Teach for America Case

Professor Oliver J. Rutz
Customer Analytics
14th December ‘2022

Group Members
Yunshan Liu, Xiaofu Li, Rida Sohail, Ryland Mirano, Sadia Haque

Introduction

This report focuses on classic scoring activity (regularly carried out for customer acquisition). Teach for America is the organization in discussion. It gives us a list of 3748 names of recent graduates to whom they have made job offers. TFA's difficulty is a lack of understanding regarding which characteristics most substantially impact recent graduates' decision to matriculate when they get offers from their corps; in other words, they want to know which applicants are most likely to accept their offer. Furthermore, because TFA typically has fixed capacity in each recruiting cycle, they want to know how many recent grads they may issue offers to in order to avoid the number of applicants eventually joining the corps exceeding their opening spots. Our team opted to utilize the Binary Logit Model to examine the relationship between matriculation and a set of factors, and List Scoring to forecast each candidate's likelihood of matriculation as well as who TFA should extend an offer to.

Data Description

Teach for America: Understanding matriculation of recent graduates who received offers from their corps
TFA has sent out offers to 3747 recent graduates and has observed matriculation (Accept their Offer or Not Accept their Offer)
The list owner has given data on 30 variables listed below:

- Matriculated: the observed matriculation (0 = no, 1 = yes)
- OfferDeadline: Deadline of offer
- PositionName: Position ID
- Region: Region ID
- RegionPrefLevel: Region Preference of Application (1, 2, 3)
- RequestedRegionalReassignment: Yes/No whether assignment is in requested region
- Ethnicity: Ethnicity of list owner
- UndergraduateUniversity: Undergraduate University of list owner
- Senior2007: Grade level status in 2007 (0 = not a senior, 1 = is a senior)
- Degreedate: Date receiving the degree
- CumulativeGPA: GPA of list owner measured in 4.0 scale
- SpecialEducationPrefLevel: Special Education Pref Level of list owner (measured 1-3 scale)
- BilingualPreferred: Bilingual preference (0 = not preferred, 1 = preferred)
- Major1: Major of list owner
- Major2: Second major of list owner
- Gender: Gender of list owner (0 = female, 1 = male)

- ReceivedPellGrants: Received pell grants or not (No/ Partial/ Maximum)
- FederalStudentLoans: The amount of federal loans student receive
- PrimarySubject: Primary teaching subject of list owner
- GradeLevel: Teaching grade level of list owner (MIDDLE/ ELEM/ HIGH)
- IsMathMajorMinor: If the list owner is math major or not (0 = no, 1 = yes)
- IsScienceMajorMinor: If the list owner is science major or not (0 = no, 1 = yes)
- IsEngineeringMajorMinor: If the list owner is engineering major or not (0 = no, 1 = yes)
- IsMathSciorEngMajorMinor: If the list owner is math/sci/eng major or not (0 = no, 1 = yes)
- Contacted: If the list owner contacted with TFA or mot (0 = no, 1 = yes)
- EmailDialogue: The number of email dialogue the list owner contacted with TFA
- InpersonDialogue: The number of in person dialogue the list owner contacted with TFA
- LimitedEmail: The number of limited email dialogue the list owner contacted with TFA
- LimitedPhone: The number of limited phone dialogue the list owner contacted with TFA
- PhoneDialogue: The number of phone dialogue the list owner contacted with TFA

Methodology

1. Preparing Data

There are 30 variables and 3747 observations in the original dataset TFA provided to us. However, there are some missing values in the dataset and not all the variables are necessarily related to “Matriculated”, the variable we want to predict with models. For example, data in “EmailDialogue”, “InpersonDialogue”, “LimitedEmail”, “LimitedPhone”, “PhoneDialogue” are enough to be summarized in “contacted”. Variables like “FederalStudentLoans” have too many missing values, deleting the rows that missing value in “FederalStudentLoans” will significantly reduce the number of observations, so we just delete columns with too many missing values in the dataset. Besides, some variables have too many unique values, for example, “UndergraduateUniversity” has 128 unique values in the dataset, which makes no sense to analyze the correlation of such x factor between y, “Matriculated”, so we delete such columns as well. After the cleaning process, there are 18 variables and 2095 observations remain in the dataset.

