QF632-2025-W2

Number of participants: 45

1. When handling missing prices in a high-frequency limit-order-book time series, forward-filling (using last observed price) can introduce which primary risk?

14 correct answers out of 33 respondents

	Heteroskedasticity bias	3%	1 vote
	Look-ahead (future) leakage	55%	18 votes
✓	Artificially reduced volatility	42%	14 votes
	Overestimation of liquidity	0%	0 votes

2. Suppose you're imputing missing FX-rate returns before computing rolling-window volatility features. Why might linear interpolation between returns be inappropriate?

11 correct answers out of 34 respondents

✓	It distorts the autocorrelation structure of returns	32%	11 votes
	It assumes constant drift in log-prices	53%	18 votes
	It violates no-arbitrage constraints	12%	4 votes
	It biases the Sharpe ratio upward	3%	1 vote

3. For a credit-scoring model trained on quarterly accounting ratios, missing data occur systematically for small firms. Which approach helps mitigate selection bias due to MNAR missingness?

20 correct answers
out of 33 respondents

	Single imputation via sample mean	9%	3 votes
	Indicator (flag) for missingness in feature set	24%	8 votes
✓	Multiple imputation including a missingness indicator and auxiliary variables	61%	20 votes
	K-nearest-neighbors on complete-case subset	6%	2 votes

4. In K-fold cross-validation, as K increases (with fixed dataset size), the bias and variance of the estimated generalization error behave as:

10 correct answers out of 33 respondents

✓	Bias decreases, variance increases	30%	10 votes
	Bias increases, variance decreases	12%	4 votes
	Both bias and variance decrease	55%	18 votes
	Both bias and variance increase	3%	1 vote

5. Which procedure correctly prevents “data leakage” during hyperparameter tuning?

17 correct answers out of 32 respondents

	Performing a single K-fold CV on the full dataset and selecting the best hyperparameters	9%	3 votes
✓	Using nested cross-validation: an inner loop for hyperparameter tuning and an outer loop for performance estimation	53%	17 votes
	Repeating random train–test splits until performance stabilizes	25%	8 votes
	Applying leave-one-out CV both for tuning and for final evaluation	13%	4 votes

6. For time-series data (e.g.\ stock prices), which cross-validation scheme properly respects temporal order?

27 correct answers
out of 30 respondents

	Standard random K-fold CV	0%	0 votes
	Stratified K-fold CV	10%	3 votes
✓	Forward-chaining (rolling) CV, where each fold trains on $[1 : t]$ and tests on $(t : t + h]$	90%	27 votes
	Leave-one-out CV	0%	0 votes

7. You suspect an interaction: "Strategy" (Trend, Mean-Reversion) × "Market Regime" (Bull, Bear) on monthly returns. Which side-by-side boxplot layout would most clearly reveal whether the effect of Strategy changes with Regime?

0 correct answer
out of 0 respondent

	Two boxes per Strategy (grouped by Regime within each Strategy)	0%	0 votes
✓	Two boxes per Regime (grouped by Strategy within each Regime)	0%	0 votes
	Four boxes in arbitrary order	0%	0 votes
	A single box per Strategy, ignoring Regime	0%	0 votes

8. You compute a correlation matrix for five factors and display it as a heatmap ("correlation plot"). You observe two factors with $|r| > 0.95$. What is the most serious concern for a linear-regression factor model including both?

0 correct answer
out of 0 respondent

	Overfitting due to too many observations	0%	0 votes
✓	Multicollinearity leading to unstable coefficient estimates	0%	0 votes
	Inability to compute pairwise scatter plots	0%	0 votes
	Reduced predictive power	0%	0 votes

9. How does the strength of the (linear) correlation between two variables affect the slope of the best-fit line in a simple regression?

0 correct answer
out of 0 respondent

✓	A stronger (absolute) correlation yields a steeper slope.	0%	0 votes
	The slope is the same regardless of correlation.	0%	0 votes
	The slope equals the correlation coefficient exactly.	0%	0 votes
	A weaker correlation yields a steeper slope.	0%	0 votes

10. Which technique is an example of feature selection (not feature extraction)?

0 correct answer
out of 0 respondent

	Principal Component Analysis	0%	0 votes
✓	Removing predictors whose correlation with the target is below a threshold	0%	0 votes
	Applying a log-transform to a skewed variable	0%	0 votes
	Creating pairwise products of existing features	0%	0 votes

11. When you add an interaction term between two continuous features, why might you center each first (subtract its mean)?

0 correct answer
out of 0 respondent

	To speed up tree-based models	0%	0 votes
✓	To eliminate multicollinearity between main effects and their product	0%	0 votes
	To ensure the interaction term is always positive	0%	0 votes
	To reduce the number of features	0%	0 votes

12. You're using k -nearest-neighbors for classification. Which transformation of your numeric features is most critical before fitting the model?

0 correct answer
out of 0 respondent

✓	Standardization (z-score)	0%	0 votes
	One-hot encoding	0%	0 votes
	Principal Component Analysis	0%	0 votes
	Binning into quartiles	0%	0 votes