Group Discussion: Model Selection

Objective:

This activity aims to deepen your understanding of key concepts in model selection, such as performance trade-offs, validation techniques, and model complexity.

You will work in groups to analyse real-world scenarios and explore how various factors influence the choice of an optimal model.

Discussion Points for Each Group:

- **Outcome Discussion**: What does the outcome imply about the model's performance and suitability?
- **Influencing Factors**: What factors influenced the model selection? Why was one model preferred over another?
- **Concept Impact**: How do overfitting, generalization, computational cost, and operational requirements affect the decision?
- **Trade-off Evaluation**: What compromises are made in terms of accuracy, speed, and usability when choosing one model over another?

Expected Outcomes:

- You will better understand how to make informed decisions about model selection based on performance, complexity, and application-specific needs.
- You will learn to articulate the reasoning behind choosing specific machine learning models and strategies in varied real-world contexts.

Examples to illustrate key concepts in Model Selection

Each of these examples illustrates different aspects of model selection, emphasizing how different scenarios might require different priorities and considerations.

1. Performance Trade-offs

Scenario: Predicting Customer Churn

A telecommunications company wants to predict which customers are likely to churn based on their usage patterns, demographics, and customer service interactions. The data science team builds two models:

- A complex deep neural network (DNN) with multiple layers and parameters.
- A simpler logistic regression model.

Outcome:

- The DNN performs exceptionally well on the training dataset, achieving near-perfect accuracy.
- However, when deployed on new, unseen data, the DNN's performance drops significantly, indicating overfitting.
- The logistic regression model, while slightly underperforming on the training data compared to the DNN, shows much better generalization on new data.

2. Validation Techniques

Scenario: Developing a Diagnostic Tool

A healthcare company is developing a diagnostic tool to detect a specific type of cancer from patient test results. The team decides to use k-fold cross-validation to evaluate the performance of three different models: a support vector machine (SVM), a random forest classifier, and a gradient boosting machine (GBM).

Procedure:

- The dataset is divided into 'k' subsets. Each model is trained on 'k-1' subsets and tested on the remaining subset, repeated 'k' times with each subset used as the test set once.
- Each model's accuracy, sensitivity, and specificity are recorded across all 'k' folds.

Outcome:

- The SVM shows high variance in performance across different folds.
- The random forest provides the best balance of sensitivity and specificity across all folds, with lower variance compared to the SVM.
- The GBM shows good performance but at a higher computational cost.

3. Model Complexity

Scenario: Real-Time Fraud Detection

A financial institution implements a real-time system to detect fraudulent transactions. The team explores several models ranging from simple logistic regression to complex ensemble methods and neural networks.

Considerations:

- The simpler logistic regression model can be implemented and run with lower latency, which is critical in real-time applications.
- More complex models like ensemble methods and neural networks offer higher accuracy but require significantly more computational resources and result in slower response times.

Outcome:

- The logistic regression model, while slightly less accurate, provides sufficiently good performance and can evaluate transactions in real-time without causing delays.
- The complex models, despite their higher accuracy, are deemed unsuitable for real-time processing due to their computational demands.