Investigating the Distributions of SMOTE-Augmented Datasets

Statistics 98

Christian Wagner & Masaoud Haidar

Harvard University

March 23, 2022

Outline

What is an unbalanced dataset?

Harvard College Admissions, Class of 2025

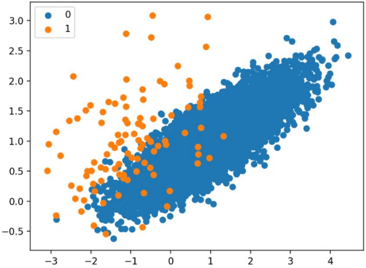
- 57,786 applicants
- 2,320 admitted
- Only 4.01% of our data represent admitted students

A General Framework

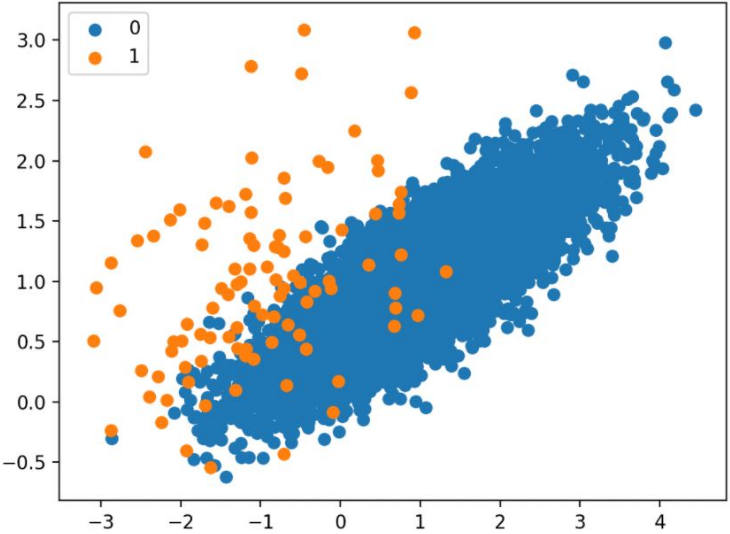
- Binary outcomes:
  - "positive" $Y = 1$
  - "negative" $Y = 0$

A General Framework

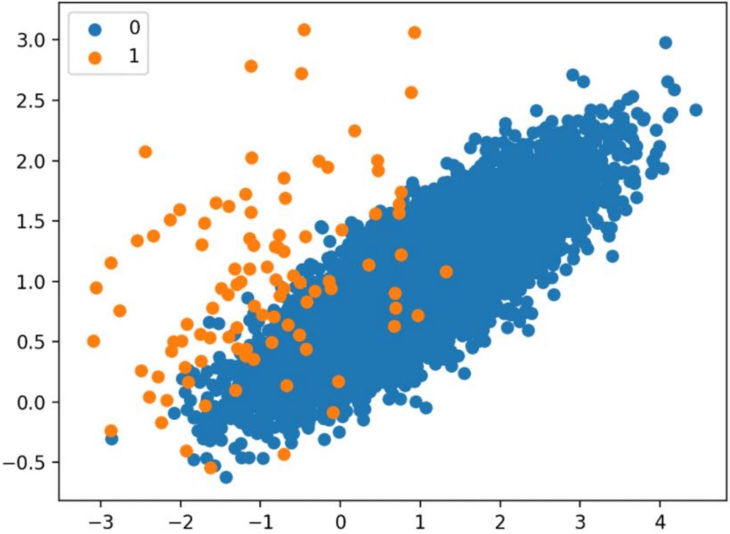
- Binary outcomes:
  - "positive" $Y = 1$
  - "negative" $Y = 0$
- One outcome is a "majority class" while the other is a "minority class"



Majority Under-sampling

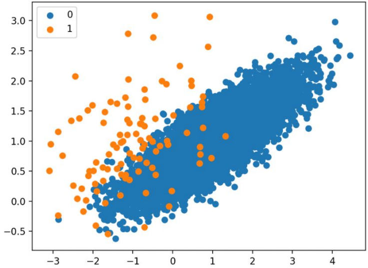


Minority Over-sampling



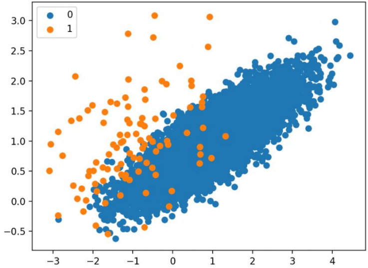
What is SMOTE?

