Of Reason, Faith, and Models: An Efficient Machine Learning Approach to Predicting Poverty in Colombia

Juan J. Rojas, Francisco Soler, Jesús Yancy

Universidad de los Andes
Big Data and Machine Learning 2025-2

Best Model Summary XGBoost

-  Kaggle F1 Score of 0.75
- In-sample performance:

  •F1 Score: 0.74 •Precision: 0.71 • Recall: 0.78
- Cross-validation: 5-fold CV, Out-of-Fold threshold optimization.
- Hyperparameters (best tune):
  - *max_depth*: 4 | *min_child_weight*: 10
  - *eta*: 0.05 | *nrounds*: 1000
  - *subsample*: 0.7 | *colsample_bytree*: 0.7
  - *gamma*: 0
- Key predictors: labor-skill & labor-intensity interactions, household structure, income sources and structural determinants (L_p, N_{people}, *P5130*).
- Achieved on a standard laptop (8GB RAM, 4 cores).

Table 4: Trained Models - Results Metrics Summary

Model	Main Optimization	In-sample F1	Out-of-sample F1
Logit 1	Raw Variables Training	0.35	0.36
Logit 2	Additional Features (1)	0.55	0.57
Logit 3	Additional Features (2)	0.61	0.62
Elastic Net 1	Regularization/TH Optimization	0.65	0.65
CART 1	Baseline Tree with Dataset (2)	0.55	0.48
Random Forest 1	Baseline RF with Dataset (2)	0.82	0.65
Random Forest 2	TH optimization	0.84	0.68
Random Forest 3	Better Imputation (3)	0.85	0.69
Gradient Boosting 1	Baseline GB with Dataset (3)	0.70	0.70
Random Forest 4	Additional Features (4)	0.86	0.71
Random Forest 5	Variable Importance	0.80	0.70
Gradient Boosting 2	Additional Features (4)	0.72	0.72
Extreme Gradient Boosting 1	Baseline XGB with Dataset (4)	0.74	0.74
Extreme Gradient Boosting 2	Weights Balance Strategy	0.74	0.74

Best Model II Hyperparameters and Features

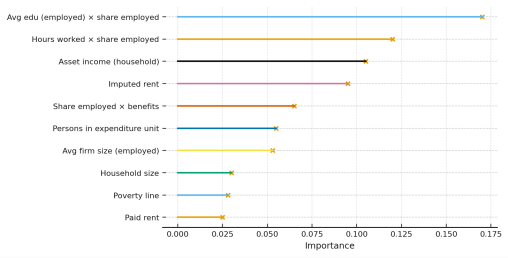

Figure 1: Feature importance (XGBoost)

Main hyperparameters:

Depth	2–4
Min child weight	10–25
Learning rate (η)	0.01–0.05
Trees (nrounds)	100–1000
Subsample	0.7
Colsample_bytree	0.4–0.7
Gamma (γ)	0–1

- Increasing the number of features played a key role in improving the F1 score throughout the experiment. Feature importance was analyzed using Random Forest variations, which revealed a decline in performance after applying a relevance-based threshold for feature selection.

Best Model III Class Imbalance

Table 5: Model Comparison Precision, Recall, F1

Model	Precision	Recall	F1
XGBoost 1	0.71	0.78	0.74
XGBoost 2	0.70	0.79	0.74
Δ (2 − 1)	−0.01	+0.01	0.00

1. Threshold optimized on OOF predictions to align classifier with F1 instead of default 0.5 cutpoint. Applied weights to increase the importance of the positive class (poor).
2. F1 remained at 0.74; In-Sample F1 suggested better performance without weights, confirmed by final Kaggle score (F1= 0.75).
3. Precision-recall trade-off: weights increased total correct poor predictions but reduced precision, worsening overall performance.
4. Threshold optimization was key to balancing performance. Tuning the threshold aligns the decision rule with F1 and compensates class imbalance, achieving the best precision–recall result, outperforming class weights.

Best Model IV Insights

- XGBoost identified labor intensity, human capital, and household structure as the primary drivers of poverty.
- Why it worked:
  1. Shallow, regularized trees to control variance and curb overfitting.
  2. OOF-based threshold calibration to optimize F1 under class imbalance.
  3. Economically grounded features rather than statistical artifacts.
- Demonstrated that income-free poverty prediction is feasible and accurate in our setting.
- Provides a foundation for low-cost monitoring and better policy targeting.
- *Our approach illustrates that the machine learning problem should not be reduced to optimization. It requires a deep understanding of the problem and domain knowledge before brute-force experimentation.*