

Predicting Youth Unemployment Persistence with Machine Learning: Evidence from World Development Indicators

Jiawei Xia^{1*}, Yihe Rong²

¹*School of Economics, Zhejiang Gongshang University, Hangzhou, China*

²*Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou, China*

**Corresponding Author. Email: 2997304721@qq.com*

Abstract. Youth unemployment is not only affected by contemporaneous macroeconomic and structural conditions but also shaped by persistence over time. Using World Development Indicators data, this paper predicts youth unemployment across 185 economies and 2,590 country-year observations from 2011 to 2024. The main predictor is the one-year lag of youth unemployment, while GDP growth, GDP per capita, inflation, tertiary enrollment, industry share, service share, and urbanization are used as controls. OLS results show a strong positive association between lagged and current youth unemployment, with a coefficient of 0.9720. Under a random country-year split, LightGBM achieves an RMSE of 2.1053 and an R-squared of 0.9725, substantially outperforming models without the lagged variable. Time holdout and country holdout tests also show stable predictive performance. Formal interaction tests find no statistically significant overall difference in persistence across income groups, whereas persistence differs significantly across baseline youth-unemployment groups. After adjustment for multiple comparisons, the high-baseline group shows significantly stronger persistence than the medium-baseline group. SHAP interpretation supports the central role of lagged youth unemployment while emphasizing that the results should be understood as predictive associations rather than causal effects.

Keywords: youth unemployment, persistence, lagged unemployment, machine learning, World Development Indicators, SHAP

1. Introduction

Youth unemployment matters as both a macroeconomic indicator and a measure of labor-market inclusion. Young workers enter the labor market with limited experience and weaker firm-specific human capital, so economic slowdowns and structural change can leave durable traces in youth employment outcomes. This paper therefore focuses on the annual persistence of youth unemployment: whether the youth unemployment rate in one year provides strong predictive information about the rate in the following year.

Okun's framework links output fluctuations with unemployment [1], while human-capital theory connects schooling and experience with labor-market outcomes [2]. A cross-country analysis should therefore retain macroeconomic, education, industrial-structure, and urbanization variables as

controls, but the main empirical question is whether youth unemployment displays cross-year inertia after these conditions are considered.

Traditional linear models estimate average conditional relationships, but cross-country youth unemployment may also involve nonlinearities, thresholds, and interactions. This study therefore uses OLS as an interpretable benchmark for persistence and compares it with Linear Regression, Lasso, Random Forest, XGBoost, and LightGBM for predictive performance.

The paper contributes in three ways. It builds a WDI modeling sample for 2011-2024 and uses the previous year's youth unemployment rate to predict the next year's rate. It compares OLS, Linear Regression, Lasso, Random Forest, XGBoost, and LightGBM, with no-lag, time-holdout, and country-holdout checks. It also examines heterogeneity by income level and baseline youth unemployment.

2. Literature review and hypotheses

The first strand concerns macroeconomic cycles. In the Okun tradition, output growth and labor demand move together, so GDP growth remains a relevant control when predicting youth unemployment pressure.

The second strand concerns education and school-to-work transition. Education raises human capital, but rapid tertiary expansion may also coincide with delayed labor-market entry or mismatch if skilled jobs do not expand at the same pace [2].

The third strand concerns industrial structure. Industry and services describe the productive structure in which young workers search for jobs, but in this paper they are treated as controls rather than as the central source of predictive power.

The fourth strand concerns machine learning in economic prediction. Random Forests, XGBoost, and LightGBM can capture nonlinear patterns in structured data [3-6], while SHAP links predictions to feature contributions [7]. Here, SHAP explains prediction, not causal effects.

Against this background, the empirical analysis evaluates three persistence expectations. Youth unemployment is expected to show strong annual persistence, so lag_youth_unemployment should be positively associated with current youth unemployment. Adding this variable should improve prediction relative to models using only macroeconomic, education, industrial-structure, and urbanization variables. Persistence may also differ across income groups and baseline youth-unemployment groups; these differences are evaluated with formal interaction tests.

3. Data and methodology

The data come from the World Bank World Development Indicators. The dependent variable is youth unemployment, measured by SL.UEM.1524.ZS [8], the unemployment rate for the labor force aged 15-24 under the modeled ILO estimate. The raw data cover 217 countries/economies, 2010-2024, and 3,270 country-year observations.

Table 1. Sample overview

Item	Value
Raw observations	3270
Raw countries/economies	217
Raw years	2010-2024

Table 1. (continued)

Final modeling observations	2590
Final countries/economies	185
Modeling years	2011-2024
Core lagged variable	lag_youth_unemployment
Baseline unemployment grouping	2010-2012 country mean, tertile groups

Table 1 summarizes the sample construction. Because lag_youth_unemployment is constructed from the previous year's youth unemployment rate, 2010 is not included in the final modeling sample. After cleaning and lag construction, the final sample contains 2,590 observations for 185 countries/economies from 2011 to 2024. Figure 1 presents the distribution and time trend of youth unemployment in the sample.