However, this refined dataset still has some categorical variables that contain text coded values unable to conduct the binary logit model later. Thus, we convert categorical variables into numerical ones by assigning a number to each text value. For instance, there are 7 unique values in “Ethnicity”: "EUROPEAN", "ASIAN", "AFRICAN", "MULTI", "HISPANIC", "OTHER" "NATIVE" . We assigned each of them a number from 1-7, respectively, and each number stands for an ethnicity. Another example is“ReceivedPellGrants”. We code the answer “No” as 0, "Yes, I received a partial Pell Grant" as 1, and "Yes, I received the maximum Pell Grant" as 2. After we finished changing the type of our data, we downloaded and named the refined dataset "tfa_clean.csv", and read in the data in our list scoring RMD script. A brief view of our data is as following:

```
'data.frame':   2570 obs. of  18 variables:
 $ PersonId                    : int  1583124 1593734 1591628 1578563 1347694 1547629 1582267
1580322 1563800 1591501 ...
 $ Matriculated                : int  0 1 0 1 0 1 1 1 1 1 ...
 $ RegionPrefLevel             : int  2 1 1 1 1 1 1 1 1 1 ...
 $ RequestedRegionalReassignment: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Ethnicity                   : int  1 2 2 3 1 1 1 1 1 1 ...
 $ Senior2007                  : int  1 1 1 0 1 1 0 1 1 1 ...
 $ CumulativeGPA               : num  3.9 3 3.9 2.9 4 3.5 3.2 3.7 3.4 3.9 ...
 $ SpecialEducationPrefLevel   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ BilingualPreferred          : int  1 0 0 0 0 0 1 0 0 0 ...
 $ Gender                      : int  0 0 1 1 1 0 1 0 0 1 ...
 $ ReceivedPellGrants          : int  0 0 0 0 0 0 1 0 0 0 ...
 $ PrimarySubject              : int  1 2 1 3 1 1 1 4 5 6 ...
 $ GradeLevel                  : int  2 2 1 2 1 1 1 2 2 3 ...
 $ IsMathMajorMinor            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ IsScienceMajorMinor         : int  0 1 0 0 0 0 0 1 0 1 ...
 $ IsEngineeringMajorMinor     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ IsMathSciorEngMajorMinor    : int  0 1 0 0 0 0 0 1 0 1 ...
 $ Contacted                   : int  1 0 1 0 0 1 0 0 1 0 ...
```

2. Exploring Data

```
    PersonId       Matriculated    RegionPrefLevel RequestedRegionalReassignment   Ethnicity       Senior2007     CumulativeGPA
 Min.   :1126729   Min.   :0.0000   Min.   :1.000   Min.   :0.00000               Min.   :1.000   Min.   :0.0000   Min.   :2.500
 1st Qu.:1530876   1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:0.00000               1st Qu.:1.000   1st Qu.:1.0000   1st Qu.:3.400
 Median :1589811   Median :1.0000   Median :1.000   Median :0.00000               Median :1.000   Median :1.0000   Median :3.600
 Mean   :1566475   Mean   :0.8152   Mean   :1.067   Mean   :0.04708               Mean   :1.892   Mean   :0.8918   Mean   :3.556
 3rd Qu.:1641297   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.00000               3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:3.800
 Max.   :1696428   Max.   :1.0000   Max.   :3.000   Max.   :1.00000               Max.   :7.000   Max.   :1.0000   Max.   :4.000
 SpecialEducationPrefLevel BilingualPreferred     Gender       ReceivedPellGrants PrimarySubject     GradeLevel
IsMathMajorMinor
 Min.   :1.000             Min.   :0.0000     Min.   :0.0000   Min.   :0.0000     Min.   : 1.000   Min.   :1.000   Min.   :0.00000
 1st Qu.:1.000             1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:0.00000
 Median :1.000             Median :0.0000     Median :0.0000   Median :0.0000     Median : 5.000   Median :2.000   Median :0.00000
 Mean   :1.517             Mean   :0.1887     Mean   :0.2728   Mean   :0.3525     Mean   : 6.572   Mean   :1.699   Mean   :0.03113
 3rd Qu.:2.000             3rd Qu.:0.0000     3rd Qu.:1.0000   3rd Qu.:0.7500     3rd Qu.:12.000   3rd Qu.:2.000   3rd Qu.:0.00000
 Max.   :3.000             Max.   :1.0000     Max.   :1.0000   Max.   :2.0000     Max.   :21.000   Max.   :3.000   Max.   :1.00000
 IsScienceMajorMinor IsEngineeringMajorMinor IsMathSciorEngMajorMinor   Contacted
 Min.   :0.0000      Min.   :0.00000         Min.   :0.0000           Min.   :0.0000
 1st Qu.:0.0000      1st Qu.:0.00000         1st Qu.:0.0000           1st Qu.:0.0000
 Median :0.0000      Median :0.00000         Median :0.0000           Median :0.0000
 Mean   :0.1346      Mean   :0.02101         Mean   :0.1763           Mean   :0.1856
 3rd Qu.:0.0000      3rd Qu.:0.00000         3rd Qu.:0.0000           3rd Qu.:0.0000
 Max.   :1.0000      Max.   :1.00000         Max.   :1.0000           Max.   :1.0000
```

