IDENTIFYING KEY SOCIOECONOMIC DETERMINANTS OF INCOME INEQUALITY

A Machine Learning Approach Using World Bank Data

Benoit R. Goye

Data Science and Advanced Programming | December 2025

University of Lausanne

Why it matters

- **Rising inequality → reduced social mobility**
- **Political polarization**
- **Slower economic growth**
- **GINI coefficient: 0-100**

Key questions

- **Which policy levers reduce inequality?**
- **Are drivers universal or context-dependent?**
- **Can we predict inequality accurately?**

Traditional challenges

Non-linear relationships | High-dimensional data | Missing values (developing countries)

RQ1: Feature Importance

Which socioeconomic indicators are the **strongest predictors** of GINI coefficients?

RQ2: Cross-Model Consistency

Are predictor rankings **consistent across different ML algorithms**?

RQ3: Predictive Performance

How well can we **predict inequality** across different contexts?

Contribution

Data-driven approach using 5 ML algorithms to explore inequality drivers while implementing advanced optimization techniques

Data: World Bank Database (2000-2023)

Data & Methodology

Unil.

Target Variable

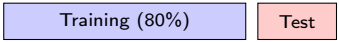
GINI coefficient | ∼1,800 country-year observations

∼60 Predictor Variables

1. **Economic (GDP, trade, inflation)**
2. **Demographics (population, urbanization)**
3. **Human development (health, education)**
4. **Labor market (employment rates)**
5. **Infrastructure (electricity, internet)**
6. **Governance (gender parity)**

Preprocessing

- **KNN imputation (k=5)**
- **Exclude features >50% missing**
- **80-20 train-test split**
- **5-fold cross-validation**



Model	Type	Key Feature
Decision Tree	Baseline	Single tree, recursive partitioning
Random Forest	Ensemble (Bagging)	Bootstrap + random features Reduces variance
Gradient Boosting	Ensemble (Boosting)	Sequential fitting of residuals Reduces bias
XGBoost	Advanced Boosting	Regularized objective 2nd-order optimization
LightGBM	Advanced Boosting	Histogram-based + GOSS Computational efficiency

Hyperparameters: 200 estimators, learning rate 0.05, max depth 5-10, regularization

1. Bootstrap Confidence Intervals

Train each model 100 times on bootstrap samples. Compute 95% CI. Parallel processing (6× speedup).

2. Permutation Importance Tests

Randomly permute features (50 permutations). Measure drop in R^2. One-sample t-test ($\alpha = 0.05$).

3. Cross-Model Consistency

Spearman rank correlations between models. High $\rho > 0.7 \rightarrow$ robust rankings.

Goal

Ensure feature importance rankings are statistically robust and consistent

Unil.

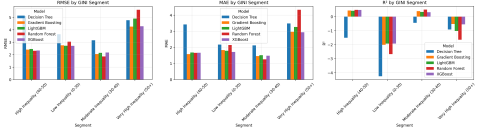
Testing context-dependence:

By Income Level:

- Low income
- Lower-middle income
- Upper-middle income
- High income

By Geographic Region:

- **East Asia & Pacific**
- **Europe & Central Asia**
- **Latin America & Caribbean**
- **Middle East & North Africa**
- **South Asia**
- **Sub-Saharan Africa**



Key Question

Are inequality drivers universal or context-specific?

Model Performance: Ensemble Methods Excel

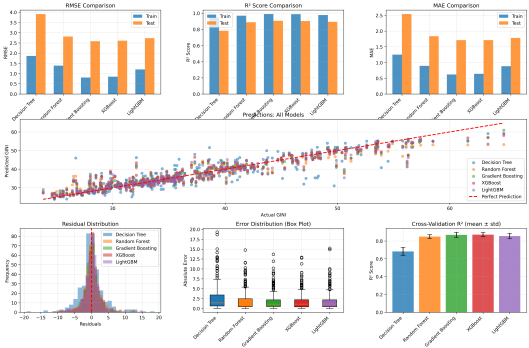
Results

Unil.

Model	Test R^2	RMSE
Decision Tree	0.787	3.89
Random Forest	0.890	2.79
Gradient Boosting	**0.905**	2.58
XGBoost	**0.902**	2.62
LightGBM	0.895	2.72

Key Findings

Ensembles outperform single tree. GB & XGBoost: $R^2 > 0.90$. Good generalization.

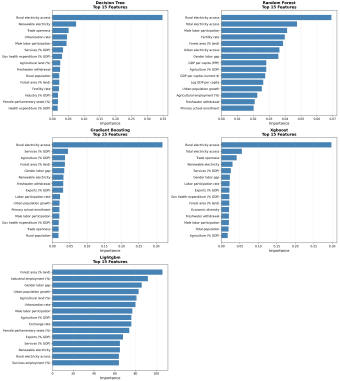


Top 10 Features: Infrastructure Dominates

Results

Unil.

