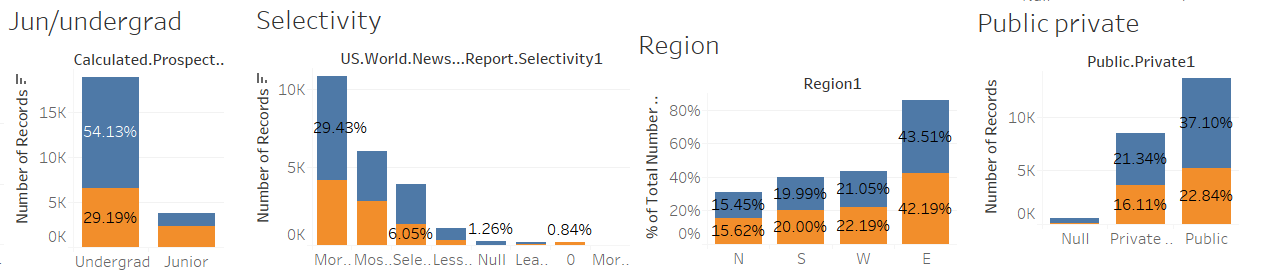
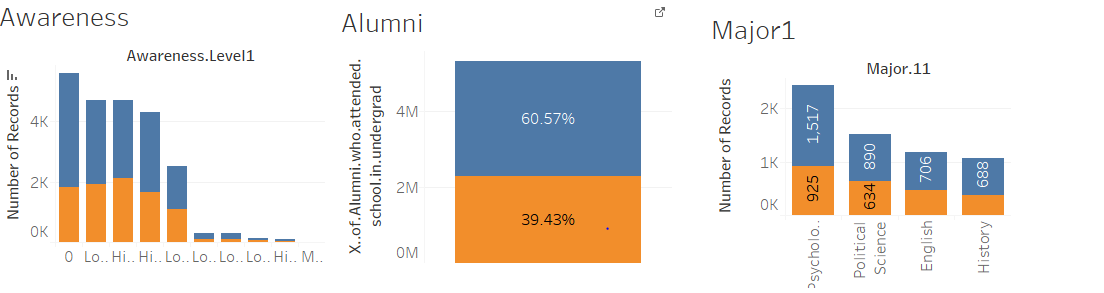
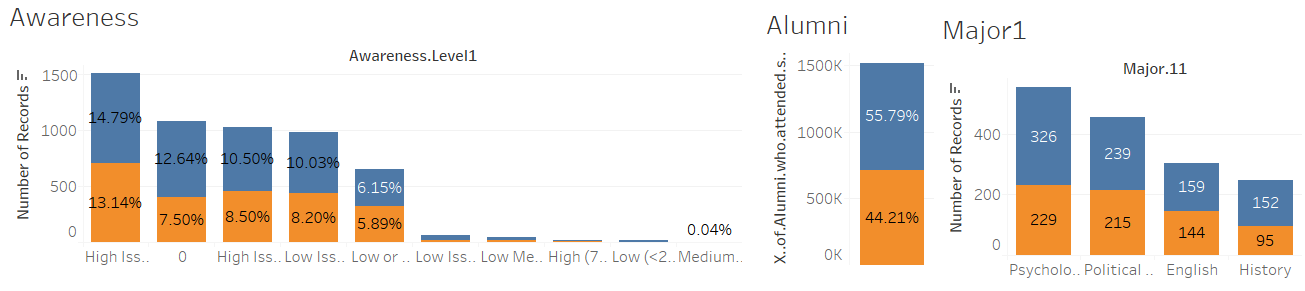


Fig 5 : Dashboard for predictor variables

So, for these parameters we utilized the percentage distribution for generating weightage.



We incorporated our business sense as well while giving weightage to these parameters like as by now we already know that size of the university is much more important attribute compare to other parameters like alumni or GPA etc. for increase in acceptance count.



