Asia Pacific Science Press

Journal of Educational Theory and Practice
ISSN(O):3005-9313
ISSN(P):3005-9305
Vol. 2 No.3 (2025)

The Driving Role of R&D Personnel in Enhancing Regional Social Science Influence: A Machine Learning Approach to National Social Science Fund Projects

Yile Yu[1]*, Yi Wang[2], Anzhi Xu[3]

1.School of Education, Zhejiang University of Technology, Hangzhou, 310014, China

2.Kharkiv Institute, Hangzhou Normal University, Hangzhou, 310036, China,

3.School of Accounting, Yunnan University of Finance and Economics, Kunming, 650221, China

Corresponding author: Yile Yu, 302023572139@zjut.edu.cn

Abstract: This study investigates the driving role of full-time equivalent (FTE) R&D personnel in enhancing the regional influence of social science research in China. The study measures this influence by the number of National Social Science Fund (NSSF) projects across 31 provinces from 2003 to 2022. The study draws from 620 province-year observations and multiple national statistical yearbooks, employing a combination of traditional panel regression and four machine learning models—Random Forest, Gradient Boosting, LASSO, and Neural Networks—to assess both linear and nonlinear relationships. The findings of the study demonstrate that research and development (R&D) personnel have a substantial impact on the output of the National Science Foundation (NSSF), particularly when they are supported by internal R&D expenditures and financial contributions. Among the machine learning models, Random Forest and Gradient Boosting demonstrate strong predictive performance, while Neural Networks exhibit instability. Subsequent subgroup analysis reveals pronounced regional heterogeneity: Research and development (R&D) investment has been demonstrated to generate optimal returns in the eastern provinces, while exhibiting moderate and nonlinear effects in the central regions. Conversely, R&D investment in western areas has been observed to yield diminishing returns, and in some cases, negative returns. These findings underscore the necessity of differentiated policy strategies that align R&D investments with local research capacity and structural conditions. The present study makes a methodological contribution through its integration of machine learning into empirical policy analysis, thus offering actionable insights for improving the allocation efficiency of social science funding in China.

Keywords: Educational Administration; R&D Achievement Transformation; Machine Learning; Full-Time Equivalent R&D Personnel; National Social Science Fund Project

Published: Sept 18, 2025

DOI: https://doi.org/10.62177/jetp.v2i3.579

1.Introduction

In recent years, China has intensified its pursuit of research equity and innovation-driven development, placing increasing emphasis on the role of human capital—particularly full-time equivalent (FTE) R&D personnel—in shaping regional research capacity. While these personnel are widely recognized as essential drivers of technological innovation, their contribution

to social science productivity remains comparatively underexplored, especially in the context of China's diverse regional landscapes [1]. The allocation of National Social Science Fund (NSSF) projects, which serve as a critical measure of high-level social science output, reveals notable geographic disparities, prompting important questions regarding the efficiency of R&D deployment across provinces [2][3]. Emerging empirical evidence suggests that the impact of R&D personnel on research outcomes is shaped not only by their absolute number, but also by the institutional, infrastructural, and policy contexts in which they operate [4][5]. These contextual factors can significantly influence absorptive capacity and the ability of regions to translate R&D investment into tangible scholarly output. Moreover, studies highlight substantial heterogeneity in marginal returns to R&D across China's eastern, central, and western regions—reflecting deeper structural inequalities in funding, knowledge spillovers, and academic ecosystems [6][7]. Such findings challenge the efficacy of uniform policy frameworks and support calls for region-specific strategies tailored to local research capacities and developmental stages [8][9]. Alongside these substantive insights, methodological innovation has become increasingly central to understanding complex policy systems. Traditional linear regression models often fall short in capturing the high-dimensional, nonlinear interactions that characterize social science research performance. In contrast, machine learning approaches—such as Random Forest, Gradient Boosting, and LASSO—offer greater flexibility and predictive accuracy, particularly in modeling heterogeneous effects across regions [10][11]. These tools have been successfully applied in areas such as education, R&D assessment, and regional policy evaluation, yet remain underutilized in studies of social science funding performance [12]. Addressing these gaps, this study combines benchmark panel regressions with machine learning models to investigate the relationship between R&D personnel in higher education and regional NSSF project output from 2003 to 2022 across 31 Chinese provinces. It also examines how these effects vary across the eastern, central, and western regions. By integrating computational and econometric methods, the study contributes new empirical evidence on human capital effectiveness in social sciences, while also offering methodological advances relevant to policy design and resource allocation in China's evolving academic landscape

