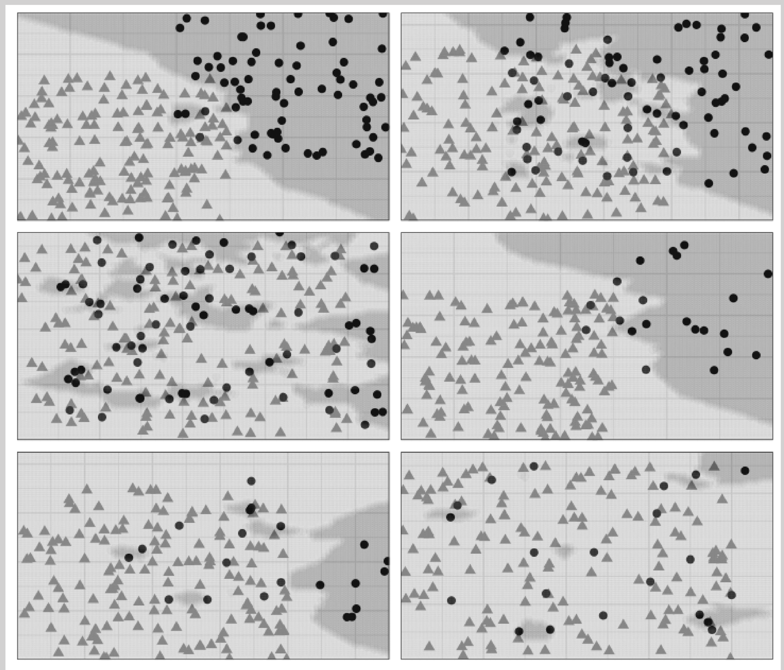
Hyperparameter-Tuned Oversampling of Imbalanced Data with Overlapping Features

Njesa Totty [1] Claudio Fuentes [2] James Molyneux [2]

[1]Framingham State University [2]Oregon State University

Imbalance and Overlap

- *Imbalanced data* - a dataset with an unequal distribution of classes (e.g. Li et al. 2021)
- *Class overlap* - when regions of the feature space contain both classes (e.g. Denil and Trappenberg 2010)
- Ohter Examples: spam filters, credit card fraud, object detection
- Overlap inhibits classifier performance more than imbalance (e.g. Vuttipittayamongkol et al. 2021).



Literature on Imbalance and Overlap

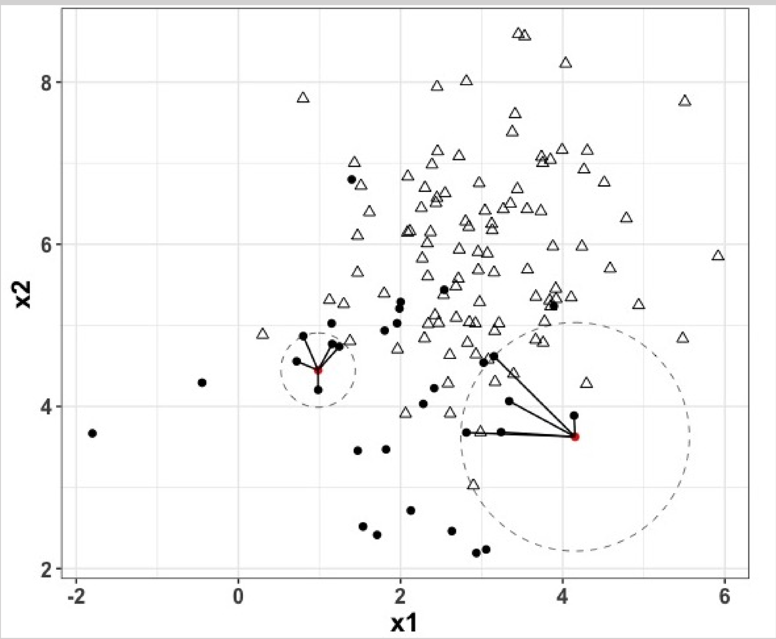
- Research about imbalanced learning is ever growing (Krawczyk 2016; Fernández et al. 2018)
- Many solutions pertaining to imbalance and overlap
  - (e.g. Chawla et al. 2002; Xiong et al. 2010; Denil and Trappenberg 2010; Oh 2011; Borsos et al. 2018; Li et al. 2021)
- Algorithm-level methods adapt how a classifier learns
  - Examples: cost-sensitive learning (e.g. C. Zhang et al. 2018), and threshold selection (e.g. Johnson and Khoshgoftaar 2019), weighted loss functions (e.g. Shahee and Ananthakumar 2021)
- Data-level methods modify the data set that the classifier trains on
  - Examples: feature selection (Borsos et al. 2018; Omar et al. 2021), random undersampling (RUS) (Koziarski 2020; Hoyos-Osorio et al. 2021), random oversampling (ROS) (Fajardo et al. 2021; Liu et al. 2023)

