HPV Prevalence Data – Exploration & Missing Data Analysis

SAINATH R

NIT Raipur

May 23, 2025

Project Background & Replication

- Previous intern: Daljeet used XGBoost on HPV prevalence data.
- Target Variables: `ncc_combined`, `high_cin_combined`, `ncc_16_prevalence`, etc.
- I successfully replicated Daljeet's results for all target variables.
- Daljeet dropped many prevalence, pregnancy, and case-related features.

Daljeet's Setup (ncc_combined)

- Model: XGBoost Regressor
- Dropped Features: All prevalence & cases
- related Parameters:
  - eta=0.05, max_depth=3, gamma=0.3
  - min_child_weight=4, subsample=1.0
  - colsample_bytree=0.7, lambda=2.0, alpha=1.0
- Test R^2 = 0.5281

My Replication: ncc_combined

- Dropped only target-related columns.
- Model: XGBoost
- Test R^2: 0.7416

High CIN Combined: Comparison:

Metric	Daljeet	Me
Train MSE	33.3991	0.0024
Train R^2	0.4095	1.0000
Test MSE	172.7375	5.6040
Test R^2	-0.1423	0.9412

My Model Parameters (High CIN Combined)

- colsample_bytree=1.0, learning_rate=0.1
- max_depth=4, min_child_weight=3
- n_estimators=200, reg_alpha=0, reg_lambda=1
- subsample=0.7

NCC-16 Prevalence: Comparison:

Metric	Daljeet	Me
Train R^2	0.8687	0.9932
Test R^2	0.8033	0.7416

My Parameters (ncc_16_prevalence)

- colsample_bytree=1.0, learning_rate=0.1
- max_depth=4, min_child_weight=3
- n_estimators=100, reg_alpha=1, reg_lambda=1
- subsample=0.9

Daljeet Parameters (ncc_16_prevalence)

- gamma = 0.9773, min child weight = 4
- subsample = 0.9999, colsample bytree =
- 0.5012 colsample bylevel = 0.9995
- lambda = 9.9672, alpha = 0.0050

