Naming of files

2021-04-28 Updated by Yan

1. Resampled data files: [Data]\_[X|Y]\_[Train]\_[No].csv
   1. Each dataset contains four csv files. Each csv contains headers but not row numbers.
   2. [Data]: describing the resampling method used to generate this dataset.

Undersampling:

1.1\_RandomUnderSampler: Random majority under-sampling with replacement

1.2\_One-SidedSelection: One-Sided Selection

1.3\_NeighbourhoodCleaningRule: Neighbourhood Cleaning Rule

Oversampling:

2.1\_RandomOverSampler: Random minority over-sampling with replacement

2.2\_SMOTENC: SMOTE for Nominal and Continuous

2.3\_ADASYN: Adaptive synthetic sampling approach for imbalanced learning

* 1. [No] represents the number of resampling for each type.
  2. Example:

1.1\_RandomUnderSampler\_X\_Train\_1.csv

2.3\_ADASYN\_Y\_Train\_5.csv

* 1. Variables in X dataset:

['trip\_id', 'household\_id', 'person\_n, 'trip\_n'] +

['age', 'male', 'driving\_license', 'car\_ownership', 'distance', 'dur\_walking', 'dur\_cycling', 'dur\_pt\_access', 'dur\_pt\_inv', 'dur\_pt\_int\_total', 'pt\_n\_interchanges', 'dur\_driving', 'cost\_transit', 'cost\_driving\_total']

* 1. Value in Y dataset: 'walk', 'pt', 'cycle', 'drive'