Striking a Balance

Performance vs. Fairness Considerations when Rebalancing Data

Nathan Einstein and Anthony Rentsch

Harvard John A. Paulson
School of Engineering
and Applied Sciences

Background

Resampling techniques are one frequently-adopted approach for improving model performance when there is a strong class imbalance in the available data. But what are the implications of purposefully altering the distribution of the data on the fairness of the model's predictions?

We empirically assessed the impact of various sampling techniques for correcting for class imbalance on the fairness (defined in several ways) of predictions generated by several types of models.

This research question arose while examining an open-source predictive policing algorithm based on a patented combined over-/under-sampling technique.[1] We therefore chose to draw on the same type of data this platform was designed to ingest: open-access crime records from the Boston Police Department.[2]

Approach

We decomposed the effect of resampling on algorithmic fairness into two components:

1. the systematic differences in sampling distributions that result from different rebalancing techniques, and
2. how those differences in distribution would impact the predictions of a particular algorithm (requires testing how different model classes respond to the same change in distribution)

Procedure:

1. **Data preprocessing**: Prepared (i) a feature set including the count of vandalism, drug-crime, and aggravated assault cases from 07/2015-04/2019 per census tract; (ii) the target variable: presence of a murder/non-negligent manslaughter by census tract, per month; and (iii) the implicit sensitive attribute: % non-hispanic white by tract (not included in model)[3].
2. **Model-fitting**: Reserving the last year of data as a test set, we fit four classes of commonly-used classifiers (SVM, kNN, Random Forest, MLP) to the training set after rebalancing the data using one of the following techniques:
   - Oversampling: random oversampling, SMOTE, ADASYN, SMOTE-EEN (mixed over/under)
   - Undersampling: random undersampling, cluster centroids, Tomek Links
3. **Performance assessment**: We tested the performance of the fitted models on the held-out test set.
4. **Fairness assessment**: We considered *where* the fitted models made errors, in particular with respect to the sensitive attribute. To simplify the analysis, we labeled tracts as "high-" or "low-minority" by thresholding their % white share at the city average (47% white).

Performance and Fairness Assessment

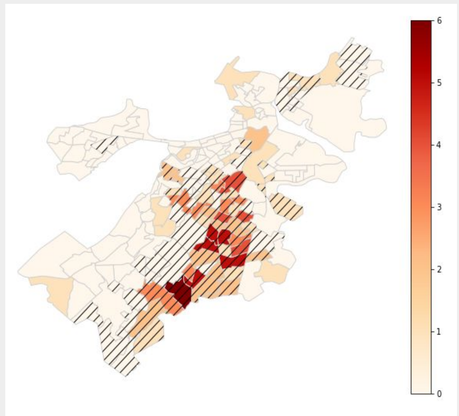

Figure 1. Co-occurrence of minority-heavy census tracts (cross-hatched) and murders. Shading indicates number of months with a murder.

Predictors and sensitive attribute are closely correlated:

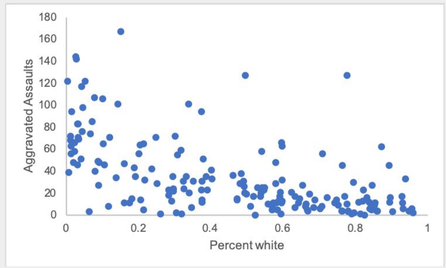

Figure 2. Relationship between tract-level racial demographics and number of reported aggravated assaults.

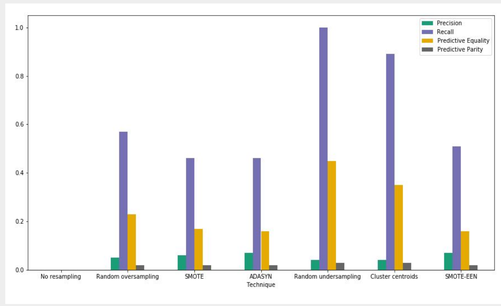

Figure 3. Comparison of several model performance and fairness metrics across different resampling techniques.

Change in predictive fairness:

- Oversampling techniques, especially SMOTE and ADASYN, in general produced fairer results with respect to both predictive parity and equality than the undersampling techniques.
- Undersampling tended to cause excessive false positives, and therefore produced results that fared markedly worse with respect to predictive equality.

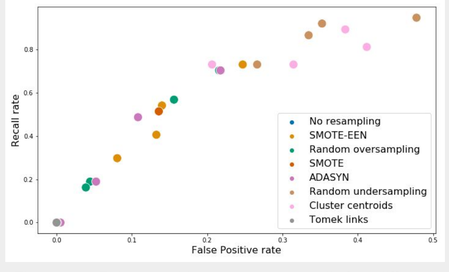

Figure 4. False positive rate vs. recall for different resampling techniques, all classifiers.

Broad takeaway:

For hard-to-predict, rare-occurring events, it is nearly impossible to improve a model's recall without also increasing false positives.

Especially when there is sparsity (little variation) in the covariates used to predict the outcome, resampling techniques cannot prevent the inevitable fact that the model will systematically commit false positives with observations that are similar to the positive examples it has been trained on.

Distributional Effects

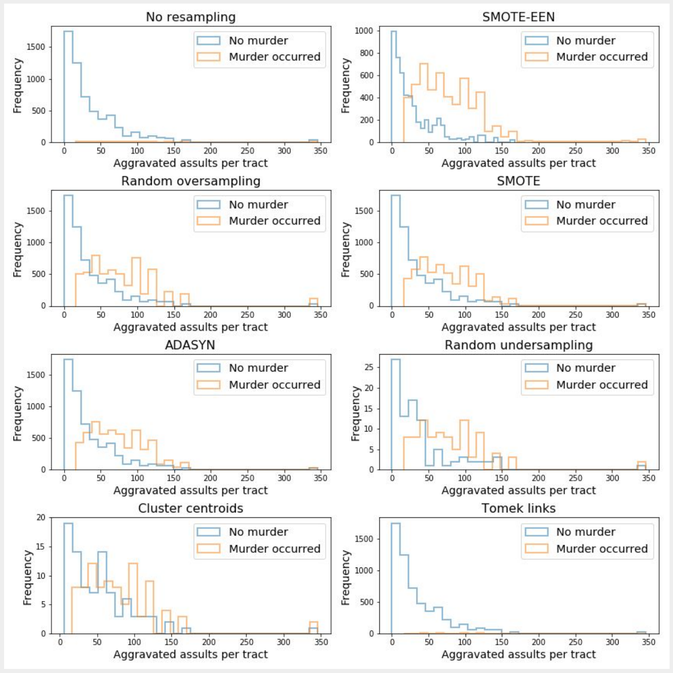

Figure 5. Distribution of aggravated assaults per tract in training data before and after applying various resampling techniques.

- SMOTE and ADASYN increased the spread of the minority-class distribution (tracts with murders) with respect to the proxy feature (aggravated assaults).
- The "cluster centroids" undersampling technique similarly increased the spread of the majority (no-murder) distribution, while random undersampling merely made that group's distribution less even.

References

[1] CivicScape, "A New Standard for Real-Time Policing." www.civicscape.com.
[2] Boston Police Dept. *Crime Incident Reports (August 2015 - To Date)*. Accessed at data.boston.gov.
[3] U.S. Census Bureau, 2013-2017 American Community Survey 5-Year Estimates.