SMOGN-based Regression Modeling for Cervical Lesion Prediction

Sainath Raja

National Institute of Technology, Raipur

June 27, 2025

Overview

- Small, imbalanced datasets reduce the predictive power of ML
- models. SMOGN helps balance these datasets for regression tasks.
- We applied SMOGN + XGBoost to predict cervical lesion prevalence.

What is SMOGN?

- **SMOGN** stands for *Synthetic Minority Over-sampling Technique for Regression with Gaussian Noise*.
- It extends SmoteR by:
  - Splitting the dataset into **minority** and **majority** bins.
  - Applying both **oversampling** and **undersampling**.
  - Adding **Gaussian noise** to generate more *diverse synthetic samples*.
- Especially effective in small, skewed clinical datasets.

SMOGN Binning via Relevance Threshold

- SMOGN performs binning using a **relevance function** that maps the target variable to a score between 0 and 1.
- Observations with relevance scores above a set threshold are considered **rare (minority)**.
- The default threshold is **0.8**.
- After tuning, the optimal threshold was found to be **0.7**, which improved model performance.
- Binning and sampling were more effective after threshold tuning.

Model Used: XGBoost

- Applied **XGBoost Regressor** with optimized hyperparameters.
- Included pipeline steps:
  1. Preprocessing (Imputation, Scaling, Encoding)
  2. PCA (95% Variance Retention)
  3. Regression (XGBoost)

Used Hyperparameters

- n_estimators = 200
- max_depth = 4
- learning_rate = 0.0104
- subsample = 0.8587
- colsample_bytree = 0.7470
- gamma = 2.8597
- reg_alpha = 2.4260
- reg_lambda = 4.9061

Results: Low CIN Combined (XGBoost)

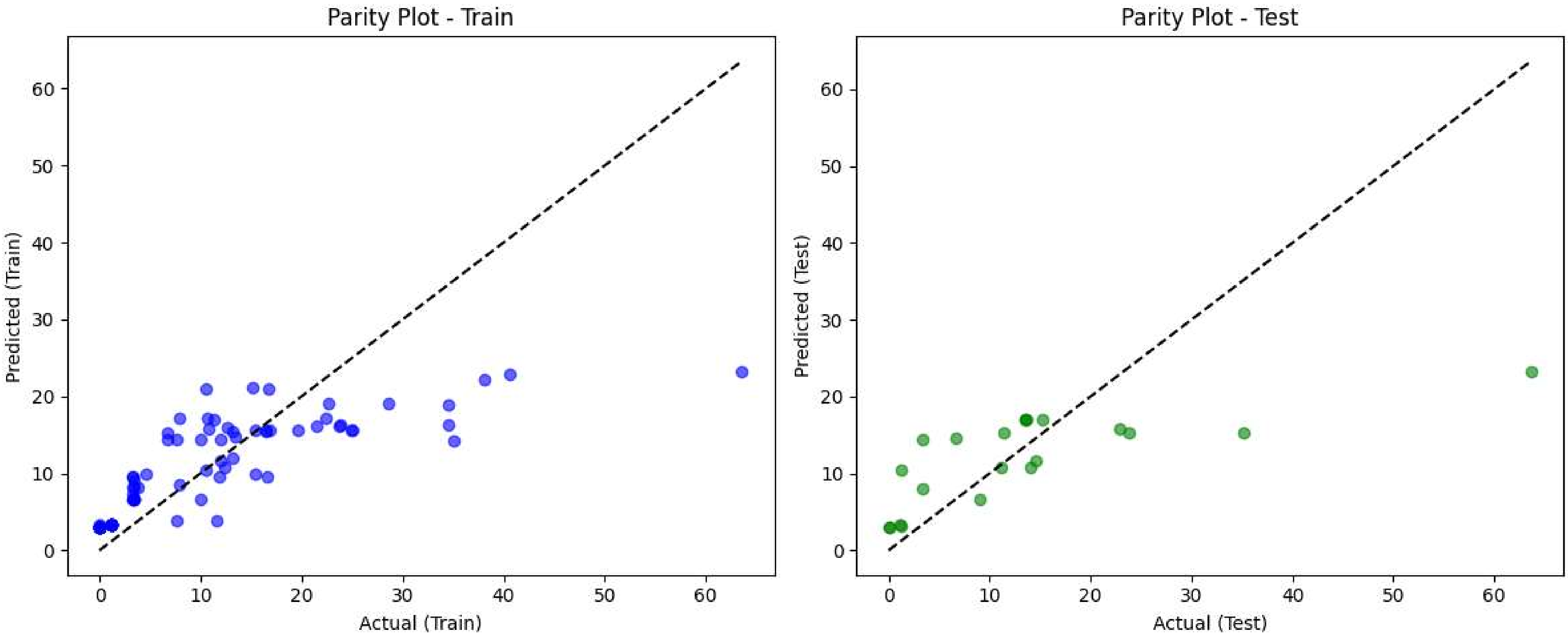
Before Threshold Tuning (Threshold = 0.8)

- **Train R^2:** 0.5385
- **Test R^2:** 0.4095
- **Train Relative RMSE:** 0.6098
- **Test Relative RMSE:** 0.6262

After Threshold Tuning (Threshold = 0.7)

- **Train R^2:** 0.5648
- **Test R^2:** 0.4035
- **Train Relative RMSE:** 0.7129
- **Test Relative RMSE:** 0.8308

Results: Low CIN Combined (XGBoost)



Results: High CIN Combined (XGBoost)

