SMOGN-based Regression Modeling for Cervical Lesion Prediction

Sainath Raja

National Institute of Technology, Raipur

June 23, 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.

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: High CIN Combined (XGBoost)

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

Results: High CIN Combined (XGBoost)

Results: Low CIN Combined (XGBoost)

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

Results: Low CIN Combined (XGBoost)

Results: CIN Combined (XGBoost)

- **Train R^2:** 0.3305
- **Test R^2:** -0.1219
- **Train Relative RMSE:** 0.2753
- **Test Relative RMSE:** 0.3608

Results: CIN Combined (XGBoost)

Conclusion

- SMOGN improved diversity in training samples.
- XGBoost worked well for Low/High CIN regression.
- **CIN combined task needs further investigation** (Test R^2 was negative).
- Future work: Try other models and advanced imbalance methods.