HPV Prevalence Data – Exploration & Missing Data Analysis

Feature Engineering and Next Steps

Daljeet Singh

IIT MADRAS

8 Feb

Initial Data Overview I

Initial Variables:

Variable Names
1. Country
2. economy
3. Anemia prevalence among women of reproductive age (% of women ages 15-49)
4. Anemia prevalence among non-pregnant women (% of women ages 15-49)
5. Physicians per 1,000 people
6. Anemia prevalence among pregnant women (%)
7. Continent
8. Mean targeted age
9. Population estimate
10. Sample size studied
11. NCC-16-cases
12. NCC-18-cases
13. Low CIN-16-cases
14. Low CIN-18-cases
15. High CIN-16-cases
16. High CIN-18-cases
17. ICC - 16 - any - cases
18. ICC - 16 - SCC - cases
19. ICC - 16 - ADC - cases
20. ICC - 18 - any - cases
21. ICC - 18 - SCC - cases
22. ICC - 18 - ADC - cases
23. Low CIN prevalence
24. High CIN prevalence
25. ICC prevalence
26. Smoking Prevalence (Current smoking prevalence females, 2016)
27. Total Fertility rate (2017)
28. Contraception use (updated 2019)
29. HIV Prevalence (in adults)
30. Sexual Initiation age
dots... (total 67 Variables)

Data Set

- **Initial Data Shape:** (244 rows, 67 columns)
- **Number of Variables (excluding targets):** 55

Target Variables

12 HPV-related prevalence targets:

1. NCC-16-prevalence
2. NCC-18-prevalence
3. Low CIN-16-prevalence
4. Low CIN-18-prevalence
5. High CIN-16-prevalence
6. High CIN-18-prevalence
7. ICC - 16 - any - prevalence
8. ICC - 16 - SCC - prevalence
9. ICC - 16 - ADC - prevalence
10. ICC - 18 - any - prevalence
11. ICC - 18 - SCC - prevalence
12. ICC - 18 - ADC - prevalence

Dataset Generation

Columns dropped:

- Continent
- NCC-16-cases, NCC-18-cases, Low CIN-16-cases, Low CIN-18-cases
- High CIN-16-cases, High CIN-18-cases
- ICC - 16 - any - cases, ICC - 16 - SCC - cases, ICC - 16 - ADC - cases
- ICC - 18 - any - cases, ICC - 18 - SCC - cases, ICC - 18 - ADC - cases
- ICC prevalence, High CIN prevalence, Low CIN prevalence

Objective: Simplify data to focus on relevant predictors and targets.

Non-Null Counts (Before Country Aggregation)

- NCC-16-prevalence: 99
- NCC-18-prevalence: 94
- Low CIN-16-prevalence: 63
- Low CIN-18-prevalence: 53
- High CIN-16-prevalence: 71
- High CIN-18-prevalence: 65
- ICC - 16 - any - prevalence: 89
- ICC - 16 - SCC - prevalence: 69
- ICC - 16 - ADC - prevalence: 42
- ICC - 18 - any - prevalence: 87
- ICC - 18 - SCC - prevalence: 68
- ICC - 18 - ADC - prevalence: 42

Non-Null Counts (After Country Aggregation)

- NCC-16-prevalence: 85
- NCC-18-prevalence: 80
- Low CIN-16-prevalence: 60
- Low CIN-18-prevalence: 52
- High CIN-16-prevalence: 66
- High CIN-18-prevalence: 62
- ICC - 16 - any - prevalence: 88
- ICC - 16 - SCC - prevalence: 69
- ICC - 16 - ADC - prevalence: 42
- ICC - 18 - any - prevalence: 86
- ICC - 18 - SCC - prevalence: 68
- ICC - 18 - ADC - prevalence: 42

Number of relevant features (post-cleaning): 37

Data Shape After Aggregation

- **Final Data Shape:** (192, 49)
- This includes:
  - 37 features
  - 12 target variables

Missing Value Counts (After Aggregation) I

Column	Missing
Reported STIs (in%)	184
Multiple pregnancies (%age)	184
Sexual Initiation age	175
Sample size studied	152
ICC - 18 - ADC - prevalence	150
ICC - 16 - ADC - prevalence	150
Low CIN-18-prevalence	140
Low CIN-16-prevalence	132
High CIN-18-prevalence	130
High CIN-16-prevalence	126
ICC - 18 - SCC - prevalence	124
ICC - 16 - SCC - prevalence	123
NCC-18-prevalence	112
NCC-16-prevalence	107
ICC - 18 - any - prevalence	106
ICC - 16 - any - prevalence	104
Physicians per 1,000 people	97
HPV vaccination introduction	83
Population estimate	65
HPV Vaccine	62
Mean targeted age	57
HIV Prevalence (in adults)	47
Smoking Prevalence (Current smoking prevalence females, 2016)	45
Male circumcision category	28
Age adjusted incidence (standardized rates)	18
Number of deaths (all ages, 2021)	18
Mortality rates (age standardized)	18
Condom Use	11
Total Fertility rate (2017)	11
Contraception use (updated 2019)	9
Anemia prevalence among non-pregnant women (% of women ages 15-49)	6
Anemia prevalence among pregnant women (%)	6
Coverage in last 3 years of women 25 - 65 years (%)	6
Coverage in last 5 years of women 25 - 65 years (%)	6
Coverage ever screened of women 25 - 65 years (%)	6
Coverage in last 3 years of women 30 - 49 years (%)	6
Coverage in last 5 years of women 30 - 49 years (%)	6
Coverage ever screened of women 30 - 49 years (%)	6
Anemia prevalence among women of reproductive age (% of women ages 15-49)	6
Gross national income (GNI) per capita	3
Human Development Index (HDI)	3
Mean years of schooling	3
Diabetes Prevalence	2
Hypertension	2
Expected years of schooling	2
Life expectancy at birth	2
Mean marital age	1
Incidence of TB	0
Start of Screening coverage (year)	0

Missing ness Highlights & Mechanism

- **Highest Missingness Targets:**
  - ICC - 16 - ADC - prevalence ($\sim 82.8\%$)
  - ICC - 18 - ADC - prevalence ($\sim 82.8\%$)
- **Moderate Missingness:**
  - NCC-16-prevalence ($\sim 59\%$)
  - NCC-18-prevalence ($\sim 61\%$)
- **Systematic Patterns:**
  - Strong correlations among related targets (e.g., Low CIN-16 vs. Low CIN-18: 0.89)
  - Suggests data are likely **Missing at Random (MAR)**, not purely MCAR

Correlation of Missingness with Features

- **NCC-16-prevalence**: correlated with
  - Reported STIs (in%): 0.4565
  - HDI: 0.3236
- **Low/High CIN-16 or CIN-18**: strong correlation with
  - Multiple pregnancies (%age): 0.54–0.68
- **ICC - 16 - ADC** and **ICC - 18 - ADC**:
  - Hypertension, HPV vaccination introduction, Population estimate

Reinforces MAR assumption; missingness relates to socioeconomic/health factors.

Implications & Recommendations

1. **Potential Bias:** Missing data correlated with covariates can bias naive methods.
2. **Advanced Imputation:** Use multiple imputation or model-based methods for MAR data.

Feature Engineering: Aggregated Scores I

(a) Screening Coverage Score

- Columns:
  - Coverage ever screened of women 25–65 years (%)
  - Coverage in last 5 years of women 25–65 years (%)
  - Coverage in last 3 years of women 25–65 years (%)
  - HPV Vaccine coverage
- Method: Mean (skip NA)

(b) Healthcare Access Score

- Columns:
  - Physicians per 1,000 people
  - Mean targeted age for screening programs
- Method: Mean (skip NA)

(c) Anemia Prevalence Score

- Columns:
  - Anemia prevalence among pregnant women
  - Anemia prevalence among non-pregnant women

Feature Engineering: Aggregated Scores II

- Method: Mean (skip NA)

(d) Disease Incidence Score

- Columns:
  - Incidence of TB
  - Hypertension
  - Diabetes Prevalence
- Method: Mean (skip NA)

Model Results For Various Targets and Parity Plots

15 Feb 2025

Modeling Continue 15 Feb

- Four target variables:
  1. NCC_combined
  2. Low_CIN_combined
  3. High_CIN_combined
  4. ICC_combined
- 15% gold (hold-out) set, 85% training set
- 20-fold cross-validation on the training set

Dataset Shapes & Features

Shapes:

- NCC_combined: (99, 41)
- Low_CIN_combined: (63, 41)
- High_CIN_combined: (71, 41)
- ICC_combined: (90, 41)

Columns:

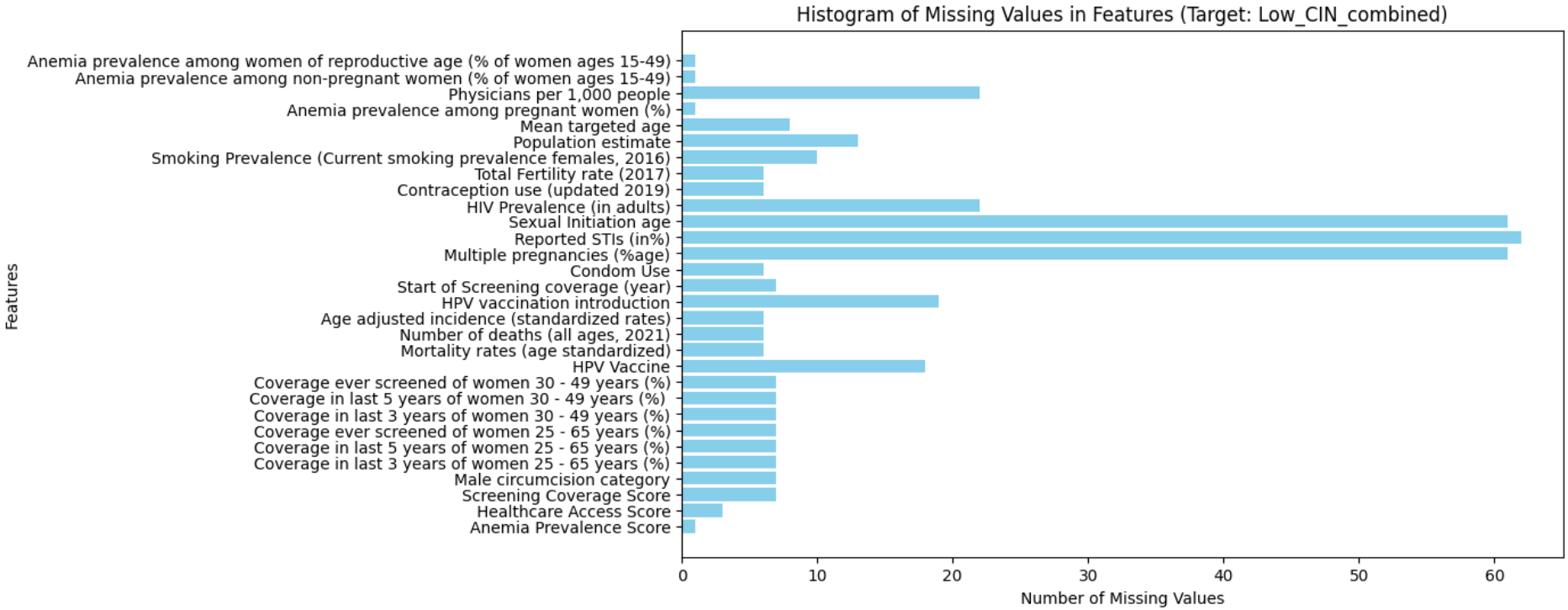
- Numeric: 40 columns (e.g., Anemia prevalence, Physicians per 1,000, etc.)
- Categorical: `Male circumcision category`
- Number of features: 40 (36 original + 4 aggregate ones)

Features List

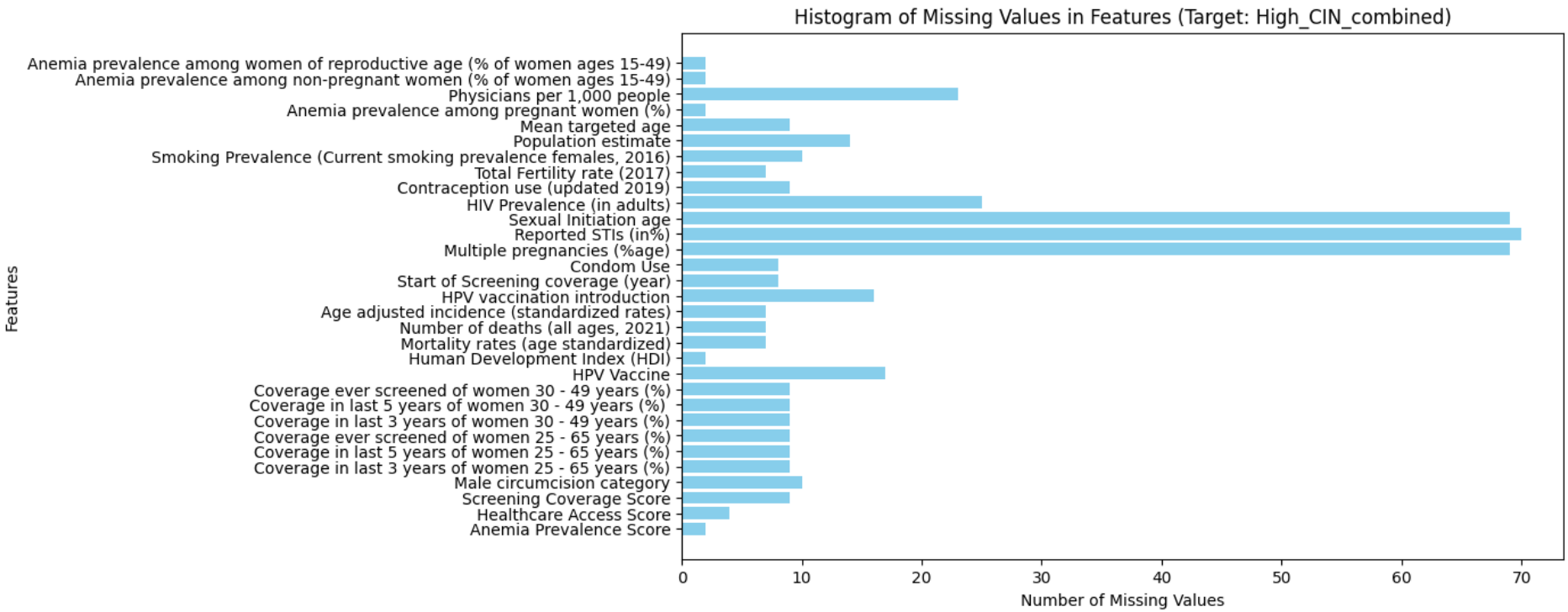
1. Anemia prevalence among women of reproductive age (% of women ages 15-49)
2. Anemia prevalence among non-pregnant women (% of women ages 15-49)
3. Physicians per 1,000 people
4. Anemia prevalence among pregnant women (%)
5. Mean targeted age
6. Population estimate
7. Smoking Prevalence (Current smoking prevalence females, 2016)
8. Total Fertility rate (2017)
9. Contraception use (updated 2019)
10. HIV Prevalence (in adults)
11. Sexual Initiation age
12. Reported STIs (in%)
13. Mean marital age
14. Multiple pregnancies (%age)
15. Condom Use
16. Start of Screening coverage (year)
17. HPV vaccination introduction
18. Age adjusted incidence (standardized rates)
19. Number of deaths (all ages, 2021)
20. Mortality rates (age standardized)
21. Human Development Index (HDI)
22. Life expectancy at birth
23. Expected years of schooling
24. Mean years of schooling
25. Gross national income (GNI) per capita
26. Incidence of TB
27. Diabetes Prevalence
28. HPV Vaccine
29. Hypertension
30. Coverage ever screened of women 30 - 49 years (%)
31. Coverage in last 5 years of women 30 - 49 years (%)
32. Coverage in last 3 years of women 30 - 49 years (%)
33. Coverage ever screened of women 25 - 65 years (%)
34. Coverage in last 5 years of women 25 - 65 years (%)
35. Coverage in last 3 years of women 25 - 65 years (%)
36. Male circumcision category
37. Screening Coverage Score
38. Healthcare Access Score
39. Anemia Prevalence Score
40. Disease Incidence Score

NCC_combined: Missingness Histogram

Low_CIN_combined: Missingness Histogram



High_CIN_combined: Missingness Histogram



ICC_combined: Missingness Histogram

Train/Gold Splits & Cross-Validation

- 15% Gold set, 85% Train set
- Example: NCC_combined (99 rows)
  - Gold set: 14 rows
  - Train set: 85 rows
- 20-fold cross-validation on each training set

Preprocessing Steps

1. **Imputation**: Median for numeric columns
   - Some columns had no valid data for certain targets, automatically skipped
2. **Encoding**: One-hot encoding of `Male circumcision category`
3. **Scaling**: Standard scaling for numeric features (post-imputation and encoding)

Models Considered

- LinearRegression
- Ridge
- Lasso
- RandomForest
- GradientBoosting
- SVR
- PCR (Principal Component Regression)
- XGBOOST

Hyperparameter Tuning:

- 20-fold cross-validation on the training set
- Best parameters selected by grid search

Number of Observations

Target	Total	Train (before outlier)	Train (after outlier)	Gold
NCC_combined	99	85	83	14
Low_CIN_combined	63	54	48	9
High_CIN_combined	71	61	57	10
ICC_combined	90	77	76	13

NCC_combined Results

Model	CV MSE	Train MSE	Train R^2	Gold MSE	Gold R^2
LinearRegression	58.3162	13.3315	0.5579	41.6454	0.1071
Ridge	31.7203	26.0629	0.1356	43.9126	0.0585
Lasso	31.0699	30.1526	0.0000	47.8523	-0.0259
RandomForest	33.6328	16.5868	0.4499	41.5206	0.1098
GradientBoosting	30.2400	13.8388	0.5410	38.0189	0.1849
SVR	30.8153	18.2792	0.3938	49.0972	-0.0526
PCR	30.0062	28.1699	0.0658	42.7456	0.0835
XGBoost (RMSE=4.2198)	–	7.7413	0.7433	24.9298	0.4655

Low_CIN_combined Results

Model	CV MSE	Train MSE	Train R^2	Gold MSE	Gold R^2
LinearRegression	285.2500	7.5527	0.7295	502.7585	-0.9768
Ridge	28.4951	22.5849	0.1911	333.6145	-0.3117
Lasso	29.2811	23.1528	0.1708	320.6109	-0.2606
RandomForest	29.2032	12.7729	0.5425	319.5467	-0.2564
GradientBoosting	29.2559	21.5348	0.2287	305.6075	-0.2016
SVR	26.3177	22.2788	0.2021	327.7201	-0.2886
PCR	28.5114	25.0385	0.1032	347.6682	-0.3670
XGBoost (RMSE=4.3295)	–	7.5608	0.7292	301.1251	-0.1840

