Predicting HDL Cholesterol from NHANES Variables

MAS 651 Final Submission — 2026 ASA South Florida Student Data Challenge

Daniel Regalado & Miguel Rocha | *University of Miami* | *February 2026*

1. Motivation & Project Overview

Low HDL cholesterol is a well-established risk factor for cardiovascular disease, the leading cause of death worldwide. Predicting HDL from routinely collected survey data — without requiring laboratory lipid panels — can support early screening, risk stratification, and targeted lifestyle interventions in population health settings where direct measurement is unavailable or costly.

This project predicts Direct HDL-Cholesterol (LBDHDD_outcome, mg/dL) from 95 NHANES variables spanning demographics, anthropometrics, dietary intake, and behavioral indicators. The training dataset contains 1,000 observations and the test set contains 200 observations. No missing values were present in either dataset. The target variable is approximately normally distributed (mean = 54.73, median = 54.16, skewness = 0.376), supporting standard regression approaches without target transformation.

2. Key Findings from Exploratory Analysis

Waist circumference (r = –0.596) and BMI (r = –0.484) emerged as the strongest predictors, confirming that central adiposity is the primary driver of HDL variation. Gender showed a strong positive correlation (r = +0.523), with females exhibiting systematically higher HDL levels (mean 59.2 vs. 49.8 mg/dL for males). Dietary variables displayed weak individual correlations ($|r| < 0.19$), suggesting their effects operate through interactions rather than additive relationships. These findings motivated the use of interaction-based feature engineering and nonlinear models.

3. Feature Engineering

A custom sklearn-compatible transformer (FeatureEngineer) was built with leakage-safe design: all thresholds are computed in fit() on training data only. The pipeline generates engineered features across four categories: (1) **Sex interactions** — Waist × Sex, BMI × Sex, Age × Sex, Alcohol × Sex, Race × Sex, Income × Sex, Food Diversity × Sex, and Fish × Sex, capturing the strong gender-dependent effects identified in EDA; (2) **Body composition interactions** — Waist × BMI and Age × Waist; (3) **Polynomial terms** — Waist^2, BMI^2, Age^2 for nonlinear effects; (4) **Log transforms** applied to highly skewed dietary variables. Permutation importance analysis identified 34 raw features and 3 engineered features with zero or negative contribution. These were systematically removed, reducing the pipeline from 107 to 73 input features — improving both generalization and training efficiency.

4. Modeling Approach & Validation Strategy

For the final submission, all 1,000 training observations were used for model fitting (no held-out validation split) to maximize the information available to each model. Performance estimates are obtained exclusively through 5-fold cross-validation with out-of-fold predictions, providing unbiased error estimates while allowing all data for final training. Hyperparameters for tree-based models (CatBoost, XGBoost, LightGBM) were optimized using Optuna Bayesian optimization with 5-fold CV. Elastic Net used GridSearchCV, and neural networks used Optuna with early stopping on an internal 85/15 holdout. Eight model configurations were compared:

Model	CV RMSE	CV MAE	CV R^2
CatBoost	4.5339	3.6680	0.7466
Stacking_2 (Cat+XGB)	4.5394	3.6698	0.7460
Stacking_3 (Cat+XGB+LGB)	4.5510	3.6774	0.7446
XGBoost	4.5869	3.7059	0.7406
LightGBM	4.6255	3.7123	0.7362
Elastic Net	5.4302	4.2123	0.6356
NN with Dropout	5.7620	—	—
NN Standard	6.8587	—	—

Table 1: Model comparison sorted by CV RMSE. Best model highlighted in green.

CatBoost achieved the best overall performance (CV RMSE = 4.5339, R^2 = 0.7466), outperforming even the stacking ensembles. The Stacking_2 (Cat+XGB) and Stacking_3 (Cat+XGB+LGB) configurations performed slightly worse (4.5394 and 4.5510, respectively), indicating that combining boosting models through a Ridge meta-learner did not capture additional complementary signal — likely because all three gradient boosting variants extract very similar predictive patterns from the engineered feature space. CatBoost's ordered boosting mechanism and internal regularization proved particularly effective for this dataset size and structure. The model hierarchy reflects the classical bias-variance tradeoff: Elastic Net (high bias, low variance) establishes the linear ceiling; gradient boosting methods progressively reduce bias; and neural networks exhibit high variance given the moderate sample size (1,000 observations, 73 features), explaining their weaker performance on structured tabular data despite dropout regularization improving generalization (5.76 vs. 6.86 RMSE).

