A Deep-Dive of UC Berkeley Enrollment Data

Suraj Rampure
suraj.rampure@berkeley.edu

Sukrit Arora
sukrit.arora@berkeley.edu

Rafael Calleja
rafael.calleja@berkeley.edu

Abstract

In this paper, we present an analysis of University of California, Berkeley enrollment data from several perspectives. First, we build binary classifiers to predict the location of a school (in-state or out-of-state) given multiple gender and ethnic features. In particular, we explore the performance of neural network-based models as they compare to more traditional models for this task.

We then build and interpret regression models that use all but one of the aforementioned gender and ethnic features to predict the held-out feature, and again contrast traditional and neural network approaches.

Finally, we interpret enrollment data from the lens of established fairness criterion, and make conclusions about the separation and sufficiency of various attributes.

Code: `https://github.com/surajrampure/281a-final-project`

1 Data Preparation

1.1 Sources

As former UC Berkeley undergraduates, we were interested in analyzing admissions statistics at the school level; specifically, we wanted to look at the relationships between gender, ethnicity, out-of-state status, and acceptance rates. We sourced data from the University of California's Infocenter, under "Admissions by source school" [1]. There, we generated two datasets by using the following sets of filters:

- `gender`: Tab "FR Gnd by Yr". All high schools, Fall 2019 admissions term, Berkeley campus.
- `ethnicity`: Tab "FR Eth by Yr". All high schools, Fall 2019 admissions term, Berkeley campus.

Below, we show sample outputs from the `gender` dataset. The structure of the `ethnicity` dataset is very similar.

Preprint. Under review.

Name	Location	Count	Gender	Value
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	Adm	Male	3
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	App	Male	52
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	Enr	Female	6
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	Adm	Female	6
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	App	Female	57
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	Enr	All	7
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	Adm	All	9
ABRAHAM LINCOLN HIGH SCHOOL52910	San Francisco	App	All	109

All rows for one particular high school in `gender`.

Name	Location	Count	Gender	Value
SHANGHAI SMIC PRIVATE SCHOOL694240	NaN	App	Male	18
HOLLYWOOD HIGH SCHOOL51615	Los Angeles	App	Female	29
INTERNATIONAL HIGH SCHOOL FAIS52943	San Francisco	App	Female	27
PIONEER HIGH SCHOOL230088	MI	Adm	Female	5
OAKRIDGE INT SCH NEWTON CAMPUS671040	NaN	Adm	All	3

A random sample of all rows of `gender`. Of note, in-state schools have the name of a California city as their `Location`, while out-of-state domestic schools have a two-letter state code and international schools have `NaN`.

1.2 Issues

Unfortunately, not all schools reported all relevant data. For instance, many schools were listed in the dataset as having applicants, but not students admitted. Many other schools only listed the total number of admitted students, without a gender/ethnicity breakdown. As such, we filtered to only include schools that reported the number of white applicants and admitted students, and the number of male applicants and admitted students.

1.3 Cleaned Results

The first few rows of our final cleaned and merged dataset is shown below. School names are truncated for brevity.

Name	AppMale	AdmMale	AppWhite	AdmWhite	Location
SAINT FRAN	0.496063	0.520000	0.188976	0.120000	INS
NORTH ALLE	0.518519	0.300000	0.259259	0.400000	OOS
DOS PUEBLO	0.461538	0.363636	0.472527	0.454545	INS
SIERRA CAN	0.394737	0.500000	0.342105	0.400000	INS
THOUSAND O	0.518519	0.500000	0.574074	0.500000	INS

Specifically:

- `AppMale` and `AdmMale` are the proportion of students who applied and were admitted that were male, respectively.
- `AppWhite` and `AdmWhite` are the proportion of students who applied and were admitted that were white, respectively.
- `Location` is whether or not the given school was in-state. Both domestic out-of-state and international schools were given the same value, `OOS`; `INS` means in-state.