High_CIN_combined Results

Model	CV MSE	Train MSE	Train R^2	Gold MSE	Gold R^2
LinearRegression	111.3750	10.8439	0.8083	412.2227	-1.7259
Ridge	48.8438	38.1793	0.3250	172.5993	-0.1413
Lasso	46.4194	34.6328	0.3877	189.9922	-0.2564
RandomForest	54.2275	23.3681	0.5869	163.9817	-0.0844
GradientBoosting	51.7815	18.4334	0.6741	163.3070	-0.0799
SVR	51.4975	42.4727	0.2491	176.0807	-0.1644
PCR	48.0390	44.9614	0.2051	166.5483	-0.1013
XGBoost (RMSE=6.6719)	–	33.3991	0.4095	172.7375	-0.1423

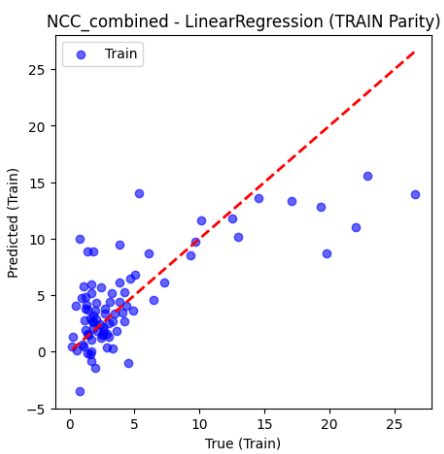
ICC_combined Results

Model	CV MSE	Train MSE	Train R^2	Gold MSE	Gold R^2
LinearRegression	97.9678	18.6618	0.5599	85.1995	-2.9786
Ridge	40.7402	31.6631	0.2533	31.5767	-0.4745
Lasso	37.5741	33.2730	0.2153	29.2705	-0.3668
RandomForest	34.6455	12.8582	0.6968	44.2402	-1.0659
GradientBoosting	35.4696	14.0746	0.6681	39.4455	-0.8420
SVR	39.0530	10.1275	0.7612	27.7302	-0.2949
PCR	41.0111	35.0558	0.1733	37.9919	-0.7741
XGBoost (RMSE=5.5502)	–	12.0076	0.7168	47.6089	-1.2232

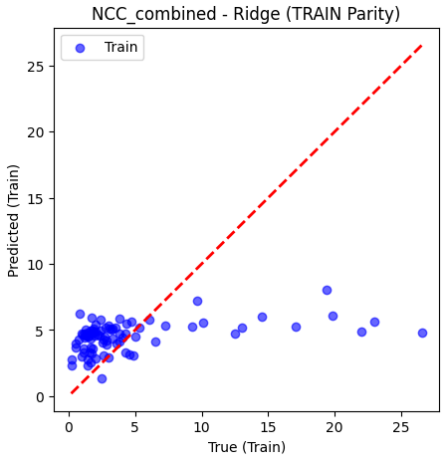
Best Models Summary for All Targets (Based on Gold MSE)

Target	Best Model	Gold MSE	Key Parameters
NCC_combined	XGBoost	24.93	eta=0.1, max_depth=5, gamma=0, min_child_weight=2, subsample=0.8, colsample_bytree=0.8, lambda=2.0, alpha=0.0
Low_CIN_combined	XGBoost	301.13	eta=0.1, max_depth=3, gamma=0.1, min_child_weight=2, subsample=1.0, colsample_bytree=1.0, lambda=1.0, alpha=0.0
High_CIN_combined	GradientBoosting	163.31	learning_rate=0.01, max_depth=3, n_estimators=100
ICC_combined	SVR	27.73	C=10, kernel=rbf

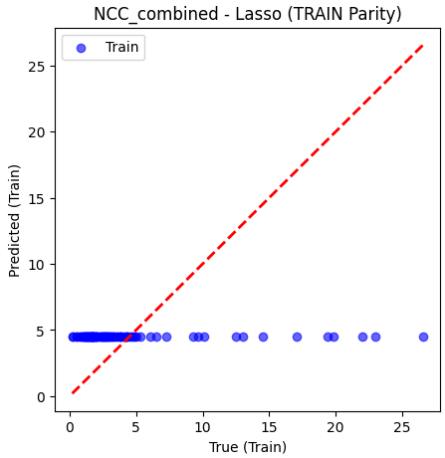
NCC_combined: Train Parity Plots



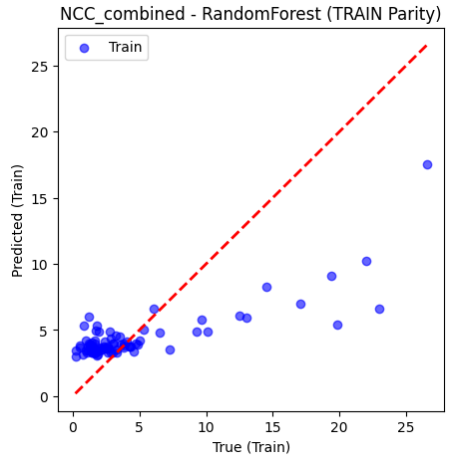
(a) LinearRegression



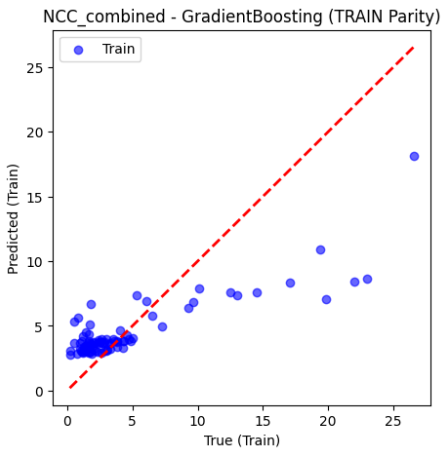
(b) Ridge



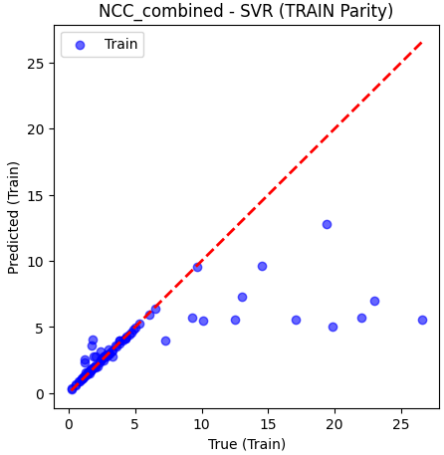
(c) Lasso



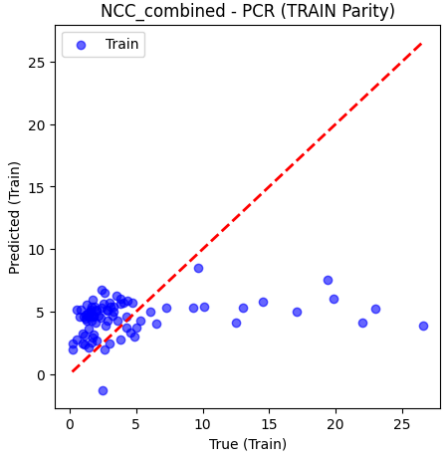
(d) RandomForest



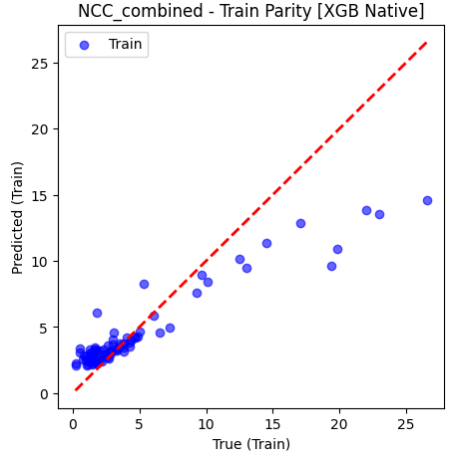
(e) GradientBoosting



(f) SVR

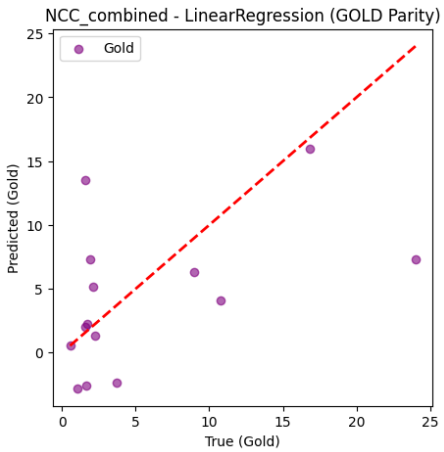


(g) PCR

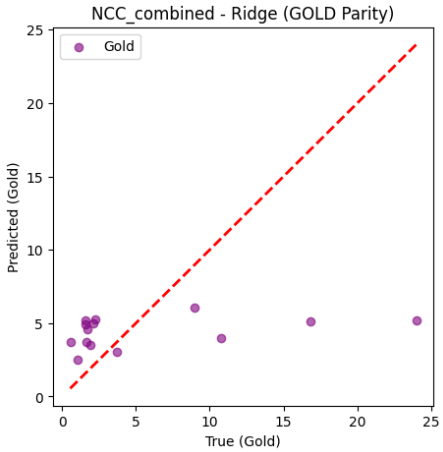


(h) XGBoost

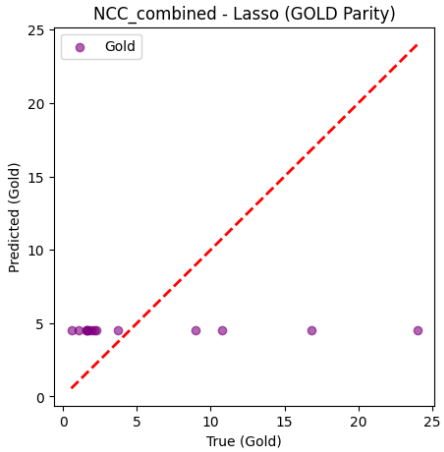
NCC_combined: Gold Test Parity Plots



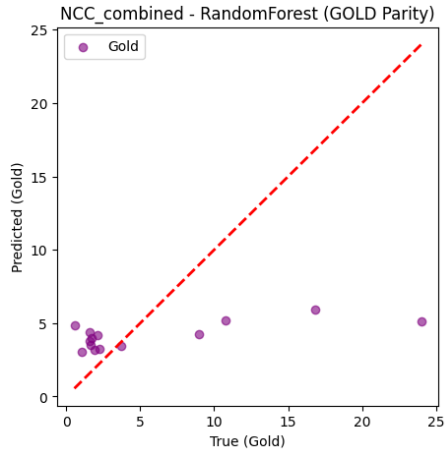
(a) LinearRegression



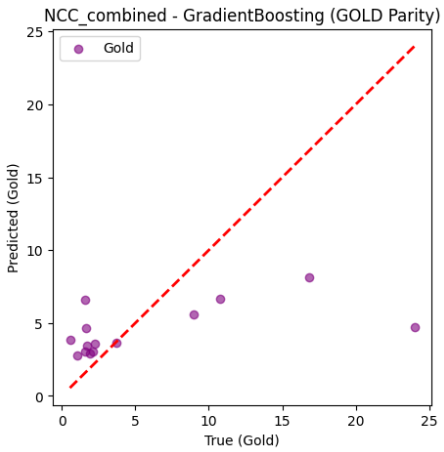
(b) Ridge



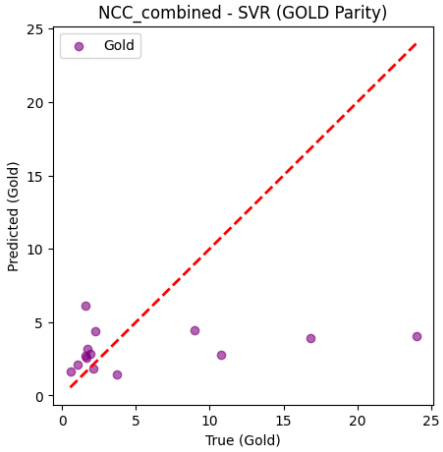
(c) Lasso



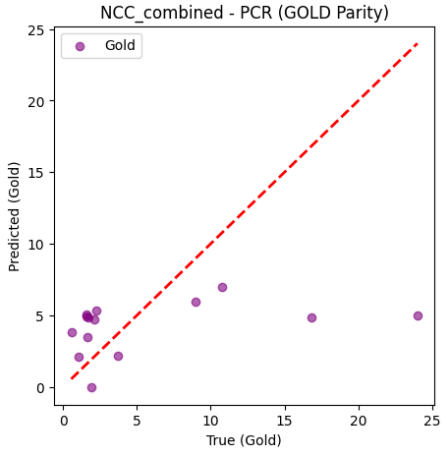
(d) RandomForest



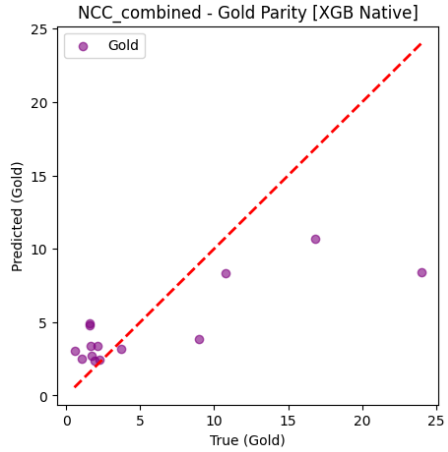
(e) GradientBoosting



(f) SVR

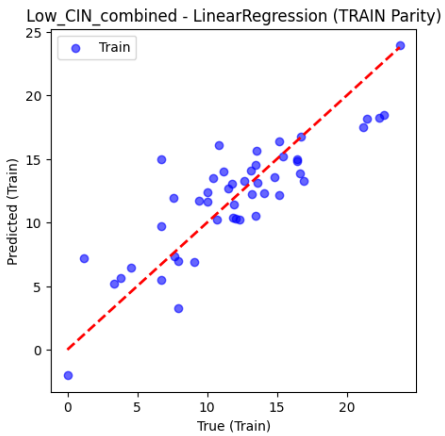


(g) PCR

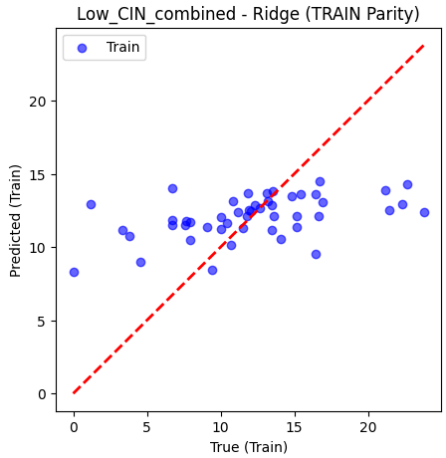


(h) XGBoost

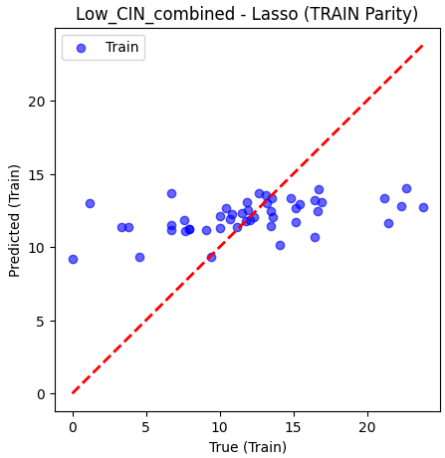
Low_CIN_combined: Train Parity Plots



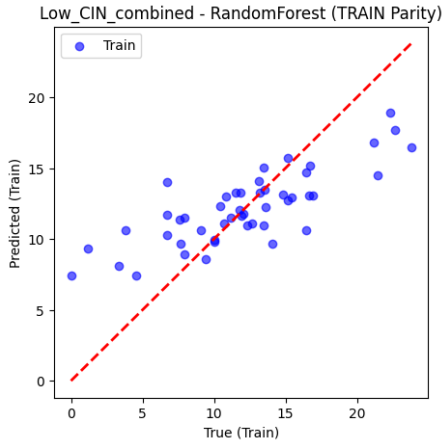
(a) LinearRegression



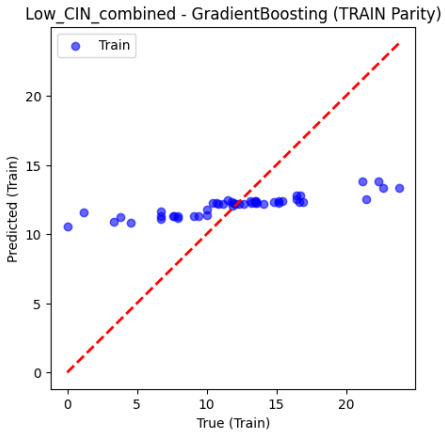
(b) Ridge



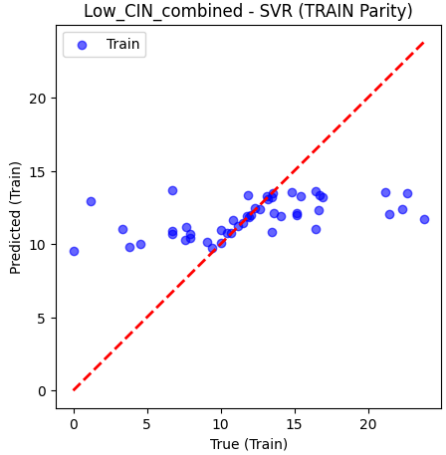
(c) Lasso



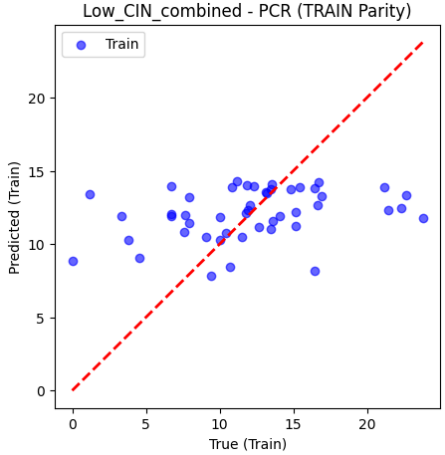
(d) RandomForest



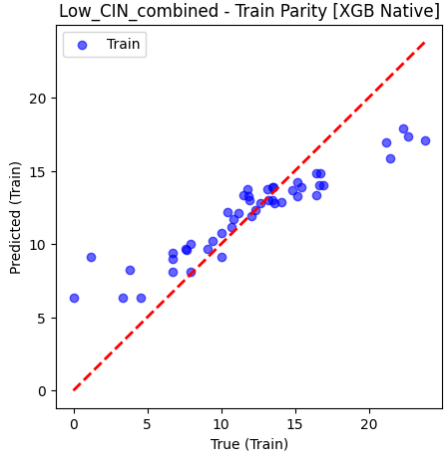
(e) GradientBoosting



(f) SVR

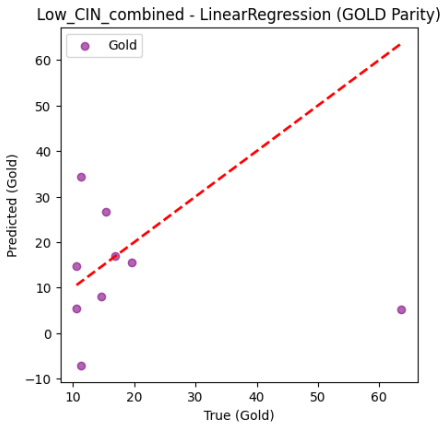


(g) PCR

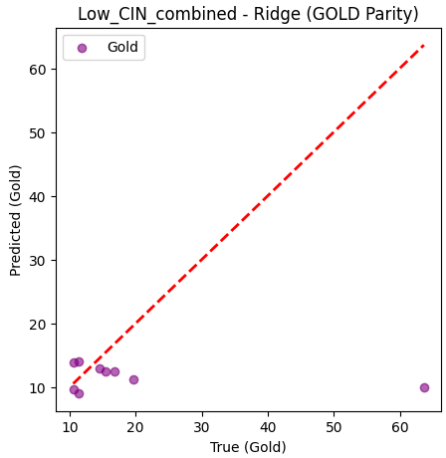


(h) XGBoost

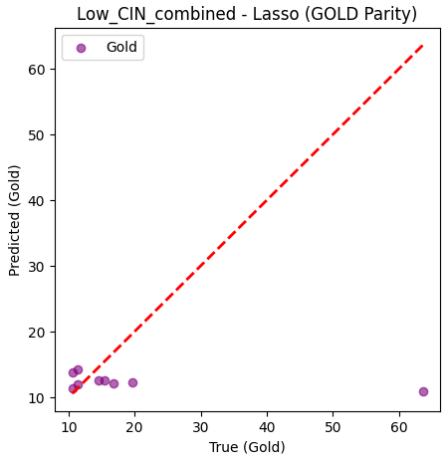
Low_CIN_combined: Gold Test Parity Plots