The summary function in R provides us a way to see some statistics of each variable. Since "Matriculated" is coded in 0/1 indicator, which 1 stands for accepted TFA's offer while 0 stands for didn't accept TFA's offer, the mean of this binary variable is simply the proportion of 1's. We can see the mean of "Matriculated" is 0.8152, so this means 81.52% of people who received TFA's offer in our dataset joined the corps. Same as variable "Gender", we can easily see 27.28% of applicants are male from the mean of gender. Besides mean, min, max and mode also provide valuable insights to our data. For instance, we can know the cumulative GPA of people who received TFA's offer ranging between 2.5 and 4.0, and 3.6 was obtained by the largest number of people.

3. Running BINARY LOGIT MODEL on the TRAINING Data

We then run the Binary Logit Model in R studio on the TRAINING Data (we defined the first 80% IDs as training data, last 20% IDs as testing data). The dependent variable we defined in the model is "Matriculated", and independent variables are all 17 variables in the dataset besides "Matriculated".

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   4.92741    0.86468   5.699 1.21e-08 ***
RegionPrefLevel              -0.55854    0.17225  -3.243 0.001185 **
RequestedRegionalReassignment -0.04206    0.26749  -0.157 0.875062
Ethnicity                    -0.03409    0.04509  -0.756 0.449624
Senior2007                    0.13082    0.18287   0.715 0.474379
CumulativeGPA                -0.74711    0.21162  -3.530 0.000415 ***
SpecialEducationPrefLevel    -0.04530    0.06995  -0.648 0.517262
BilingualPreferred            0.18632    0.15744   1.184 0.236610
Gender                       -0.19412    0.13226  -1.468 0.142178
ReceivedPellGrants           -0.10819    0.09448  -1.145 0.252193
PrimarySubject               -0.01302    0.01257  -1.036 0.300243
GradeLevel                    0.04031    0.08748   0.461 0.644932
Contacted                     0.13837    0.14491   0.955 0.339657
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1873.0  on 2055  degrees of freedom
Residual deviance: 1845.3  on 2043  degrees of freedom
AIC: 1871.3