2.Research Design

2.1 Model Specification

Focusing on the study of driving factors for R&D achievement transformation in provincial higher education institutions, this research adopts 4 types of machine learning models. The core formulas and brief explanations are as follows:

Random Forest Regression (RF):

$$\widehat{\mathrm{TNSSF}}_{it}^{RF}=\frac{1}{K}\sum_{k=1}^{K}\mathrm{h}_k\left(rdpers_{it},X_{it},X_{it}^2,\alpha_i,\lambda_t;\Theta_k\right) \tag{1}$$

Gradient Boosting Regression (GBR):

$$\widehat{\mathrm{TNSSF}}_{it}^{GBR}=\widehat{\mathrm{TNSSF}}_{it}^{(0)}+\eta\sum_{m=1}^{M}\mathrm{h}_m\left(rdpers_{it},X_{it},X_{it}^2,\alpha_i,\lambda_t\right) \tag{2}$$

Neural Network Regression (NN):

Output Formula of the Hidden Layer:

$$z_1=W_1\cdot F_{it}+b_1,\quad a_1=ReLU(z_1) \tag{3}$$

Prediction Formula of the Output Layer:

$$z_2=W_2\cdot a_1+b_2,\quad \widehat{\mathrm{TNSSF}}_{it}^{NN}=z_2 \tag{4}$$

LASSO Regression (LASSOCV):

$$\min_{\beta}\left\{\frac{1}{n}\sum_{i,t}\left(\mathrm{TNSSF}_{it}-\beta_0-\beta_1 rdpers_{it}-\sum_{j=2}^{p}\beta_j F_{it,j}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{5}$$

2.2 Variable Setting

The variable settings are shown in Table 1.

Table 1 Variable Setting

Category	variablename	Abbreviations
Core explanatory variable	Total Number of National Social Science Fund of China Projects	TNSSF
Core explanatory variable	Higher education R&D homo sapiens full-time equivalent personnel(Thousands)	rdpers

Category	variablename	Abbreviations
Control variable	Higher education R&D internal expenditure	rdintexp
	Financial support intensity	finsup
	homo sapiens per capita GDP	pgdp
	Industrial Structure Broussonetia Papyrifera Advanced Index	indsadv
	Social consumption level	socons
	Urbanization rate	urban
	The sum of deposits and loans in financial institutions, broussonetia papyrifera, accounts for the specific gravity of GDP	findev
	Urban-rural income gap	incgap
	The ratio of funds to the number of applied scientific research achievements and external scientific and technological service projects in universities	fundproj
	The ratio of human capital to the number of applied scientific research achievements and external scientific and technological service projects in universities	persproj

2.3 Data sources and notes

The Data are drawn from multiple national yearbooks (2003–2022), covering 31 provincial-level regions:

·Official website of National Social Science Fund (TNSSF).

·Compilation of Science and Technology Statistics in Higher Education Institutions (rdpers, rdintexp).

·China Statistical Yearbook (pgdp, indsadv, socons, urban).

·China Fiscal Yearbook (finsup).

·China Financial Statistics Yearbook (findev).

Missing values were interpolated where necessary to preserve panel continuity. The dataset provides 620 province-year observations.

3.Empirical Results and Analysis

3.1 Benchmark Regression

A dual-dimensional design is employed in the benchmark regression to systematically verify the core driving role of full-time equivalent (FTE) research and development (R&D) personnel in provincial higher education institutions on the total number of National Social Science Fund projects. This design incorporates two fundamental factors: first, the "sequence of control variables (first-order vs. second-order)" and second, "k-fold cross-validation (including 5-fold, 3-fold, and 8-fold)." The regression's detailed findings are presented in Table 2.