The Synthetic Minority Oversampling TEchnique (SMOTE) by Chawla et al. (2002)

- $z = \mathbf{X}_0 + w(\mathbf{X} - \mathbf{X}_0)$, $w \sim$ Uniform$(0, 1)$
- $\mathbf{X}$ = nearest minority neighbor; $\mathbf{X}_0$ = point of interest
- Most popular and influential solution to date (Garcia et al. 2016)

Issues to consider:

- Non-numeric data (Mukherjee and Khushi 2021)
- Introducing overlap (R. Zhang et al. 2023)
- Within-class imbalance (Douzas et al. 2018)
- Compact synthetic data (Elreedy and Atiya 2019)



SMOTE-Based Methods

There are at least 85 variants of SMOTE (e.g. Kovács 2019) including:

- Non-cluster based approaches: Borderline-SMOTE (Han et al. 2005), Adaptive Synthetic Sampling Approach (ADASYN) (He et al. 2008)
- Cluster-based approaches: Cluster-SMOTE (Cieslak et al. 2006), k-means SMOTE (Douzas et al. 2018)
- Adaptations to w's distribution: Gaussian (Lee et al. 2017), Gamma (Kamalov and Denisov 2020)
- Handling nominal data: SMOTE-NC proposed by Chawla et al. (2002) for mixed-type data

Research Gap

- There are very few synthetic oversampling techniques for mixed data (e.g. Limanto et al. 2024; Fonseca and Bacao 2023; Mukherjee and Khushi 2021) and none address the issues of generating more overlap, within-class imbalance, or compact examples simultaneously.

The Strategic SMOTE (S-SMOTE)

The following factors may be selected in S-SMOTE:

- $\eta = (\eta_1, \dots, \eta_\ell)$ = vector of dominance thresholds to approve points for oversampling, decreasing by η_{inc}
- $\rho = (\rho_1, \rho_2, \rho_3, \rho_4)$ = weights used to select approved points for oversampling
- F = distribution of w for interpolation/extrapolation
- k_{max} = maximum number of neighbors to use for any given point
- p_{min} = minimum number of minority points to use for oversampling

Additionally,

- Gower's distance is used to include categorical variables in distance calculations.
- Majority vote is used to select levels for categorical variables of synthetic examples.

SMOTE is a special case of S-SMOTE with $\eta = 0$, $\rho = 1/n_{min}$, $k_{max} = k$, $p_{min} = 1$, F = Uniform$(0, 1)$.
The R implementation of S-SMOTE allows the user to tune these hyperparameters.

Algorithm

Set $\eta_{now} = \eta_1$, $E^* = \{\emptyset\}$, $E = \{1, \dots, n_{min}\}$, n_{min} = number of minority points.