5. Final Model Evaluation

Residual diagnostics for CatBoost are computed using cross-validated out-of-fold predictions across all 1,000 training observations, providing an unbiased estimate of model performance. Errors are centered near zero (mean = +0.054) with approximate normality (skewness = 0.171) and no systematic pattern in residuals vs. predicted values. Only 3.3% of observations (33/1,000) exceed the ±2σ threshold, consistent with a well-specified model. The residual standard deviation of 4.534 indicates stable prediction uncertainty. Error analysis reveals that predictions are most accurate in the normal range (HDL 50–60) but deteriorate at extremes (HDL > 70), consistent with regression-to-the-mean effects and the absence of genetic and medication variables in the dataset.

6. Interpretability (SHAP & LIME)

SHAP analysis of the CatBoost model reveals that the top 10 features explain over 60% of total model impact. The dominant predictors are:

Rank	Feature	Mean |SHAP|	Cum. %
1	Gender (RIAGENDR)	1.516	12.8%
2	Waist Circumference	1.121	22.3%
3	Waist²	1.039	31.1%
4	Waist × BMI	1.009	39.6%
5	Age × Sex	0.745	45.9%
6	Alcohol × Sex	0.722	52.0%
7	Alcohol (DR1TALCO)	0.718	58.1%
8	Income × Sex	0.714	64.1%
9	Alcohol Freq × Amount	0.557	68.8%
10	BMI (BMXBMI)	0.515	73.2%

Table 2: Top 10 features by SHAP importance from CatBoost.

The SHAP ranking directly confirms the EDA findings: waist circumference and BMI, identified as the strongest linear predictors during exploratory analysis (r = −0.596 and −0.484), remain the dominant nonlinear drivers. Crucially, engineered interaction terms (Age × Sex, Alcohol × Sex, Income × Sex) rank among the top 10, demonstrating that the feature engineering strategy successfully encoded the conditional effects observed during EDA into features the model can exploit — and explaining why tree models substantially outperform the linear baseline. LIME local explanations are consistent with SHAP: accurate predictions arise when feature contributions balance around typical training patterns, while large errors occur when multiple strong contributors reinforce each other near distribution extremes.

7. Test Predictions & Submission

The final CatBoost model, trained on all 1,000 observations, was used to generate test predictions directly. The predicted distribution closely matches the training distribution: mean = 54.41 mg/dL (training: 54.73), std = 7.60 (training: 9.01), range = [40.15, 75.31] mg/dL. The slightly lower standard deviation in predictions is expected, as tree-based models produce predictions bounded by observed training values. The absence of extreme predictions confirms stable generalization within the learned feature space. Predictions were saved as pred.csv with 200 rows and a single column named ‘pred’.

8. Conclusion

Tree-based ensemble methods, particularly **CatBoost**, provided the best predictive performance for HDL cholesterol (CV RMSE = 4.53, R^2 = 0.75). Notably, stacking ensembles did not improve upon CatBoost’s standalone performance, suggesting that the boosting variants capture overlapping rather than complementary predictive patterns. Neural networks underperformed due to the moderate sample-to-feature ratio, confirming that gradient boosting remains the preferred approach for structured tabular datasets of this size. Feature engineering guided by EDA — especially sex-based interaction terms — proved critical, with engineered features occupying 6 of the top 10 SHAP positions. Permutation importance-based feature selection further improved generalization by removing 37 noise features. Remaining prediction errors are concentrated at HDL extremes (<40 and >70 mg/dL) and are attributable to unobserved factors (genetics, medication, exercise) rather than model misspecification. The analysis demonstrates that clinically-motivated feature engineering combined with Bayesian-optimized ensemble methods offers an effective and interpretable framework for health outcome prediction from observational data.

