TIKTOK
DATA SCIENCE
INTERVIEW
QUESTIONS

WHICH MODEL WOULD YOU CHOOSE BETWEEN ONE WITH 85% ACCURACY AND ONE WITH 82% ACCURACY?

The choice depends on the business context and the model's trade-offs. Consider -

Cost of Errors: If errors have high costs (e.g., medical predictions), prioritize higher accuracy.

Performance Metrics Beyond Accuracy: Check precision, recall, F1-score, or AUC for nuanced performance insights - if the dataset is imbalanced or business goals align with these metrics better it would make more sense to change metrics for evaluation models.

Model Complexity: Prefer the simpler model if the impact of the accuracy difference is marginal.

Deployment Needs: Choose the model better suited for scalability, interpretability, or resource constraints.

HOW WOULD YOU IMPROVE A CLASSIFICATION MODEL SUFFERING FROM LOW PRECISION IN PREDICTING CUSTOMER PURCHASES ON AN E-COMMERCE PLATFORM?

Refine Data Quality

- Ensure accurate labeling and remove noisy data.
- Balance the dataset to mitigate bias towards non-purchases.

Feature Engineering

- Incorporate more features (e.g., session duration, cart size).
- See if you can engineer new features - interaction terms or embeddings for user and product behaviors.

Adjust Model Threshold

- Increase the decision threshold to prioritize precision over recall.

Model Tuning

- Optimize hyperparameters using grid or random search.
- Use regularization to reduce overfitting.

Algorithm Selection

- Experiment with precision-focused algorithms like XGBoost or SVM.

Post-Processing

- Apply cost-sensitive learning or precision-weighted loss functions.

IS THERE A PROBLEM WITH RUNNING AN A/B TEST WITH 20 DIFFERENT VARIANTS?

A A/B test with 20 variants introduces challenges -

Statistical Issues

- Multiple Comparisons Problem: Increases the risk of Type I errors (false positives).
- Smaller Sample per Variant: Dilutes statistical power, requiring a larger total sample size.

Complexity in Interpretation

- Comparing many variants complicates understanding and deriving actionable insights.

Resource Constraints

- Increased development, monitoring, and deployment efforts.

Mitigation:

- Use Multi-Arm Bandit Testing to dynamically allocate traffic to better-performing variants.
- Apply statistical corrections (e.g., Bonferroni) to control for false positives.
- Consider a pre-test analysis to determine feasibility based on traffic and desired confidence levels.

HOW DO YOU DECIDE BETWEEN USING XGBOOST AND RANDOM FOREST FOR MACHINE LEARNING PROBLEMS?

Decision Drivers are. - Dataset size, problem complexity, need for speed, and tolerance for tuning effort. If accuracy is key and resources allow, choose XGBoost. For interpretability and simplicity, use Random Forest.

Performance and Use Case:

- XGBoost: Preferred for structured/tabular data where boosting improves predictive accuracy. It handles interactions better and excels in competitions like Kaggle.
- Random Forest: Works well for simpler problems or when interpretability and robustness to noisy data are more critical.

Speed and Scalability:

- XGBoost: Faster with optimized gradient boosting and parallelization
- Random Forest: Slower for large datasets but simpler to set up

Overfitting:

- XGBoost: More prone to overfitting if not tuned properly.
- Random Forest: Less likely to overfit due to feature bagging.

Hyperparameter Tuning:

- XGBoost: Requires fine-tuning parameters like learning rate, tree depth, and regularization terms for optimal performance.
- Random Forest: Needs fewer hyperparameters; mainly focuses on the number of trees and features to split.

WAS THIS HELPFUL?

Be sure to save it so you can come back to it later!

@karunt