**Binning Of Predictor variables:**

Binning of Predictor variables is done for Alumni and Major1. As Alumni is categorical variable it may constitute to more variation while using PCA so we bin them accordingly.

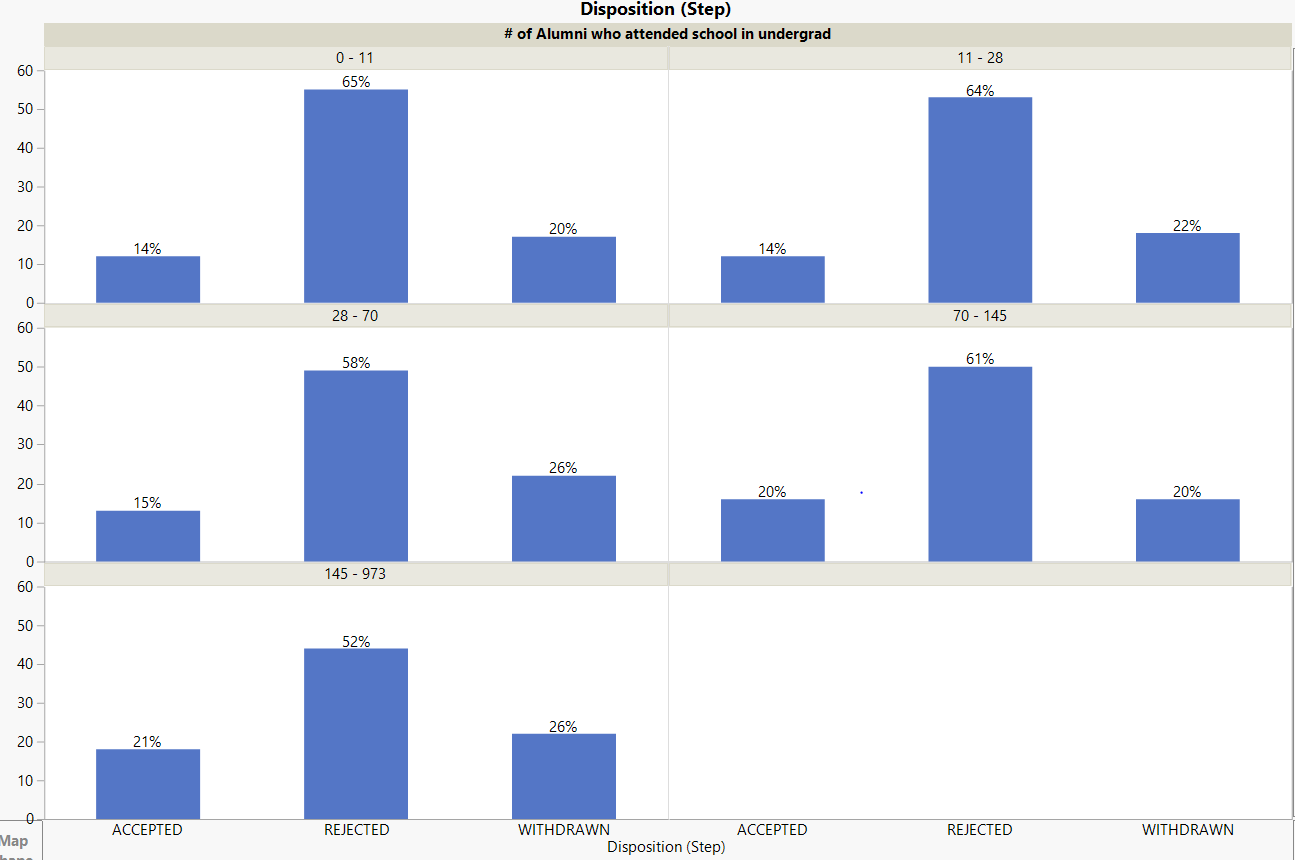


Fig 6: Binning Of Alumni variable

As seen from the diagram Alumni <70 and Alumni < 70 constitute maximum variation in percentage acceptance so we binned them accordingly. Other similar predictor variables(Current CMs who attended school in undergrad, Current TFA Staff who attended school in undergrad) constitute same kind of variation so we excluded them for simplicity.

