

Robust and Reproducible Prediction of HDL Cholesterol Under Outcome Perturbation: An NHANES Machine Learning Study

Francis Okyere | Sherif Mohammed
Department of Statistics, Florida State University

Abstract

This study evaluates the robustness of predictive models for HDL cholesterol using NHANES data under outcome perturbation. Linear, regularized, and ensemble methods were compared using cross-validation. XGBoost achieved the strongest accuracy and robustness, while anthropometric and demographic predictors remained stable, supporting reliable and reproducible health prediction.

1. Introduction

High-density lipoprotein cholesterol (HDL-C) is a key biomarker for cardiovascular health and metabolic risk (Gordon et al., 1989). Accurate prediction of HDL-C using large-scale observational health data supports risk stratification and personalized interventions (Hastie, Tibshirani, & Friedman, 2009).

In practice, outcomes may be corrupted by random noise arising from laboratory variability, reporting error, or privacy-preserving perturbations (Dwork & Roth, 2014). This noise can degrade predictive accuracy and destabilize model selection. This study investigates the robustness of multiple statistical and machine learning models to predict HDL-C using NHANES data under increasing levels of outcome perturbation. Our analysis goes beyond leaderboard optimization by emphasizing model stability, interpretability, and reproducibility under realistic outcome perturbations.

2. Data and Preprocessing

The training data were derived from the National Health and Nutrition Examination Survey (NHANES), which provides nationally representative health and nutrition data for the U.S. population (National Center for Health Statistics [NCHS], 2024). The outcome variable was LBDHDD_outcome (mg/dL). The predictors included dietary intake, demographics, and anthropometric measures.

Missing values were imputed using median imputation for continuous variables and mode imputation for categorical variables. Categorical variables were encoded with one-hot encoding, and the numeric predictors were standardized to support stable fitting and fair regularization of the model.

Additional Gaussian noise was injected to assess sensitivity:

$$Y^{(\sigma)} = Y + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

with $\sigma \in \{0, 0.5, 1, 2, 3, 5\}$.

3. Methodology: Modeling, Validation, and Stability

3.1 Candidate Models

We evaluated:

- Ordinary Least Squares (OLS)
- Ridge Regression (Hoerl & Kennard, 1970)
- Lasso Regression (Tibshirani, 1996)

- Elastic Net (Zou & Hastie, 2005)
- Random Forest (Breiman, 2001)
- XGBoost (Chen & Guestrin, 2016)

3.2 Validation and Tuning Strategy

A fixed K -fold cross-validation scheme was used for out-of-fold (OOF) evaluation. Hyperparameters were tuned at baseline noise ($\sigma = 0$) and reused across noise levels to isolate outcome-noise effects (Hastie et al., 2009). Performance was evaluated using RMSE, MAE and R^2 , with RMSE serving as the official competition ranking metric.

3.3 Feature Stability

Elastic Net stability was assessed using bootstrap resampling to quantify how consistently predictors were selected across repeated perturbations. Selection frequency and coefficient magnitude were aggregated into a stability score.