2 Binary Classification of In-State Status

Our original goal was to model whether or not students would be accepted to Berkeley given certain features. However, the University of California's public dataset only provides data at the school level, not the applicant level – presumably for privacy reasons – so we had to broaden our scope.

2.1 Two-Feature Model

2.1.1 Non-Neural Models

We begin by using `AppMale` and `AdmMale` to predict whether a school is in-state. After performing an 80-20 train/test split, we fit four models and evaluated their accuracies, which are shown below:

Model	Training Accuracy	Testing Accuracy
Logistic Regression	0.804	0.800
Balanced Logistic Regression	0.617	0.662
Decision Tree	0.994	0.675
Random Forest	0.994	0.725

After fitting all models but the balanced logistic regression model, we were weary of the strong performance of the logistic regression model as compared to the two tree-based models. After investigating, we discovered significant class imbalance in our data: across both the training and testing sets, there were 318 observations that were in-state and 78 that were out-of-state. Our logistic regression model was skewed by this, as its predictions were all in-state:

	Predicted OOS	Predicted INS
Actual OOS	0	0.2
Actual INS	0	0.8

To combat this, we fit the balanced logistic regression model. The balanced logistic regression model has the same architecture, however the in-state and out-of-state classes in the training data are given different weights in the formulation for cross-entropy loss. Specifically, the weight for class i in balanced logistic regression is

$$w_i = \frac{n}{2 \cdot n_i}$$

where n is the number of training samples overall and n_i is the number of training samples belonging to class i. The expression for mean cross-entropy loss L_{CE} is then

$$L_{CE} = -\frac{1}{n} \sum_{i=1}^{n} \left[w_1 \cdot y \log \hat{y} + w_0 \cdot (1 - y) \log(1 - \hat{y}) \right]$$

For our training data, since $n_1 > n_0$, we have that $w_1 < w_0$; in fact, $w_0 = \frac{316}{2 \cdot 62} \approx 2.55$ and $w_1 = \frac{316}{2 \cdot 254} \approx 0.62$.

The confusion matrix for our balanced logistic regression model shows that it made varied predictions, unlike the unbalanced model:

	Predicted OOS	Predicted INS
Actual OOS	0.1375	0.0625
Actual INS	0.275	0.525

2.1.2 Neural Models

For the two feature input data, we first created a 2 Hidden Layer, Fully-Connected (FC) Neural Network (NN) with hidden dimensions of 50 and 25 respectively [2]. We then trained the neural network using stochastic gradient descent (SGD) for 200 epochs, evaluating the output of the network with our true labels using cross-entropy Loss. However, we didn't see any improvement over the logistic regression for this model.

In order to try to improve accuracy, we weighted the cross-entropy loss with the inverse class frequencies to try to compensate for the imbalance in the training dataset. However, even after accounting for the class imbalance, we still didn't see any improvement over logistic regression.

To see if we could push the accuracy using a NN model, we tried to push the capacity of the network, and trained a 4 Hidden Layer FC NN, with hidden dimensions of $(1000, 500, 500, 250)$ respectively. We trained this network in the same fashion as above. Even with this larger model, we saw the exact same accuracy as the logistic regression model.

After further investigation, it seems that in all models, the network learns to simply output a positive prediction regardless of the test data given. We hypothesize that the NN learns to do this because of the imbalance in the training dataset, small dataset size, and the lack of information available to it with just the two features it trained on.

2.2 Four-Feature Model

2.2.1 Non-Neural Models

Moving beyond just `gender` data, we decided to use our two ethnicity features (`AppWhite` and AdmWhite) in addition to our two gender features to classify schools as in-state.

Once again, we fit several different models on a subset of our data, and evaluated their accuracies on held-out training sets:

Model	Training Accuracy	Testing Accuracy
Logistic Regression	0.901	0.947
Balanced Logistic Regression	0.576	0.632
Decision Tree	1.0	0.947
Random Forest	1.0	0.947

There are some notable differences between the accuracy of our four-feature models and two-feature models. Chiefly among them being, all models but balanced logistic regression performed better on both the training and testing set when using four features compared to when using two features. More concretely, this means that the proportion of white students who applied and were admitted from a given school tell us something about whether or not that school is in-state.

