Hands-On Session on Model Selection

Finalize Model

Finalize Model

- Cross validation
- Data Imbalance
- Data Leakage
- Underfitting
- Overfitting
- Hyperparameter Tuning
- Model evaluation criteria

cross-validation

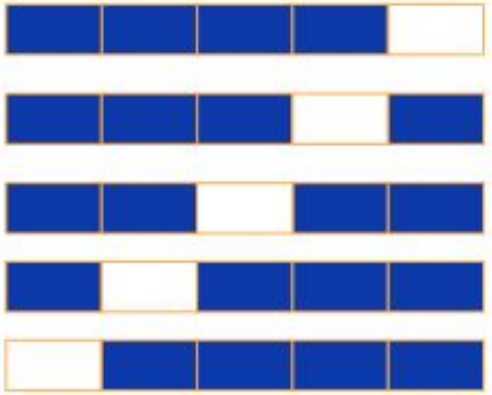
cross-validation

K fold cross-validation will divide data into k-folds

Train model on k-1 folds and test its performance on the last fold

K fold cross-validation will generate k models and k performance scores

Instead of getting only 1 score, here we'll get k scores, which will give a better picture of the variance in model performance



Handle imbalanced data

Datasets used in banking, health and market analytics usually have imbalance i.e. one class is in majority and one is in minority (less than 5%)

During training on such datasets, the model gives more weightage to the majority class and gets biased

To avoid such situations, we can use oversampling or undersampling techniques on data

- Oversampling will create artificial data points for the minority class
- Undersampling will remove data points from the majority class

We can't afford to lose data points in case of small data size, so oversampling is preferred in such cases

Data leakage

- Data leakage is the situation where the model, while it is being created, is influenced by test data
- Due to data leakage, model performance on test data is not trustworthy as the sanctity of test data is compromised

Data leakage can happen in multiple ways.

- Standardizing data before splitting into training and testing data. For e.g. using z-score
- Imputing missing values for the entire data before splitting into training and testing data
- Hyper parameter tuning to improve performance on test data
- Best way to avoid data leakage is to keep a portion of the sample data away before doing any processing

Underfitting and Overfitting

Hyperparameter Tuning

- GridSearchCV
- RandomizedsearchCV

Model Evaluation Criteria