Train In Data

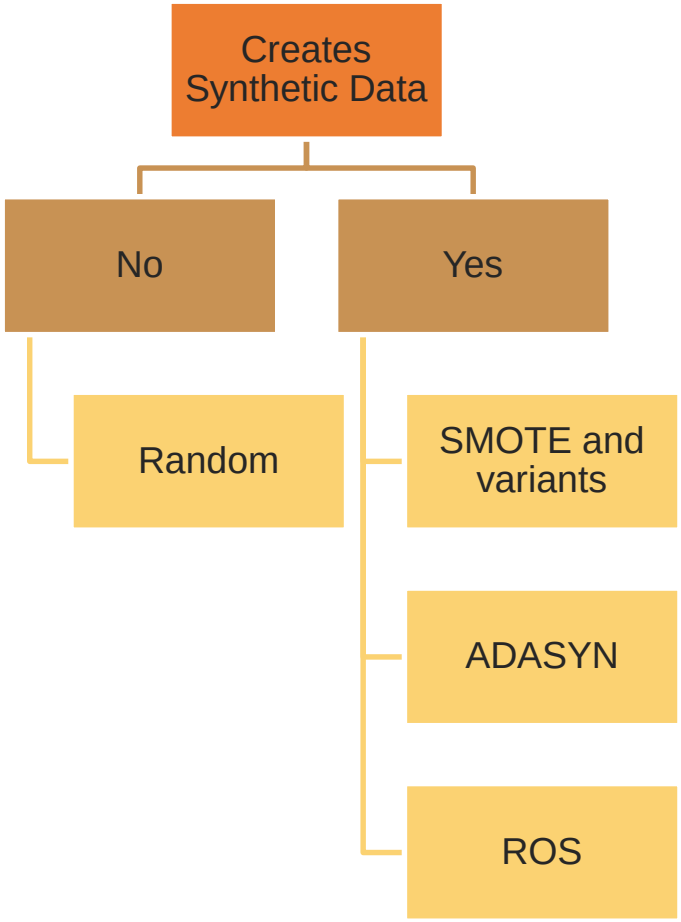
Wrap-up
Over-sampling

In summary

- There is no consensus in the community regarding which technique should be used with imbalanced datasets

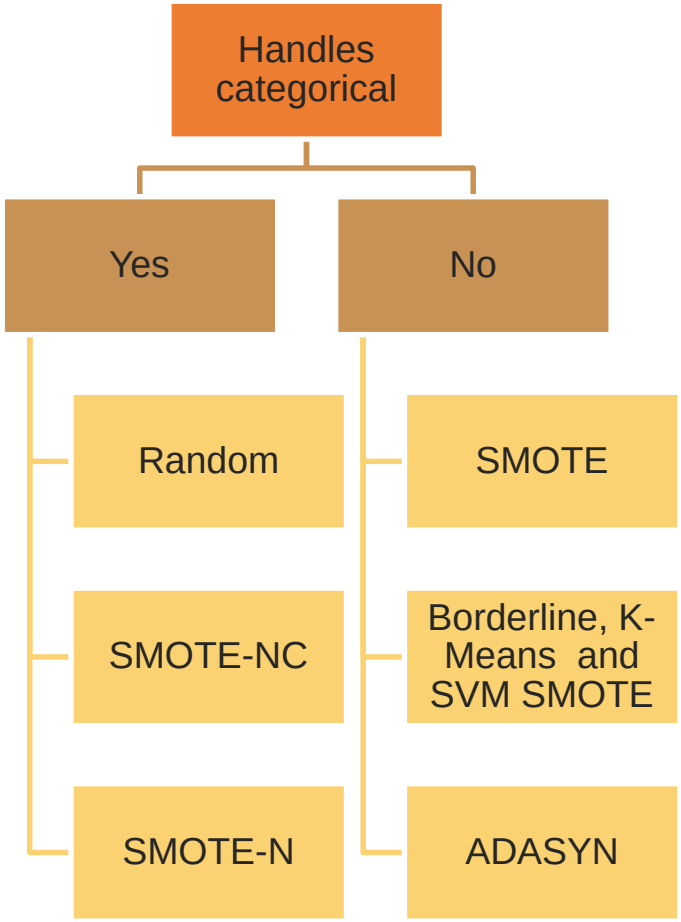
- Trial and test

Duplication vs Synthetic data



- Random Over-sampling "duplicates" examples from the minority class
- All other algorithms "create" new synthetic data, different from the original observations
- Probably, the latter is better