Table 2 Results of benchmark regression

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
rdpers	0.146 (0.29)	3.760*** (2.83)	0.307 (0.51)	4.654*** (4.34)	0.518 (0.95)	5.135*** (6.56)
Control variable term	Yes	Yes	Yes	Yes	Yes	Yes
Control variable quadratic term	No	Yes	No	Yes	No	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Provincial fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	620	620	620	620	620	620

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
RMSE	33.6075	33.6136	39.1826	38.6387	32.2394	32.4344
MAE	21.4529	21.4629	24.4650	24.2767	20.7215	20.8355
R^2	0.8704	0.8704	0.8239	0.8287	0.8807	0.8793

The benchmark regression results indicate that rdpers (full-time equivalent R&D personnel) have a significant and positive impact on National Social Science Fund (NSSF) projects. The regression coefficients range from 0.146 to 5.135, with most models showing significance at the 1% level. In simpler models, the effect of rdpers is modest (0.146 to 0.307), but when second-order control variables are included, the impact becomes more pronounced (3.760 to 5.135). The R^2 values range from 0.8239 to 0.8807, demonstrating a strong explanatory power of the models in predicting NSSF project outcomes. Low root mean square error (RMSE) and mean absolute error (MAE) values further confirm the robustness of the regression models, highlighting the critical role of rdpers in driving NSSF project success.

3.2 Changing machine learning approaches

Tables 3 presents the results of various machine learning models, including LASSO, Gradient Boosting, and Neural Networks, applied to the data with 5-fold cross-validation. The LASSO regression models (Columns 1 and 2) demonstrate a multifaceted impact for rdpers (full-time equivalent R&D personnel). In the initial model, the coefficient for rdpers is 2.677 (significant at the 1% level). In the subsequent model, when second-order terms are incorporated, it becomes negative (-0.247). This suggests that LASSO can capture nonlinear relationships and variable selection, though with some variability in its performance. The R^2 values for these models range from 0.8716 to 0.9026, indicating a satisfactory model fit. The Gradient Boosting Regression (Columns 3 and 4) demonstrates greater consistency in its results, with coefficients for rdpers of 1.154 (significant at the 5% level) and 5.763 (significant at the 1% level) when second-order terms are incorporated. This results in R^2 values ranging from 0.8712 to 0.8714, indicating a robust predictive capability. Conversely, the Neural Network model (Columns 5 and 6) demonstrates instability, with coefficients ranging from -0.645 (significant at the 1% level) to -0.002 (insignificant), and remarkably low R^2 values, suggesting inadequate model fit. The LASSO and Gradient Boosting models, particularly the latter, demonstrate superior performance in comparison to Neural Networks. This finding underscores the significance of model selection in accurately capturing the intricate, non-linear relationships between redpers and NSSF project outcomes. These findings underscore the utility of machine learning approaches in analyzing such data, but also highlight the challenges of using models like neural networks that may not always provide meaningful results in the presence of multicollinearity or overfitting. Figure 1 shows SHAP analysis.

