UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Sampling Techniques

Activity: How to sample data

- Please read the following file:
- 9 - How_to_sample_data.ipynb

Discussion

What techniques are used to sample data to train models?

Picking random points and training models on those points is not as straightforward as it initially may seem. You will now carry out independent research into the literature and report on five best-in-class techniques that are used in data analytics.

Identify three techniques that are used to sample data to train models. As a guide, you should look at sampling and resampling techniques that are used to create training, validation and test sets that can be plugged into the predictive process.

For each technique you find, make a note of whether it is a sampling or resampling technique, and whether it is part of the evaluation or not.

Share your findings on our discussion forum.

Sampling strategies

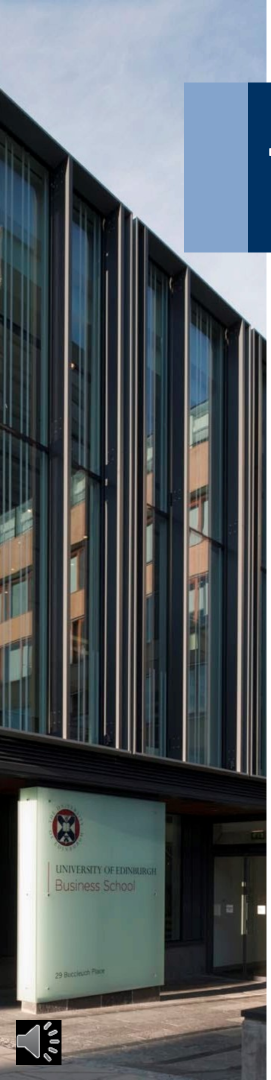
- Please watch the video through this link
- https://media.ed.ac.uk/media/Sampling/1_g0klxb7x/114521421

- Please read and try the following exercise:
- 10 - Assessment_Sampling_strategies.ipynb

Activity: Sampling strategies

Training, test and validation sets

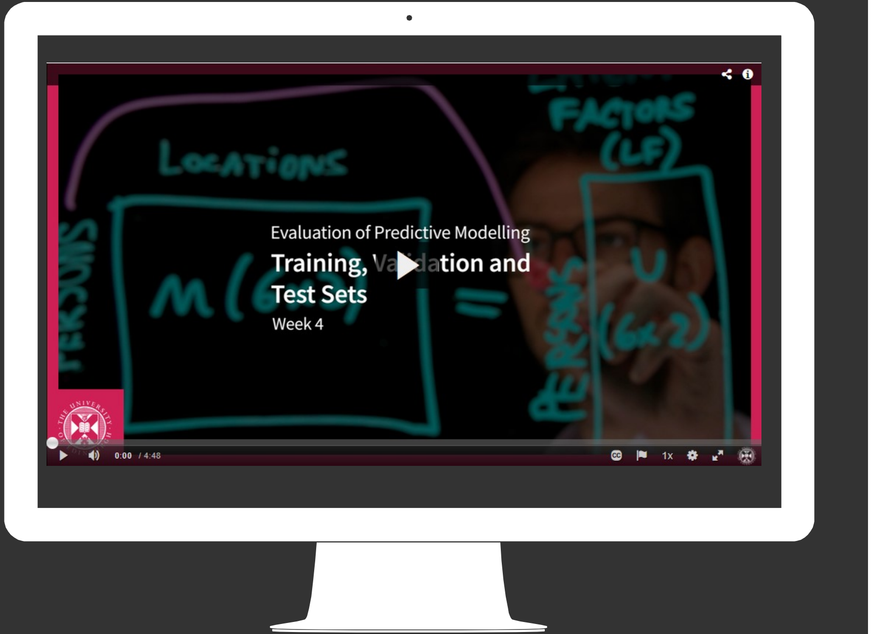
- let’s use the example of customer churn prediction.
- The main goal of prediction is to be able to use the model on different customers at a later stage.
- Now, if we use all the points in our dataset, we might make a good model and we can test the error measures. However, if we adjust our model based on the same data over and over again, we might run into the situation that if we test on new data, we get high error rate.
➔ over-fitting
- **To avoid this, we typically isolate a part of the data to train the model, the training set, and a part to measure the performance, the validation set (or test set).**



Training, test and validation sets

- **The training set** to build various models, or models with numerous parameters. We typically build a number of models with an array of parameter settings, which we apply to the observations in our validation set. **The validation set** gives us our well-known performance metrics, which we use to decide on the performance of a model with certain parameter settings.
- The observations that are part of the validation set are called the **holdout samples**.
- Nevertheless, we can still run into the same problem we had before: the optimisation of parameters and choosing the model might still reflect the patterns that are present in the validation set. Therefore, we validate the final performance measures, such as accuracy, on yet another set, the test set.
- For this final evaluation, we often recreate the model on both the training and validation sets.
- Sometimes, when we only want to build a single model, we only use a training and a test set, as we are not improving the model in any way, just running it.

Training, test and evaluation sets



Evaluation of Predictive Modelling

Training, Validation and Test Sets

Week 4

- Please watch the video through this link

https://media.ed.ac.uk/media/Training%2C+Validation+and+Test+Sets/0_awqefnoa/114521421

Addressing Imbalance

Measuring skewness

- Skewness of the distribution towards a certain label can have a negative influence on classification results.
- For continuous variables, there are metrics that measure the skewness, e.g. sample skewness:

$$\text{Sample skewness} = \frac{\text{sample third central moment}}{\text{sample sd}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$$

$$\text{Pearson's second skewness coefficient} = \frac{3(\text{mean} - \text{median})}{\text{sd}}$$

- For nominal variables:
A heuristic to see whether a distribution is skewed is by looking at the ratio between the proportion of the most and least occurring classes.
If this ratio is greater than 20, the skewness is significant.