- Synthetic Minority Over-sampling Technique



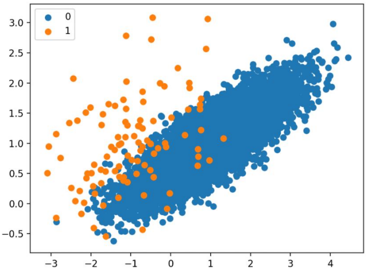
What is SMOTE?

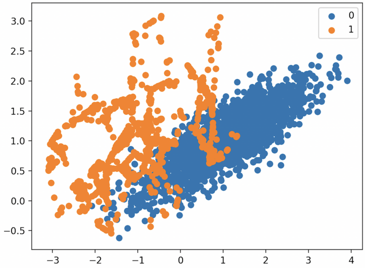
- Synthetic Minority Over-sampling Technique
- Algorithm:
  1. Select a point
  2. Randomly select from among its k nearest neighbors
  3. Draw a line segment between the two
  4. Add a randomly chosen point along this line segment
  5. Repeat until desired balance is achieved



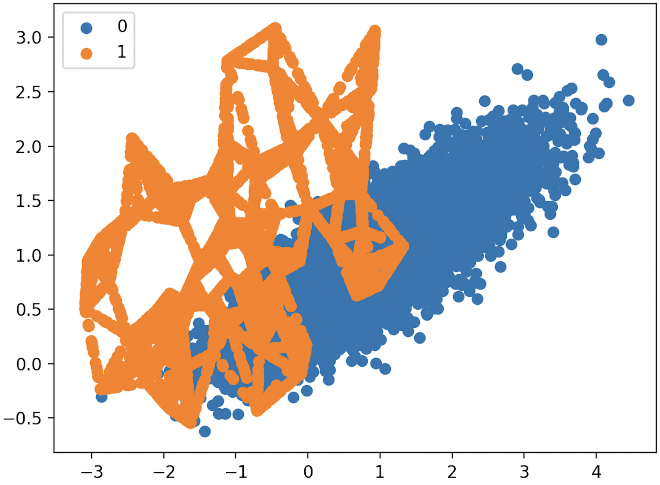
What is SMOTE?

- Synthetic Minority Over-sampling Technique
- Algorithm:
    1. Select a point
    2. Randomly select from among its k nearest neighbors
    3. Draw a line segment between the two
    4. Add a randomly chosen point along this line segment
    5. Repeat until desired balance is achieved





Too much SMOTE... IMPOSSIBLE!



Research Question

How does data augmentation with SMOTE affect the distribution of the predictors, and in turn affect the classification Model?

- Model:
$$\operatorname{logit}(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
Set:
$$P(Y=1) = 0.05$$
- Over-sample using SMOTE, then check how does this affects the distribution of X_1 and X_2 and the logistical model predicting Y.

Randomized Data Generation

- 10000 replicates of sample size $n = 1000$

Randomized Data Generation

- 10000 replicates of sample size $n = 1000$
- $X_{1,i} \sim \text{Expo}(1)$.