Given the strong class imbalance, as mentioned above, the coefficients of the balanced logistic regression model are the most reliable. We present the coefficients from the balanced two-feature and four-feature logistic regression models.

Feature	Two-Feature Balanced	Four-Feature Balanced
`AppMale`	-2.331	-0.687
`AdmMale`	-0.493	-0.364
`AppWhite`	NA	-0.108
`AdmWhite`	NA	-0.388
Intercept	1.374	0.697

It appears that the weight of `AppMale` in the two-feature model was spread across all other features in the four-feature model, though of all features `AppMale` still had the greatest weight in the four-feature model.

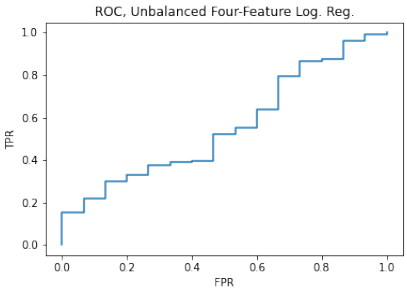
As a final metric, we show the ROC curves for both the unbalanced and balanced four-feature models (Figure 1). They are almost identical. Their AUCs are similar, at 0.558 and 0.563, respectively; unsurprisingly the AUC for the four-feature model is slightly higher.

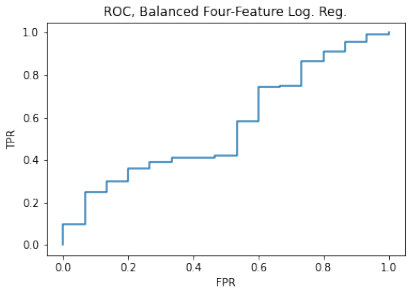
2.2.2 Neural Models

For the four feature input data, we used the largest network that we used on the two feature NN classifier: A 4 Hidden Layer FC NN, with hidden dimensions of $(1000, 500, 500, 250)$ respectively, trained using SGD for 200 epochs, using weighted cross-entropy Loss.

With this larger capacity network, we saw an improvement in both train and test accuracy, with a train accuracy of 98.01% and a test accuracy of 95.21%.

We hypothesize that this network was able to do perform better than both our two feature model and our non-neural, four feature models because the input data had more features to learn from, and our





NN was more expressive, with more learned weights to explain the relationship between our input and targets.

3 Regression of Admitted White Male Proportions

In addition to trying to classify using the UC admissions data, we also attempted to do some regression on the data. Specifically, we tried to predict `AdmMale` using the other 5 features using a variety of regression techniques.

First, we needed to modify our data in order to use the categorical feature (INS/OOS/INT) in our regressor. We used a one-hot encoding scheme in order to represent this feature. Because our models all have a learned bias term, we drop the INS column in order to avoid rank deficiency in our data matrix. Our transformed data looked like this:

Name	AppMale	AdmMale	AppWhite	INT	OOS	Target
ACALANES HIGH	0.450000	0.500000	0.675000	0	0	0.600000
ADOLFO CAMAR	0.553191	0.500000	0.191489	0	0	0.375000
ADRIAN C WIL	0.393939	0.285714	0.196970	0	0	0.285714
AGOURA HIGH	0.542857	0.545455	0.585714	0	0	0.454545
ALAMEDA SCI	0.620690	0.625000	0.206897	0	0	0.375000

We tried 7 different regression models:

1. Ordinary Least Squares (OLS)
2. Ridge Regression (L_2 Regularized Least Squares)
3. Ridge Regression with Hyperparameter Cross Validation
4. Lasso (L_1 Regularized Least Squares)
5. Lasso with Hyperparameter Cross Validation
6. KNeighborsRegressor, a non-parametric regression model
7. Fully-Connected NN Regressor (same architecture as our 4 Feature FC NN, but trained with an MSE loss)

We got the following MSE on our training and test sets, respectively:

Model	Training Mean Squared Error	Testing Mean Squared Error
OLS	0.015	0.023
Ridge	0.016	0.023
RidgeCV	0.016	0.023
Lasso	0.041	0.040
LassoCV	0.016	0.023
KNeighbors	0.012	0.027
FCNN	0.012	0.022

We see here that our FC NN performed best in terms of minimizing the Mean Squared Error loss in both train and test error. This makes sense, as the FCNN has many degrees of freedom that it can use to represent our data.

