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).

- Adopted an incremental experimental approach: change one element at a time (features, model, or hyperparameters).
- Measured performance using out-of-sample F1 score.
- All models trained with 5-fold cross-validation.
- Focus: Interpret how each design decision affects predictive power.

Table 2: Model's F1 Score and Dataset Specs

Model	F1	∅	(1)	(2)	(3)	(4)
Logit 1	0.36	X				
Logit 2	0.57		X			
Logit 3	0.62			X		
Elastic Net 1	0.60			X		
CART 1	0.48			X		
RF 1	0.65			X		
RF 2 (TH OP)	0.68			X		
RF 3 (TH OP)	0.69				X	
GB 1	0.70				X	
RF 4 (TH OP)	0.71					X
RF 5 (VI)	0.70					X
GB 2	0.72					X
XGB 1	0.74					X
XGB 2 (B)	0.74					X

Notes: ∅ = 10 household vars. (1) = +5 individual aggregates. (2) = +12 vars + 5 interactions. (3) = fixed imputation. (4) = +28 vars. TH OP = threshold optimization, VI = variable importance, B = balancing.

Modeling II From Linear to Complex Trainings - Other models trained

Sequential performance evolution:

- Logistic Regression: Baseline F1 = 0.36 → 0.62 after feature expansion.
- Logistic Regression + Elastic Net: Including regularization + threshold tuning → F1 = 0.65.
- CART: single tree, prone to overfitting → F1 = 0.48.
- Random Forest: from initial F1= 0.65 to F1 = 0.71 after threshold optimization, optimized imputation and feature expansion.
- Gradient Boosting: Bernoulli loss + OOF thresholding → F1 = 0.70. F1 = 0.72 reached after feature expansion

Insight: Combination of Feature quality, Methodology progression and Hyperparameters calibration (such as thresholds optimization) improves predictions more than algorithmic complexity and computational power.

Logistic regression

- Raw variables reached 0.36, multiple features creation moved F1 to 0.57 and then 0.62.
- Main optimization = Feature Creation (thoughtful creation pending).

CART

- Single tree training with tuned depth of 7. F1 reached 0.48. Overfitting and Generalization problems.
- Main optimization = Basic tree training for baseline comparison.

Logistic Regression with Elastic Net

- Optimal hyperparameters set at $\alpha = 0.1$ and $\lambda = 0.001$. Threshold optimized with PR curve reached F1 of 0.65.
- Main optimization = Model regularization, threshold optimization.

Random Forest

- Started at 0.65 and achieved F1 = 0.71 with min.node.size 30, mtry 9 and 200 trees
- Main optimizations = threshold tuning, fixed imputations, expanded variables.
- Variable importance applied, elbow point set around 18 most important variables, but training performance decreased with F1= 0.70.

Gradient Boosting Models

- Tuned parameters: depth = 4, shrinkage = 0.01, 2500 trees, min.node.size=20. Used stochastic sampling (bag fraction = 0.5).
- OOF threshold optimization maximized F1 = 0.70 and features additions increased F1 to 0.72 .
- Main optimizations = Same optimizations and variables as RF (4), Boosting outperformed it by capturing nonlinear feature complementarities.

Table 3: Trained Models - Results Metrics

Model	Precision	Recall	In-sample F1	Out-of-sample F1
Logit 1	0.65	0.24	0.35	0.36
Logit 2	0.70	0.46	0.55	0.57
Logit 3	0.71	0.53	0.61	0.62
Elastic Net 1	0.57	0.73	0.65	0.65
CART 1	0.66	0.47	0.55	0.48
Random Forest 1	0.90	0.75	0.82	0.65
Random Forest 2	0.79	0.91	0.84	0.68
Random Forest 3	0.77	0.92	0.85	0.69
Gradient Boosting 1	0.64	0.77	0.70	0.70
Random Forest 4	0.80	0.92	0.86	0.71
Random Forest 5	0.75	0.87	0.80	0.70
Gradient Boosting 2	0.68	0.79	0.72	0.72
Extreme Gradient Boosting 1	0.71	0.78	0.74	0.74
Extreme Gradient Boosting 2	0.70	0.79	0.74	0.74

Final private score in Kaggle reached F1= 0.75 with Extreme Gradient Boosting 1.