Second Binning is done for Major one As (> 5 % Acceptance in most Common Major, 3-5 % Acceptance, 1-3 % Acceptance in most Common Major, < 1 % Acceptance in most Common Major). While assigning weight to each bin no of element in each bin is taken in consideration to average the weight. We have excluded other Majors and minors as they constitute very less to variability.

Percentage Distribution of Majors of Accepted Applicants:

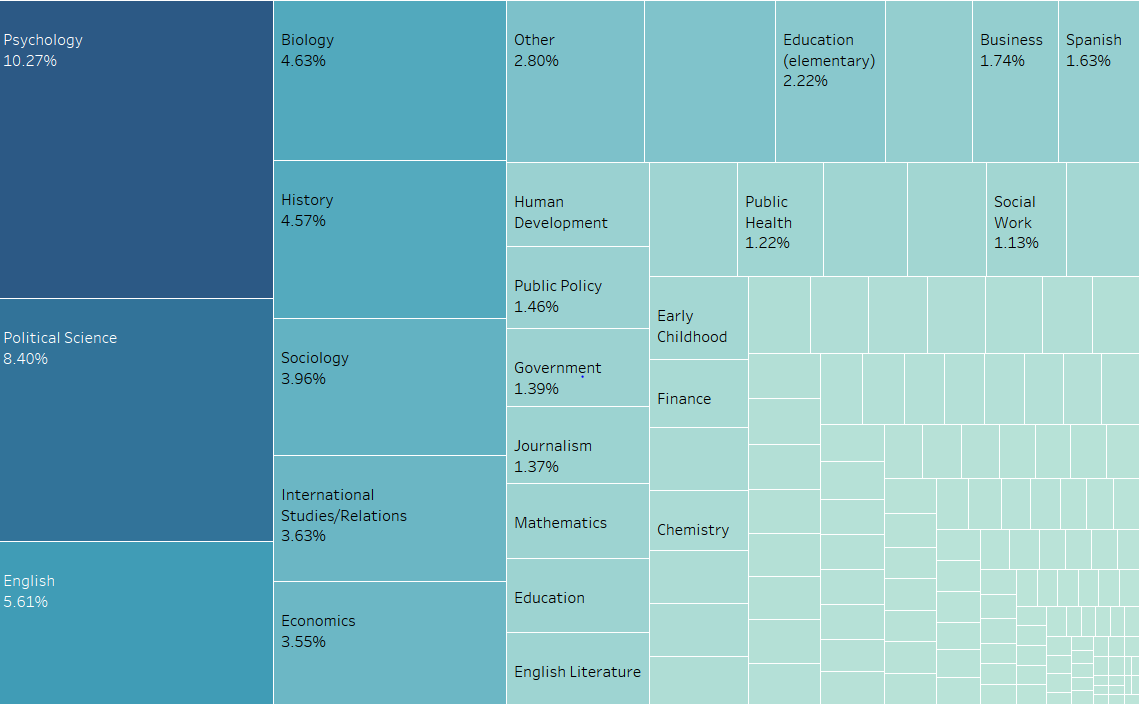


Fig 7: Binning of Major1 Variable

# Conclusion:

The TFA data was subjected to exploratory analysis and necessary modifications were made based on the exploratory analysis. Different insights were gained and further analysis is performed utlising the different Classification and Regression models built on the sample data. Based on acquired information and business sense different key parameter for admittance improvement were identified.We have found that the 3 tier system more effective in optimizing the resource allocation and can prove more efficent than previous 2 tier system in long run.Also we have built a recommender ranking system utilising the weightage created using the models and distribution graphs with the aim of categorising the university in recommended 3 tier system.Proposed Ranking system will improve the flexibility in adding and categorizing new and present universities in coming future.