Rank	Feature	Imp.
1	Rural electricity	0.288
2	Total electricity	0.067
3	Trade openness	0.034
4	Renewable energy	0.031
5	Forest area	0.027
6	Agriculture % GDP	0.026
7	Exports % GDP	0.024
8	Gender labor gap	0.023
9	Rural population	0.023
10	Services employ.	0.022



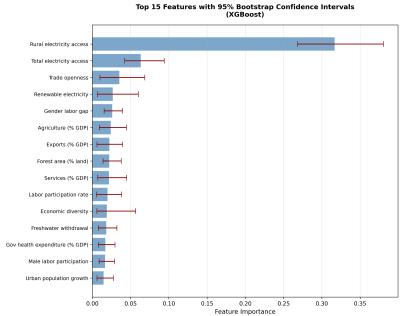
Implication

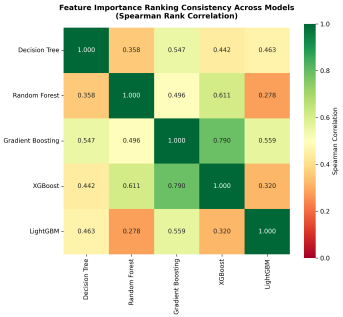
Inequality driven by **small subset** of critical factors

Statistical Significance Confirmed

Results

Unil.





Bootstrap CI

All top 10: 95% CI excludes zero. High stability.

Cross-Model

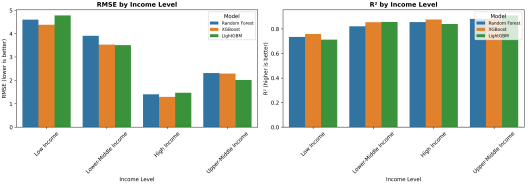
GB vs XGBoost: $\rho = 0.84$**. Core drivers consistent.**

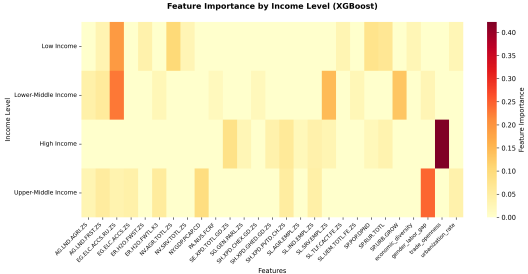
Conclusion

Top predictors are **robust** across resampling, algorithms, and tests

Context Matters: Income-Level Heterogeneity

Results





Performance

High: R^2 = **0.86-0.88** | **Upper-mid:** **0.85-0.91** | **Lower-mid:** **0.81-0.86** | **Low:** **0.69-0.73**

Top Drivers

Low: **Infrastructure** | **Mid:** **Urbanization** | **High:** **Trade & labor**

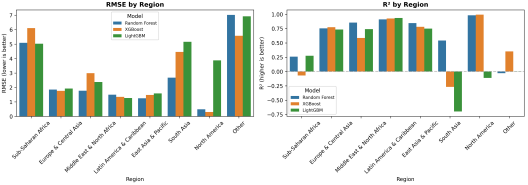
Policy Implication

Tailored interventions needed by development stage

Regional Heterogeneity: Geography Matters

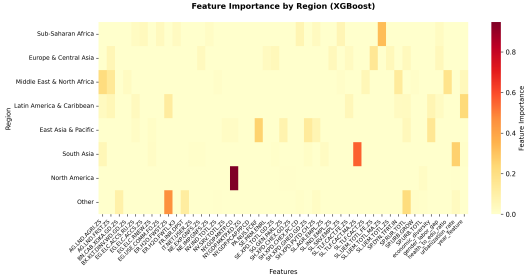
Results

Unil.



Performance by Region

Europe & Central Asia: Highest R^2 | Latin America: Moderate | Sub-Saharan Africa: Lower



Different Drivers

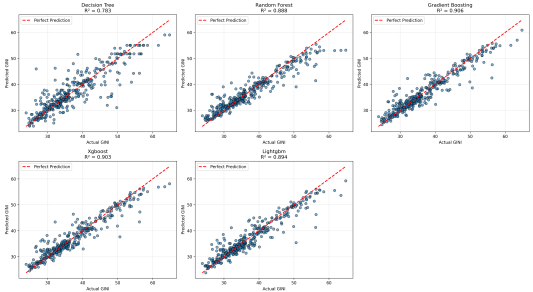
Region-specific factors. Historical & institutional differences matter.