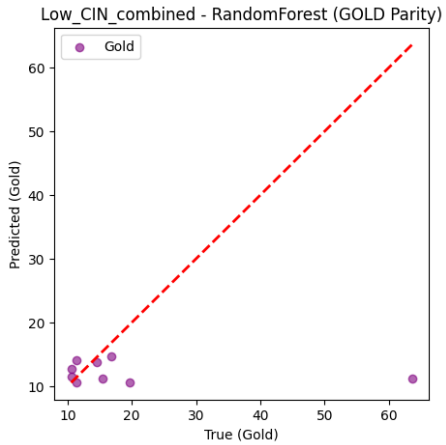
(a) LinearRegression



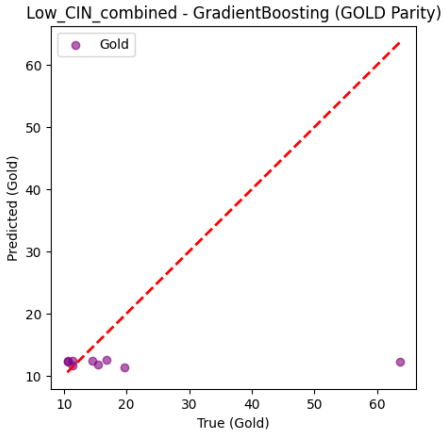
(b) Ridge



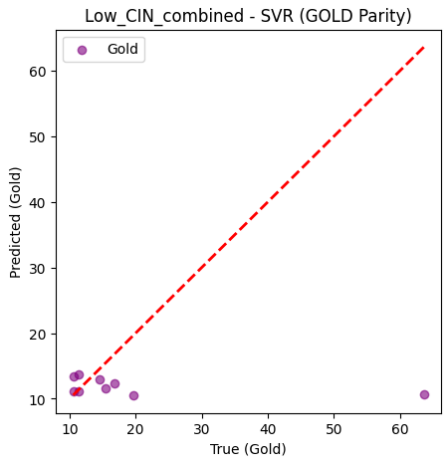
(c) Lasso



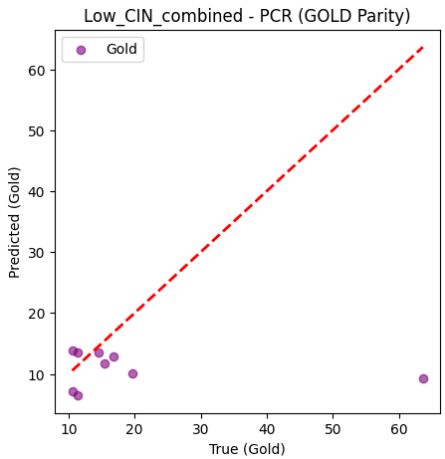
(d) RandomForest



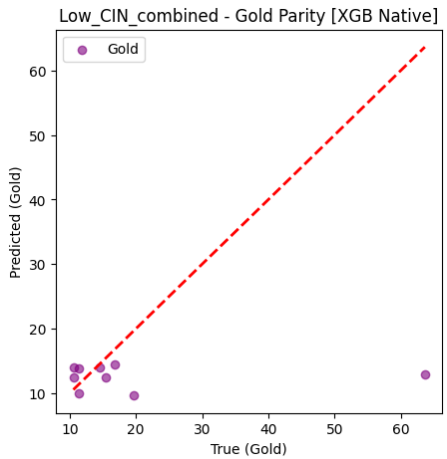
(e) GradientBoosting



(f) SVR

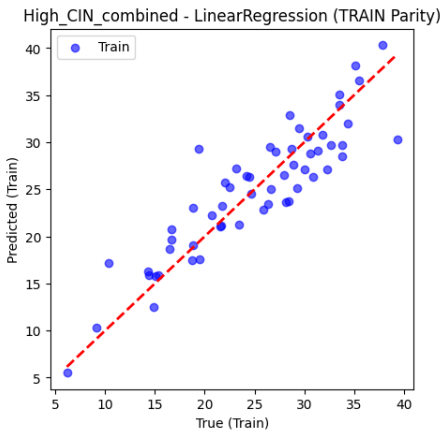


(g) PCR

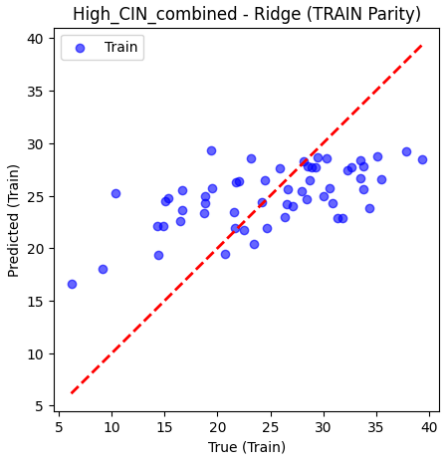


(h) XGBoost

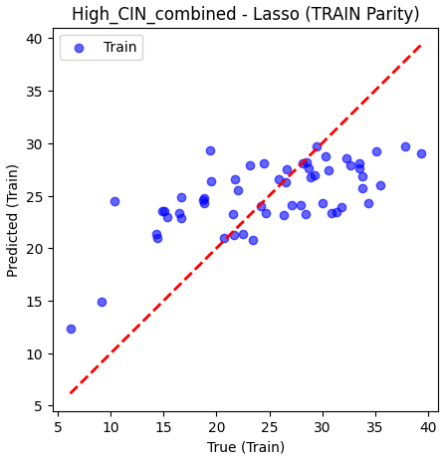
High_CIN_combined: Train Parity Plots



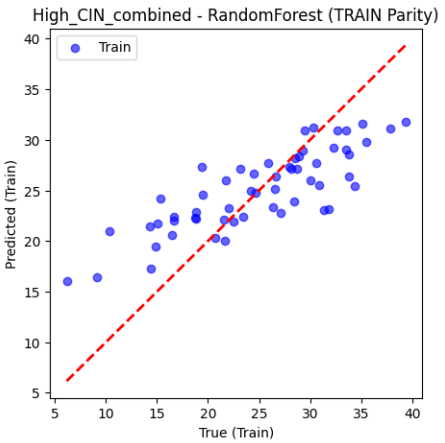
(a) LinearRegression



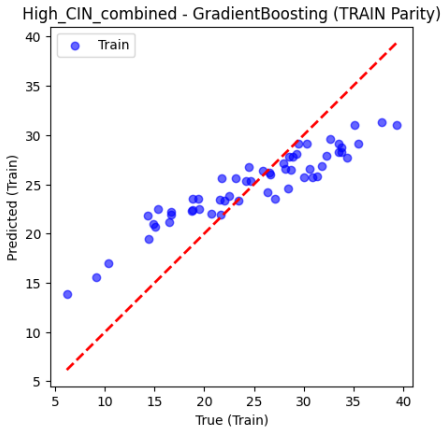
(b) Ridge



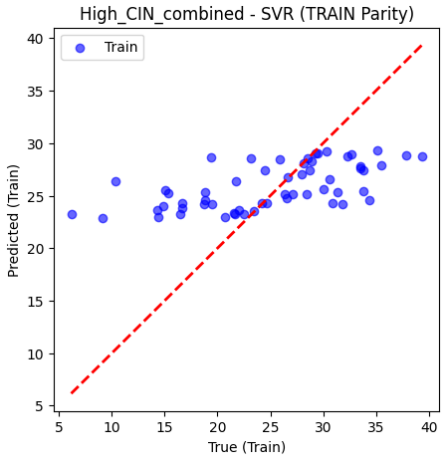
(c) Lasso



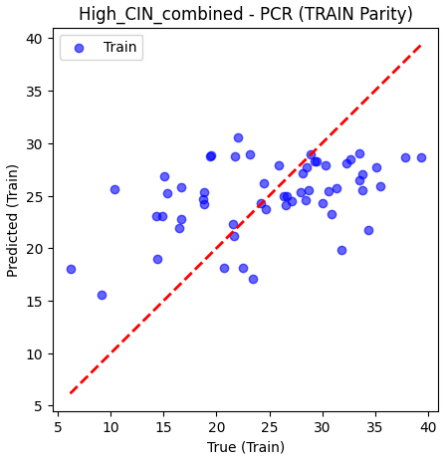
(d) RandomForest



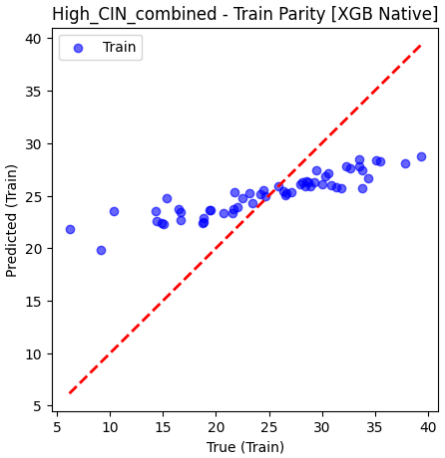
(e) GradientBoosting



(f) SVR

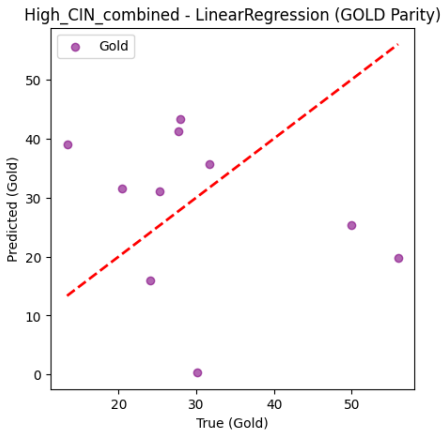


(g) PCR

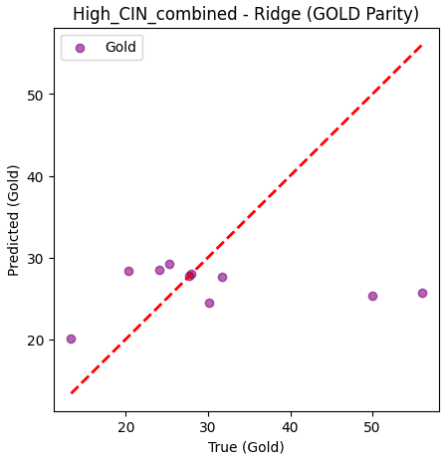


(h) XGBoost

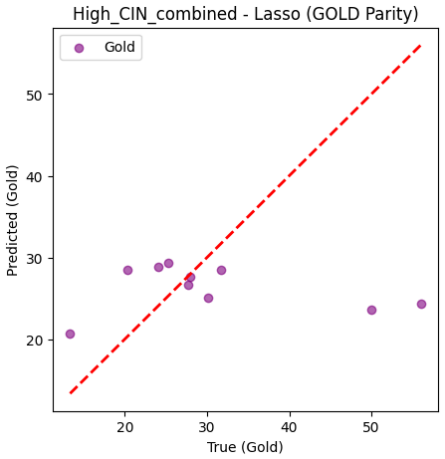
High_CIN_combined: Gold Test Parity Plots