Appendix — Supplementary Tables & Figures

A.1 Optuna Hyperparameter Optimization

Bayesian optimization via Optuna replaced traditional GridSearchCV for tree-based models, enabling efficient exploration of continuous parameter spaces. The table below summarizes the best hyperparameters found:

Model	Best Hyperparameters (Optuna)	CV RMSE
CatBoost	iterations=1596, depth=6, lr=0.0098, l2_leaf_reg=0.09	4.5339
XGBoost	n_estimators=1969, max_depth=4, lr=0.0040, colsample=0.52	4.5869
LightGBM	n_estimators=439, max_depth=6, lr=0.0320, num_leaves=24	4.6255
NN Dropout	2 layers (64, 256 units), dropout=0.4, lr=0.0028, batch=32	5.4529

A.2 Permutation Importance & Feature Selection

Permutation importance identified 34 raw features and 3 engineered features (High Waist, Is Obese, Waist-to-BMI ratio) with zero or negative contribution. All 37 features were removed, reducing dimensionality from 107 to 73 features.

Category	Features Removed	Count
Dietary details	DR1EXMER, DR1_300, DR1_330Z, DR1HELP, DR1TS060, DR1TS100, DR1STY, DR1TP184, DR1TFDFE, DR1TLYCO, DR1TBCAR, DR1TACAR, DR1TCHL, DR1TP226, DR1TMAGN, DR1TCRYP, DR1TZINC, DR1TFF, DR1TIRON, DR1TFOLA, DR1TNIAC, DR1TM221, DR1TVB12, DR1TVD, DR1TM201, DR1TTHEO, DR1TCHOL	27
Survey / behavioral	WTDR2D, DRQSPREP, DBQ095Z, DR1DAY, DRQSDIET, DMDMARTZ, DRD340, ALQ111, DRABF, DR1MRESP	10
Engineered (harmful)	High_Waist, Is_Obese, Waist_to_BMI	3

A.3 Retained Engineered Features

Rank	Feature	Type	Rank	Feature	Type
1	Waist × Sex	Interaction	9	Waist × BMI	Interaction
2	BMI × Sex	Interaction	10	Age × Waist	Interaction
3	Age × Sex	Interaction	11	Alc Freq × Amount	Interaction
4	Alcohol × Sex	Interaction	12	Waist²	Polynomial
5	Race × Sex	Interaction	13	BMI²	Polynomial
6	Income × Sex	Interaction	14	Age²	Polynomial
7	Food Div × Sex	Interaction	15	Log transforms	Transform
8	Fish × Sex	Interaction			

A.4 Residual Diagnostics — CatBoost (Cross-Validated Out-of-Fold)

Statistic	Value	Statistic	Value
Mean Residual	+0.054	CV RMSE	4.5339
Std Deviation	4.534	CV MAE	3.6680
Skewness	0.171	CV R²	0.7466
Outliers (|r| > 2σ)	33/1000 (3.3%)		

A.5 SHAP Feature Importance — Extended Ranking (Top 20)

Rank	Feature	Mean |SHAP|	Rank	Feature	Mean |SHAP|
1	Gender (RIAGENDR)	1.516	11	BMI²	0.447
2	Waist Circumference	1.121	12	Drinking Freq (ALQ121)	0.406
3	Waist²	1.039	13	Food Diversity × Sex	0.388
4	Waist × BMI	1.009	14	Fish × Sex	0.375
5	Age × Sex	0.745	15	Race × Sex	0.268
6	Alcohol × Sex	0.722	16	Waist × Sex	0.257
7	Alcohol (DR1TALCO)	0.718	17	Carbohydrate (DR1TCARB)	0.255
8	Income × Sex	0.714	18	Age (RIDAGEYR)	0.245
9	Alc Freq × Amount	0.557	19	Lutein+Zeaxanthin	0.243
10	BMI (BMXBMI)	0.515	20	Age²	0.230

Engineered features (interaction terms and polynomials) account for 12 of the top 20 positions, validating the EDA-guided feature engineering strategy.