Number of Fisher Scoring iterations: 4
```

From the Binary Logit Model summary report, we can see "RegionPrefLevel" and "Cumulative GPA" both have the P value less than 0.05, which indicates that region preference of application and cumulative GPA are significant to the matriculation. Some other variables like "Gender", "BilingualPreferred", and "Received PellGrants", although having the P value larger than 0.05, their P values are still relatively low, which indicates that they are considerably significant to matriculation, too.

4. Scoring the TESTING Data

To understand which applicants TFA should give offers to, we can use the model we obtained from training data and conduct List Scoring using Logit on TESTING Data (last 20% IDs in the dataset). First, we computed for each person on the list the predicted response probability (we named it "BinaryLogitProbability"), rounded result of predicted response probability (we named it "BinaryLogitPredict"), predicted lift (we named it "lift"). We added these variables along with the real response (we named it "Matriculated") to the response probability chart. Then, we sorted the prospects in decreasing order of lift. The sorted response probability chart allows us to predict the accept/not accept offer decision for each of the 514 IDs in the TESTING sample, and TFA should be more interested in prospects with higher values of the above measures.

	ID	BinaryLogitProbability	BinaryLogitPredict	Matriculated	lift
2129	1531258	0.9295512	1	1	1.119600
2473	1693582	0.9112292	1	1	1.097532
2558	1662443	0.9103465	1	1	1.096469
2316	1677825	0.9069278	1	1	1.092351
2269	1661306	0.9068470	1	1	1.092254
2428	1695510	0.9022735	1	1	1.086745
2523	1437720	0.9013921	1	1	1.085684
2117	1693201	0.8963424	1	0	1.079602
2508	1696415	0.8948723	1	1	1.077831
2450	1544459	0.8945494	1	1	1.077442
2163	1640935	0.8939319	1	1	1.076698
2407	1685783	0.8928962	1	1	1.075451
2148	1571994	0.8915621	1	1	1.073844

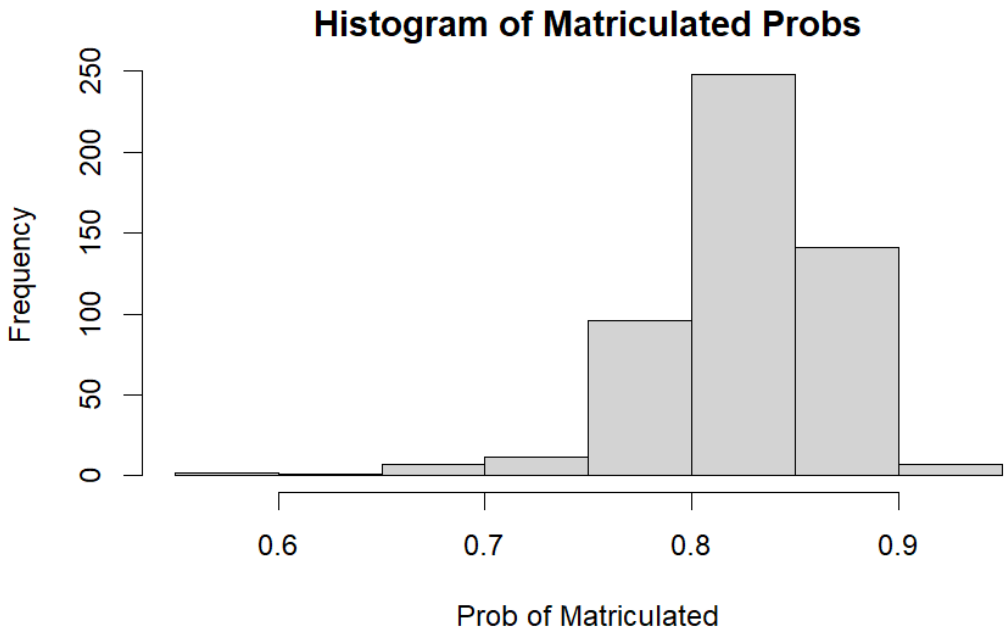
Based on the calculated values in this response probability chart, we then created a series of histograms, confusion matrix, and plots to better visualize the result as shown in the key findings section. We also created some bar graphs to analyze the relationship between matriculation and some particular factors.

Key Findings

1. **Marginal matriculated rate vs. number of applicants Chart**

