Recap of Lecture on Data Preprocessing, Class imbalance and ROC –AUC curve

Data Preprocessing steps

- Data Cleansing:
  - Tackle missing data
  - For text applications, remove stop words, convert all verbs to root forms
  - Remove inconsistencies (rows with same attribute values but different labels)
- Feature engineering – Converting raw data to a usable form for ML. Ex
  - Feature rejection (drop unrelated features) and feature selection
  - Feature Transformation: transform raw data with PCA and reduce dimensionality by using only important components
  - Feature extraction: Generate new features by combining existing features

Tackling missing data

- Missing values
  - Delete rows with missing data
  - Delete columns if it has multiple empty entries
  - Assign all possible values if categorical
  - Assign average value of column, if values are continuous. Assign median value of column if categorical
  - Decision relative mean/median: Use only rows with same label as that of row with missing value
  - Closest Assign feature value with closest matched row with same label or with centroid of closest matched cluster with same label

Handling class imbalance

- Use stratified k fold cross validation – each fold has same ratio of two or more classes
- Use micro-averaged recall / precision if you want that class with greater representation should be given more priority (Microaverage gives equal priority to each instance)
- Use macro-averaged recall / precision if you want that class with less presentation also gets equal priority as other classes. (Macroaverage gives equal priority to all classes)

ROC – AUC Curve

Receiver Operating Characteristics - Area Under the Curve

AUC of Curve 1: 0.92 (best)
AUC of curve 2: 0.75
AUC of Curve 3: 0.5