Randomized Data Generation

- 10000 replicates of sample size $n = 1000$
- $X_{1,i} \sim \text{Expo}(1)$.
- $X_{2,i} | X_{1,i} \sim \mathcal{N}(X_{1,i}, 1)$, which results in $\text{Corr}(X_1, X_2) = 1/\sqrt{2} \approx 0.707$

Randomized Data Generation

- 10000 replicates of sample size $n = 1000$
- $X_{1,i} \sim \text{Expo}(1)$.
- $X_{2,i} | X_{1,i} \sim \mathcal{N}(X_{1,i}, 1)$, which results in $\text{Corr}(X_1, X_2) = 1/\sqrt{2} \approx 0.707$
- $\beta_1 = \beta_2 = 1$

Randomized Data Generation

- 10000 replicates of sample size $n = 1000$
- $X_{1,i} \sim \text{Expo}(1)$.
- $X_{2,i}|X_{1,i} \sim \mathcal{N}(X_{1,i}, 1)$, which results in $\text{Corr}(X_1, X_2) = 1/\sqrt{2} \approx 0.707$
- $\beta_1 = \beta_2 = 1$
- $\beta_0 = \ln(0.05/0.95) - 4.15$. This results in $\mathbb{E}(P(Y = 1)) = 0.05$

Randomized Data Generation

- 10000 replicates of sample size $n = 1000$
- $X_{1,i} \sim \text{Expo}(1)$.
- $X_{2,i}|X_{1,i} \sim \mathcal{N}(X_{1,i}, 1)$, which results in $\text{Corr}(X_1, X_2) = 1/\sqrt{2} \approx 0.707$
- $\beta_1 = \beta_2 = 1$
- $\beta_0 = \ln(0.05/0.95) - 4.15$. This results in $\mathbb{E}(P(Y = 1)) = 0.05$
- $Y_i \sim \text{Bernolli}(P(Y_i = 1))$ where

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

SMOTE

- We split our data into train and test. The same test data will be used for all the following cases and we don't oversample from it.

SMOTE

- We split our data into train and test. The same test data will be used for all the following cases and we don't oversample from it.
- (X_1, X_2, Y) from the train data will make up our first dataset. This will be used as a control group and we call it our "level 0".

SMOTE

- We split our data into train and test. The same test data will be used for all the following cases and we don't oversample from it.
- (X_1, X_2, Y) from the train data will make up our first dataset. This will be used as a control group and we call it our "level 0".
- We oversample the training data using SMOTE to achieve a rate of $1:9$ of positive to negative data points. This will be our "level 0.1".

SMOTE

- We split our data into train and test. The same test data will be used for all the following cases and we don't oversample from it.
- (X_1, X_2, Y) from the train data will make up our first dataset. This will be used as a control group and we call it our "level 0".
- We oversample the training data using SMOTE to achieve a rate of $1 : 9$ of positive to negative data points. This will be our "level 0.1".
- We oversample the training data using SMOTE to achieve a rate of $1 : 1$ of positive to negative data points. This will be our "level 0.5".

Recording Results

For each of the three levels, we record the following:

- **Distribution of the Predictors:** Record the mean and variance of X_1 and X_2

Recording Results

For each of the three levels, we record the following:

- **Distribution of the Predictors:** Record the mean and variance of X_1 and X_2
- **Correlation between the Predictors**

Recording Results

For each of the three levels, we record the following:

- **Distribution of the Predictors:** Record the mean and variance of X_1 and X_2
- **Correlation between the Predictors**
- **Parameters of the Model:** We will record the estimated β_0, β_1, and β_2.

Recording Results

For each of the three levels, we record the following:

- **Distribution of the Predictors:** Record the mean and variance of X_1 and X_2
- **Correlation between the Predictors**
- **Parameters of the Model:** We will record the estimated β_0, β_1, and β_2.
- **The Model's perfromance:** We will evaluate the accompanying logistic regression model on the test data by calculating the F1 score.