   From the marginal matriculation rate chart, we are looking at the numbers of applicants (prospects) vs. their matriculation probability. From the graph, we found that the matriculation probability decreases as the number of applicants increases. Our findings show that matriculation probability is at its highest between 0 and 100 applicants, where probability hovers around 0.85-0.90. Furthermore, matriculation probability steadily declines from 100 to 500 applicants, before plummeting from 500 applicants and onward. The probability from 500+ applicants goes from about 0.75 to under 0.60 and presumably lower. We can conclude that as the number of applicants increase, there are less openings within Teach For America and more competition to join the teaching corps.

2. **Histogram of Matriculated Probability**

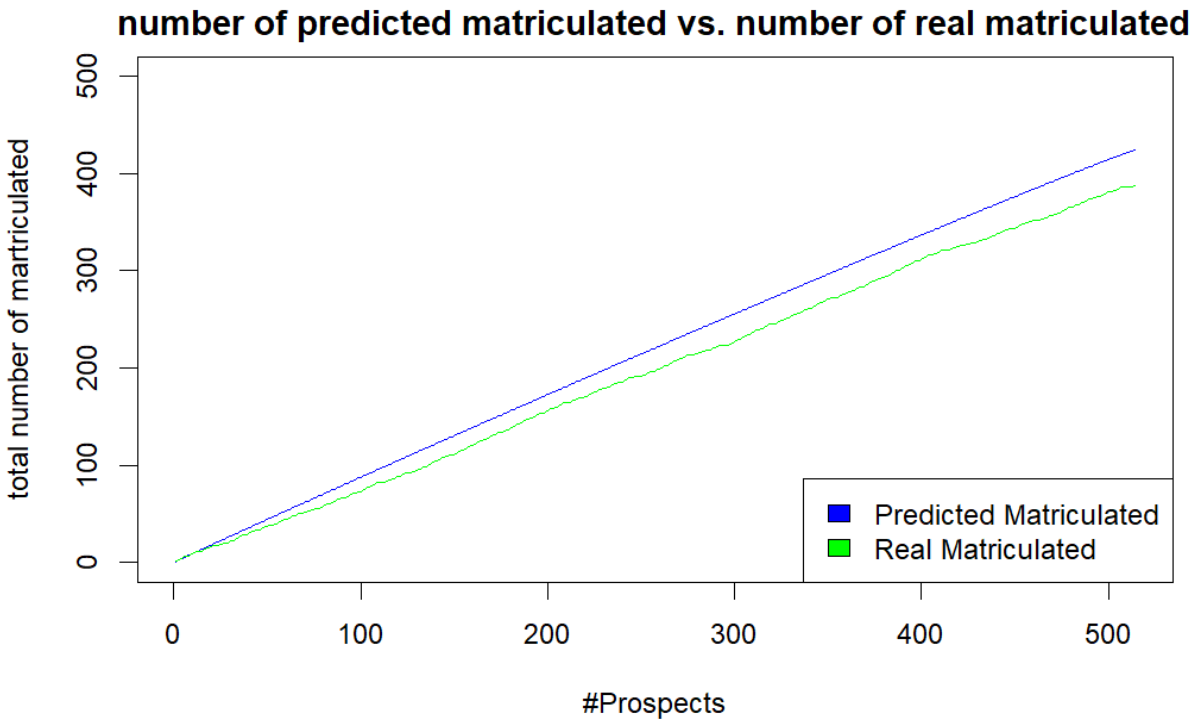


From the above histogram, it is observed that the overwhelming majority matriculation probability falls between 0.80 and 0.85. We can contribute this to the marginal matriculation rate graph, as the probability hovers around 0.8 to 0.85 at the 200 to 250 applicant prospect mark. The findings from the histogram can tell us that a higher number of applicants have a high probability rate of being matriculated. The highest number of applicants have a matriculation probability between 0.75 and 0.90.

3. Number of predicted matriculated vs. number of real matriculated

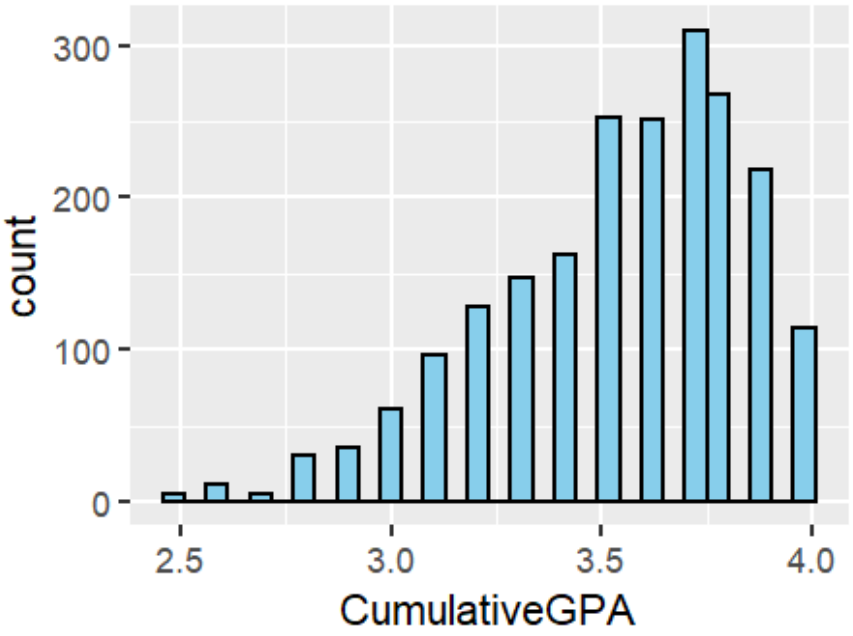
```
[1] 514
[1] 423.8503
```

data.testing$Matriculated	prediction.testing$BinaryLogitPredict 1	Row Total
0	126	126
1	388	388
Column Total	514	514



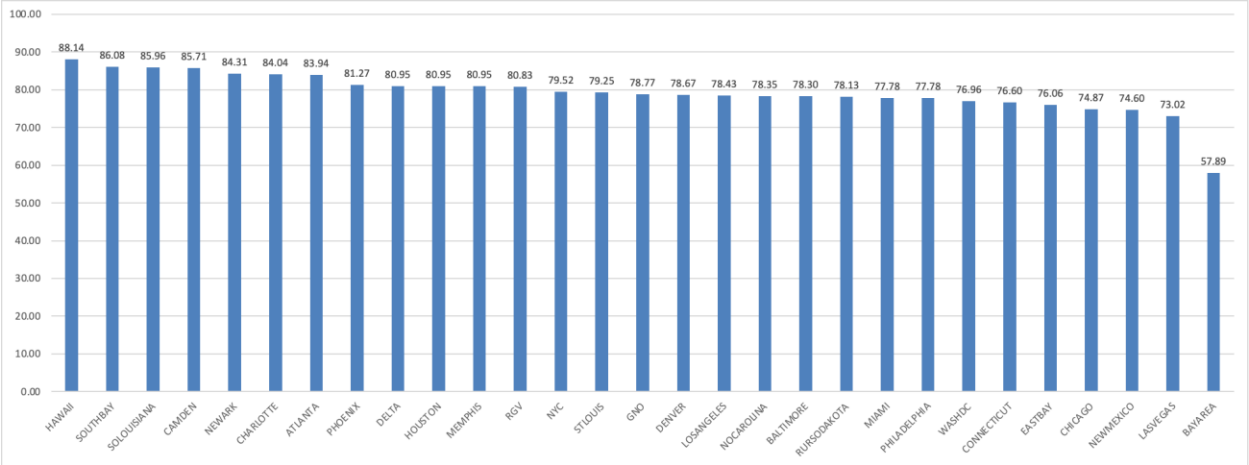
By summing up predicted matriculation and real matriculation, we received the result that among 514 people in the TESTING Data, the predicted matriculation is 423.85 people while the real matriculation is 388 people. The line graph provides a more intuitive view to show that there are more people that are predicted than the number of people who actually end up matriculating. This phenomenon could be due to a range of factors such as last minute changes in priorities of applicants or applicants receiving other offers. If we take a scenario with a limit of 300 applicants to be matriculated, TFA can consider giving offers to more than 300 applicants to ensure a more accurate number of people accepting their offers.

4. Cumulative GPA Distribution of Matriculated Applicants