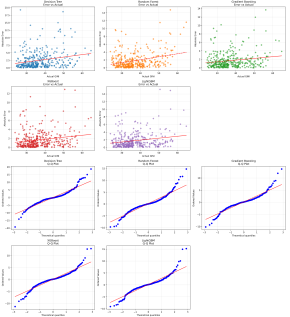
Key Insight

Inequality determinants vary by **income level** and **geography**

Model Diagnostics: Strong Performance

Results





Predicted vs Actual

Strong $R > 0.90$. GINI 20-45: excellent. GINI > 50: underprediction.

Error Analysis

Errors near zero. Few outliers. Consistent across models.

Conclusion

Models generalize well but struggle with extreme outliers

Computational Performance: 3-5× Speedup

Implementation & Optimization

Unil.

Parallel Processing Strategies:

Component	Method	Before	After
Data collection	ThreadPoolExecutor (6 workers)	12 min	**2 min**
Bootstrap tests	joblib Parallel (n_jobs=-1)	60 sec	**10 sec**
Model training	scikit-learn parallelization	25 sec	**5 sec**
Total pipeline	Combined optimizations	**25-35 min**	**5-7 min**

Intelligent Caching:

- SHA256 hash validation
- Auto cache invalidation
- 60-100× speedup on reruns

Impact:

- **6× faster** data collection
- **6× faster** bootstrap
- **5× faster** training

Modular Architecture

9-stage pipeline: Data collection → Preprocessing → Training → Prediction → Evaluation → Comparison → Segmentation → Statistical tests → Report generation

Execution Modes

- Quick: **Full pipeline (default)**
- Fast: **2015-2023 only**
- Optimized: **Hyperparameter tuning**
- Custom: **Fine-grained control**

Quality Assurance

- **Type annotations**
- **Comprehensive docstrings**
- **Data hash validation**
- **Deterministic results (seed=42)**
- **Dependency management**

Reproducibility

- **Git version control**
- **Automated caching**
- **Complete documentation**
- **Public code repository**

Key Findings Summary

Conclusions & Policy Implications

Unil.

1. Predictive Power: Ensemble methods achieve $R^2 > 0.90$
   - **Inequality is highly predictable from socioeconomic factors**
   - **Gradient Boosting & XGBoost perform best**
2. Infrastructure Dominates: Rural electricity access (imp. 0.304)
   - **5× more important than any other feature**
   - **Composite measure of development, institutions, connectivity**
3. Robust Core Drivers: Trade, economic structure, gender gaps
   - **Consistent across all five models**
   - **Statistically significant (bootstrap & permutation tests)**
4. Context-Dependent: Different drivers at different income levels
   - **Low income: Basic infrastructure | Middle: Urbanization | High: Trade & labor markets**

Three priority areas for reducing inequality:

1. Expand Infrastructure Access

- Particularly rural electricity connectivity
- Enables economic opportunities & education access
- Signals institutional capacity for service delivery

2. Promote Inclusive Labor Markets

- Reduce gender labor gaps
- Increase labor force participation (especially women)
- Address youth unemployment

3. Manage Structural Transformation

- Progressive taxation & social insurance
- Education access that scales with industrial demands
- Minimum wage floors to prevent wage compression

Limitations & Future Directions

Unil.

Limitations:

- **Prediction not causation**
  - Need IV, DiD, RDD for causal effects
- Missing data imputation uncertainty
- Stationarity assumption
  - Relationships may change over time
- Omitted variables
  - Institutional quality, tax progressivity

Future Research:

1. **Causal inference**
   - Double ML, causal forests
2. **Additional data**
   - Satellite imagery
   - Institutional quality measures
   - Alternative inequality metrics
3. **Methodological advances**
   - SHAP values for interactions
   - Panel methods with time-varying coefficients

Bottom Line

ML complements traditional econometrics: prediction + pattern recognition vs. causal identification

Methodological Contributions

- **First systematic comparison of 5 tree-based ML algorithms**
- **Robust statistical validation framework**
- **Advanced optimization (3-5× speedup)**
- **Fully reproducible pipeline**

Substantive Contributions

- **Infrastructure as dominant predictor**
- **Context-dependence quantified**
- **High predictability ($R^2 > 0.90$)**
- **Tailored policy recommendations**

Innovation

Combines ML prediction power with rigorous statistical validation

Impact

Data-driven evidence for targeted inequality reduction

Data & Code: Publicly available for replication

Unil.

THANK YOU

Any Questions?

APPENDIX

Txt | TxT

Unil.

Txt