4. Results and Robustness Analysis

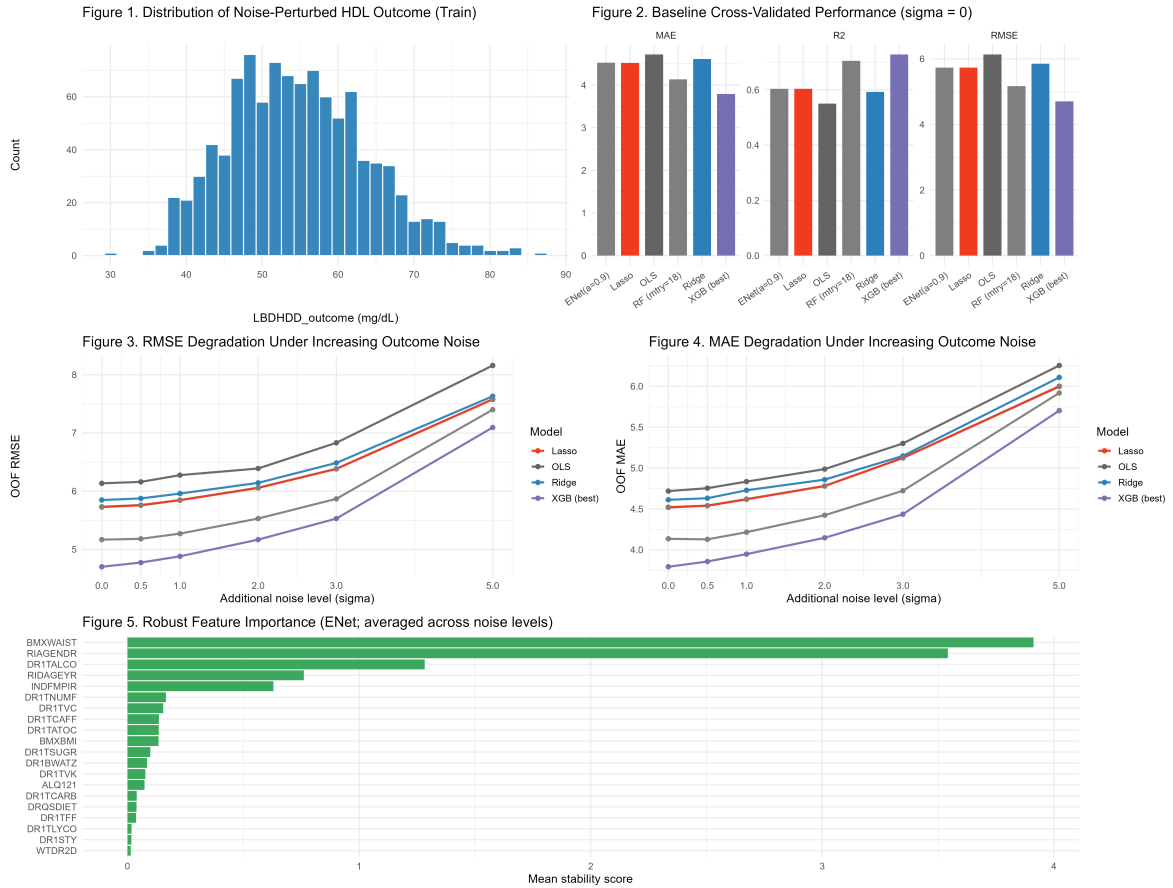


Figure 1: Combined visualization of outcome distribution, baseline performance, degradation under noise, and feature stability.

Figures 1–5 summarize the outcome distribution, baseline predictive performance, robustness to outcome noise, and feature stability of the evaluated models. The HDL-C outcome exhibits an approximately unimodal distribution centered near 55 mg/dL, with moderate dispersion and limited skewness, supporting the use of regression-based modeling approaches.

At the baseline noise level ($\sigma = 0$), XGBoost achieved the strongest predictive accuracy, attaining the lowest RMSE and MAE and the highest R^2 , followed by Random Forest and regularized linear models. Ordinary Least Squares exhibited inferior performance, reflecting its sensitivity to multicollinearity and the high-dimensional predictor structure. These results highlight the importance of regularization and ensemble learning in biomedical prediction tasks.

As outcome noise increased, all models showed monotonic degradation in predictive performance. However, the rate of deterioration differed substantially across methods. OLS was most sensitive to perturbation, exhibiting the steepest increases in RMSE and MAE. In contrast, XGBoost and Random Forest degraded more gradually, demonstrating superior robustness under noisy conditions. Regularized linear models exhibited intermediate behavior, remaining stable under moderate noise but deteriorating under severe perturbation.

Across all noise levels, ensemble methods consistently outperformed linear approaches, particularly under high outcome perturbation. Although Ridge and Elastic Net reduced variance relative to OLS, their performance gap relative to ensemble methods widened as noise increased, indicating limited capacity to capture complex nonlinear relationships under severe perturbation.

Feature stability analysis revealed that waist circumference, gender, alcohol intake, age, and income-to-poverty ratio were selected with high frequency and consistent signs across resampling iterations and noise levels. These predictors exhibited strong stability scores, supporting their interpretability and reliability. In contrast, many dietary micronutrients displayed low selection frequencies and unstable coefficients, suggesting weaker or context-dependent associations with HDL-C.

Overall, these results demonstrate that ensemble-based models, particularly XGBoost, provide superior accuracy, robustness, and reliability for predicting HDL-C under outcome perturbation, while stability analysis enhances interpretability and supports reproducible biomedical modeling.

5. Final Model and Prediction Generation

Based on cross-validation and robustness analysis, XGBoost was selected as the final model. The model was trained on the full training dataset and used to generate predictions for the unlabeled test set. The submission file `pred.csv` contains exactly one column named `pred`, with rows aligned to the test dataset order.

5.1 Code Availability and Reproducibility

All data preprocessing, model training, validation, and prediction steps were implemented in a fully automated and documented pipeline. The complete source code and instructions for reproducing the results and generating the submission file are publicly available on GitHub:

<https://github.com/KojoOkyere/hdl-noise-robust-prediction>.

Running the main script (`main_pipeline.R`) from the project root directory reproduces all reported analyses and outputs.

6. Discussion and Conclusion

Ensemble-based models substantially outperformed classical linear models, reflecting their ability to capture nonlinear interactions and complex predictor dependencies (Breiman, 2001; Chen & Guestrin, 2016). Regularized regression demonstrated improved robustness relative to OLS, consistent with bias–variance trade-off theory (Hastie et al., 2009). Shrinkage stabilizes coefficient estimates under perturbation, though at the cost of

increased bias. The strong stability of anthropometric and demographic predictors is consistent with established relationships between adiposity, age, lifestyle, and lipid metabolism. The reduced stability of many micronutrient variables suggests that single-day dietary recalls may be insufficiently precise for stable prediction. These findings support the deployment of ensemble and shrinkage methods for predictive modeling in noisy or privacy-preserving health environments (Dwork & Roth, 2014). All analyzes were conducted using a fully reproducible and publicly available workflow to ensure transparency and independent verification.

Key findings:

- Best baseline model: XGBoost
- Most robust under noise: XGBoost (RF competitive)
- Stable predictors: waist circumference, gender, alcohol intake, age, income

References

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD*, 785–794.
- Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.
- Gordon, D. J., et al. (1989). High-density lipoprotein cholesterol and cardiovascular disease. *Circulation*, 79(1), 8–15.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression. *Technometrics*, 12(1), 55–67.
- National Center for Health Statistics. (2024). *National Health and Nutrition Examination Survey (NHANES), 2024* [Data set]. U.S. Department of Health and Human Services. <https://github.com/luminwin/ASASF/>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1), 267–288.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2), 301–320.