While $\frac{|E^*|}{|E|} < p_{min}$ and $\eta_{now} \geq \eta_\ell$:

For each minority point $i \notin E^*$,

1. Set k_i = furthest minority neighbor, out of k_{max}, that is as close as at least η_{now}*100% minority points, if it exists. If such a k_i does not exist exit the loop.
2. Set p_i = the proportion of minority points as far as k_i.
3. Set $E^* = E^* \cup \{i\}$.

Set $\eta_{now} = \eta_{now} - \eta_{inc}$.

Return:

- E^* = the set of minority points approved for oversampling
- $\mathbf{p}$ = a vector of dominance proportions for each point in E^*
- $\mathbf{k}$ = a vector giving the number of nearest neighbors to use for oversampling each point in E^*

Then:

- Assign points in E^* to sets Q_1, Q_2, Q_3, Q_4 using med($\mathbf{p}$/max($\mathbf{p}$)) and med($\mathbf{k}$) as cutoffs.
- Select points from Q_1, Q_2, Q_3, Q_4 with probabilities $\rho = (\rho_1, \rho_2, \rho_3, \rho_4)$, respectively.
- For each selected point generate $Z = X_0 + w(X - X_0)$ where X is an approved nearest neighbor of X_0 and $w \sim F$.

Simulation Studies

Data difficulties:

- Simulated data with varying characteristics and compared performance of various models trained on data oversampled in different ways.
- High overlap, small `ncol`, categorical and missing data had primary impact before choice of oversampling method and distribution of w.

Hyperparameters:

- Applied S-SMOTE to data with different amounts of imbalance and overlap while varying ρ, η, and k_{max}.
- Use of smaller values for ρ_1 and ρ_4 (e.g. 0.05) improved minority class accuracy and use of $\eta = (0.60, 0.58, \dots, 0.20)$ led to less variable performance in certain cases.

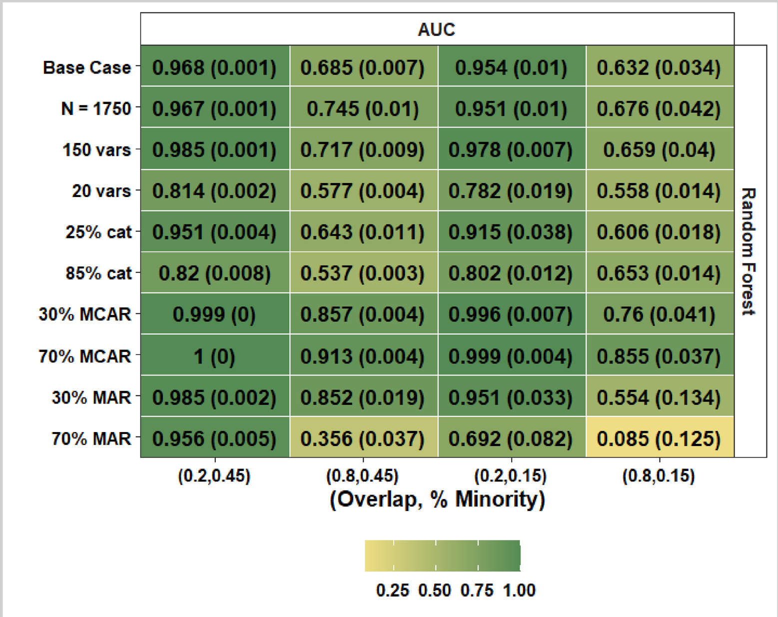


Figure 1. Median performance taken over medians of all oversampling methods.