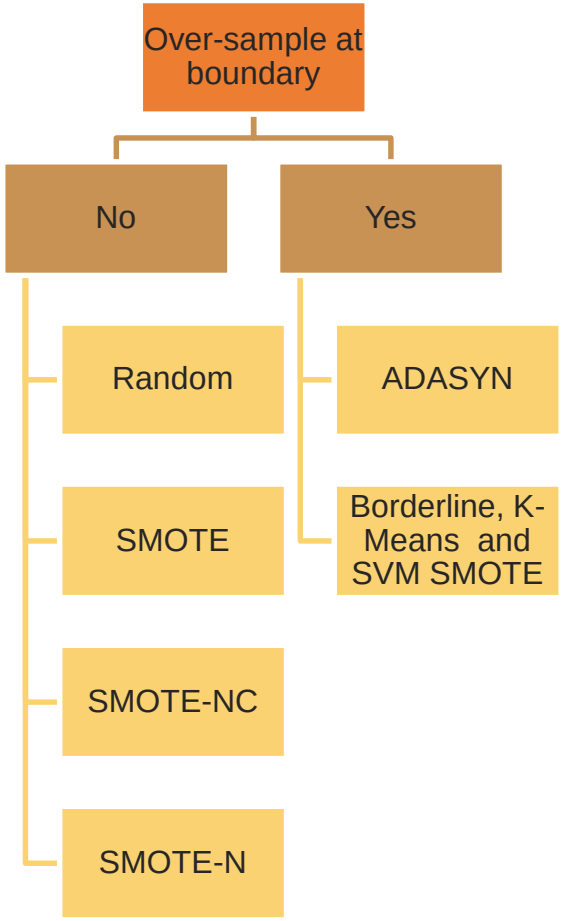
Over- sampling categorical variables



- **Random over-sampling** and the SMOTE variants NC and N, handle categorical variables out of the box

- For all the rest, we need to encode the variables first, and potentially, use alternative distance metrics.

Over-sample at the boundary



- **NO:** All observations from minority are used as templates, so new examples can be created in "safe" zones

- **Yes:** new examples are created from samples of the minority at the boundary with other classes

Expanding the decision boundary

- **ADASYN and Borderline SMOTE:** create new examples by interpolation between a template from the minority and a neighbours from either class, so they expand the decision boundary

SMOTE and ADASYN rely on KNN

- KNN is distance based ➔ scale the variables
- For categorical and discrete variables the traditional distance metrics (i.e., Euclidean, Manhattan) are not suitable ➔ consider using alternative metrics, or alternative options for imbalanced data
- Some methods involve training several KNNs (Borderline, SVM SMOTE), thus, they may scale poorly

SVM SMOTE relies on SVMs

- We need to know if we need a linear or other kernel ➔ not super straightforward

- SVMs take long to train with big datasets

K-Means SMOTE – specific use

- Suitable when there are intra-class clusters
- Nice in idea, but has too many hyperparameters to adjust

Over vs under-sampling

- In practice, over-sampling is used more frequently than under-sampling (i.e, lack of appetite to reduce the dataset size)
- SMOTE is very popular among data practicioners (not necessarily the best solution)
- Cleaning techniques do not reduce the data size dramatically, some of them may be computationally costly.

Re-sampling vs other methods

- A re-sampled data set can then be used to train a variety of algorithms
- Cost-sensitive learning needs to be introduced to each algorithm
- Ensemble algorithms are specific algorithms, we may want something else.

Train In Data

THANK YOU

www.trainindata.com