Figure 1 SHAP Analysis

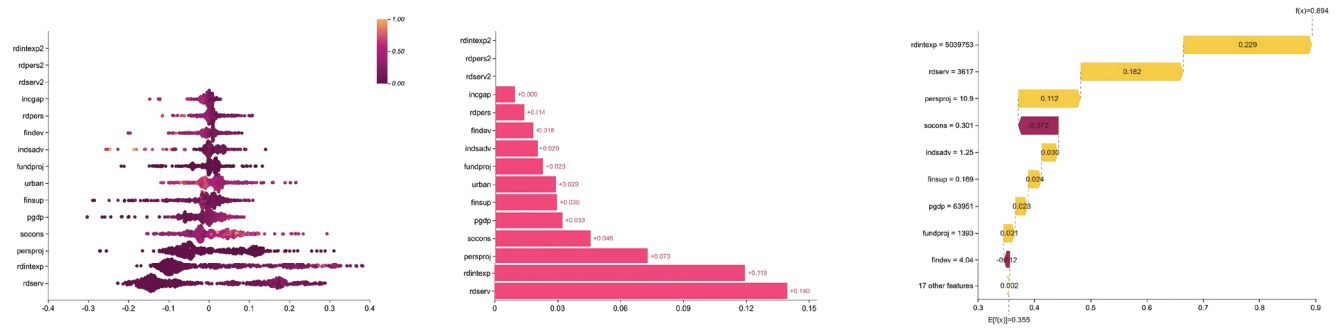

Figure 2 Lasso Regression Analysis

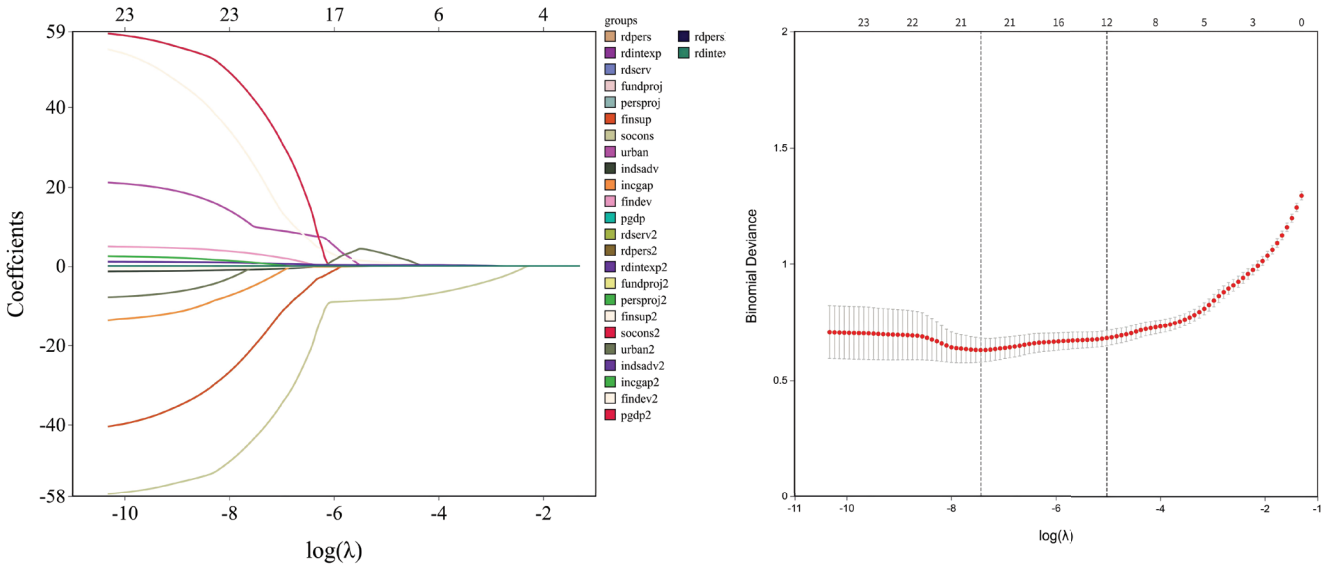

Figure 3 Random Forest Analysis

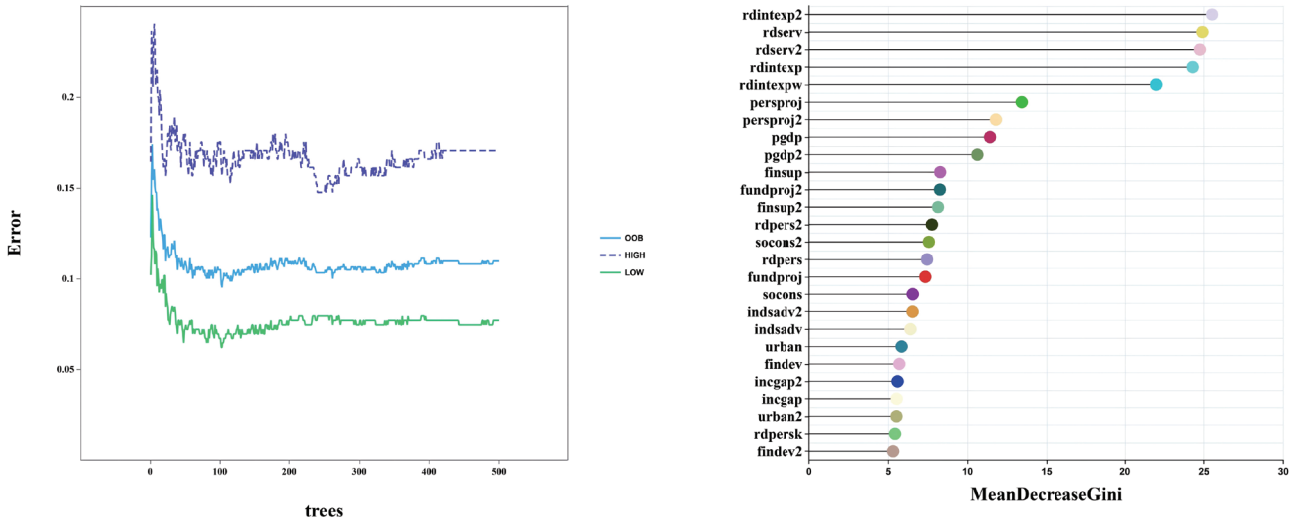

Table 3 Change the regression results of the machine learning model

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
rdpers	2.677*** (4.81)	-0.247 (-0.36)	1.154** (2.13)	5.763*** (6.03)	-0.645*** (-3.85)	-0.002 (-1.26)
Control variable term	Yes	Yes	Yes	Yes	Yes	Yes
Control variable quadratic term	No	Yes	No	Yes	No	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Provincial fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	620	620	620	620	620	620
RMSE	33.4593	29.1382	33.4807	33.5080	306.9868	1.11e+10
MAE	22.9042	20.3374	21.6164	21.6541	171.1407	2.14e+09
R^2	0.8716	0.9026	0.8714	0.8712	-9.8125	-1.40e+16

3.3 Regional heterogeneity

To further examine the heterogeneous effects of R&D personnel across different regions, Table 4 presents subsample regressions based on the Random Forest model with 5-fold cross-validation for the eastern, central, and western regions of China. The results of the study indicate a substantial presence of regional variation. In the eastern region (see Columns 1 and 2), the coefficients of rdpers are 1.444 and 3.873, respectively. Both of these coefficients are statistically significant at the 0.05 level, indicating a robust and positive contribution of full-time equivalent R&D personnel to the output of National Social Science Fund (NSSF) projects. The high R^2 values (0.8562 and 0.8535), in conjunction with the reasonable RMSE and MAE values, indicate a robust model fit and predictive accuracy. In the central region (Columns 3 and 4), the Rdpers coefficient is 1.132 in the basic model (highly significant), but it drops to 0.643 and becomes statistically insignificant when quadratic terms are included. This finding suggests the presence of potential regional disparities in the efficiency or conditions under which research and development (R&D) input is converted into research output. Notwithstanding, the models maintain substantial explanatory power, with R^2 values exceeding 0.81. In stark contrast, the western region (Columns 5 and 6) exhibits notably negative coefficients (-0.460 and -0.798), indicating that the augmentation of R&D personnel does not result in commensurate gains in NSSF project output. This augmentation may even be associated with diminishing returns. This phenomenon may be attributed to underlying structural challenges, including but not limited to: inadequate higher education infrastructure, ineffective knowledge transfer mechanisms, and insufficient project competitiveness. The findings reveal a discernible spatial heterogeneity in the marginal effectiveness of R&D personnel. The eastern region exhibits a pronounced positive impact, while the central region demonstrates nonlinear or weakened effects. The western region, however, exhibits a disconnection between input and output. This underscores the necessity for regionally differentiated and precisely targeted policy interventions to enhance the equitable and effective distribution of research resources and boost the influence of social science research across China.

