greatlearning
Learning for Life

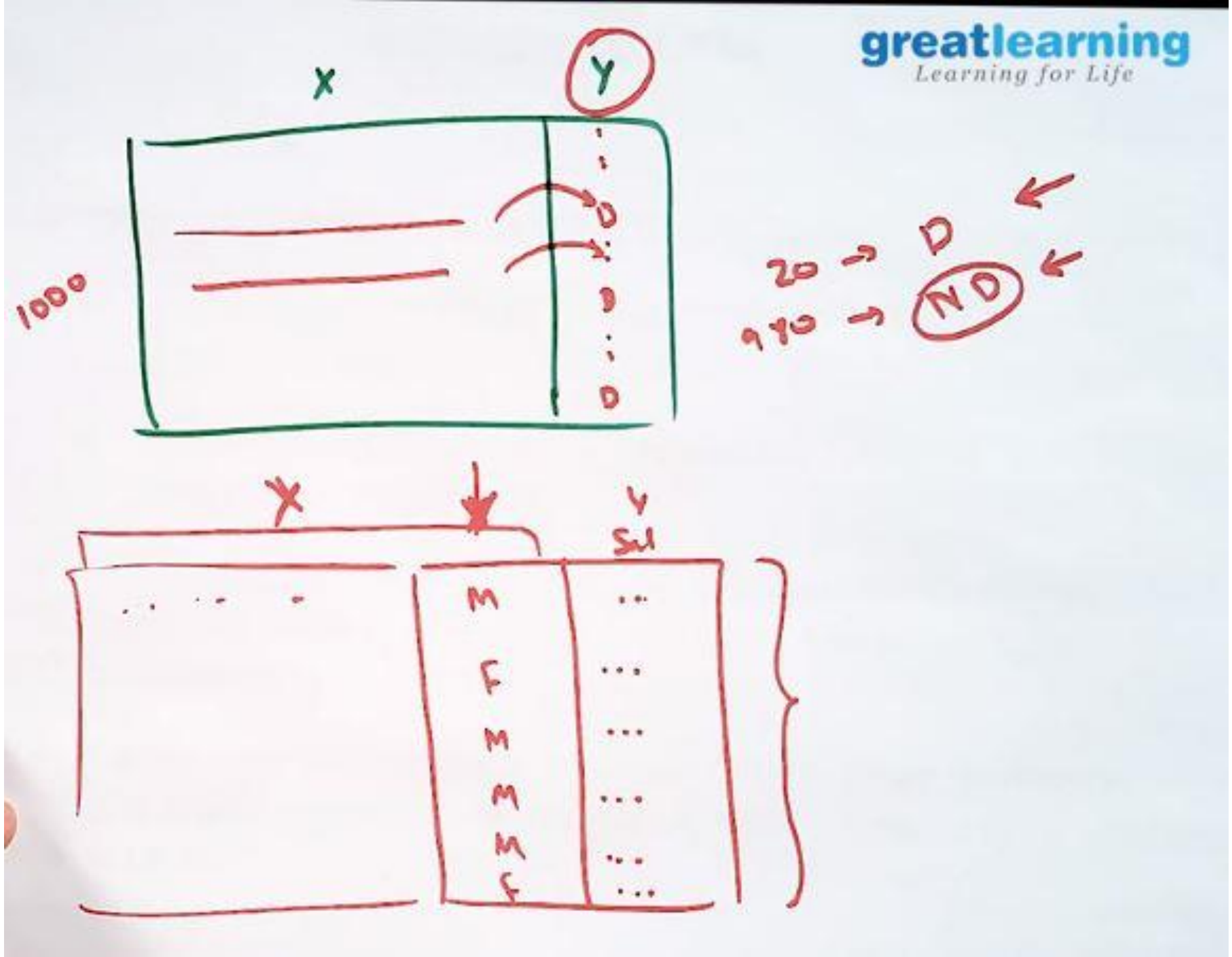
Over/Under-Sampling, Cross Validation

Machine Learning

Imbalanced Data

greatlearning
Learning for Life

- Data Imbalance can be of the following types:
  - Under-representation of a class in one or more important predictor (independent) variables
  - Under-representation of one class in the outcome (dependent) variable
- Many machine-learning techniques, such as neural networks, make more reliable predictions from being trained with balanced data.
- Certain analytical methods, however, notably linear regression and logistic regression, do not benefit from a balancing approach.



Imbalanced Data

greatlearning
Learning for Life

- Data Imbalance can be of the following types:
  - Under-representation of a class in one or more important predictor (independent) variables
  - Under-representation of one class in the outcome (dependent) variable
- Many machine-learning techniques, such as neural networks, make more reliable predictions from being trained with balanced data.
- Certain analytical methods, however, notably linear regression and logistic regression, do not benefit from a balancing approach.

Over-sampling and Under-sampling

- to compensate for an imbalance that is already present in the data
- More complex oversampling techniques, include the creation of artificial data points
- The end-result of over-/under-sampling is the creation of a balanced dataset
- Over-sampling is generally employed more frequently than under-sampling

- Over-sampling
  - Random sampling
  - SMOTE: Synthetic Minority Over-sampling Technique
    - generates synthetic samples from the minority class
    - data created by drawing new samples along lines drawn between existing minority data points
- Under-sampling
  - Random sampling
  - Cluster Centroids