From the cumulative GPA distribution, we can find that a heavy number of matriculated applicants have between a 3.5 and 3.9 GPA. These are typically those that would be on the Dean's List/Honor Roll. Based upon the GPA distribution data, TFA tends to bring on applicants with a higher GPA (3.5+). We can also conclude that applicants with higher GPAs tend to accept TFA's offer more than those with a GPA less than 3.5.

5. Matriculation Rates by Regions



The bar graph above plots the matriculation rates by regions and allows us to compare which regions have the highest probabilities of accepting their offers from TFA. We can observe that certain regions have higher matriculation rates, with the top 5 regions being Hawaii, Southbay, Louisiana, Camden and Newark. At the same time, certain regions have relatively lower matriculation rates such as the Bay Area which has only a 57.89%. A reason behind this could be due to the Bay Area being a huge business hub and applicants may be swayed by offers from more lucrative

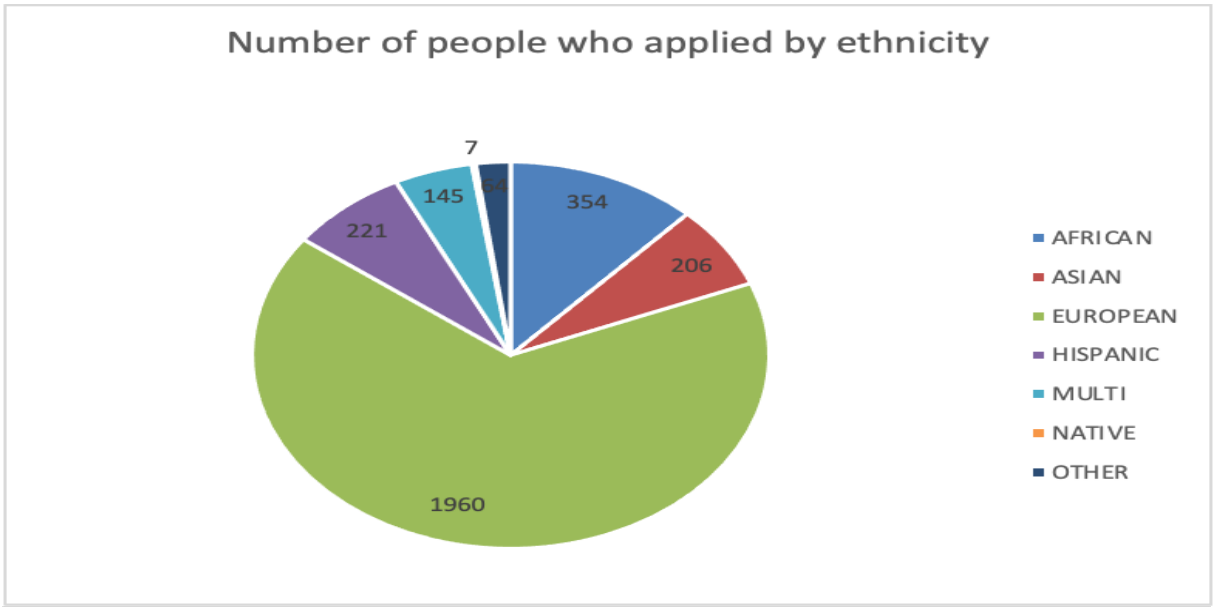
organizations. In response to this, TFA could either focus on hiring more from the top 5 regions or increase branding efforts in less popular areas like Las Vegas and the Bay Area.

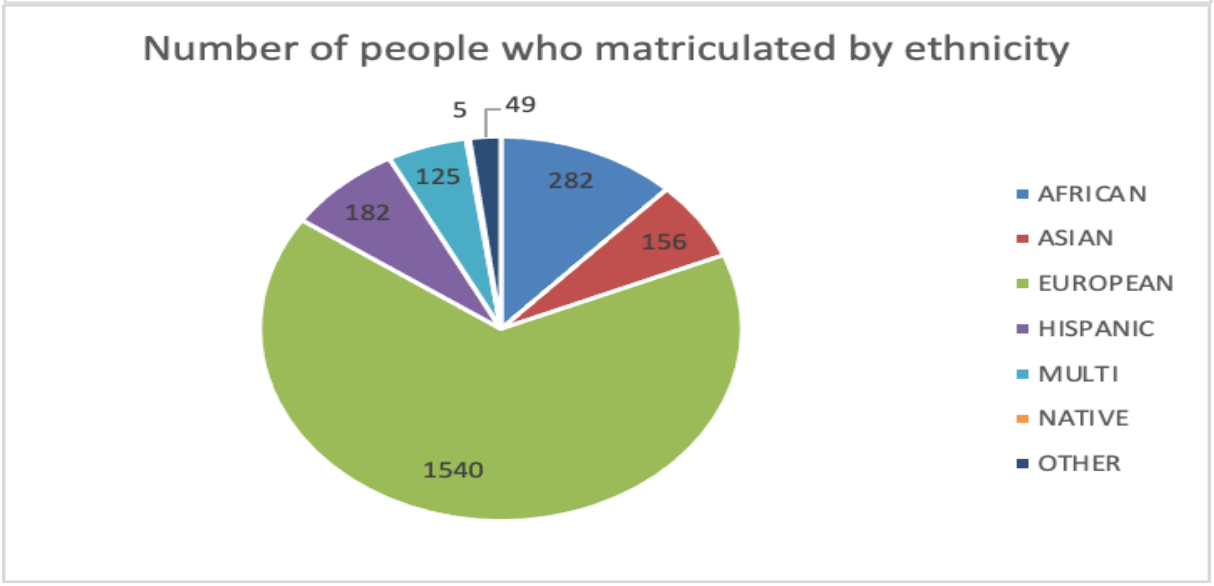
6. Diversity in hiring

	Number of people matriculated		
Gender	No	Yes	Grand Total
F	533	2164	2697
M	229	789	1018
(blank)	9	23	32
Grand Total	771	2976	3747

Gender

When looking at the gender ratio among applicants and people matriculated, it can be seen that for both cases, there are almost 3 times as many women as men. This indicates a large gender imbalance. In order to have a more appropriate gender balance, TFA could try to encourage more men to apply.





Race

There is a staggering lack of racial diversity among the talent pool encountered by the TFS. Majority of the applicants and people who matriculated were Europeans, forming approximately 65% of the total pool. The lowest percentages came from Native Americans and Europeans at 0.2% and 7% respectively. It is especially important to have a racially diverse workforce as it directly affects the performance of students in the long run. When minority students have the same race teachers, they are proven to perform better on standardized tests, have improved attendance and have better discipline (https://www.brookings.edu/research/the-importance-of-a-diverse-teaching-force/)

Recommendations