4 Fairness Criterion

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Separation

$$P(R = r|Y = y, A = a) \tag{1}$$

Sufficiency

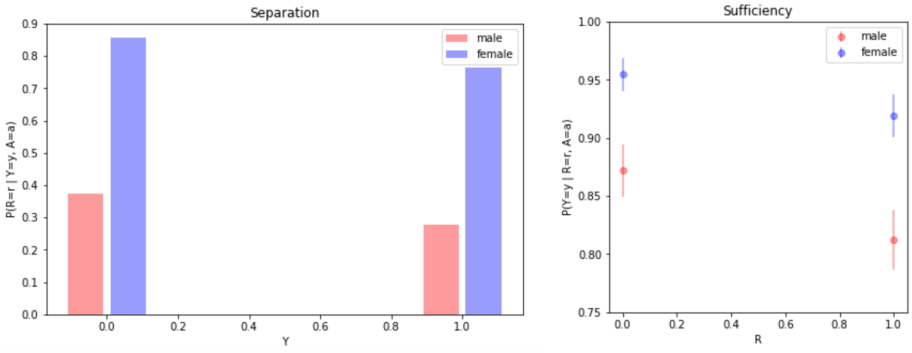
$$P(Y = y|R = r, A = a) \tag{2}$$

In our problem, Y is whether or not a school is in-state (ground truth), R is whether our model predicted a school is in-state (prediction), and A is whether or not a school is predominantly male or predominantly female (sensitive attribute).

We analyzed the predictions made by our 4-feature model to look at the fairness criterions described in "Fairness and machine learning" [3]. As described there are three points at which we would want to consider fairness: pre-processing, at training time, and post-processing. Due to the nature of the problem we are analyzing, it was difficult to examine fairness during the first two time frames and so we analyzed the fairness of our model in post-processing.

For both separation and sufficiency we looked at the rate or "probability" for both our model's prediction and the ground truth of whether or not a school was in-state or out-of-state.

4.1 Separation and Sufficiency



We calculated values according to the separation and sufficiency equations above and found that neither were satisfied by our model. With regards to separation, we can see that given the predominant gender of the school's application base, there is a big difference in whether our model predicts the school is in-state (1) vs. out-of-state(0) given the school's true locational status and predominant gender. With regards to sufficiency swe see something very similar in that the ground truth has a wide variation given our model's prediction and the school's predominant gender.

5 Conclusion

Despite our extensive analysis of the dataset, we were unable to infer any strong conclusions about the fairness of UC Berkeley admissions process. While the dataset itself is large, its lack of granularity means that even after significant pre-processing, meaningful information is still difficult, if not impossible, to extract.

Many of the schools we analyzed were missing data necessary for comprehensive analysis, and the inconsistency between data provided by different schools meant that comparing them meaningfully was challenging. If, instead, UC Berkeley released anonymized school district, gender, race, test score, and GPA data per student, omitting districts from which too few students of a certain race/gender applied to ensure anonymity, the data would be significantly more useful. As it stands, this dataset appears to be more of a gesture than a meaningful effort toward transparency in the admissions process.

References

[1] University of California. *Admissions by Source School*. URL: https://www.universityofcalifornia.edu/infocenter/admissions-source-school. (accessed: 12.05.2020).

[2] Jonathan Shewchuck. *Concise Machine Learning*. URL: https://people.eecs.berkeley.edu/~jrs/papers/machlearn.pdf. (accessed: 12.05.2020).

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2019.