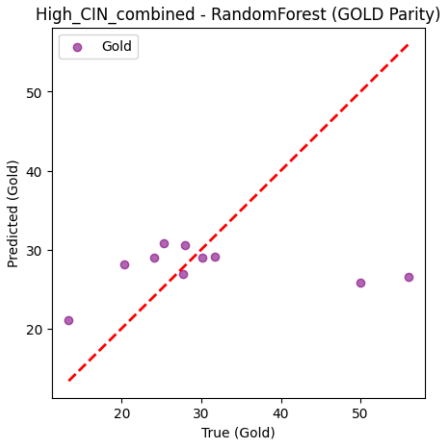
(a) LinearRegression



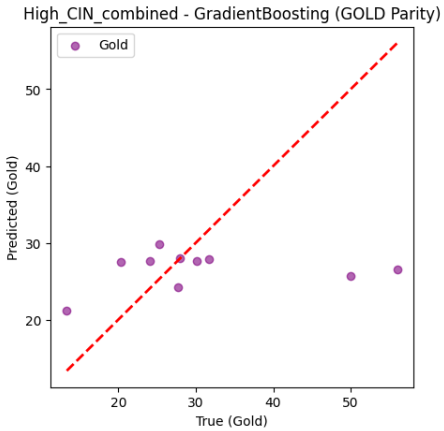
(b) Ridge



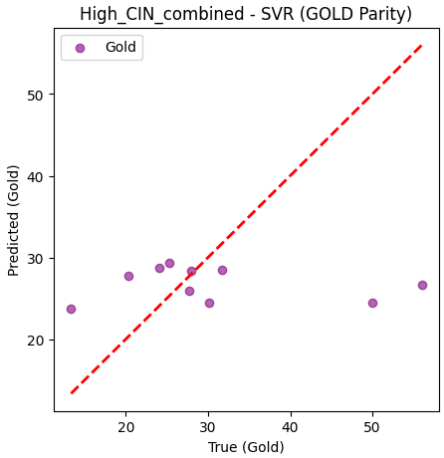
(c) Lasso



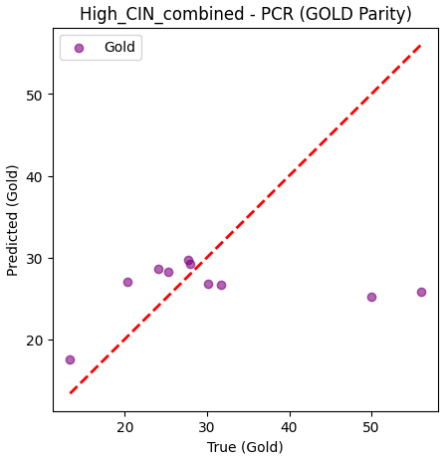
(d) RandomForest



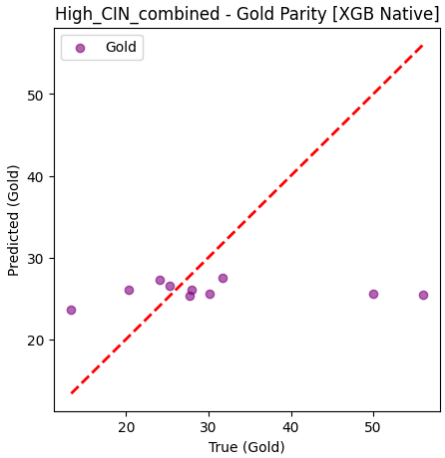
(e) GradientBoosting



(f) SVR

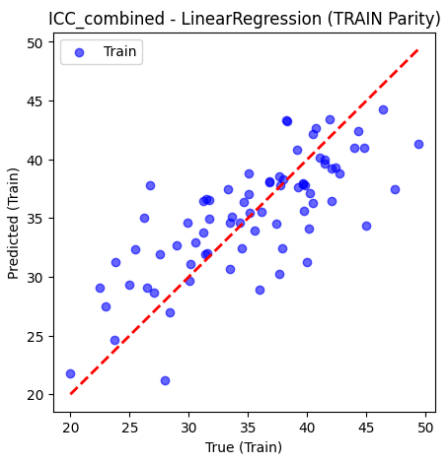


(g) PCR

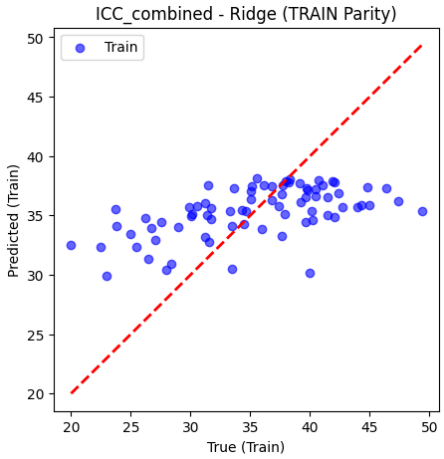


(h) XGBoost

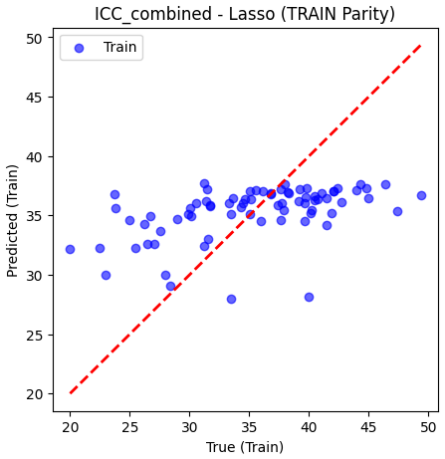
ICC_combined: Train Parity Plots



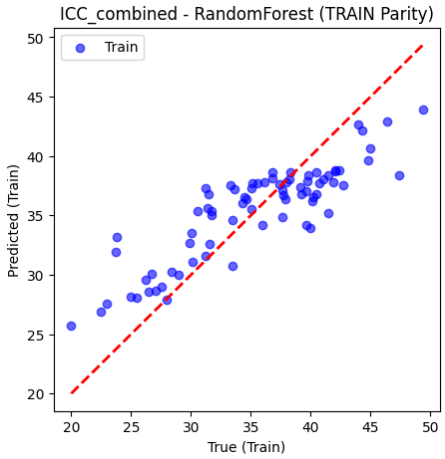
(a) LinearRegression



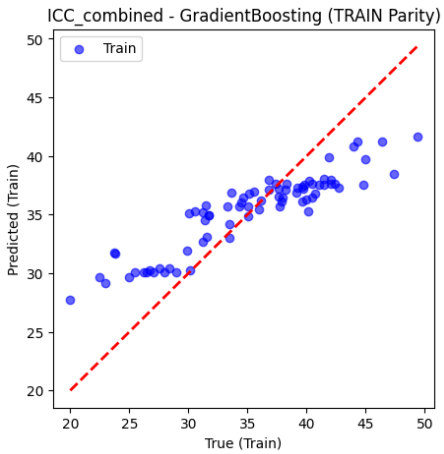
(b) Ridge



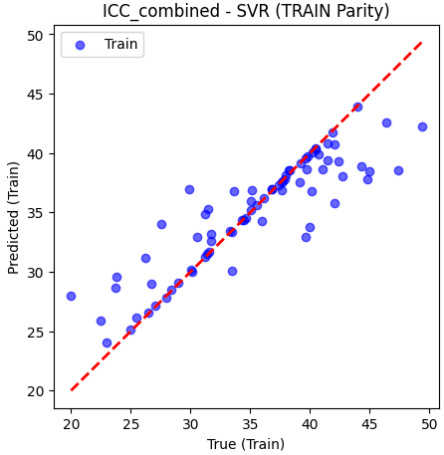
(c) Lasso



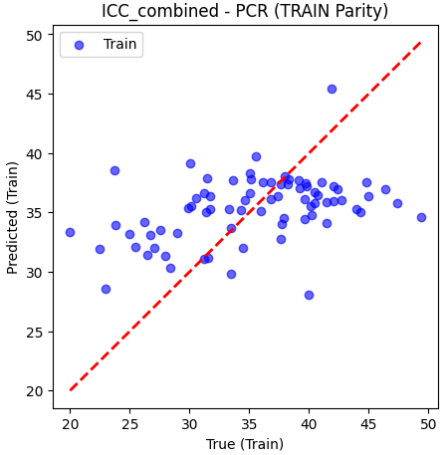
(d) RandomForest



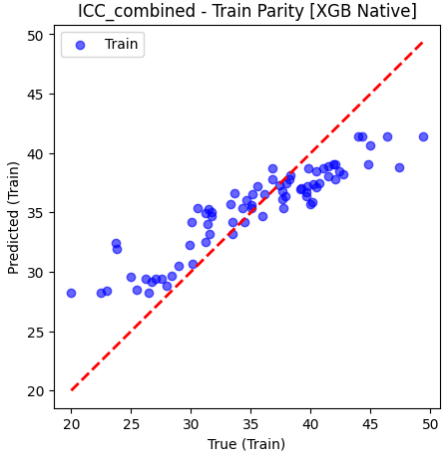
(e) GradientBoosting



(f) SVR

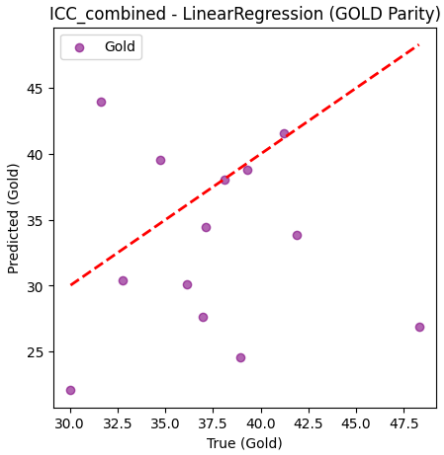


(g) PCR

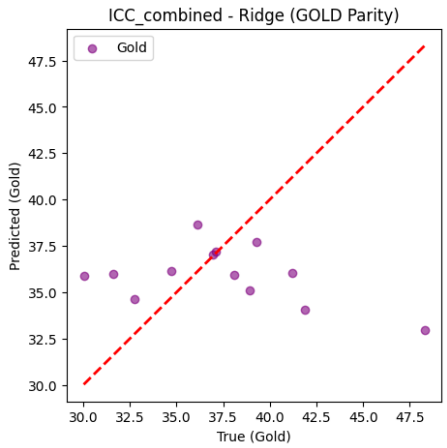


(h) XGBoost

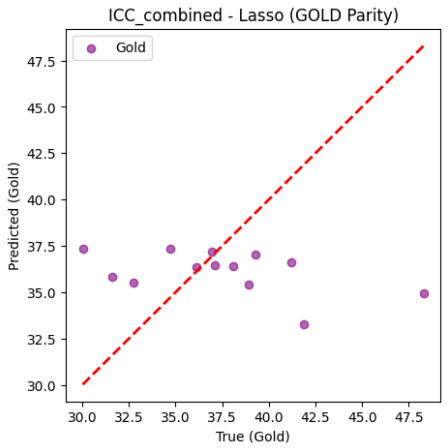
ICC_combined: Gold Test Parity Plots



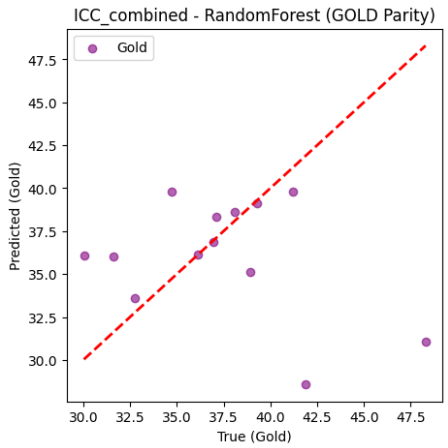
(a) LinearRegression



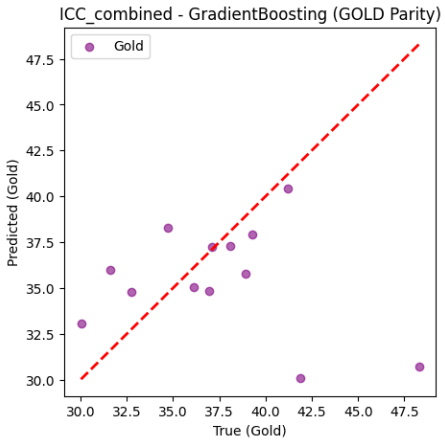
(b) Ridge



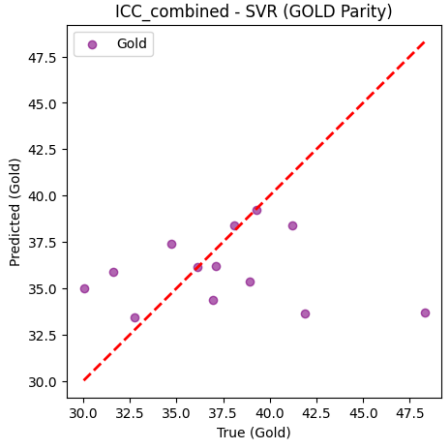
(c) Lasso



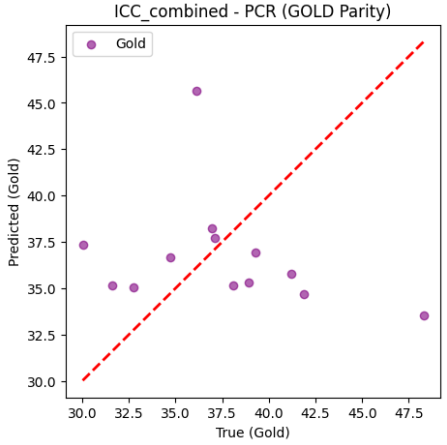
(d) RandomForest



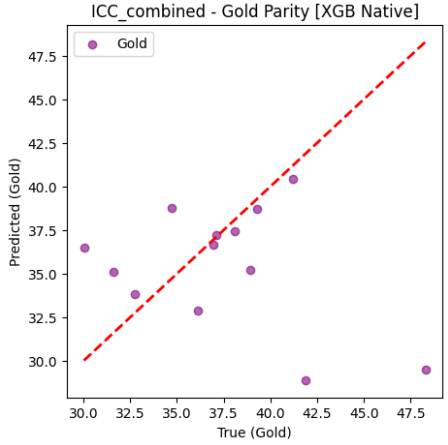
(e) GradientBoosting



(f) SVR



(g) PCR



(h) XGBoost

Checking Train and Test Distributions and further steps

22 Feb 2025

Kolmogorov–Smirnov (KS) Test: Overview

What is it?

A nonparametric test to check if two samples come from the same distribution.

Mechanics:

- Compute *empirical cumulative distribution functions* (CDFs).
- The **KS statistic** is the maximum difference between these CDFs.
- Larger difference implies more dissimilar distributions.

Hypotheses:

- H_0: Samples are from the same continuous distribution.
- H_a: They come from different distributions.

Interpretation:

- If p ¡ 0.05, we typically say there's a **significant difference**.
- If p $\geq$ 0.05, no strong evidence of difference by this test.

KS Test: NCC_combined

Skipped columns: Reported STIs (in%) (All NaNs in gold)

Feature	KS Stat	p-value
Anemia prevalence among women of reproductive age (% of women ages 15-49)	0.3058	1.9789e-01
Anemia prevalence among non-pregnant women (% of women ages 15-49)	0.3058	1.9789e-01
Physicians per 1,000 people	0.3981	1.0177e-01
Anemia prevalence among pregnant women (%)	0.3039	2.0216e-01
Mean targeted age	0.3758	9.9905e-02
Population estimate	0.4987	6.1812e-02
Smoking Prevalence (Current smoking prevalence females, 2016)	0.3387	1.5570e-01
Total Fertility rate (2017)	0.1492	9.7227e-01
Contraception use (updated 2019)	0.3182	2.5679e-01
HIV Prevalence (in adults)	0.1770	9.6918e-01
Sexual Initiation age	0.2727	1.0000e+00
Mean marital age	0.2530	3.6623e-01
Multiple pregnancies (%age)	0.6000	1.0000e+00
Condom Use	0.3249	2.2423e-01
Start of Screening coverage (year)	0.2307	6.1541e-01
HPV vaccination introduction	0.1404	9.9367e-01
Age adjusted incidence (standardized rates)	0.2390	6.3067e-01
Number of deaths (all ages, 2021)	0.2593	5.2506e-01
Mortality rates (age standardized)	0.2051	7.9288e-01
Human Development Index (HDI)	0.2838	3.0638e-01
Life expectancy at birth	0.2334	4.6161e-01
Expected years of schooling	0.3084	1.6393e-01
Mean years of schooling	0.1969	6.7044e-01
Gross national income (GNI) per capita	0.2944	2.0454e-01
Incidence of TB	0.2582	3.3383e-01
Diabetes Prevalence	0.2289	4.8743e-01
HPV Vaccine	0.1694	9.6434e-01
Hypertension	0.1583	8.7570e-01
Coverage ever screened of women 30 - 49 years (%)	0.1912	8.5494e-01
Coverage in last 5 years of women 30 - 49 years (%)	0.2211	7.1443e-01
Coverage in last 3 years of women 30 - 49 years (%)	0.2509	5.6469e-01
Coverage ever screened of women 25 - 65 years (%)	0.1912	8.5494e-01
Coverage in last 5 years of women 25 - 65 years (%)	0.2386	6.2806e-01
Coverage in last 3 years of women 25 - 65 years (%)	0.2439	6.0610e-01
Screening Coverage Score	0.3386	2.2296e-01
Healthcare Access Score	0.2326	4.9984e-01
Anemia Prevalence Score	0.3058	1.9789e-01
Disease Incidence Score	0.2659	3.0999e-01
NCC_combined	0.1971	6.6302e-01

KS Test: Low_CIN_combined

Skipped columns:

- Sexual Initiation age (All NaNs in gold)
- Reported STIs (in%) (All NaNs in train)
- Multiple pregnancies (%age) (All NaNs in gold)

KS Test Results:

Low_CIN_combined

Feature	KS Stat	p-value
Anemia prevalence among women of reproductive age (% of women ages 15-49)	0.1939	8.7893e-01
Anemia prevalence among non-pregnant women (% of women ages 15-49)	0.1939	8.7893e-01
Physicians per 1,000 people	0.2796	7.3439e-01
Anemia prevalence among pregnant women (%)	0.1915	8.8967e-01
Mean targeted age	0.4762	1.0263e-01
Population estimate	0.2895	6.0474e-01
Smoking Prevalence (Current smoking prevalence females, 2016)	0.2724	5.9583e-01
Total Fertility rate (2017)	0.2584	6.0932e-01
Contraception use (updated 2019)	0.5685	8.8796e-03
HIV Prevalence (in adults)	0.3362	3.9228e-01
Mean marital age	0.3472	2.5823e-01
Condom Use	0.3540	2.3931e-01
Start of Screening coverage (year)	0.3988	1.7992e-01
HPV vaccination introduction	0.3229	3.6995e-01
Age adjusted incidence (standardized rates)	0.2868	4.7396e-01
Number of deaths (all ages, 2021)	0.2791	5.1890e-01
Mortality rates (age standardized)	0.3514	2.4708e-01
Human Development Index (HDI)	0.2847	4.8849e-01
Life expectancy at birth	0.3056	4.0151e-01
Expected years of schooling	0.2014	8.6538e-01
Mean years of schooling	0.1389	9.9382e-01
Gross national income (GNI) per capita	0.2222	7.7748e-01
Incidence of TB	0.2847	4.8849e-01
Diabetes Prevalence	0.2431	6.7874e-01
HPV Vaccine	0.3125	4.0290e-01
Hypertension	0.3403	2.7747e-01
Coverage ever screened of women 30 - 49 years (%)	0.2063	8.4912e-01
Coverage in last 5 years of women 30 - 49 years (%)	0.2302	7.4763e-01
Coverage in last 3 years of women 30 - 49 years (%)	0.2381	7.1801e-01
Coverage ever screened of women 25 - 65 years (%)	0.3016	4.2979e-01
Coverage in last 5 years of women 25 - 65 years (%)	0.2143	8.2453e-01
Coverage in last 3 years of women 25 - 65 years (%)	0.2857	5.0125e-01
Screening Coverage Score	0.3016	4.2979e-01
Healthcare Access Score	0.3859	2.0198e-01
Anemia Prevalence Score	0.1939	8.7893e-01
Disease Incidence Score	0.3056	4.0151e-01
Low_CIN_combined	0.3542	2.3763e-01

KS Test: High_CIN_combined

Skipped columns:

- Reported STIs (in%) (All NaNs in gold)

KS Test Results:

High_CIN_combined

Feature	KS Stat	p-value
Anemia prevalence among women of reproductive age (% of women ages 15-49)	0.3000	3.6497e-01
Anemia prevalence among non-pregnant women (% of women ages 15-49)	0.2818	4.4115e-01
Physicians per 1,000 people	0.6667	1.6681e-03
Anemia prevalence among pregnant women (%)	0.2636	5.2517e-01
Mean targeted age	0.2917	4.0701e-01
Population estimate	0.3696	2.4591e-01
Smoking Prevalence (Current smoking prevalence females, 2016)	0.5809	1.0477e-02
Total Fertility rate (2017)	0.3844	1.9667e-01
Contraception use (updated 2019)	0.3578	2.6570e-01
HIV Prevalence (in adults)	0.2115	9.8571e-01
Sexual Initiation age	0.0000	1.0000e+00
Mean marital age	0.3035	3.3594e-01
Multiple pregnancies (%age)	1.0000	1.0000e+00
Condom Use	0.4519	8.5895e-02
Start of Screening coverage (year)	0.2667	5.5726e-01
HPV vaccination introduction	0.3270	4.3860e-01
Age adjusted incidence (standardized rates)	0.3726	2.2528e-01
Number of deaths (all ages, 2021)	0.2665	6.0514e-01
Mortality rates (age standardized)	0.3231	3.7463e-01
Human Development Index (HDI)	0.3909	1.1699e-01
Life expectancy at birth	0.2667	4.9271e-01
Expected years of schooling	0.4614	3.5633e-02
Mean years of schooling	0.5018	1.7294e-02
Gross national income (GNI) per capita	0.4088	8.4195e-02
Incidence of TB	0.3667	1.5725e-01
Diabetes Prevalence	0.5316	9.4856e-03
HPV Vaccine	0.4894	1.6735e-01
Hypertension	0.2368	6.3896e-01
Coverage ever screened of women 30 - 49 years (%)	0.6789	1.3409e-03
Coverage in last 5 years of women 30 - 49 years (%)	0.6789	1.3409e-03
Coverage in last 3 years of women 30 - 49 years (%)	0.5025	4.0386e-02
Coverage ever screened of women 25 - 65 years (%)	0.6397	3.2251e-03
Coverage in last 5 years of women 25 - 65 years (%)	0.6005	7.1850e-03
Coverage in last 3 years of women 25 - 65 years (%)	0.4436	9.4837e-02
Screening Coverage Score	0.5809	1.0477e-02
Healthcare Access Score	0.2226	7.1578e-01
Anemia Prevalence Score	0.2818	4.4115e-01
Disease Incidence Score	0.4193	7.3493e-02
High_CIN_combined	0.2211	7.1443e-01

KS Test: ICC_combined

Skipped columns:

- Sexual Initiation age (All NaNs in train)
- Reported STIs (in%) (All NaNs in train)
- Multiple pregnancies (%age) (All NaNs in train)

KS Test Results:

ICC_combined

Feature	KS Stat	p-value
Anemia prevalence among women of reproductive age (% of women ages 15-49)	0.1622	9.0416e-01
Anemia prevalence among non-pregnant women (% of women ages 15-49)	0.1486	9.4768e-01
Physicians per 1,000 people	0.2391	7.5423e-01
Anemia prevalence among pregnant women (%)	0.1509	9.4017e-01
Mean targeted age	0.1812	8.1001e-01
Population estimate	0.2164	6.1939e-01
Smoking Prevalence (Current smoking prevalence females, 2016)	0.2981	2.4307e-01
Total Fertility rate (2017)	0.2603	3.7243e-01
Contraception use (updated 2019)	0.2752	3.1510e-01
HIV Prevalence (in adults)	0.4439	4.8085e-02
Mean marital age	0.1923	7.2771e-01
Condom Use	0.1408	9.5626e-01
Start of Screening coverage (year)	0.2098	6.6478e-01
HPV vaccination introduction	0.2818	5.3453e-01
Age adjusted incidence (standardized rates)	0.2297	5.2424e-01
Number of deaths (all ages, 2021)	0.2434	4.5242e-01
Mortality rates (age standardized)	0.2244	5.5300e-01
Human Development Index (HDI)	0.2879	2.5838e-01
Life expectancy at birth	0.3310	1.3696e-01
Expected years of schooling	0.3553	9.1777e-02
Mean years of schooling	0.1731	8.3348e-01
Gross national income (GNI) per capita	0.2763	3.0298e-01
Incidence of TB	0.2318	5.0791e-01
Diabetes Prevalence	0.3047	2.0391e-01
HPV Vaccine	0.2281	7.7428e-01
Hypertension	0.2763	3.0298e-01
Coverage ever screened of women 30 - 49 years (%)	0.2429	4.5774e-01
Coverage in last 5 years of women 30 - 49 years (%)	0.2714	3.2667e-01
Coverage in last 3 years of women 30 - 49 years (%)	0.2286	5.3297e-01
Coverage ever screened of women 25 - 65 years (%)	0.2714	3.2667e-01
Coverage in last 5 years of women 25 - 65 years (%)	0.2571	3.8897e-01
Coverage in last 3 years of women 25 - 65 years (%)	0.2143	6.1293e-01
Screening Coverage Score	0.2714	3.2667e-01
Healthcare Access Score	0.2029	6.7985e-01
Anemia Prevalence Score	0.1486	9.4768e-01
Disease Incidence Score	0.2186	5.7947e-01
ICC_combined	0.2389	4.7333e-01

Key Observations

- **NCC_combined**:
  - Many p-values > 0.05; no strong evidence of distribution differences.
- **Low_CIN_combined**:
  - One difference: `Contraception use (updated 2019)`, $p \approx 0.0089$.
- **High_CIN_combined**:
  - Multiple features with $p < 0.05$: e.g., `Physicians per 1,000 people`, `Smoking Prevalence`, `Coverage in last 5 years of women 30 - 49%`, etc.
  - Suggests distribution shift between Train and Gold sets in these features.
- **ICC_combined**:
  - Mostly consistent; borderline difference for `HIV Prevalence` ($p \approx 0.048$).