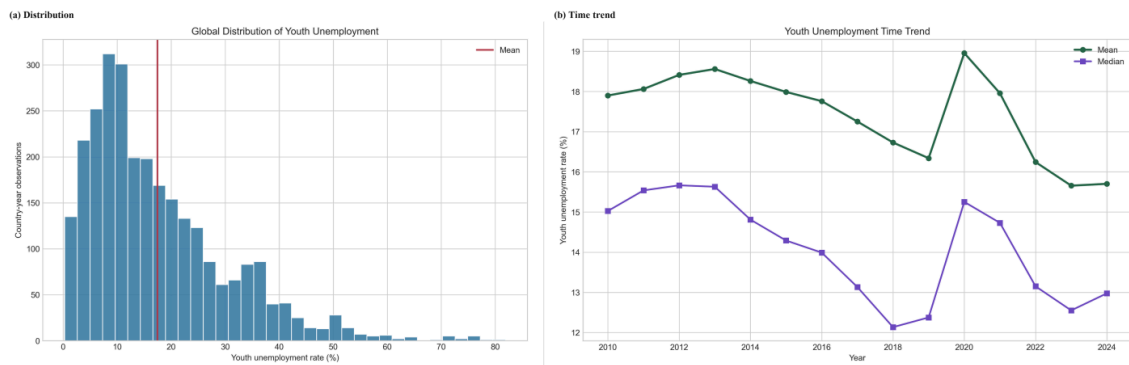


Figure 1. Descriptive patterns in youth unemployment

The main specification uses lag_youth_unemployment as the core predictor. The control variables are GDP growth, GDP per capita, inflation, tertiary enrollment, industry share, service share, and urbanization. The variables employment_ratio and labor_force_participation are not included in the main specification because they are conceptually close to unemployment outcomes and may raise concerns about outcome leakage.

The empirical strategy has four steps. OLS benchmarks persistence; Linear Regression, Lasso, Random Forest, XGBoost, and LightGBM are compared under a random country-year split; no-lag, 2021-2024 time-holdout, and country-holdout tests check generalization; and heterogeneity is assessed with country and year fixed-effects interaction models. The interaction models include the original controls and country-clustered standard errors. Omnibus Wald tests assess joint heterogeneity, and pairwise p-values are adjusted by Holm's method. Baseline-group regressions begin in 2013.

4. Results

Table 2. Selected OLS benchmark results for youth unemployment persistence

Variable	Coefficient	p-value	Interpretation
lag_youth_unemployment	0.9720	0.000	Strong positive persistence across years

Table 2. (continued)

gdp_growth	-0.1107	0.000	Higher growth is associated with lower youth unemployment
tertiary_enrollment	-0.0102	0.000	Negative conditional association
service_share	0.0124	0.052	Weak positive association

Table 2 reports the OLS benchmark for youth unemployment persistence. The coefficient on `lag_youth_unemployment` is 0.9720 and is statistically significant, indicating a strong positive association between lagged and current youth unemployment.

Selected coefficients are reported for brevity, with other control variables included in the specification.

The lag coefficient is the central OLS result. It indicates that youth unemployment has strong annual persistence, but it should be read as a predictive association rather than a causal effect.

Table 3 compares machine-learning performance under the random country-year split. LightGBM performs best, with an RMSE of 2.1053, an MAE of 1.3384, and an R-squared of 0.9725. Other models perform closely once the lagged variable is included.

Table 3. Model performance under random country-year split

Model	Test RMSE	Test MAE	Test R-squared
LightGBM	2.1053	1.3384	0.9725
XGBoost	2.1393	1.4015	0.9716
Random Forest	2.1996	1.3565	0.9700
Linear Regression	2.2096	1.3931	0.9698
Lasso	2.2124	1.3922	0.9697

Compared with the no-lag model, adding `lag_youth_unemployment` substantially improves prediction. The best no-lag model is LightGBM, with RMSE = 5.3561 and R-squared = 0.8223; the best lag model reaches RMSE = 2.1053 and R-squared = 0.9725. This is a key machine-learning finding.

Table 4. Prediction performance under different holdout settings

Split setting	Feature set	Best model	RMSE	MAE	R-squared
random country-year	lag model	LightGBM	2.1053	1.3384	0.9725
time holdout 2021-2024	lag model	Lasso	2.1521	1.4348	0.9688
country holdout	lag model	Linear Regression	2.1827	1.3803	0.9617

Table 4 reports prediction performance under different holdout settings. The lag-based model remains accurate not only under the random country-year split, but also when future years and unseen countries are held out. These results are prediction evidence, not causal evidence.

