Data Science Quiz (Medium)

1. Handling class imbalance in classification. Common solutions are oversampling, undersampling, SMOTE, or class weights. The scenario could be about a marketing campaign with low conversion rate. The correct answer might be using SMOTE since it's a synthetic method and better than just oversampling the minority class.

2. Choosing evaluation metrics. For a medical diagnosis model where false negatives are critical, recall is more important. The options could include accuracy, precision, recall, F1. Correct is recall.

3. Missing data in customer age. MCAR, MAR, MNAR. If missingness is random, deletion is okay. The scenario says missing completely at random, so listwise deletion is acceptable. Correct answer is a.

4. Ethical consideration in model development. GDPR compliance involves data anonymization. The correct answer would be anonymizing customer data before processing.

5. Model interpretability for stakeholders. Using SHAP values is a good method. The options might include LIME, SHAP, PCA, t-SNE. Correct is SHAP.

6. Cross-validation for time series. TimeSeriesSplit is appropriate. The answer is a.

7. Feature engineering for transactional data. Creating a 'days since last purchase' feature is temporal aggregation. Correct answer is a.

8. Hyperparameter tuning. Bayesian Optimization is more efficient than grid or random search. Correct is c.

9. Data leakage. Scaling before train-test split causes leakage. The answer is c.

10. Big data processing. Spark is suitable for handling large datasets across clusters. Answer is c.

1. SMOTE is the right approach for handling imbalance. The options are correct.

2. Recall is correct for minimizing false negatives.

3. MCAR and listwise deletion is correct.

4. Anonymizing data aligns with GDPR.

5. SHAP provides global interpretability.

6. TimeSeriesSplit is correct for time series.

7. 'Days since last purchase' is a temporal feature.

8. Bayesian Optimization is more efficient.

9. Scaling before split causes data leakage.

10. Spark handles distributed processing.

1. A data scientist is building a binary classification model for a marketing campaign with a 5% conversion rate. Which technique is MOST appropriate to address class imbalance while preserving information from both classes?

a) Random undersampling of the majority class

b) Oversampling the minority class with duplication

c) Using SMOTE to generate synthetic minority class samples

2. When developing a medical diagnosis model where false negatives are critical, which evaluation metric should take priority?

a) Accuracy

b) Precision

c) F1-Score

3. During data preprocessing, 15% of customer age values are missing from a retail dataset. Preliminary analysis shows the missingness is completely random. What is the BEST immediate approach?

b) Impute using median age to preserve data volume

c) Develop a predictive model to estimate missing ages

d) Exclude the age feature entirely from analysis

4. A team needs to ensure GDPR compliance when processing EU user data. Which practice

is REQUIRED during the data collection stage?

a) Storing raw IP addresses for geolocation analysis

c) Using third-party cookies without explicit consent

d) Retaining all user data indefinitely for model retraining

5. Which technique provides BOTH global and local model interpretability for stakeholders?

a) Principal Component Analysis (PCA)

c) t-SNE visualization

d) Logistic regression coefficients

6. When working with time-series sales data containing seasonal patterns, which cross-validation method is MOST appropriate?

b) Stratified K-Fold validation

c) Leave-One-Out cross-validation

d) Random permutation shuffling

7. A feature engineering task for transactional data should prioritize:

b) Removing all correlated features using VIF threshold

c) Applying one-hot encoding to numerical variables

d) Normalizing all features before interaction terms

8. Which hyperparameter tuning method provides better efficiency for large parameter spaces compared to grid search?

a) Manual trial-and-error selection

b) Random search with fixed iterations

d) Genetic algorithms with 100 generations

9. A model shows 98% accuracy during development but performs poorly in production. What is the MOST likely cause of this discrepancy?

a) Insufficient model complexity

b) Lack of feature scaling

d) Underfitting to training patterns

10. When processing a 50GB dataset on a single machine with limited RAM, which tool is MOST suitable?

a) Pandas with chunk processing

b) MySQL database queries

d) Excel Power Query filters