Further Steps: Stratified Sampling for Regression I

Motivation:

- In regression problems, the target variable is continuous.
- Random splits may not preserve the distribution, especially in the tails.

Proposed Approach:

- **Bin the Target:** Divide the continuous target variable into several bins (e.g., using quantiles).
- **Stratified Sampling:** Perform train-test splitting within each bin to maintain the target distribution in both sets.
- **Benefits:**
  - Ensures that all ranges of the target variable are properly represented.
  - Reduces potential bias and improves model generalizability.

Next Steps:

Further Steps: Stratified Sampling for Regression II

- Evaluate how well the stratified split preserves the overall target distribution.
- Compare model performance using random vs. stratified splits.

Pseudo Huber Error Objective Function

- The Pseudo Huber Error function is used as an objective function in regression models.
- It provides a smooth approximation to the Huber loss, combining the sensitivity of the squared error for small residuals with the robustness of the absolute error for large residuals.
- **Robustness to Outliers:**
  - Reduces the influence of outliers by behaving like the absolute error when errors are large.
  - Maintains differentiability, which helps in gradient-based optimization.
- **Mathematical Formulation:**

$$L_\delta(a) = \delta^2 \left(\sqrt{1 + \left(\frac{a}{\delta}\right)^2} - 1 \right)$$

where a is the residual and δ is a parameter that controls the transition between quadratic and linear loss.

Enhanced Model with Huber Error & Categorical Year Feature

- **Model Enhancements:**
  - Adopted Huber error as the loss function.
  - Converted *Start of Screening Coverage (Year)* to a categorical feature.
- **Optimized XGBoost Parameters:**
  - `eta`: 0.1, `max_depth`: 5, `gamma`: 0.2, `min_child_weight`: 3
  - `subsample`: 1.0, `colsample_bytree`: 0.8, `lambda`: 2.0, `alpha`: 0.0
- **Performance Metrics:**
  - **Best CV RMSE**: 5.6288 (Early stopped at 110 iterations)
  - **Training Set**: MSE $= 20.0005$, $R^2 = 0.6299$
  - **Gold Set**: MSE $= 22.0083$, $R^2 = 0.5281$
- **Observations:**
  - Improved results for NCC.
  - Outliers were retained, yet performance improved.
  - Lower performance for low CIN, high CIN, and ICC cases.

High_CIN_combined Model Training (After Feature Removal)

- **Data Overview:**
  - Total observations: 71
  - **Before Outlier Removal**: Train = 61, Gold = 10
  - **After Outlier Removal (High_CIN_combined ¿ 40)**: Clean Train = 57, Gold = 10
- **Optimized XGBoost Parameters:**
  - `eta`: 0.1, `max_depth`: 5, `gamma`: 0.2, `min_child_weight`: 4
  - `subsample`: 0.8, `colsample_bytree`: 0.8, `lambda`: 1.0, `alpha`: 0.0
- **Performance Metrics:**
  - **Best CV RMSE**: 6.5353 (Early stopped at 12 iterations)
  - **Training Set**: MSE = 18.8819, $R^2 = 0.6662$
  - **Gold Set**: MSE = 146.5856, $R^2 = 0.0307$
- **Observation:** Removal of the *Start of Screening Coverage (Year)* feature significantly affected performance on the Gold set.

Future Work: Investigating Huber Error Performance

- **Objective:**
  - Analyze why the Pseudo Huber error objective function is not delivering favorable results for certain variables.
- **Focus Areas:**
  - Examine the performance discrepancies observed in the low CIN and ICC combined targets.
  - Investigate the causes behind the negative R^2 values for these targets.
- **Planned Analyses:**
  - Conduct in-depth error analysis to understand the behavior of the loss function across different data segments.

1. Outlier Detection (NCC_combined)

- **Outlier Countries:** Angola, Papua New Guinea
- **Before Removal:** 85 rows (train), 14 rows (gold)
- **After Removing** (`NCC_combined > 30`): 83 rows remain (train)

1. Outlier Detection (Low_CIN_combined)

- **Outlier Countries:** Belarus, Colombia, Ecuador, Ethiopia, Latvia, Russian Federation, Serbia
- **Before Removal:** 54 rows (train), 9 rows (gold)
- **After Removing** (`Low_CIN_combined > 25`): 48 rows remain (train)

1. Outlier Detection (High_CIN_combined)

- **Outlier Countries:** Belarus, Belize, Greece, Latvia, Russian Federation
- **Before Removal:** 61 rows (train), 10 rows (gold)
- **After Removing** (`High_CIN_combined > 40`): 57 rows remain (train)

1. Outlier Detection (ICC_combined)

- **Outlier Countries:** Belarus
- **Before Removal:** 77 rows (train), 13 rows (gold)
- **After Removing** (`ICC_combined > 50`): 76 rows remain (train)

2. Forward Selection and Model Performance

- **Outcome of Interest:** `High_CIN_combined`
- **Feature Engineering:** Original variables, ratios, polynomials, and interactions.
- **Positive-Score Set:** A large number of features yielded a strictly positive forward-selection score.
- **Model Performance:** Using only these positively scored features still gave a **negative** R^2 of -0.1638 on the gold (holdout) set.

3. Comprehensive List of Positively Scored Features I

Total Features in List: 384

Below is the **entire set** of features that were scored positively in the forward selection procedure (including original, polynomial, interaction, and ratio features).

- Anemia prevalence among women of reproductive age (% of women ages 15-49)
- Anemia prevalence among non-pregnant women (% of women ages 15-49)
- Physicians per 1,000 people
- Mean years of schooling
- Diabetes Prevalence
- Coverage ever screened of women 30 - 49 years (%)
- Coverage ever screened of women 25 - 65 years (%)
- Anemia Prevalence Score
- Unnamed: 0 Mortality rates (age standardized)
- Unnamed: 0 Total_BothSex2019

3. Comprehensive List of Positively Scored Features II

- Unnamed: 0 Total_Female2019
- Unnamed: 0 RatePer100k_Female2019
- Anemia prevalence among women of reproductive age (% of women ages 15-49)[2]
- Anemia prevalence among women of reproductive age (% of women ages 15-49) Anemia prevalence among non-pregnant women (% of women ages 15-49)
- Anemia prevalence among women of reproductive age (% of women ages 15-49) Physicians per 1,000 people
- Anemia prevalence among women of reproductive age (% of women ages 15-49) Anemia prevalence among pregnant women (%)
- Anemia prevalence among women of reproductive age (% of women ages 15-49) Multiple pregnancies (%age)
- Anemia prevalence among women of reproductive age (% of women ages 15-49) Diabetes Prevalence
- Anemia prevalence among women of reproductive age (% of women ages 15-49) Total_BothSex2019

3. Comprehensive List of Positively Scored Features III

- Anemia prevalence among women of reproductive age (% of women ages 15-49) Anemia Prevalence Score
- Anemia prevalence among non-pregnant women (% of women ages 15-49)^2
- Anemia prevalence among non-pregnant women (% of women ages 15-49) Physicians per 1,000 people
- Anemia prevalence among non-pregnant women (% of women ages 15-49) Anemia prevalence among pregnant women (%)
- Anemia prevalence among non-pregnant women (% of women ages 15-49) Multiple pregnancies (%age)
- Anemia prevalence among non-pregnant women (% of women ages 15-49) Diabetes Prevalence
- Anemia prevalence among non-pregnant women (% of women ages 15-49) Total_BothSex2019
- Anemia prevalence among non-pregnant women (% of women ages 15-49) Total_Female2019

3. Comprehensive List of Positively Scored Features IV

- Anemia prevalence among non-pregnant women (% of women ages 15-49) Anemia Prevalence Score
- Physicians per 1,000 people[2]
- Physicians per 1,000 people Smoking Prevalence (Current smoking prevalence females, 2016)
- Physicians per 1,000 people Multiple pregnancies (%age)
- Physicians per 1,000 people Number of deaths (all ages, 2021)
- Physicians per 1,000 people Human Development Index (HDI)
- Physicians per 1,000 people Expected years of schooling
- Physicians per 1,000 people Mean years of schooling
- Physicians per 1,000 people HPV Vaccine
- Physicians per 1,000 people Screening Coverage Score
- Physicians per 1,000 people Healthcare Access Score
- Physicians per 1,000 people Anemia Prevalence Score
- Anemia prevalence among pregnant women (%) Mean targeted age

3. Comprehensive List of Positively Scored Features V

- Anemia prevalence among pregnant women (%) Smoking Prevalence (Current smoking prevalence females, 2016)
- Anemia prevalence among pregnant women (%) Contraception use (updated 2019)
- Anemia prevalence among pregnant women (%) Diabetes Prevalence
- Anemia prevalence among pregnant women (%) RatePer100k_Female2019
- Anemia prevalence among pregnant women (%) Anemia Prevalence Score
- Mean targeted age Smoking Prevalence (Current smoking prevalence females, 2016)
- Mean targeted age Mean years of schooling
- Population estimate Incidence of TB
- Smoking Prevalence (Current smoking prevalence females, 2016) Contraception use (updated 2019)
- Smoking Prevalence (Current smoking prevalence females, 2016) Coverage ever screened of women 25 - 65 years (%)
- Smoking Prevalence (Current smoking prevalence females, 2016) Coverage in last 5 years of women 25 - 65 years (%)

3. Comprehensive List of Positively Scored Features VI

- Smoking Prevalence (Current smoking prevalence females, 2016) Coverage in last 3 years of women 25 - 65 years (%)
- Smoking Prevalence (Current smoking prevalence females, 2016) Screening Coverage Score
- Total Fertility rate (2017) Contraception use (updated 2019)
- Total Fertility rate (2017) Total_BothSex2019
- Contraception use (updated 2019) Age adjusted incidence (standardized rates)
- Contraception use (updated 2019) Hypertension
- Contraception use (updated 2019) Disease Incidence Score
- Multiple pregnancies (%age) Mean years of schooling
- Multiple pregnancies (%age) Coverage ever screened of women 30 - 49 years (%)
- Multiple pregnancies (%age) Coverage ever screened of women 25 - 65 years (%)
- Multiple pregnancies (%age) RatePer100k_BothSex2019
- Multiple pregnancies (%age) Anemia Prevalence Score
- Condom Use Human Development Index (HDI)

3. Comprehensive List of Positively Scored Features VII

- Condom Use Life expectancy at birth
- Condom Use Expected years of schooling
- Condom Use Gross national income (GNI) per capita
- Age adjusted incidence (standardized rates) Incidence of TB
- Age adjusted incidence (standardized rates) Diabetes Prevalence
- Age adjusted incidence (standardized rates) RatePer100k_Female2019
- Age adjusted incidence (standardized rates) Disease Incidence Score
- Human Development Index (HDI) Mean years of schooling
- Human Development Index (HDI) Gross national income (GNI) per capita
- Human Development Index (HDI) Hypertension
- Mean years of schooling^2
- Mean years of schooling HPV Vaccine
- Mean years of schooling Coverage ever screened of women 30 - 49 years (%)
- Mean years of schooling Coverage in last 5 years of women 30 - 49 years (%)

3. Comprehensive List of Positively Scored Features VIII

- Mean years of schooling Coverage in last 3 years of women 30 - 49 years (%)
- Mean years of schooling Coverage ever screened of women 25 - 65 years (%)
- Mean years of schooling Coverage in last 5 years of women 25 - 65 years (%)
- Mean years of schooling Coverage in last 3 years of women 25 - 65 years (%)
- Mean years of schooling Screening Coverage Score
- Mean years of schooling Healthcare Access Score
- Gross national income (GNI) per capita Coverage ever screened of women 30 - 49 years (%)
- Gross national income (GNI) per capita Coverage in last 5 years of women 30 - 49 years (%)
- Gross national income (GNI) per capita Coverage ever screened of women 25 - 65 years (%)
- Gross national income (GNI) per capita Coverage in last 5 years of women 25 - 65 years (%)
- Gross national income (GNI) per capita Screening Coverage Score

3. Comprehensive List of Positively Scored Features IX

- Gross national income (GNI) per capita Healthcare Access Score
- Incidence of TB RatePer100k_BothSex2019
- Incidence of TB RatePer100k_Female2019
- Diabetes Prevalence^2
- Diabetes Prevalence Hypertension
- Diabetes Prevalence Coverage ever screened of women 30 - 49 years (%)
- Diabetes Prevalence RatePer100k_BothSex2019
- Diabetes Prevalence Anemia Prevalence Score
- Diabetes Prevalence Disease Incidence Score
- HPV Vaccine Coverage ever screened of women 25 - 65 years (%)
- HPV Vaccine Screening Coverage Score
- Coverage ever screened of women 30 - 49 years (%)^2
- Coverage ever screened of women 30 - 49 years (%) Coverage ever screened of women 25 - 65 years (%)
- Coverage ever screened of women 30 - 49 years (%) Screening Coverage Score

3. Comprehensive List of Positively Scored Features X

- Coverage ever screened of women 30 - 49 years (%) Healthcare Access Score
- Coverage ever screened of women 25 - 65 years (%)^2
- Coverage ever screened of women 25 - 65 years (%) RatePer100k_BothSex2019
- Coverage ever screened of women 25 - 65 years (%) Healthcare Access Score
- Total_BothSex2019 Anemia Prevalence Score
- RatePer100k_BothSex2019 Disease Incidence Score
- Total_Female2019 Anemia Prevalence Score
- RatePer100k_Female2019 Disease Incidence Score
- Total_Male2019 RatePer100k_Male2019
- RatePer100k_Male2019 Disease Incidence Score
- Screening Coverage Score Healthcare Access Score
- Anemia Prevalence Score^2
- ratio_Unnamed: 0_over_Physicians per 1,000 people
- ratio_Unnamed: 0_over_Population estimate

3. Comprehensive List of Positively Scored Features XI

- ratio_Unnamed: 0_over_Mean marital age
- ratio_Unnamed: 0_over_Expected years of schooling
- ratio_Unnamed: 0_over_Coverage in last 5 years of women 30 - 49 years (%)
- ratio_Unnamed: 0_over_Coverage in last 3 years of women 30 - 49 years (%)
- ratio_Unnamed: 0_over_Coverage in last 5 years of women 25 - 65 years (%)
- ratio_Unnamed: 0_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_Unnamed: 0_over_Disease Incidence Score
- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_Contraception use (updated 2019)
- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_Sexual Initiation age
- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_Reported STIs (in%)
- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_Multiple pregnancies (%age)

3. Comprehensive List of Positively Scored Features XII

- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_Mean years of schooling
- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Anemia prevalence among women of reproductive age (% of women ages 15-49)_over_RatePer100k_Female2019
- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Contraception use (updated 2019)
- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Sexual Initiation age
- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Reported STIs (in%)
- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Multiple pregnancies (%age)
- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Mean years of schooling

3. Comprehensive List of Positively Scored Features XIII

- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Anemia prevalence among non-pregnant women (% of women ages 15-49)_over_Anemia Prevalence Score
- ratio_Physicians per 1,000 people_over_Unnamed: 0
- ratio_Physicians per 1,000 people_over_Anemia prevalence among pregnant women (%)
- ratio_Physicians per 1,000 people_over_Mean targeted age
- ratio_Physicians per 1,000 people_over_Sexual Initiation age
- ratio_Physicians per 1,000 people_over_Reported STIs (in%)
- ratio_Physicians per 1,000 people_over_Mean marital age
- ratio_Physicians per 1,000 people_over_Multiple pregnancies (%age)
- ratio_Physicians per 1,000 people_over_Mean years of schooling
- ratio_Physicians per 1,000 people_over_Diabetes Prevalence

3. Comprehensive List of Positively Scored Features XIV

- ratio_Physicians per 1,000 people_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Physicians per 1,000 people_over_Healthcare Access Score
- ratio_Anemia prevalence among pregnant women (%)_over_Physicians per 1,000 people
- ratio_Anemia prevalence among pregnant women (%)_over_Mean years of schooling
- ratio_Anemia prevalence among pregnant women (%)_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_Mean targeted age_over_Physicians per 1,000 people
- ratio_Mean targeted age_over_Incidence of TB
- ratio_Mean targeted age_over_Diabetes Prevalence
- ratio_Population estimate_over_Unnamed: 0
- ratio_Population estimate_over_Gross national income (GNI) per capita
- ratio_Population estimate_over_Coverage ever screened of women 30 - 49 years (%)
- ratio_Population estimate_over_Coverage in last 5 years of women 30 - 49 years (%)

