UNIVERSITY OF EDINBURGH
Business School

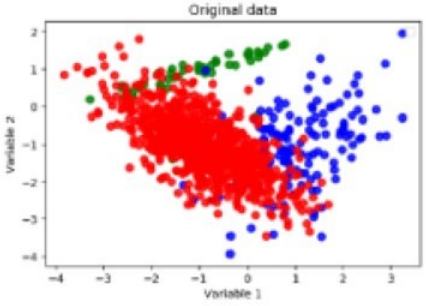
Predictive Analytics and Modelling of Data

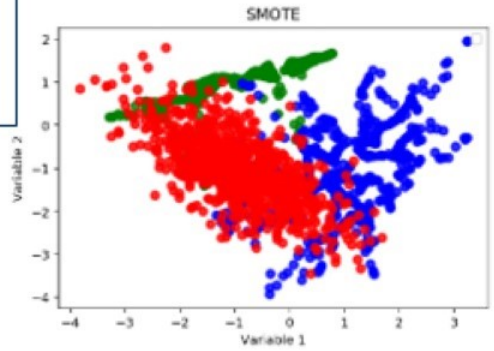
CMSE11428 (2020-2021)

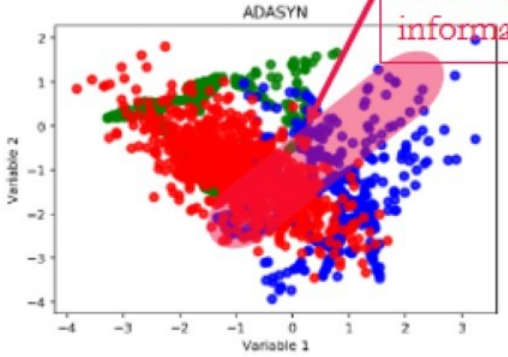
Dr Xuefei Lu
The University of Edinburgh Business School

Under and Over sampling

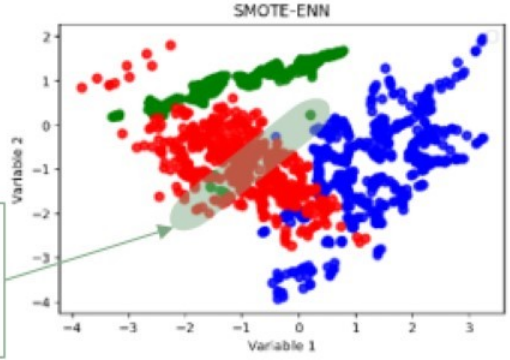
Create minorities based on N.neighbours



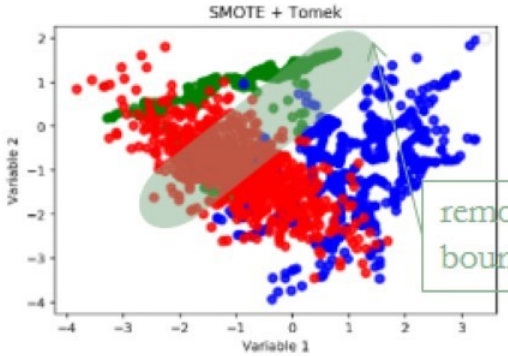




increase points at the boundaries→ more information



reduce points at the boundaries→ less ambiguity



remove points at the boundaries

Training, test, and validation sets

Build models

Evaluation

Build models

Model selection/
parameter tuning

Evaluation

The modelling process

Would you apply these to the whole data set, the training and/or test set?

- Missing values
- Outliers
- Normalisation
- Over/under-sampling
- PCA

The modelling process

The main idea is not to allow the model to gain insight from the test data.

We do not want information from the test/validation set "leak" into your training data!

- Normalisation, PCA:
  Same parameters used on your training data should be used on test/validation sets
- Outliers, missing values: depends how you deal with them
- Separate case:
  - Over/under-sampling: typically only done on the training set

- Some pacakges provide CV estimation of model parameters:
- The idea behind CV estimation method is to find the set of parameters,e .g. β,if apply that minimizes the cross-validation error.

$$\epsilon_{\mathrm{CV}}\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_{\mathrm{n}}^2; \mathcal{Y}\right) = \frac{1}{N} \sum_{k=1}^{K} \epsilon_{\mathrm{CV},k}\,.$$

- In our case, additional test set is necessary for a final evaluation to tune model parameters
- https://scikit-learn.org/stable/modules/cross_validation.html

UNIVERSITY OF EDINBURGH
Business School