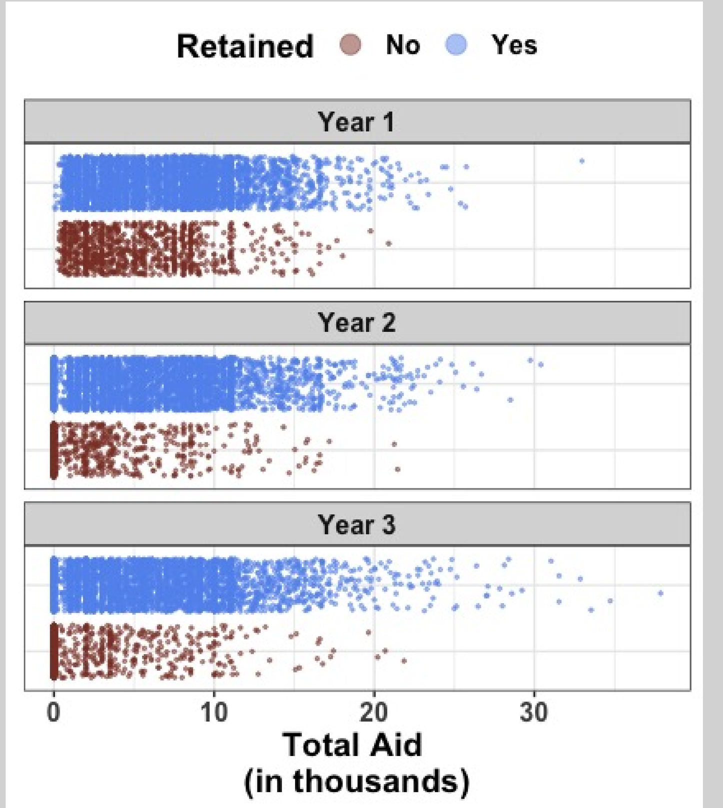
Predicting First-Year Retention

Goal: Predict retention of first time full-time (FTFT) freshmen, a vulnerable student group (e.g. Ameri et al. 2016).

Data: Obtained in collaboration with Office of Institutional Research and Analytics and other administrative offices at Oregon State University.

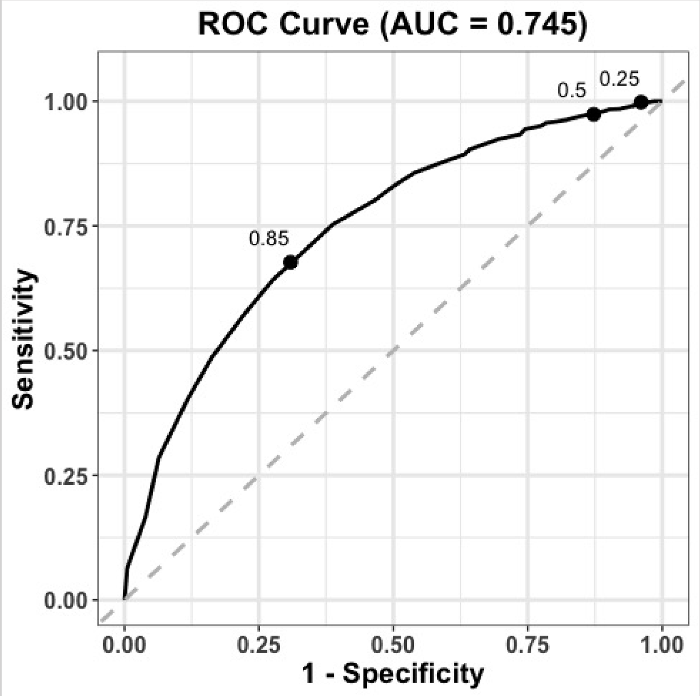
Cohort	1st yr. ret.	2nd yr. ret.	3rd yr. ret.
2011-2012	84.0%	91.2%	93.0%
2012-2013	84.7%	91.6%	94.4%
2013-2014	84.1%	91.5%	92.2%
Overall	84.2%	91.4%	93.2%