3. Comprehensive List of Positively Scored Features XV

- ratio_Population estimate_over_Coverage in last 3 years of women 30 - 49 years (%)
- ratio_Population estimate_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Population estimate_over_Coverage in last 5 years of women 25 - 65 years (%)
- ratio_Population estimate_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_Population estimate_over_Screening Coverage Score
- ratio_Smoking Prevalence (Current smoking prevalence females, 2016)_over_Contraception use (updated 2019)
- ratio_Smoking Prevalence (Current smoking prevalence females, 2016)_over_HIV Prevalence (in adults)
- ratio_Smoking Prevalence (Current smoking prevalence females, 2016)_over_Diabetes Prevalence
- ratio_Smoking Prevalence (Current smoking prevalence females, 2016)_over_Coverage ever screened of women 30 - 49 years (%)
- ratio_Smoking Prevalence (Current smoking prevalence females, 2016)_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Total Fertility rate (2017)_over_HPV Vaccine

3. Comprehensive List of Positively Scored Features XVI

- ratio_Total Fertility rate (2017)_over_Coverage ever screened of women 30 - 49 years (%)
- ratio_Total Fertility rate (2017)_over_Coverage in last 5 years of women 30 - 49 years (%)
- ratio_Total Fertility rate (2017)_over_Coverage in last 3 years of women 30 - 49 years (%)
- ratio_Total Fertility rate (2017)_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_Total Fertility rate (2017)_over_Screening Coverage Score
- ratio_Contraception use (updated 2019)_over_Anemia prevalence among women of reproductive age (% of women ages 15-49)
- ratio_Contraception use (updated 2019)_over_Smoking Prevalence (Current smoking prevalence females, 2016)
- ratio_Contraception use (updated 2019)_over_Human Development Index (HDI)
- ratio_Contraception use (updated 2019)_over_Gross national income (GNI) per capita

3. Comprehensive List of Positively Scored Features XVII

- ratio_HIV Prevalence (in adults)_over_Mean years of schooling
- ratio_HIV Prevalence (in adults)_over_Diabetes Prevalence
- ratio_Mean marital age_over_Unnamed: 0
- ratio_Mean marital age_over_Physicians per 1,000 people
- ratio_Mean marital age_over_Mortality rates (age standardized)
- ratio_Mean marital age_over_Expected years of schooling
- ratio_Mean marital age_over_Gross national income (GNI) per capita
- ratio_Mean marital age_over_HPV Vaccine
- ratio_Mean marital age_over_Hypertension
- ratio_Mean marital age_over_Total_BothSex2019
- ratio_Mean marital age_over_RatePer100k_BothSex2019
- ratio_Mean marital age_over_Total_Female2019
- ratio_Mean marital age_over_Total_Male2019
- ratio_Multiple pregnancies (%age)_over_Anemia prevalence among women of reproductive age (% of women ages 15-49)

3. Comprehensive List of Positively Scored Features XVIII

- ratio_Multiple pregnancies (%age)_over_Anemia prevalence among non-pregnant women (% of women ages 15-49)
- ratio_Multiple pregnancies (%age)_over_Physicians per 1,000 people
- ratio_Multiple pregnancies (%age)_over_Mean years of schooling
- ratio_Multiple pregnancies (%age)_over_Incidence of TB
- ratio_Multiple pregnancies (%age)_over_Diabetes Prevalence
- ratio_Multiple pregnancies (%age)_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Multiple pregnancies (%age)_over_Anemia Prevalence Score
- ratio_Condom Use_over_Anemia prevalence among pregnant women (%)
- ratio_Condom Use_over_RatePer100k_BothSex2019
- ratio_Condom Use_over_RatePer100k_Female2019
- ratio_Condom Use_over_RatePer100k_Male2019
- ratio_Condom Use_over_Disease Incidence Score
- ratio_Age adjusted incidence (standardized rates)_over_HPV Vaccine

3. Comprehensive List of Positively Scored Features XIX

- ratio_Mortality rates (age standardized)_over_Mean marital age
- ratio_Mortality rates (age standardized)_over_Expected years of schooling
- ratio_Mortality rates (age standardized)_over_Gross national income (GNI) per capita
- ratio_Mortality rates (age standardized)_over_Healthcare Access Score
- ratio_Human Development Index (HDI)_over_Contraception use (updated 2019)
- ratio_Human Development Index (HDI)_over_Gross national income (GNI) per capita
- ratio_Human Development Index (HDI)_over_Incidence of TB
- ratio_Human Development Index (HDI)_over_Diabetes Prevalence
- ratio_Human Development Index (HDI)_over_Total_BothSex2019
- ratio_Human Development Index (HDI)_over_Disease Incidence Score
- ratio_Life expectancy at birth_over_Diabetes Prevalence
- ratio_Life expectancy at birth_over_HPV Vaccine
- ratio_Life expectancy at birth_over_RatePer100k_BothSex2019

3. Comprehensive List of Positively Scored Features XX

- ratio_Life expectancy at birth_over_RatePer100k_Female2019
- ratio_Life expectancy at birth_over_Total_Male2019
- ratio_Life expectancy at birth_over_Healthcare Access Score
- ratio_Expected years of schooling_over_Unnamed: 0
- ratio_Expected years of schooling_over_Mean marital age
- ratio_Expected years of schooling_over_Mortality rates (age standardized)
- ratio_Expected years of schooling_over_Gross national income (GNI) per capita
- ratio_Expected years of schooling_over_Incidence of TB
- ratio_Expected years of schooling_over_Diabetes Prevalence
- ratio_Expected years of schooling_over_Coverage in last 3 years of women 30 - 49 years (%)
- ratio_Expected years of schooling_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Expected years of schooling_over_RatePer100k_Female2019
- ratio_Expected years of schooling_over_Screening Coverage Score

3. Comprehensive List of Positively Scored Features XXI

- ratio_Mean years of schooling_over_Anemia prevalence among women of reproductive age (% of women ages 15-49)
- ratio_Mean years of schooling_over_Anemia prevalence among non-pregnant women (% of women ages 15-49)
- ratio_Mean years of schooling_over_Physicians per 1,000 people
- ratio_Mean years of schooling_over_Anemia prevalence among pregnant women (%)
- ratio_Mean years of schooling_over_HIV Prevalence (in adults)
- ratio_Mean years of schooling_over_Sexual Initiation age
- ratio_Mean years of schooling_over_Reported STIs (in%)
- ratio_Mean years of schooling_over_Multiple pregnancies (%age)
- ratio_Mean years of schooling_over_Diabetes Prevalence
- ratio_Mean years of schooling_over_Coverage ever screened of women 30 - 49 years (%)
- ratio_Mean years of schooling_over_Coverage ever screened of women 25 - 65 years (%)

3. Comprehensive List of Positively Scored Features XXII

- ratio_Mean years of schooling_over_Coverage in last 5 years of women 25 - 65 years (%)
- ratio_Mean years of schooling_over_RatePer100k_Male2019
- ratio_Mean years of schooling_over_Healthcare Access Score
- ratio_Mean years of schooling_over_Anemia Prevalence Score
- ratio_Gross national income (GNI) per capita_over_Population estimate
- ratio_Gross national income (GNI) per capita_over_Contraception use (updated 2019)
- ratio_Gross national income (GNI) per capita_over_Mean marital age
- ratio_Gross national income (GNI) per capita_over_Mortality rates (age standardized)
- ratio_Gross national income (GNI) per capita_over_Human Development Index (HDI)
- ratio_Gross national income (GNI) per capita_over_Expected years of schooling
- ratio_Gross national income (GNI) per capita_over_Incidence of TB

3. Comprehensive List of Positively Scored Features XXIII

- ratio_Gross national income (GNI) per capita_over_HPV Vaccine
- ratio_Gross national income (GNI) per capita_over_Total_BothSex2019
- ratio_Gross national income (GNI) per capita_over_RatePer100k_Female2019
- ratio_Gross national income (GNI) per capita_over_Total_Male2019
- ratio_Incidence of TB_over_Mean targeted age
- ratio_Incidence of TB_over_Multiple pregnancies (%age)
- ratio_Incidence of TB_over_Human Development Index (HDI)
- ratio_Incidence of TB_over_Expected years of schooling
- ratio_Incidence of TB_over_Gross national income (GNI) per capita
- ratio_Diabetes Prevalence_over_Physicians per 1,000 people
- ratio_Diabetes Prevalence_over_Mean targeted age
- ratio_Diabetes Prevalence_over_Smoking Prevalence (Current smoking prevalence females, 2016)
- ratio_Diabetes Prevalence_over_HIV Prevalence (in adults)
- ratio_Diabetes Prevalence_over_Sexual Initiation age

3. Comprehensive List of Positively Scored Features XXIV

- ratio_Diabetes Prevalence_over_Reported STIs (in%)
- ratio_Diabetes Prevalence_over_Multiple pregnancies (%age)
- ratio_Diabetes Prevalence_over_Human Development Index (HDI)
- ratio_Diabetes Prevalence_over_Life expectancy at birth
- ratio_Diabetes Prevalence_over_Expected years of schooling
- ratio_Diabetes Prevalence_over_Mean years of schooling
- ratio_Diabetes Prevalence_over_Coverage ever screened of women 30 - 49 years (%)
- ratio_Diabetes Prevalence_over_Coverage in last 5 years of women 30 - 49 years (%)
- ratio_Diabetes Prevalence_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Diabetes Prevalence_over_Coverage in last 5 years of women 25 - 65 years (%)
- ratio_Diabetes Prevalence_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_Diabetes Prevalence_over_Screening Coverage Score

3. Comprehensive List of Positively Scored Features XXV

- ratio_HPV Vaccine_over_Total Fertility rate (2017)
- ratio_HPV Vaccine_over_Mean marital age
- ratio_HPV Vaccine_over_Age adjusted incidence (standardized rates)
- ratio_HPV Vaccine_over_Life expectancy at birth
- ratio_HPV Vaccine_over_Gross national income (GNI) per capita
- ratio_HPV Vaccine_over_Disease Incidence Score
- ratio_Hypertension_over_Mean marital age
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Population estimate
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Smoking Prevalence (Current smoking prevalence females, 2016)
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Total Fertility rate (2017)
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Sexual Initiation age
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Reported STIs (in%)

3. Comprehensive List of Positively Scored Features XXVI

- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Mean years of schooling
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Diabetes Prevalence
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_RatePer100k_BothSex2019
- ratio_Coverage ever screened of women 30 - 49 years (%)_over_Total_Male2019
- ratio_Coverage in last 5 years of women 30 - 49 years (%)_over_Unnamed: 0
- ratio_Coverage in last 5 years of women 30 - 49 years (%)_over_Population estimate
- ratio_Coverage in last 5 years of women 30 - 49 years (%)_over_Total Fertility rate (2017)
- ratio_Coverage in last 5 years of women 30 - 49 years (%)_over_Diabetes Prevalence
- ratio_Coverage in last 5 years of women 30 - 49 years (%)_over_Total_Male2019
- ratio_Coverage in last 3 years of women 30 - 49 years (%)_over_Unnamed: 0

3. Comprehensive List of Positively Scored Features XXVII

- ratio_Coverage in last 3 years of women 30 - 49 years (%)_over_Population estimate
- ratio_Coverage in last 3 years of women 30 - 49 years (%)_over_Total Fertility rate (2017)
- ratio_Coverage in last 3 years of women 30 - 49 years (%)_over_Expected years of schooling
- ratio_Coverage in last 3 years of women 30 - 49 years (%)_over_Total_Male2019
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Anemia prevalence among women of reproductive age (% of women ages 15-49)
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Anemia prevalence among non-pregnant women (% of women ages 15-49)
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Physicians per 1,000 people
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Population estimate
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Smoking Prevalence (Current smoking prevalence females, 2016)

3. Comprehensive List of Positively Scored Features XXVIII

- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Total Fertility rate (2017)
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Sexual Initiation age
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Reported STIs (in%)
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Multiple pregnancies (%age)
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Expected years of schooling
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Mean years of schooling
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Diabetes Prevalence
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_RatePer100k_BothSex2019
- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Total_Male2019

3. Comprehensive List of Positively Scored Features XXIX

- ratio_Coverage ever screened of women 25 - 65 years (%)_over_Anemia Prevalence Score
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Unnamed: 0
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Population estimate
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Expected years of schooling
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Mean years of schooling
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Diabetes Prevalence
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Coverage in last 5 years of women 25 - 65 years (%)_over_Total_Male2019
- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Unnamed: 0
- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Anemia prevalence among pregnant women (%)

3. Comprehensive List of Positively Scored Features XXX

- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Population estimate
- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Total Fertility rate (2017)
- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Diabetes Prevalence
- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Total_BothSex2019
- ratio_Coverage in last 3 years of women 25 - 65 years (%)_over_Total_Male2019
- ratio_Total_BothSex2019_over_Human Development Index (HDI)
- ratio_Total_BothSex2019_over_Gross national income (GNI) per capita
- ratio_Total_BothSex2019_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_RatePer100k_BothSex2019_over_Mean marital age
- ratio_RatePer100k_BothSex2019_over_Condom Use
- ratio_RatePer100k_BothSex2019_over_Life expectancy at birth
- ratio_RatePer100k_BothSex2019_over_Coverage ever screened of women 30 - 49 years (%)

3. Comprehensive List of Positively Scored Features XXXI

- ratio_RatePer100k_BothSex2019_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Total_Female2019_over_Mean marital age
- ratio_RatePer100k_Female2019_over_Anemia prevalence among women of reproductive age (% of women ages 15-49)
- ratio_RatePer100k_Female2019_over_Anemia prevalence among non-pregnant women (% of women ages 15-49)
- ratio_RatePer100k_Female2019_over_Condom Use
- ratio_RatePer100k_Female2019_over_Life expectancy at birth
- ratio_RatePer100k_Female2019_over_Expected years of schooling
- ratio_RatePer100k_Female2019_over_Gross national income (GNI) per capita
- ratio_RatePer100k_Female2019_over_Anemia Prevalence Score
- ratio_Total_Male2019_over_Smoking Prevalence (Current smoking prevalence females, 2016)
- ratio_Total_Male2019_over_Mean marital age

3. Comprehensive List of Positively Scored Features XXXII

- ratio_Total_Male2019_over_Life expectancy at birth
- ratio_Total_Male2019_over_Gross national income (GNI) per capita
- ratio_Total_Male2019_over_Coverage in last 5 years of women 30 - 49 years (%)
- ratio_Total_Male2019_over_Coverage in last 3 years of women 30 - 49 years (%)
- ratio_Total_Male2019_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Total_Male2019_over_Coverage in last 5 years of women 25 - 65 years (%)
- ratio_Total_Male2019_over_Coverage in last 3 years of women 25 - 65 years (%)
- ratio_Total_Male2019_over_Screening Coverage Score
- ratio_RatePer100k_Male2019_over_Condom Use
- ratio_RatePer100k_Male2019_over_Mean years of schooling
- ratio_Screening Coverage Score_over_Population estimate
- ratio_Screening Coverage Score_over_Total Fertility rate (2017)
- ratio_Screening Coverage Score_over_Expected years of schooling
- ratio_Screening Coverage Score_over_Mean years of schooling

3. Comprehensive List of Positively Scored Features XXXIII

- ratio_Screening Coverage Score_over_Diabetes Prevalence
- ratio_Screening Coverage Score_over_Total_Male2019
- ratio_Healthcare Access Score_over_Physicians per 1,000 people
- ratio_Healthcare Access Score_over_Mortality rates (age standardized)
- ratio_Healthcare Access Score_over_Mean years of schooling
- ratio_Anemia Prevalence Score_over_Anemia prevalence among non-pregnant women (% of women ages 15-49)
- ratio_Anemia Prevalence Score_over_Contraception use (updated 2019)
- ratio_Anemia Prevalence Score_over_Sexual Initiation age
- ratio_Anemia Prevalence Score_over_Reported STIs (in%)
- ratio_Anemia Prevalence Score_over_Multiple pregnancies (%age)
- ratio_Anemia Prevalence Score_over_Mean years of schooling
- ratio_Anemia Prevalence Score_over_Coverage ever screened of women 25 - 65 years (%)
- ratio_Disease Incidence Score_over_Unnamed: 0

3. Comprehensive List of Positively Scored Features XXXIV

- ratio_Disease Incidence Score_over_Condom Use
- ratio_Disease Incidence Score_over_Human Development Index (HDI)
- ratio_Disease Incidence Score_over_HPV Vaccine

4. Conclusion

- **Outlier Removal:** Reduced training-set sizes for `NCC_combined`, `Low_CIN_combined`, `High_CIN_combined`, and `ICC_combined`.
- **Forward Selection:** A large set of features had positive scores for `High_CIN_combined`.
- **Negative R^2:** Despite these features, the holdout (gold) set R^2 was -0.1638, indicating possible overfitting or poor generalization.