Since TFA typically has limited new jobs available in each recruiting cycle, they need to ensure the applicants who eventually join the corps won't exceed their estimated position offering too much. One strategy to fulfill this goal is sending offers to candidates who are predicted to have higher matriculation rate, thus there won't be large variation between the number of applicants who receive the offer and applicants who finally take positions in TFA.

We may consider the variables we evaluated above in recent graduates' matriculation decisions. Because RegionPrefLevel and CulmulativeGPA turn out to be the two most significant variables correlated with matriculation, TFA should count these two factors into important consideration when deciding whether to give an offer or not. There are 3 rating levels in RegionPrefLevel, which are 1,2,3, while the mean of RegionPrefLevel among candidates who matriculated is 1.058 and median is 1. This indicates that the majority of candidates who matriculated answered "1" in RegionPrefLevel question, and TFA might want to send more offers to these people. Take a closer look at the region, the median of RequestedRegionalReassignment is 0, which means most candidates would like to work at their current city instead of being reassigned to another working place. Thus TFA might want to send offers to candidates who live in the same region as where their recruiting positions are to ensure a high matriculation rate. For example, TFA might want to recruit candidates who live in NYC for their NYC opening positions. But when hiring candidates in other cities are needed, probably caused by huge demand in a specific city, TFA could consider hiring candidates from regions that have high matriculation rate. As shown in our key findings, the top 5 regions being Hawaii, Southbay, Louisiana, Camden and Newark. Candidates who matriculated in these regions are higher than 84%, so hiring candidates from these regions could also ensure a high matriculation rate. Another important indicator is GPA. Our key finding shows that a heavy number of matriculated applicants have GPA between a 3.5 and 3.9. Thus, TFA might also want to send offers to applicants with higher GPAs since they tend to accept TFA's offer more than those with a GPA less than 3.5.