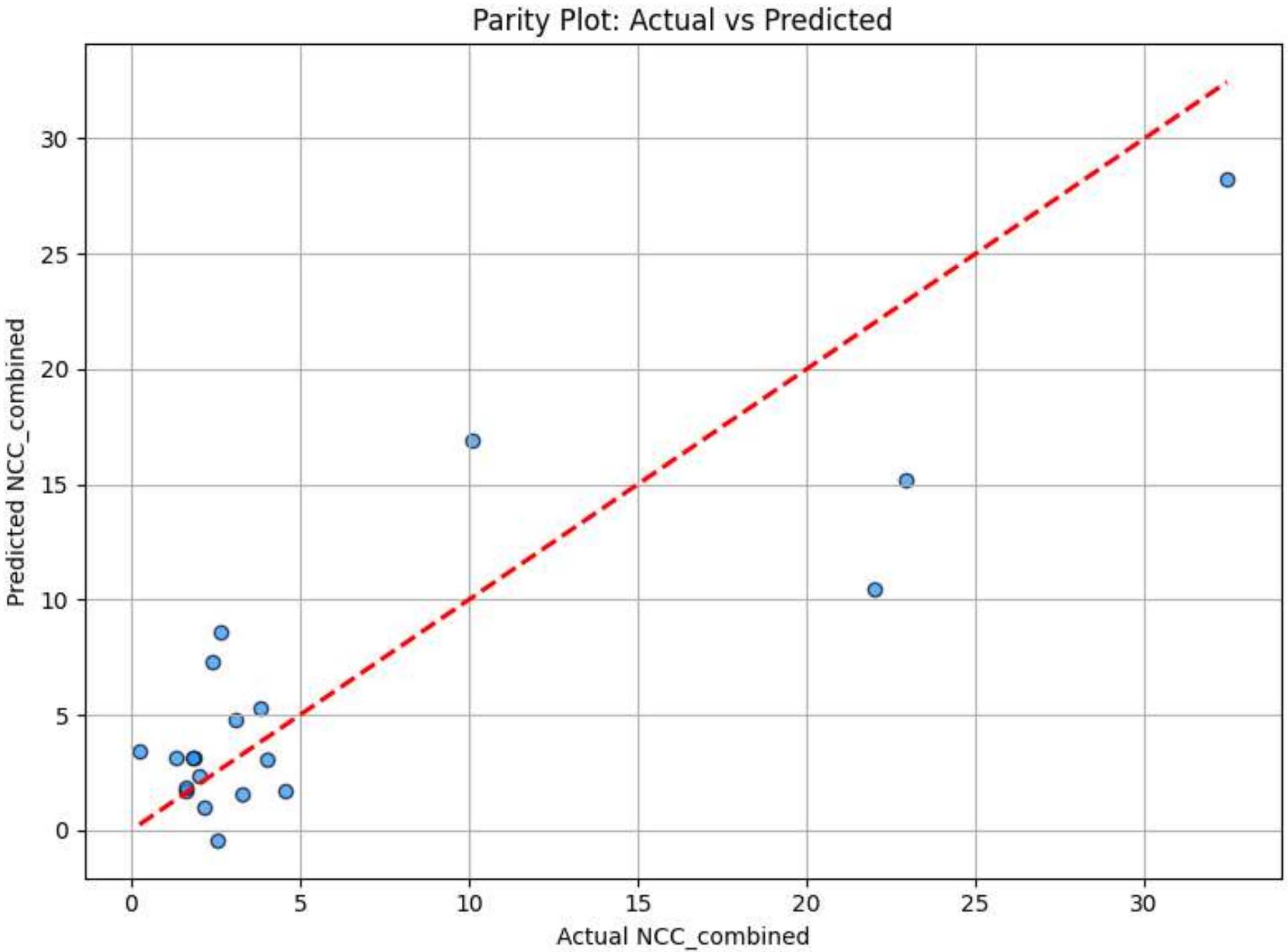
Additional Contributions

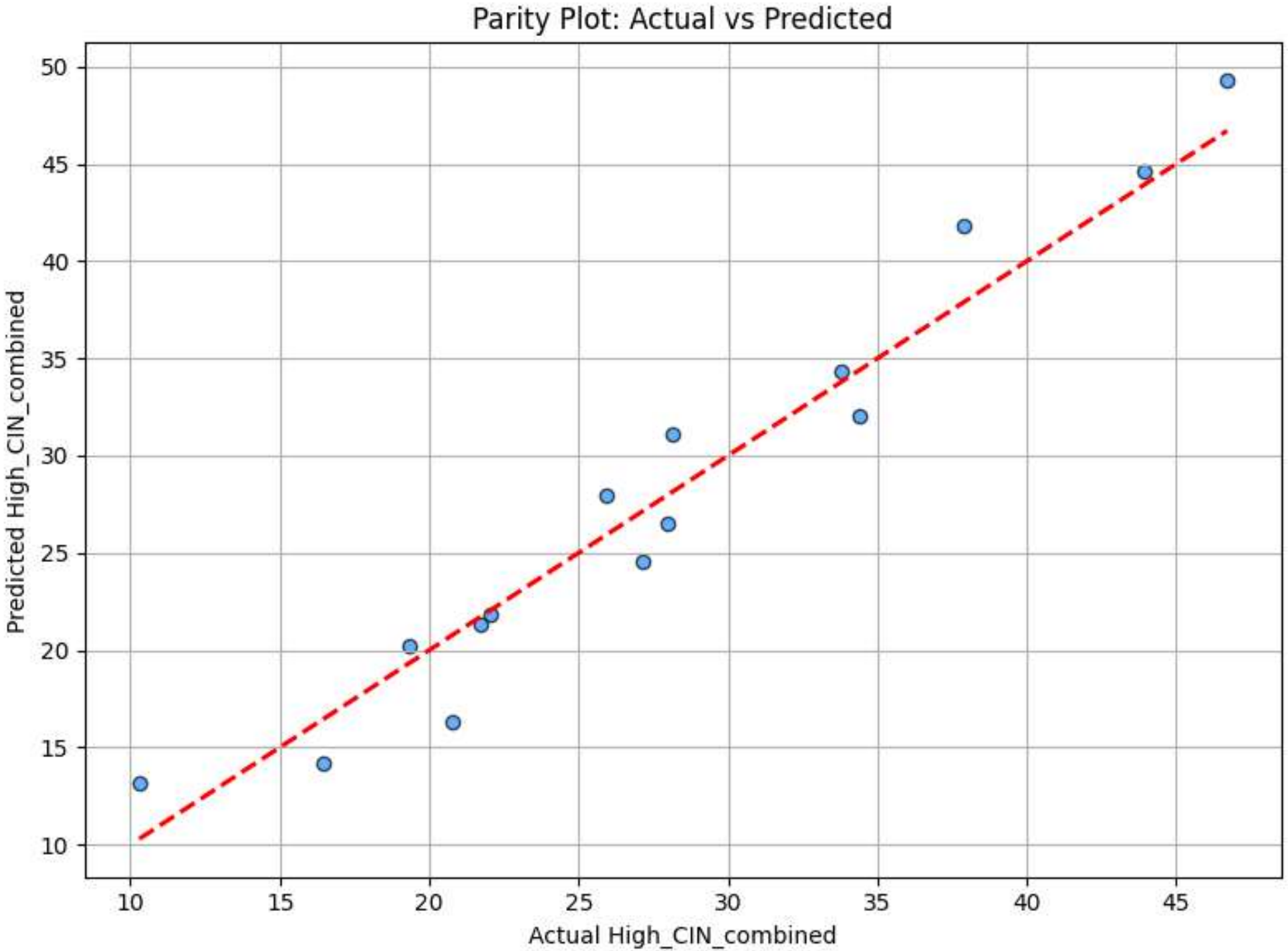
- Scraped data for:
  - Region-wise HPV prevalence in India.
  - Top cities with high cervical cancer incidence.
- This enriched the dataset and gave potential for future regional model training.