Summary of Results (Train R^2 and Gold R^2)

Model	**Train** R^2	**Gold** R^2
NCC-16-prevalence	0.6231	0.4508
High CIN-16-prevalence	0.9951	0.6522
ICC - 16 - any - prevalence	0.8783	0.3755
ICC - 18 - SCC - prevalence	0.3909	0.1098

Summary of Results (Train R^2 and Gold R^2)
(After Adding Multiple Pregnancy Feature)

Model	**Train** R^2	**Gold** R^2
NCC-18-prevalence	0.9288	0.0425
Low CIN-18-prevalence	0.3418	0.1730
High CIN-16-prevalence	0.4628	0.3346
ICC - 16 - any - prevalence	0.5277	0.2706
ICC - 16 - SCC - prevalence	0.9667	0.2778
ICC - 18 - SCC - prevalence	0.4483	0.1691

Top Features: NCC-16

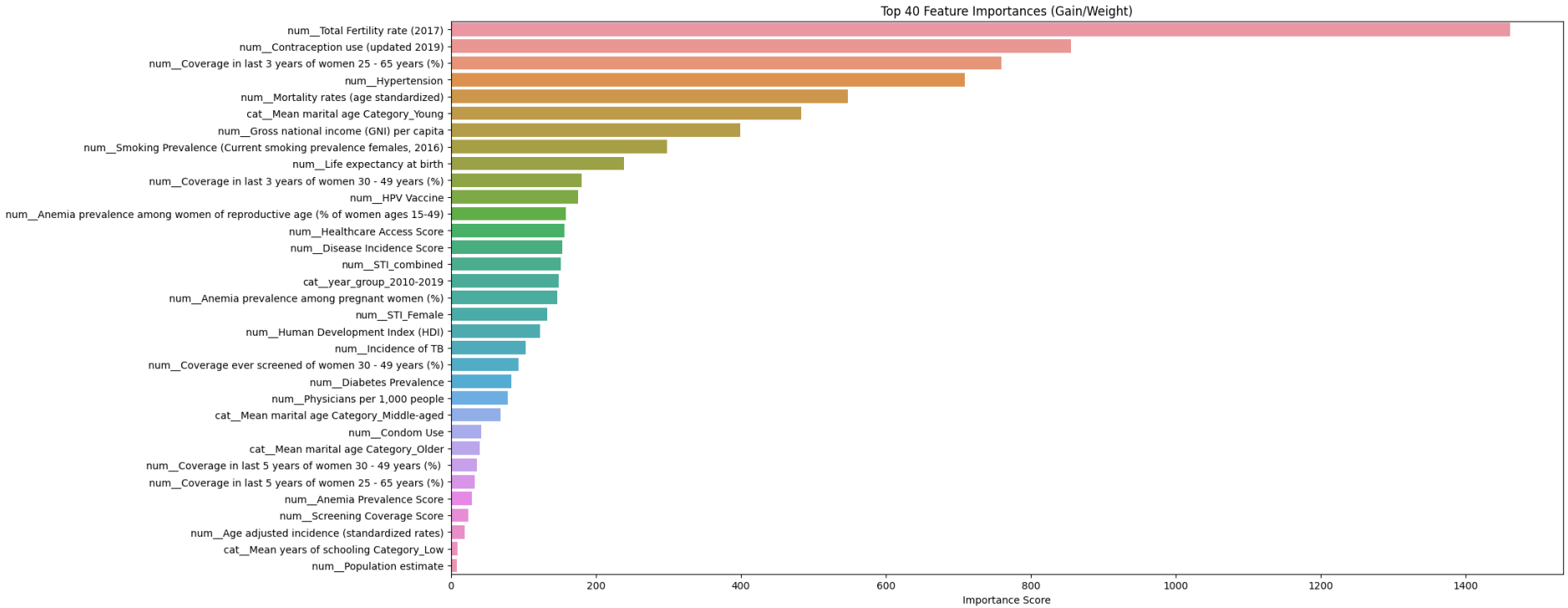


Figure: NCC-16 Feature Importance

Top Features: High CIN-16

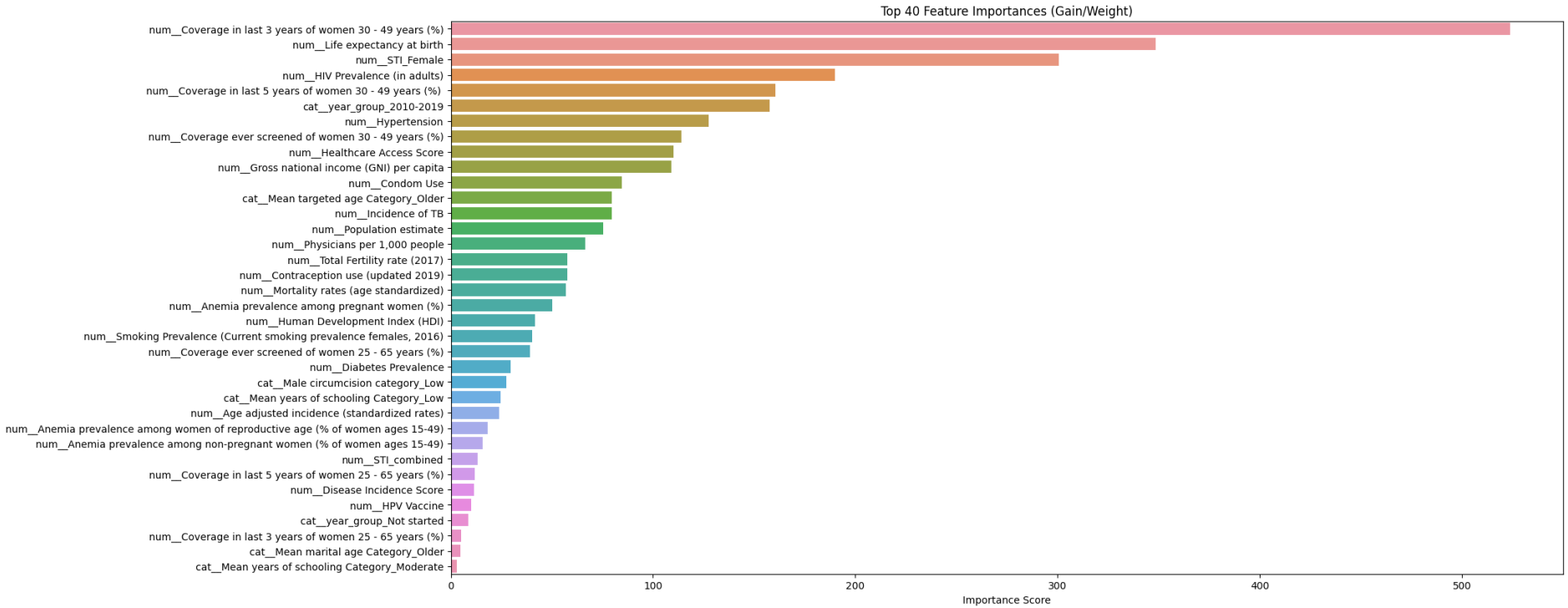


Figure: High CIN-16 Feature Importance

Top Features: ICC-16-any (Image Version)

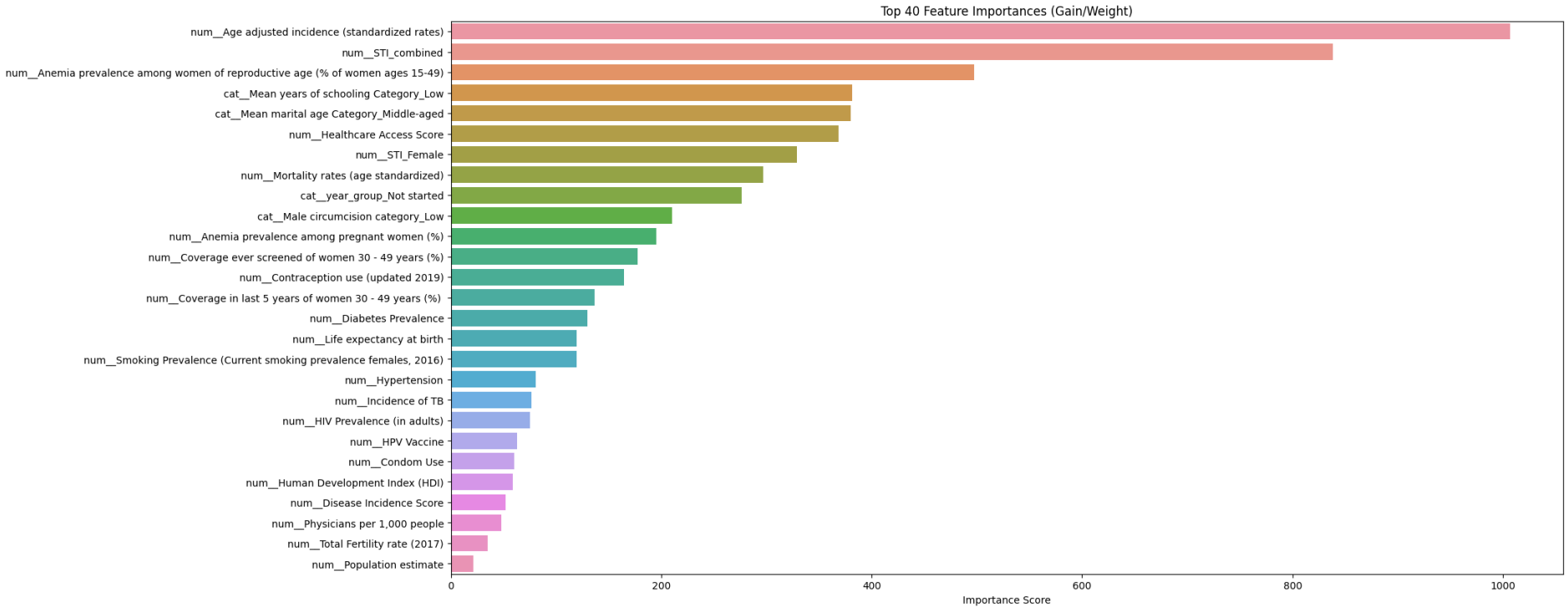


Figure: ICC-16-any Feature Importance

Predicted NCC by Country I

Country	Predicted_NCC
Afghanistan	8.842131615
Andorra	7.192936897
Antigua and Barbuda	4.418025017
Armenia	2.569822788
Azerbaijan	1.140167594
Bahamas	2.765115499
Bahamas	8.101245880
Bangladesh	1.949476242
Bolivia	2.033997297
Bosnia and Herzegovina	3.186712742
Bosnia and Herzegovina	10.61348724
Botswana	1.045061111
Brunei Darussalam	3.908594370
Burkina Faso	3.054091215
Burundi	3.837685823
Cambodia	4.350691795
Cameroon	0.972560585
Cape Verde	3.389800549
Central African Republic	7.288413525

Predicted NCC by Country II

Chad	6.742527008
Comoros	2.684973955
Congo (Republic of the Congo)	6.101756096
Cook Islands	0.716791272
Democratic Republic of the Congo	5.822137833
Djibouti	3.351704836
Dominica	5.087826729
Ecuador	3.209255934
Egypt	4.013259888
El Salvador	3.842881680
Equatorial Guinea	3.396107674
Eritrea	6.639015198
Eswatini (formerly Swaziland)	2.264927864
Ethiopia	3.011385679
Ethiopia	10.23386383
Finland	3.609902859
Gabon	3.651370764
Gambia	5.252840042
Ghana	1.727114439
Grenada	4.727834702
Grenada	5.013041496
Guinea-Bissau	5.505817890

Predicted NCC by Country III

Guyana	3.853730440
Guyana	13.13673401
Guyana	13.13673401
Haiti	2.315417290
Hungary	3.464652777
Iraq	4.632526875
Israel	2.737276793
Ivory Coast (Côte d'Ivoire)	0.000439841
Jamaica	1.581863761
Jordan	5.219561577
Kazakhstan	1.046085000
Kiribati	8.949607849
Kyrgyzstan	2.865233660
Laos	4.556898117
Latvia	3.124191999
Lesotho	3.755897999
Liberia	3.688048363
Libya	3.251317501
Luxembourg	6.692409515
Madagascar	2.851326227
Malawi	2.772884369
Maldives	0.978649676

Predicted NCC by Country IV

Mali	5.894662857
Malta	1.435806036
Marshall Islands	3.042881727
Mauritania	3.992195606
Mauritius	1.600255609
Micronesia	4.439079762
Moldova	0.952360988
Monaco	6.435379982
Montenegro	3.314433336
Myanmar	3.771939754
Namibia	2.424836159
Nauru	5.767508507
New Zealand	5.631504059
New Zealand	8.942900658
Nicaragua	9.135305405
Niger	4.519276142
Niue	3.431662321
Oman	2.848553896
Palau	0.720854998
Panama	5.051554203
Qatar	0.591746926
Rwanda	1.950275064

Predicted NCC by Country V

Rwanda	10.98786354
Saint Lucia	11.99548435
San Marino	10.40884972
Sao Tome and Principe	3.895664215
Saudi Arabia	3.233073950
Seychelles	4.956309319
Sierra Leone	3.837950468
Singapore	1.511446714
Slovakia	2.980462313
Solomon Islands	4.370803356
Somalia	2.637194395
South Sudan	6.212294579
Sri Lanka	-0.670994818
Sudan	5.138072014
Suriname	1.305367470
Suriname	10.45506573
Switzerland	3.837723494
Syrian Arab Republic	7.648045540
Tajikistan	7.018558025
Timor-Leste	2.832687378
Togo	3.521676064
Tonga	3.296543121

Predicted NCC by Country VI

Turkmenistan	4.906911850
Tuvalu	11.53431225
Ukraine	1.422181964
United Arab Emirates	2.433321714
Uzbekistan	1.748288512
Venezuela	3.714789391
Yemen	2.131803036
Zambia	3.283469200
Zimbabwe	3.151607037
Zimbabwe	12.75841236

Predictions For HPV 16 in various lesions

28 march 2025

HPV 16 Prevalence in NCC: XGBoost Results

- **Best Parameters:**
  - eta = 0.070129115210504 9
  - max-depth = 10
  - gamma = 0.977302875919318 2
  - min-child-weight = 4
  - subsample = 0.999960520340688 4
  - colsample-bytree = 0.501188943927303 4
  - colsample-bylevel = 0.999512567887682 5
  - lambda = 9.96728472274197 6
  - alpha = 0.005016440289764037
- **Best CV RMSE:** 7.9384
- **Best Iteration:** 74
- **Train Metrics:**
  - RMSE = 3.9029
  - R^2 = 0.8687
- **Gold/Test Metrics:**
  - RMSE = 3.7489
  - R^2 = 0.8033

Train Parity Plot

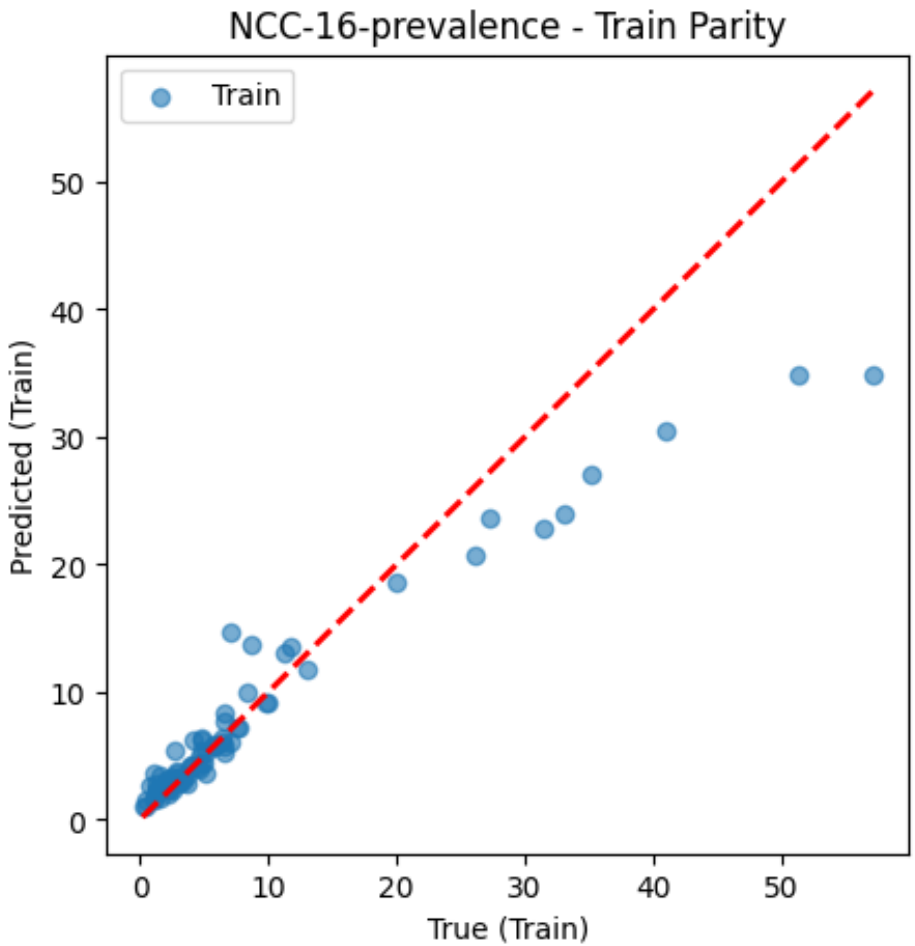


Figure: Parity plot of true vs. predicted values for the **training set**.

Test Parity Plot

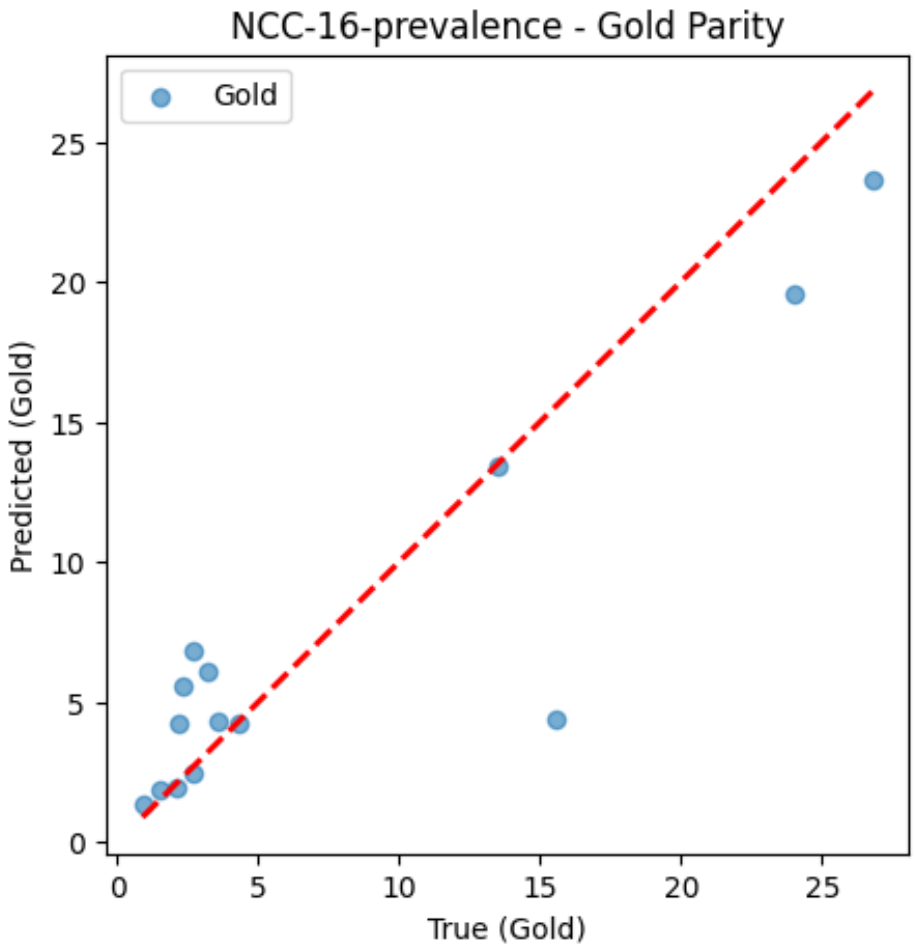


Figure: Parity plot of true vs. predicted values for the **test/gold set**.