Simulation Results: Distribution of the Predictors

- $X_1 = \text{Expo}(1)$

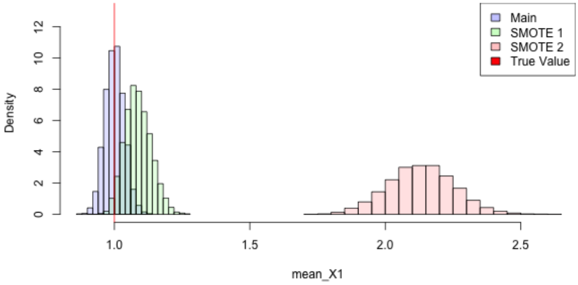


Figure: Distribution of the Mean of X_1

Simulation Results: Distribution of the Predictors

- $X_1 = \text{Expo}(1)$

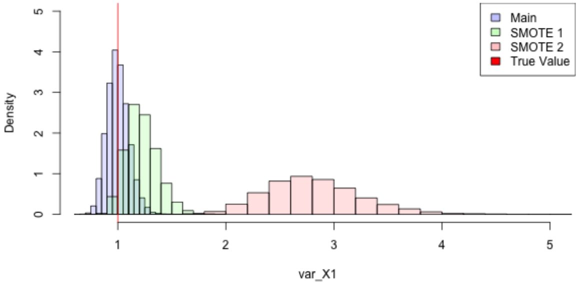


Figure: Distribution of the Variance of X_1

Simulation Results: Distribution of the Predictors

- $X_2 = \mathcal{N}(X_1, 1)$

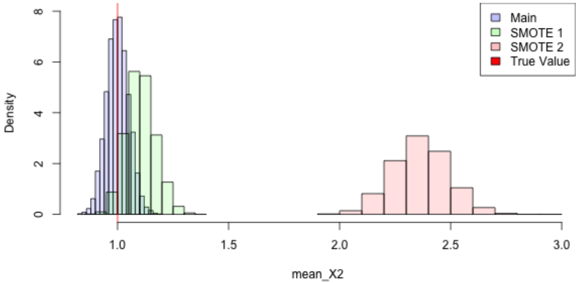


Figure: Distribution of the Mean of X_2

Simulation Results: Distribution of the Predictors

- $X_2 = \mathcal{N}(X_1, 1)$

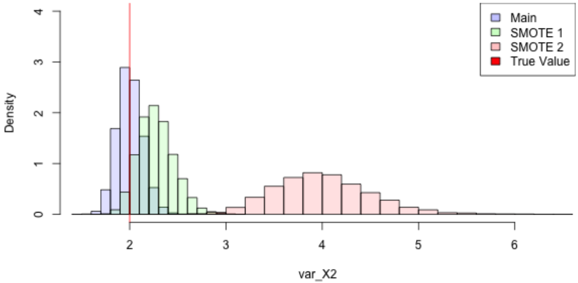


Figure: Distribution of the Variance of X_2

Simulation Results: Correlation between the Predictors

- Theoretical Correlation:

$$\mathbb{E}(X_1X_2) = \mathbb{E}(\mathbb{E}(X_1X_2|X_1)) = \mathbb{E}(X_1^2)$$

$$\text{Corr}(X_1, X_2) = \frac{\mathbb{E}(X_1X_2) - \mathbb{E}(X_1)E(X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{1}{\sqrt{2}} \approx 0.707$$

- Correlation in the Simulation:

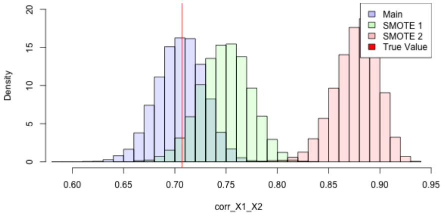


Figure: Distribution of the Correlation of X_1 and X_2

Simulation Results: Parameters of the Model

- $\beta_0 \approx -7$
- PRB (Percentage Relative Bias) for the three levels: -3.05%, 1.85%, 29.11%.