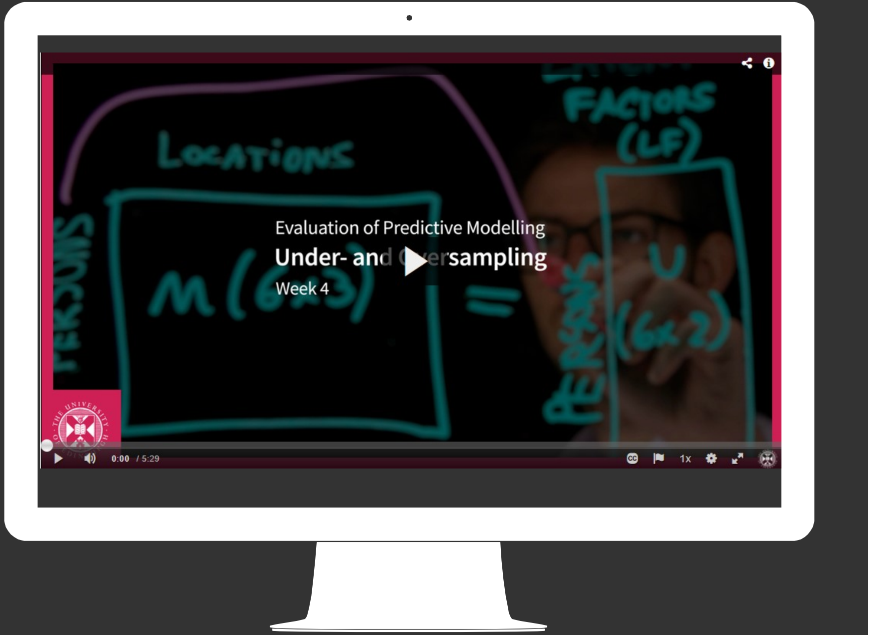
Under- and oversampling

- **Undersampling** reduces the number of samples that are present from the majority class, i.e., the one that is overrepresented. Typically, we use fixed undersampling, which aims to obtain a certain set ratio between the minority and majority classes.
- The opposite is **oversampling**, where the number of observations of the minority class is extended by duplicating them. Again, this can be done by maintaining a certain ratio.
- Synthetic Minority Oversampling TEchnique (**SMOTE)**
  SMOTE does not rely on duplicating existing observations, but rather creates similar ones to avoid any repetition of the existing observations.
  SMOTE first selects a minority class instance A at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors B at random and connecting A and B to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances A and B.
- ADAptive SYNthetic sampling technique (**ADASYN)**
  Similar to SMOTE but focuses on the samples from the minority class, which are closer to the decision boundary and are harder to learn as they are more difficult to distinguish from other classes.

Under- and oversampling

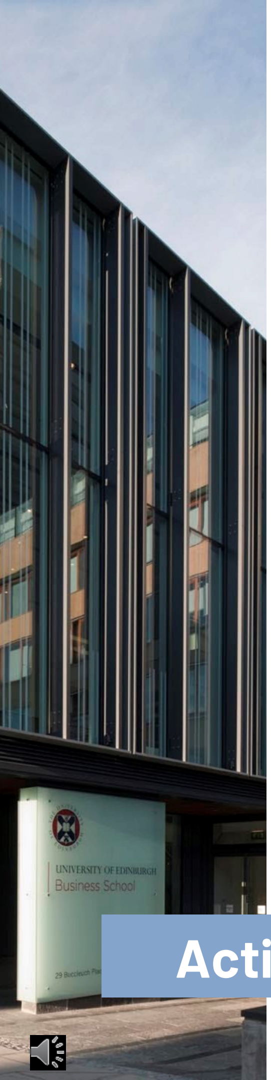
- **Hybrid sampling approaches** involve both under- and oversampling at the same time. They prevent the oversampling techniques from resulting in overfitting by increasing the number of samples of a particular class. They act as a cleaning mechanism on the dataset.
- **SMOTE-ENN**
  A combination of SMOTE and Edited Nearest Neighbour (ENN). ENN removes any observation of which the class label is different from that of at least two of its three nearest neighbours.
- **SMOTE + Tomek**
  Tomek links are a connection between two observations o_i and o_j from different classes, such that there exists no o_l such that $d(o_i, o_l) < d\left(o_i, o_j\right)$ or $d\left(o_j, o_l\right) < d\left(o_i, o_j\right)$ where $d(\cdot)$ is a distance function.
  They can be used to remove the observations overlap between classes, together with SMOTE, be used to remove both observations that are linked up through Tomek links.

Under- and Oversampling



- Please watch the video through this link

https://media.ed.ac.uk/media/Under-+and+Oversampling/1_kszcyz04/114521421



- You will now learn how to code various oversampling and SMOTE-based sampling techniques in Python. You will learn how to apply under- and oversampling and use SMOTE, SMOTE-ENN, ADASYN, as well as SMOTE + Tomek, in Python.
- Please the study the following files
- 11 - create_training_and_test_sets.ipynb
  12 - SMOTE_based_sampling_techniques.ipynb
- Then try the following exercise:
  13 - Activity_8_under_and_oversampling_techniques.ipynb + absent.csv

Activity: SMOTE-based sampling techniques

UNIVERSITY OF EDINBURGH
Business School