HPV 16 in High CIN Grade: XGBoost Results

- **Best Parameters:**
  - eta = 0.1411063477717522 4
  - max-depth = 7
  - gamma = 0.8728121488080494
  - min-child-weight— = 2
  - subsample = 0.5253763583808111
  - colsample-bytree = 0.7451855597452954
  - colsample-bylevel = 0.9883121002606654
  - lambda = 0.010459261686159542
  - alpha = 0.0021938344142665663
- **Best CV RMSE:** 11.7301
- **Best Iteration:** 8
- **Train Metrics:**
  - RMSE = 8.3179
  - R^2 = 0.6628
- **Gold/Test Metrics:**
  - RMSE = 7.3567
  - R^2 = 0.5428

Train Parity Plot

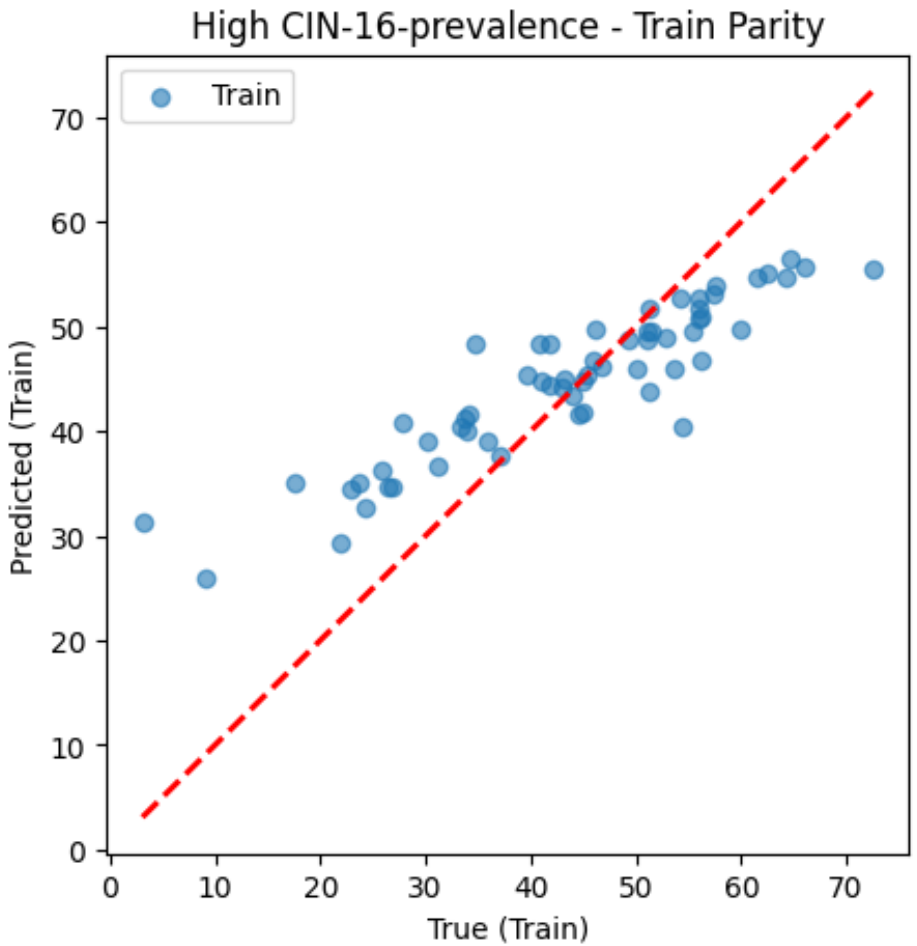


Figure: Parity plot of true vs. predicted values for the training set.

Test Parity Plot

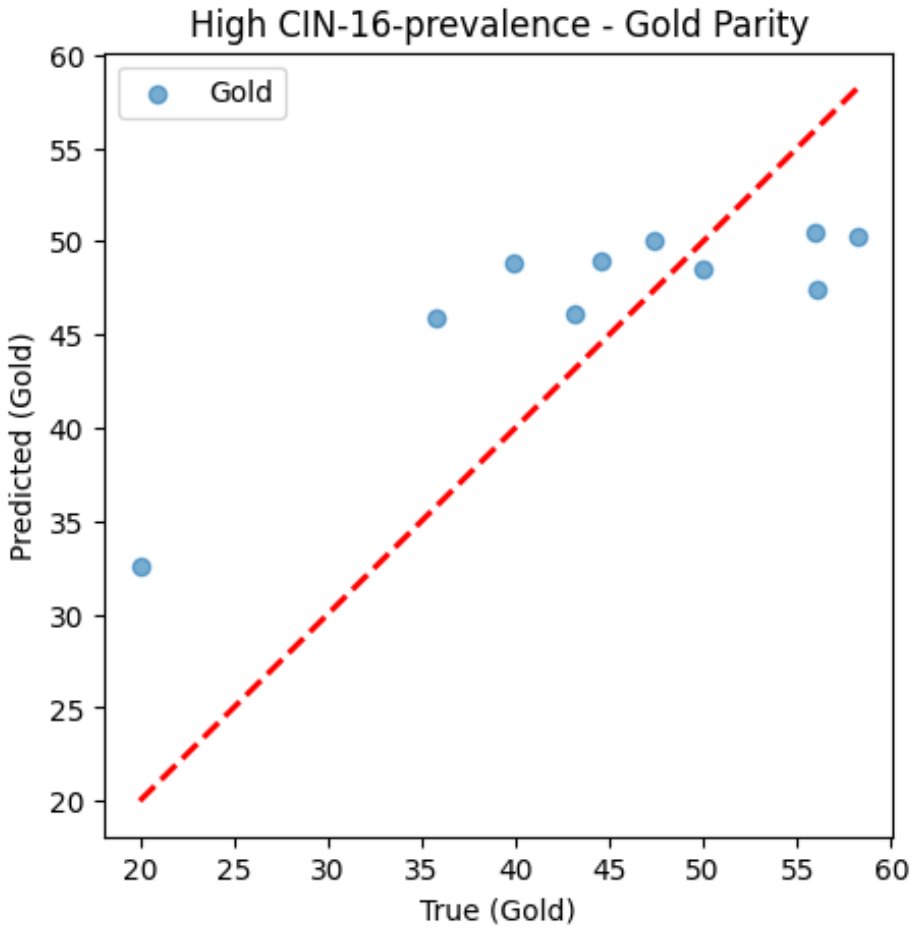


Figure: Parity plot of true vs. predicted values for the gold/test set.

Prediction Heatmap NCC Combined

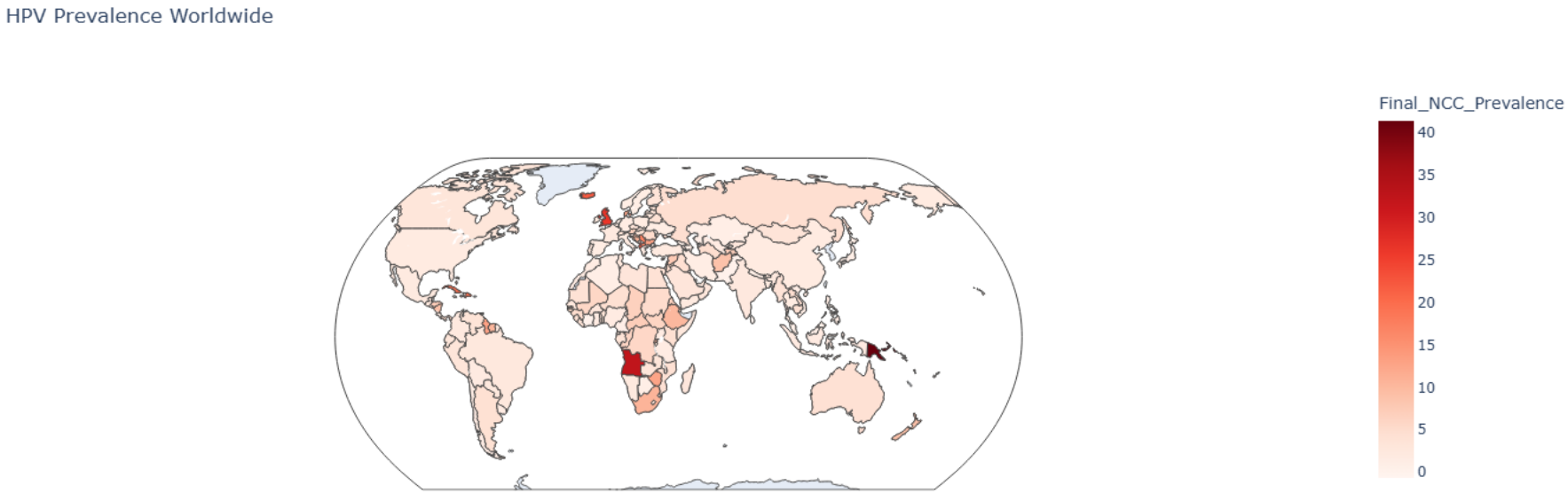


Figure: Heatmap of model predictions For NCC Combined.

Prediction Heatmap NCC 16

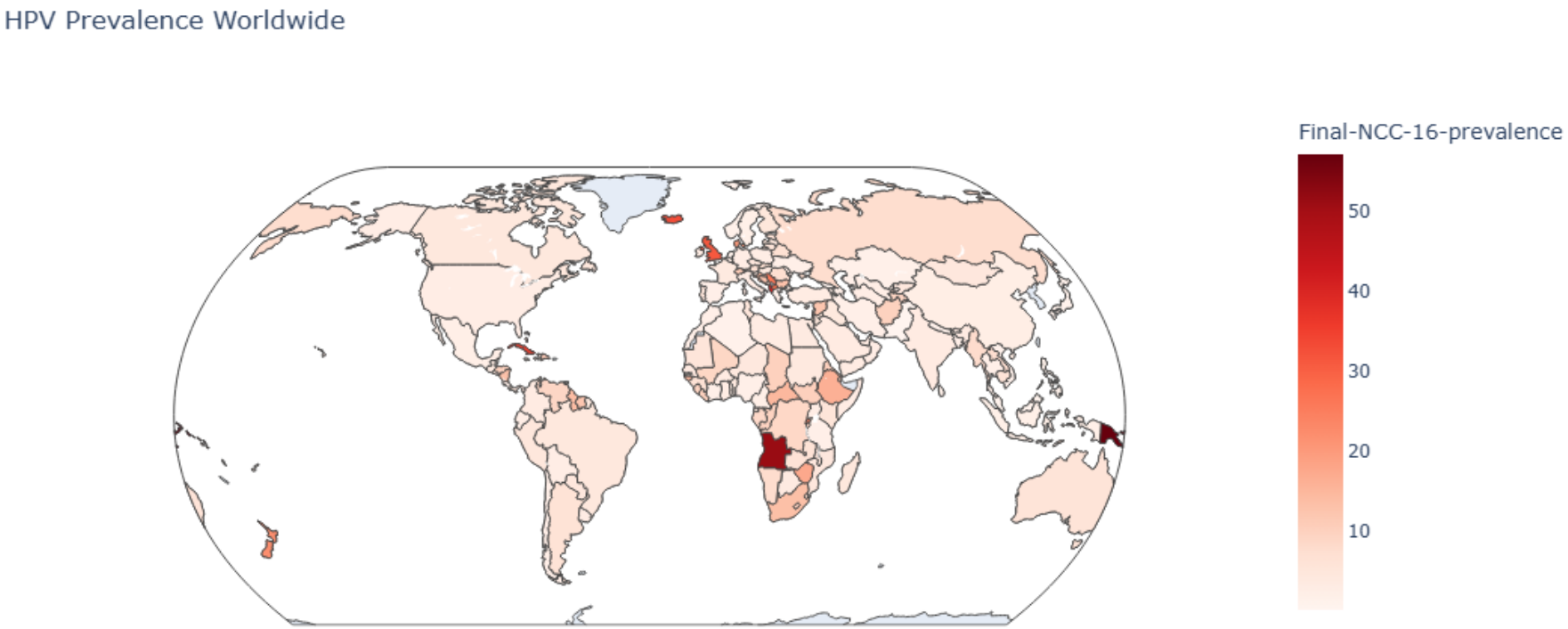


Figure: Heatmap of model predictions for NCC 16.

Prediction Heatmap High CIN 16

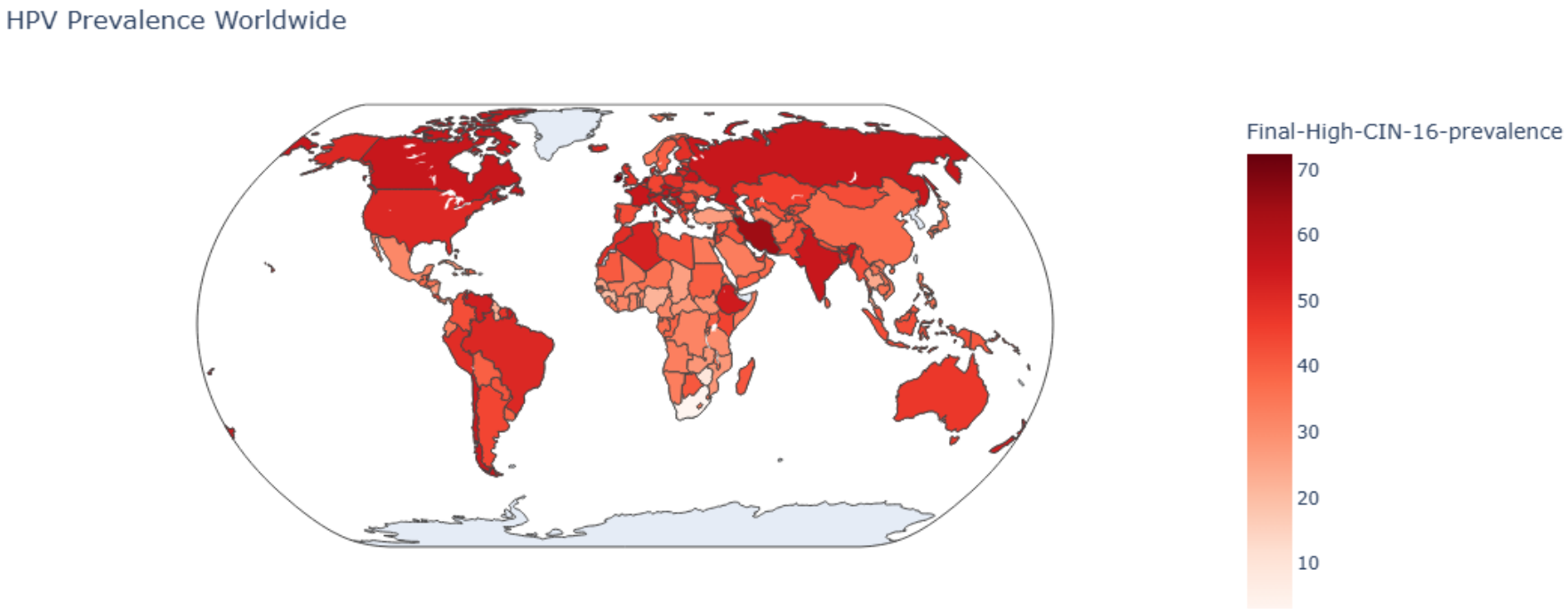


Figure: Heatmap of model predictions High CIN 16.

Data Collection Related to the indian States

Number of features:

- **17 features** are typically available.
- **3 features** are not readily available.
- **7 features** specifically relate to cervical cancer.

Features Typically Available (1/2)

- **Physicians per 1,000 people** Relevance: Reflects Health-care Access
- **Population Estimate (Females 15–49 Years)** Relevance: Essential for calculating prevalence and incidence rates, enabling targeted interventions and resource distribution.
- **Total Fertility Rate (TFR)** Relevance: Reflects reproductive behavior; inter-state differences guide family planning initiatives.
- **Contraception Use (%)** Relevance: Indicates adoption of family planning methods and impacts on maternal health outcomes.
- **HIV Prevalence (%)** Relevance: Shapes state-specific HIV/AIDS strategies and resource allocation.
- **Mean Marital Age** Relevance: Influences early pregnancy risks and broader maternal/child health outcomes.

Features Typically Available (2/2)

- **Condom Use (%)** Relevance: Key for STI/HIV prevention and birth control measures.
- **Human Development Index (HDI) (State-Level)** Relevance: Summarizes socioeconomic conditions that affect overall health and well-being.
- **Life Expectancy at Birth** Relevance: Reflects the effectiveness of healthcare systems and living conditions.
- **Expected Years of Schooling & Mean Years of Schooling** Relevance: Higher education levels correlate with better health awareness and service utilization.
- **Gross National Income (GNI) per Capita / State Domestic Product per Capita** Relevance: Economic capacity strongly influences healthcare infrastructure and service quality.

Features Typically Available (3/3)

- **Incidence of Tuberculosis (TB)** Relevance: Determines priority areas for TB control programs, critical in high-burden states.
- **Diabetes Prevalence** Relevance: Rising non-communicable disease burden requires targeted interventions in affected regions.
- **Hypertension Prevalence** Relevance: Signals growing need for cardiovascular health screening and management.
- **Non-Pregnant Women (Age 15–49) Who Are Anaemic (%)** Relevance: Indicates nutritional and health challenges, guiding supplementation programs.
- **Pregnant Women (Age 15–49) Who Are Anaemic (%)** Relevance: High anemia in pregnancy increases risks for both mother and child, necessitating focused interventions.
- **All Women (Age 15–49) Who Are Anaemic (%)** Relevance: Offers a comprehensive view of the anemia burden to prioritize state-level health strategies.

Features Not Readily Available

- **Smoking Prevalence (%) in Females** Relevance: Often underreported; vital for assessing cancer and cardiovascular risks among women.
- **Reported STIs (%)** Relevance: Reflects sexual health challenges; often requires estimation from multiple, partial data (e.g., syphilis, gonorrhea), shaping targeted STI prevention efforts.
- **Male Circumcision (WHO 2007) (%)** Relevance: Of limited but occasional importance in HIV research; sparse data impede state-level analyses.

Cervical Cancer–Related Features (1/2)

- **Start of Screening Coverage (Year)** Relevance: Shows when organized screening began; correlates with trends in early detection and potential reduction in cervical cancer rates.
- **HPV Vaccination Introduction (Year)** Relevance: Currently "Not Started" for all Indian states in national programs, indicating an unaddressed prevention gap and future policy priority.
- **Age-Adjusted Incidence of Cervical Cancer (%, standardized rates)** Relevance: Crucial for comparing the disease burden across states with differing age profiles.

Cervical Cancer–Related Features (2/2)

- **Number of Deaths (All Ages) Due to Cervical Cancer**
  Relevance: Highlights mortality impact; critical for evaluating treatment access and screening effectiveness.
- **Mortality Rates (%, Age Standardized) Due to Cervical Cancer**
  Relevance: Tracks changes in cancer outcomes over time, adjusting for demographic differences.
- **HPV Vaccine Immunisation Coverage** Relevance: Key measure of preventive intervention reach; very low or negligible in most Indian states, underscoring the need for robust vaccination drives.
- **Coverage of Cervical Cancer Screening** Relevance: Reveals the extent of screening programs and awareness; integral to detecting precancerous lesions early and reducing mortality.

Conclusion

- Many key indicators (fertility, anemia, etc.) are broadly available, guiding public health initiatives.
- Some data (female smoking, STI prevalence, detailed screening) remain sparse and scattered.
- HPV vaccination is "Not Started" in national programs across Indian states, highlighting a critical gap.
- We have to make changes accordingly in the model to use it for the indian states.