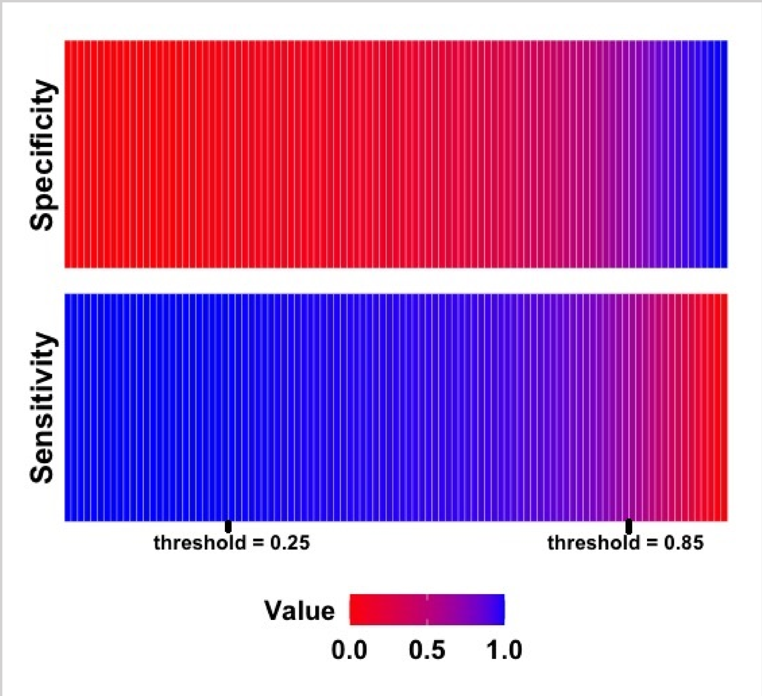
Issue: Predicting most common class automatically gave 84% accuracy and FTFT freshmen have similar data.



Performance of Models

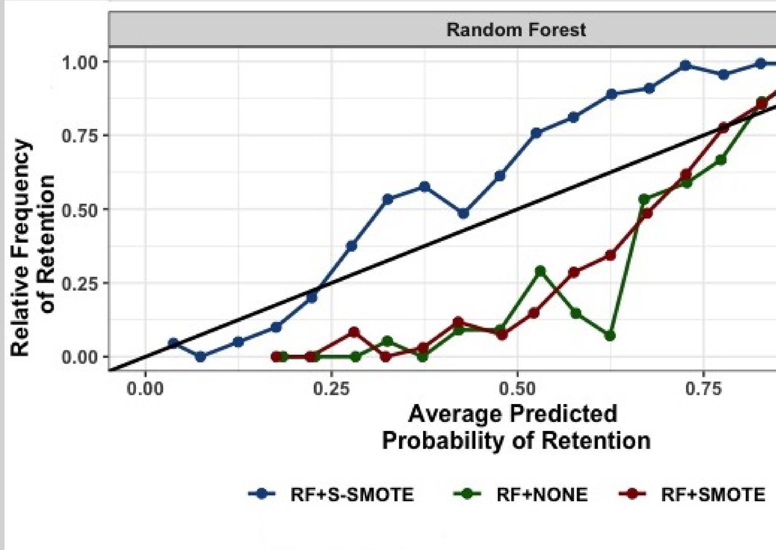
Logistic regression on unbalanced data:





Random forests:

Oversampling	Accuracy	Bal. Accuracy	Sensitivity	Specificity
None	0.909	0.710	0.996	0.424
SMOTE	0.913	0.725	0.995	0.454
S-SMOTE	0.933	0.909	0.943	0.875



- Results obtained on test data from random forests fit using 5-fold cross-validation.
- Minority class accuracy increased by 45.1% after applying S-SMOTE.
- Points below reference line in calibration plot indicate over-prediction.
- The average squared differences in (x,y) were: None = 0.073, SMOTE = 0.056, S-SMOTE = 0.026.

Conclusions

- Tuning certain factors of the oversampling process can positively impact model performance. S-SMOTE allows for this.
- Trade-off between sensitivity and specificity hard to eliminate completely
- Nuances of the data are the primary challenge