Stat 98 Simulation Proposal

Christian Wagner, Masaoud Haidar

March 2022

1 Statement of Research Question

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

2 Background and Brief Literature Review

When training predictive models, one common problem is that the training data is **unbalanced**. We will use the word unbalanced to refer to situations in which we want to predict a binary response variable with our model, which appears predominantly in the training data as only one of the two possible outcomes. There is no shortage of scenarios in which unbalanced data sets naturally arise. Here are two examples, to name a few:

- Predicting a candidate's acceptance into an elite institution that only has a very small rate of acceptance
- Predicting whether a presidential candidate will win the general election

These are all cases in which we have very imbalanced data. In the first example, we know that the vast majority of our data points will represent rejected applicants, also known as the **majority class**. The minority of our data points will represent accepted applicants, which would be known as the **minority class**. In the second, only one candidate can win, while all the rest will fail.

Notice: in both of these scenarios, we would be more interested in predicting whether or not a unit belongs to the minority class than the majority class, but we are attempting to do so with almost all of our available data in the wrong category. This poses issues to current predictive modeling procedures (Blagus & Lusa).

A few different solutions to this issue have been proposed:

1. **Undersampling the majority class**: If we have sufficient data, we might decide to resolve the imbalance in our data by using only a subset

of the majority class in our training data. However, we sacrifice precision in our predictions in the name of reducing the imbalance, a manifestation of the bias-variance trade-off.

2. **Oversampling the minority class with replacement**: If we do not wish to undersample from the majority class, we can synthetically balance our data set by over-sampling from the minority class with replacement. This method is similar to bootstrapping: a random subset of the minority class is duplicated in order to even the scales. However, this method can lead to overfitting.

3. **A hybrid approach**: The previous two methods have been employed simultaneously, in an effort to counter-balance the effect of each and to reduce the necessary level of under- and over- sampling of the majority and minority classes, meeting somewhere in between (Chawla *et. al*).

What is SMOTE?

SMOTE stands for “Synthetic Minority Over-sampling Technique.” The idea here is to avoid the imprecision associated with under-sampling the majority class, and also to avoid overfitting to the same extent as oversampling the minority class with replacement. Instead, we generate a set of synthetic minority class data based on our recorded observations, and add it in with the rest of the training data. The SMOTE algorithm operates by selecting a unit in the minority class and randomly sampling from among its k nearest neighbors and drawing line segments from the selected unit to each of the sampled points, with k specified by the user. The size of each sample is determined by how many observations we need to add to the minority class. A point is then selected at random on each line segment and added to this data. The process is then repeated for each unit in the minority class (Blagus & Lusa).

Clearly, by only choosing points between observations, we are affecting the distribution of our minority class. Additionally, this can intuitively be seen to bias our resulting data set towards the small number of observations, though to a lesser degree than oversampling with replacement. SMOTE has been used extensively in the past decade, although theoretical knowledge on its predictive effect has not been extensively explored.

3 Research Design

The project will simulate data from a logistic regression model with two predictors and a response variable with two categories (0 or 1). So, we will have the predictors X_1 and X_2 and the response Y where:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Such that $Y \sim \text{Bernoulli}(P(Y = 1))$. Let $Y = 1$ denote a "positive" response and $Y = 0$ denote a "negative" response. For the purposes of this simulation, observations with positive responses will represent the **minority class**.

To generate the data, for each data point, we generate the random variables X_1 and X_2, we calculate the probability $P(Y = 1)$, and then we sample a Bernoulli with that probability.

The parameters of the general model are:

- The sample size of each replicate is $n = 1000$.
- $X_{1,i} \sim \text{Expo}(1)$.
- $X_{2,i}|X_{1,i} \sim \mathcal{N}(X_{1,i}, 1)$, such that $\text{Corr}(X_1, X_2) = 1/\sqrt{2} \approx 0.707$.
- $\beta_1 = \beta_2 = 1$.
- $\beta_0 = \ln(0.05/0.95) - 4.15$. This results in an average of $P(Y = 1) = 0.05$.
- $Y_i \sim \text{Bernolli}(P(Y_i = 1))$ where $\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$

Then, we will split the data into 80% train data and 20% test data. Then, we will do the following two processes:

- **Original Data:** We will fit a logistic regression model on the train data and evaluate it on the test data.
- **Model with Partially Augmented Data:** We will use SMOTE to increase the number of positive data points in the train data to make the ratio of the two classes 90 : 10 (compared to the original imbalance of 95 : 5). Then, we will fit a logistic regression model on the resulting dataset and evaluate it on the test data.
- **Model with Fully Augmented Data:** We will use SMOTE to make the dataset balanced (a ratio of 1 : 1). Then again, we will fit a logistic regression model on the resulting dataset and evaluate it on the test data.

Thus, we have the true distribution and the three resulting datasets with their models to compare. For each of the three empirical datasets and their models, we will record the following:

- **Distribution of the Predictors:** We will compare the distributions of the predictors in the empirical dataset to the original true distributions. We will do this by comparing between the mean and variance of these predictors.
- **Correlation between the Predictors:** We will calculate the correlation between the two predictors in the dataset.

- **The accuracy of the Parameters of the Model:** We will record the estimated β_0, β_1, and β_2.
- **The accuracy of the Model's prediction:** We will evaluate the accompanying logistic regression model on the test data by calculating the accuracy and the F1 score.

The original data and its model serve as a control group for us, as they tell us, for example, what to expect the statistical distance to be between a generated dataset and the true distribution. The different replicates give us a full distribution of this distance. Then, the two augmented datasets tell us how does data augmentation affect the metrics we are interested in (such as the statistical distance). This allows us to figure whether the differences we find in the augmented data compared to the original data are significant or not.

As for the other design choices, we make X_1 and X_2 correlated so we can see whether that correlation remains in the augmented data. We defined β_0 such that it creates an unbalanced data set. And finally, we try both a partial data augmentation and a full data augmentation to see how the increase of the augmentation size affects the metrics we're interested in.

4 Analysis of Results

We have four main results to analyse:

- **Distribution of the Predictors:** We hypothesis that the augmented data will have the same mean as the original data but smaller variance.
- **Correlation between the Predictors:** We expect the data augmentation to preserve the correlation between the predictors.
- **The accuracy of the Parameters of the Model:** The model based on the augmented data will predict more positives than the model on the original data. As a result, we expect it to overestimate the coefficients $(\beta_0, \beta_1, \beta_2)$.
- **The accuracy of the Model's prediction:** Data augmentation introduces some bias to the model. It aims to make less false positives on the expense of making more false negatives. However, the evaluation is done on the imbalanced test data, as the evaluation must be done on "real" data, not augmented data. So, we expect the models based on augmented data to have lower accuracy but a higher F1 score than the model on the original data.

5 Sources

- Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., Kegelmeyer, Philip W. "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, no. 16 (2002), pp. 321-357.

- Blagus, Rok, and Lusa, Lara. "SMOTE for High-Dimensional Class-Imbalanced Data," *BMC Bioinformatics*, no. 14 (2013).