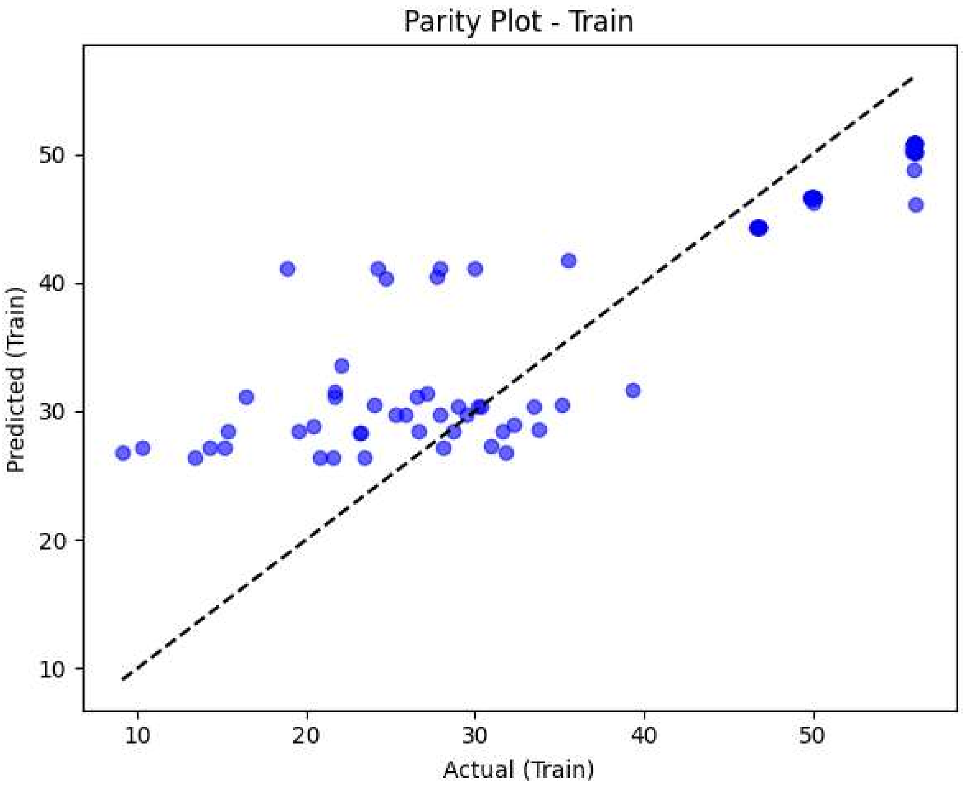
Before Threshold Tuning (Threshold = 0.8)

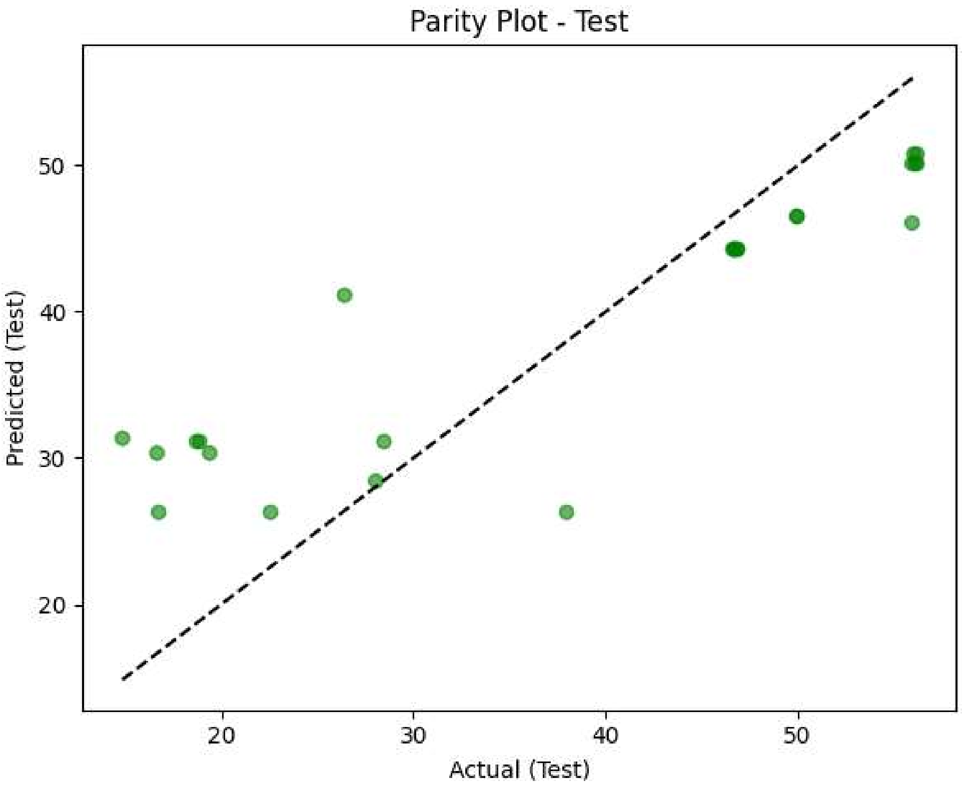
- **Train R^2:** 0.5790
- **Test R^2:** 0.3623
- **Train Relative RMSE:** 0.2468
- **Test Relative RMSE:** 0.3346

After Threshold Tuning (Threshold = 0.7)

- **Train R^2:** 0.7542
- **Test R^2:** 0.7089
- **Train Relative RMSE:** 0.1849
- **Test Relative RMSE:** 0.2177

Results: High CIN Combined (XGBoost)


Parity Plot - Train
Predicted (Train)
Actual (Train)


Parity Plot - Test
Predicted (Test)
Actual (Test)