TREATING
IMBALANCED
DATA

USING RIGHT EVALUATION METRIC

- Using inappropriate evaluation metrics for the mode which is generated using imbalanced data can be dangerous. You can use metric like
  - Precision/Specificity
  - Recall/Sensitivity
  - F1 score
  - MCC
  - AUC

RESAMPLING DATA

- Two techniques
  - Under-sampling: Under-sampling balances the dataset by reducing the size of the Majority class
  - Over-sampling: Under-sampling balances the dataset by increasing the size of the Minority class.

Generate Synthetic Samples

- A simple way to generate synthetic samples is to randomly sample the attributes from instances in the minority class.
- The most popular of such algorithms is called SMOT or the Synthetic Minority Over-sampling Technique.

Penalize Algorithms

- Use penalized learning algorithms that increase the cost of classification mistakes on the minority class.
- Penalized classification imposes an additional cost on the model for making classification mistakes on the minority class during training.
- Most used
    - penalized-SVM
    - penalized-LDA.

Tree-Based Algorithms

- Using tree-based algorithms. Decision trees often perform well on imbalanced datasets because their hierarchical structure allows them to learn signals from both classes. Most used
    - Random Forests
    - Gradient Boosted Trees
    - C4.5
    - C5.0
    - CART
    - etc....

Collect More Data (If possible)

- If it is possible to collect more data then go ahead.
- Collecting more examples for minor class may be useful later when we look at resampling your datase

Cluster the Majority class

- Instead of relying on random samples to cover the variety of the training samples, clustering the abundant class in r groups, with r being the number cases in r.
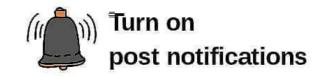
Your own ideas

- Dig deeper into the data and understand the business and develop your own ideas

If i missed any technique please mention that in the comments, so it will help others.

Thank You.

Like, Comment, Share and Save it for Later

HAPPY LEARNING
@learn.machinelearning