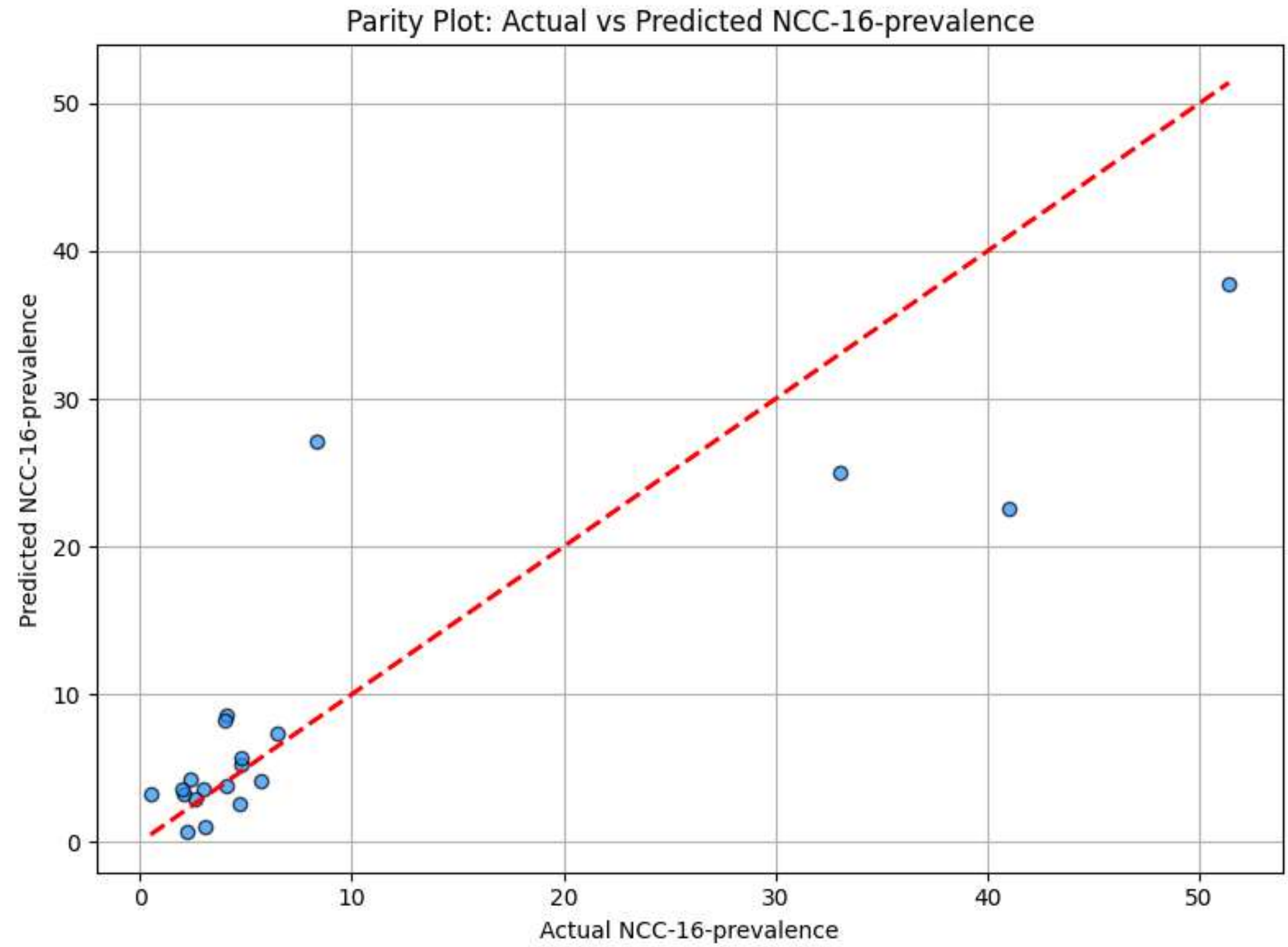
Summary

- Successfully replicated and improved upon Daljeet's models.
- Retained important features selectively (only target-specific
- dropped). Achieved better test performance on multiple target
- variables.

  Added region-specific data to expand research scope.


Parity Plot: Actual vs Predicted
Predicted NCC_combined
Actual NCC_combined
0
5
10
15
20
25
30


Parity Plot: Actual vs Predicted
Predicted High_CIN_combined
Actual High_CIN_combined
10
15
20
25
30
35
40
45
50


Parity Plot: Actual vs Predicted NCC-16-prevalence
Predicted NCC-16-prevalence
Actual NCC-16-prevalence
0
10
20
30
40
50