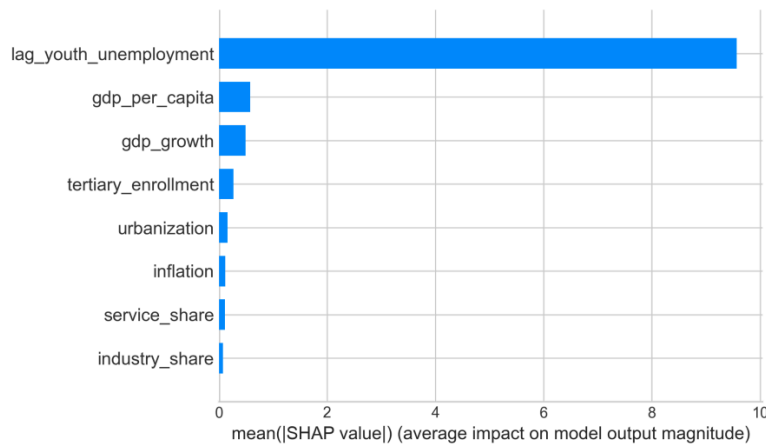


Figure 2. SHAP interpretation of the lag-based prediction model

SHAP results identify lag_youth_unemployment as the dominant predictor in the lag-based model. This indicates that the model relies heavily on the historical level of youth unemployment when predicting the next year's rate. This is consistent with the OLS coefficient and reinforces the persistence interpretation. SHAP values describe predictive contribution rather than causal marginal effects.

Table 5. Group-specific persistence slopes from formal interaction models

Grouping	Group	Lag slope	95% CI
Income	High income	0.8016	[0.7500, 0.8531]
Income	Upper middle income	0.8837	[0.8122, 0.9552]
Income	Lower middle income	0.7889	[0.7075, 0.8704]
Income	Low income	0.8522	[0.7024, 1.0021]
Baseline unemployment	Low baseline unemployment	0.7432	[0.6313, 0.8551]
Baseline unemployment	Medium baseline unemployment	0.7082	[0.6406, 0.7757]
Baseline unemployment	High baseline unemployment	0.8498	[0.8041, 0.8956]

The income-group interactions are not jointly significant (Wald $p=0.2229$), and none of the income-group pairwise differences remains significant after Holm adjustment. Baseline youth unemployment produces clearer heterogeneity. The baseline-group interactions are jointly significant (Wald $p=0.0023$).

The estimated persistence slopes are 0.7432 for the low-baseline group, 0.7082 for the medium-baseline group, and 0.8498 for the high-baseline group. After Holm adjustment, the high-baseline slope is significantly greater than the medium-baseline slope (adjusted $p=0.0022$); the other baseline-group contrasts are not significant at the 5% level.

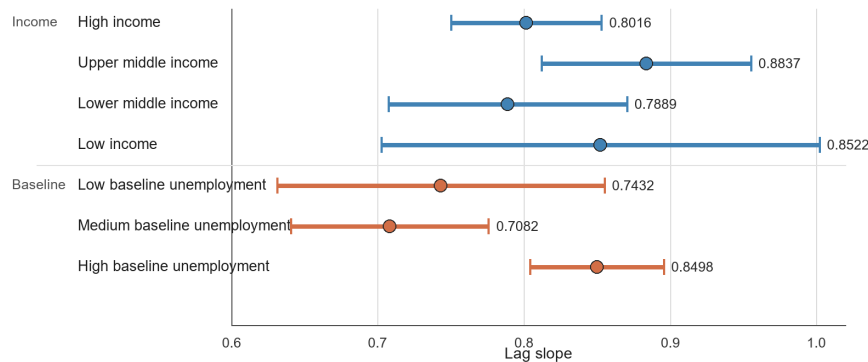


Figure 3. Formal heterogeneity estimates and 95% confidence intervals

5. Conclusion

This paper uses World Development Indicators data to predict youth unemployment persistence across countries from 2011 to 2024. The central finding is strong annual persistence: lagged youth unemployment is strongly associated with current youth unemployment, and adding the lagged variable substantially improves machine-learning accuracy.

The holdout tests support the robustness of this predictive pattern. The lag-based model remains accurate not only under a random country-year split, but also under time holdout and country holdout settings. Formal interaction tests do not identify significant overall differences across income groups. By contrast, baseline-unemployment heterogeneity is significant, with the high-baseline group showing stronger persistence than the medium-baseline group after Holm adjustment.

These findings suggest that youth unemployment should be understood not only as a cyclical or structural outcome, but also as a persistent labor-market risk. Policy discussion should emphasize early identification, continuous monitoring, education-job matching, and targeted support. The results are predictive associations rather than causal effects, and SHAP values explain predictions rather than causal mechanisms.

References

- [1] Okun, A. M. (1962). Potential GNP: Its measurement and significance. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 98-104.
- [2] Mincer, J. A. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research. <https://www.nber.org/books-and-chapters/schooling-experience-and-earnings>
- [3] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106. <https://doi.org/10.1257/jep.31.2.87>
- [4] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>
- [5] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- [6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- [7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- [8] World Bank. (n.d.). Unemployment, youth total (% of total labor force ages 15-24) (modeled ILO estimate). *World Development Indicators*. Retrieved June 1, 2026, from <https://data.worldbank.org/indicator/SL.UEM.1524.ZS>