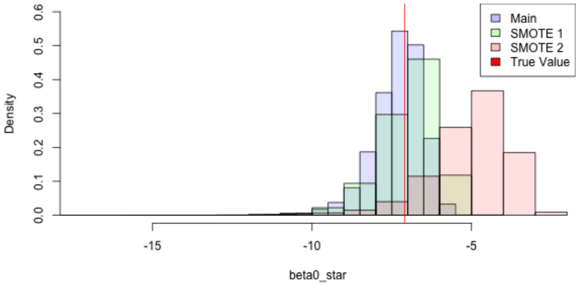


Figure: Distribution of the estimate of β_0

Simulation Results: Parameters of the Model

- Theoretical Coefficient: $\beta_1 = 1$
- Simulation:
  $\hat{\beta}_{1,0} = 1.03$, $\hat{\beta}_{1,0.1} = 1.06$, $\hat{\beta}_{1,0.5} = 1.15$.
  PRB: 3.16%, 6.49%, 15.43%.
  ANOVA test gives $F_{2,29997} = 291.2$ with $p < 2e - 16$.

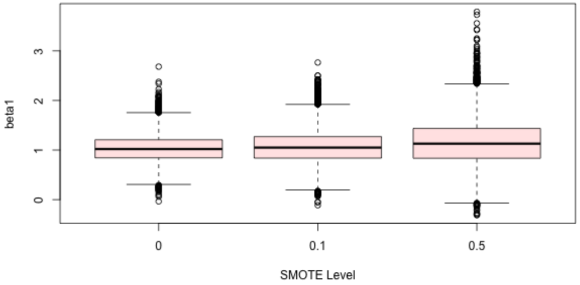


Figure: Distribution of the estimate of β

Simulation Results: Parameters of the Model

- Theoretical Coefficient: $\beta_2 = 1$
- Simulation:
  $\hat{\beta}_{2,0} = 1.03$, $\hat{\beta}_{2,0.1} = 1.08$, $\hat{\beta}_{2,0.5} = 1.18$.
  PRB: 3.41%, 8.13%, 18.35%.
  ANOVA test gives $F_{2,29997} = 463.2$ with $p < 2e - 16$.

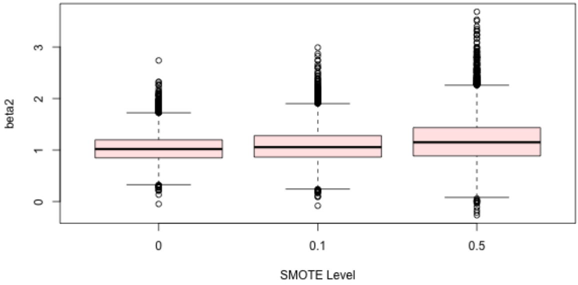


Figure: Distribution of the estimate of β

Simulation Results: The Model's Predictions

- $F1_0 = 0.603$, $F1_{0.1} = 0.629$, $F1_{0.5} = 0.460$.
  A 2-sample t-test comparing the original dataset to the 0.1 level procedures using Fisher's strict null gives $t_{19757} = 12.958$ with $p < 2e - 16$.

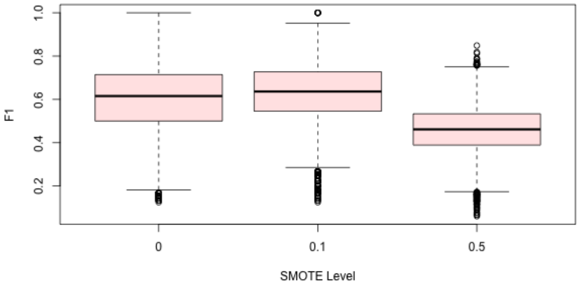


Figure: Distribution of the F1 score

Conclusion & Next Steps

- SMOTE creates a bias in the predictors and the estimated coefficients. This means we can't directly interpret the estimated coefficients.

Conclusion & Next Steps

- SMOTE creates a bias in the predictors and the estimated coefficients. This means we can't directly interpret the estimated coefficients.
- SMOTE is useful for classification up to a certain extent

Conclusion & Next Steps

- SMOTE creates a bias in the predictors and the estimated coefficients. This means we can't directly interpret the estimated coefficients.
- SMOTE is useful for classification up to a certain extent
- Possible Next Steps: How much SMOTE is too much SMOTE? Can we tune our balancing levels to optimally improve predictions?