منابع

1. Quote (Justifying Lowering Resolution due to Target Data):
"The granularity of the analysis is often constrained by the lowest resolution available for essential variables, particularly the outcome or response variable. Modeling at a finer resolution than the outcome data permits can lead to spurious precision or ecological fallacy."
— Adapted from principles in: Waller, L. A., & Gotway, C. A. (2004). Applied Spatial Statistics for Public Health Data. Wiley.

2. Quote (Justifying Aggregation Methods: Sum vs. Mean):
"When aggregating data over time or space, summation is appropriate for preserving the total magnitude of count-based phenomena (e.g., total events, total volume), while averaging is suitable for representing the central tendency or typical value of continuous measurements (e.g., average speed, mean temperature) within the aggregation window."
— Based on concepts in: Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical Modeling and Analysis for Spatial Data. CRC Press. (And incorporating the UNESCAP reference for traffic specifically).

3. Quote (Justifying Removal of Incomplete Data Points):
"Listwise deletion of cases with missing data is a necessary step when the missingness is substantial or informative within the period of analysis, or when reliable imputation is infeasible, as including incomplete records can introduce significant bias into aggregated statistics or model estimates."
— Scheffer, J. (2002). Dealing with missing data. Research Letters in the Information and Mathematical Sciences, 3(1), 153-160. (Concept widely discussed in statistical literature like Little & Rubin).

4. Quote (Justifying Train/Test Split Ratio and Random State):
"A common practice in hold-out validation is to randomly partition the data into training and testing sets, often using ratios like 80/20 or 70/30. Setting a random seed ensures the reproducibility of this split and subsequent analyses."
— Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

5. Quote (Justifying Internal Encoders - CatBoost/LGBM):
"Gradient boosting implementations like CatBoost and LightGBM incorporate specialized, efficient algorithms for handling categorical features internally, often leveraging variants of target encoding or feature hashing, which can outperform standard preprocessing encodings without manual intervention."
— Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31. (Also see LightGBM documentation).

6. Quote (Justifying Ordinal + OHE for Trees - XGB/RF/GBR):
"While tree-based models can intrinsically handle ordinally encoded features by finding optimal split points, One-Hot Encoding is generally safer for nominal variables to avoid imposing an artificial order. For high-cardinality nominal features, however, ordinal encoding might be considered pragmatically, noting trees are less sensitive than linear models. One-Hot Encoding is standard for nominal features with fewer levels like 'year'."

— Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media. (Discusses encoding trade-offs for trees).

7. Quote (Justifying OHE for Decision Tree):

"Standard implementations of algorithms like Decision Trees often require numerical input. Therefore, nominal categorical predictors must typically be converted into a numerical format, commonly achieved using One-Hot (or dummy) Encoding."

— Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Addison-Wesley.

8. Quote (Justifying OHE (drop_first=True) for Linear/KNN/GPR/SVR):

"For linear models, One-Hot Encoding must use drop_first=True (or equivalent) to avoid the 'dummy variable trap' causing perfect multicollinearity. This reduction in dimensionality can also be beneficial for distance-based (KNN) or kernel-based (GPR, SVR) methods."

— James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.

9. Quote (Justifying StandardScaler):

"Standardization, typically via StandardScaler, transforms features to have zero mean and unit variance. This is often a prerequisite or beneficial step for algorithms whose performance is sensitive to the scale and origin of input features."

— Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

10. Quote (Justifying Scaling After Split):

"To prevent data leakage and obtain an unbiased estimate of generalization performance, any data transformations, including scaling, must be fit only on the training data and then applied consistently to both the training and test sets."

— Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

11. Quote (Justifying Scaling for Specific Models - Linear/KNN/GPR/SVR):

"Feature scaling is essential for regularized linear models (Lasso, Ridge, ElasticNet) where penalty terms are scale-dependent, distance-based algorithms (KNN) where feature scales influence neighborhood definitions, and kernel methods (GPR, SVR) where dot products or distance calculations are sensitive to input magnitudes."

— Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

12. Quote (Justifying When to Address Multicollinearity - Interpretability Focus):

"Multicollinearity primarily affects the stability and interpretability of individual coefficient estimates, especially in linear models, making it critical to address when causal explanation or feature importance is a primary goal. While prediction accuracy might be less affected in some complex or regularized models, unreliable coefficients hinder interpretation."

— Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. The review of Economics and Statistics, 92-107. (Classic paper; concept echoed in modern texts like ISLR). (Note: KNN/GPR/SVR interpretation via methods like SHAP can also be affected by collinearity).

13. Quote (Justifying VIF Calculation on Numerical Features Only):

"The Variance Inflation Factor (VIF) is designed to diagnose collinearity among a set of continuous or quantitative predictors. Assessing multicollinearity involving categorical predictors often requires careful consideration or alternative methods like the Generalized VIF (GVIF)."

— Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. Journal of the American Statistical Association, 87(417), 178-183.

14. Quote (Justifying VIF > 10 as a Threshold):

"A common rule of thumb suggests that VIF values exceeding 10 indicate potentially problematic levels of multicollinearity, where the variance of the estimated regression coefficient is inflated substantially due to correlation with other predictors."

— Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis (7th ed.). Prentice Hall.

15. Quote (Justifying Variable Removal vs. PCA for Interpretability):

"While Principal Component Analysis (PCA) can effectively mitigate multicollinearity by creating orthogonal components, these components are linear combinations of the original variables, sacrificing direct interpretability. When the primary goal is explanation rather than pure prediction, variable selection or removal based on collinearity diagnostics and domain knowledge may be preferred to maintain model interpretability."

— Jolliffe, I. T. (2002). Principal Component Analysis. Springer. (Concept also central to interpretable ML literature).

16. Quote (Justifying Repeated Nested Cross-Validation):

"Repeating the entire Nested Cross-Validation process multiple times with different random partitions of the data provides a more stable and reliable estimate of the model's generalization performance and its variability, reducing the influence of any single random split."

— Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics, 6(1), 10.

17. Quote (Justifying GridSearchCV within Nested CV):

"Nested Cross-Validation employs an outer loop to estimate generalization error and an inner loop for model selection or hyperparameter tuning (e.g., using GridSearchCV). Crucially, the inner loop operates only on the training data portion defined by the outer loop's current split, preventing hyperparameter optimization bias from affecting the outer loop's performance estimate."

— Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079-2107.

18. Quote (Justifying the (mean - std) Filter and Threshold 0.40):

"Evaluating models based on a conservative metric like mean_performance - std_dev_performance from cross-validation explicitly accounts for performance variability and provides a more robust lower bound estimate. Thresholds on such metrics (e.g., $R^2 \geq 0.40$) can be employed, particularly in applied domains like public health or policy, to ensure a minimum level of practical significance and stability."

— Based on principles in: Steyerberg, E. W. (2019). Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer. (While specific mean-std formulas aren't always named, the principle of conservative estimates and practical thresholds is common).

19. Quote (Justifying Removal if Filter Fails Even Once):

"For scientific rigor and ensuring model reliability, particularly when results may inform policy or safety measures, models should demonstrate consistently acceptable performance across different data perturbations or random initializations. Failure to meet stability and performance criteria under certain valid evaluation conditions (e.g., different random seeds in repeated CV) can indicate underlying model instability or unreliability."

— Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8), e124. (While about research findings broadly, the principle of demanding consistency and robustness applies directly to model validation for reliable science).

20. Quote (Justifying full_report for Bias-Variance/Overfitting Analysis):

"Learning curves, which plot model performance on training and validation sets as a function of training set size, are essential diagnostic tools. The gap between curves indicates variance (overfitting), while convergence at a low performance level suggests high bias (underfitting). Comparing training, validation (CV), and test set errors further aids in diagnosing the bias-variance trade-off."

— Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. T. (2012). Learning from data. AMLBook.

21. Quote (Justifying Manual Tuning in Second Stage):

"Model development is often an iterative process. While automated hyperparameter search methods like GridSearch provide systematic exploration, subsequent manual adjustments guided by diagnostic tools (e.g., learning curves, residual analysis) and expert knowledge are frequently necessary to fine-tune complexity and achieve an optimal, well-generalized model."

— Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press.

22. Quote (Justifying Sufficiency of the Three Evaluation Methods):

"A comprehensive and robust model evaluation often benefits from triangulation, employing multiple complementary methods. Combining rigorous cross-validation (like Repeated Nested CV) for unbiased generalization and stability assessment, hold-out test set evaluation for final performance reporting, and techniques like bootstrapping to quantify estimate uncertainty provides a multifaceted and credible assessment of model validity."

— Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press. (The principle of combining methods for a full picture).