After TFA understands who they should send offers to while they have limited capacity, which is their priority concern, they should also take diversity of candidates and business opportunity into consideration when recruiting. Since 72.72% of their applicants are female, and 65.4% are European, TFA should encourage more male applicants and applicants in diverse ethnicities to join their corps to enrich the diversity of their teachers. To realize this, TFA can partner with local universities to encourage students to join the teaching corp and continually implement the educational justice training program that draws on scholarship by American scholars. Besides, candidates live in regions that have low matriculation rate doesn't mean they have lower value of hiring. For example the Bay Area only has a 57.89% matriculation rate, as previously stated. But teachers are still in great demand throughout the Bay Area, there is a high likelihood of success. There are 19 classes in the district without a permanent instructor. What businesses can do is launch targeted marketing efforts in such locations and collaborate with local governments on initiatives such as providing affordable housing for teachers. Teachers can be supported by a voter-approved bond and a loan secured by below-market-rate rent by collaborating with local governments. Furthermore, TFA may collaborate with local institutions to hold additional recruitment information sessions on campus, encouraging people in these areas to accept their offers. This leads to a future project into real implementation plans.

https://www.ktvu.com/news/a-small-daly-city-school-district-provides-affordable-rent
https://www.motherjones.com/politics/2016/02/teach-america-most-divisive-education-reform-group/

Possible Case Question

The primary object in our case is to assist TFA in deciding how many offers to give to applicants in order to restrict the recruitment space. Let's put this strategy to use in a setting situation. Suppose TFA is contemplating hiring 300 additional instructors around the country for the upcoming 2023 Spring recruitment. According to the Limited Supply Rule. TFA should send offers to how many prospects considering their limited spots?

A	B	C	D	E	F
n	ID	BinaryLogitProbability	SUM_p(y=1)	Matriculated	lift
354	1687973	0.811402171	299.8296827	0	0.977295
355	1694109	0.811241419	300.6409241	1	0.977102

The method we use is choosing the sum of highest predicted response rate, when the SUM_p(y=1) and k is just under 300, n is 354 and we find that SUM_p(y=1) is 299.83 , when n is 355 and the SUM_p(y=1) is 300.64. So the TFA should send offers to 354 highest prospects.

Appendix

Percentage of matriculation by Ethnicity

Ethnicity	Percentage
EUROPEAN	65.4%
AFRICAN	12.7%
HISPANIC	7.8%
ASIAN	6.3%
MULTI	5.5%
OTHER	2.1%
NATIVE	0.2%