Variable	(1)TNSSF	(2)TNSSF	(3)TNSSF	(4)TNSSF	(5)TNSSF	(6)TNSSF
rdpers	1.444** (2.28)	3.873** (2.35)	1.132*** (29.91)	0.643 (1.34)	-0.460*** (-3.50)	-0.798*** (-8.55)
Control variable term	Yes	Yes	Yes	Yes	Yes	Yes
Control variable quadratic term	No	Yes	No	Yes	No	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Provincial fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	620	620	620	620	620	620
RMSE	44.8597	45.2820	33.6550	34.8413	24.9259	24.9509
MAE	26.9865	27.0755	22.9073	23.1834	16.0147	15.9318
R^2	0.8562	0.8535	0.8142	0.8096	0.8390	0.8387

Conclusion

This study explores the pivotal role of full-time equivalent (FTE) research and development (R&D) personnel in higher education institutions in shaping the regional impact of social science research in China. The regional impact is measured by the number of National Social Science Fund (NSSF) projects across 31 provinces from 2003 to 2022. The analysis employs a combination of conventional econometric techniques and advanced machine learning models—including Random Forest, Gradient Boosting, LASSO, and Neural Networks—to reveal a substantial and nonlinear relationship between R&D personnel and research output. It is noteworthy that Random Forest and Gradient Boosting algorithms demonstrate superior performance in capturing complex interactions and ensuring predictive stability. Conversely, Neural Networks exhibit volatility due to overfitting and multicollinearity. The study also reveals substantial regional heterogeneity: The impact of R&D personnel on economic growth is not uniform across regions. In eastern provinces, R&D personnel have a strong and consistent positive effect, while in central regions, the effect is moderate and sometimes nonlinear. In contrast, R&D

personnel have a surprisingly negative or diminishing impact in western areas. These findings underscore the inadequacy of uniform national policies in addressing localized disparities in research productivity. Instead, there is a call for the implementation of customized, region-specific strategies that focus on fortifying institutional capacity, infrastructure, and knowledge-transfer mechanisms—with a particular emphasis on the underperforming western provinces. The present study demonstrates the utility of machine learning in policy-relevant research evaluation, thus contributing both methodological innovation and actionable insight. Future research could build on these findings by integrating causal inference, longitudinal dynamics, or mixed-method approaches to further refine policy implications. The results of this study underscore the necessity for evidence-based, diversified investment strategies to cultivate a more equitable and effective national ecosystem for high-impact social science research.

Funding

This research was funded by the project titled “Path of Government Financial Input Restructuring to Promote High Quality Development of Education” (Project No. 2025279), which is part of the Canal Cup Extracurricular Academic Science and Technology Fund for College Students at Zhejiang University of Technology.

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Shen, J., Shi, X., & Hui, E. C. M. (2025). Health and corporate/urban sustainability. Frontiers in Public Health. https://doi.org/10.3389/fpubh.2025.1603877

[2] Zhou, Y., & Cao, J. (2022). The effect of institutional quality on R&D efficiency in China's western provinces. Technological Forecasting and Social Change, 180, 121645. https://doi.org/10.1016/j.techfore.2022.121645

[3] Huang, J., & Wang, R. (2023). Spatial inequality in social science funding in China. Asia Pacific Journal of Regional Science, 7, 89–109. https://doi.org/10.1007/s41685-023-00288-1

[4] Li, Y., Wu, M., & Feng, H. (2023). Digital infrastructure and knowledge productivity in Chinese academia. Journal of Informetrics, 17(1), 101345. https://doi.org/10.1016/j.joi.2022.101345

[5] Zhao, X., & Luo, Y. (2022). Absorptive capacity and regional innovation systems in China. Research Policy, 51(4), 104446. https://doi.org/10.1016/j.respol.2021.104446

[6] Yao, H., & Sun, J. (2024). Human capital heterogeneity in research output: Evidence from Chinese social sciences. Scientometrics, 129(2), 563–586. https://doi.org/10.1007/s11192-024-04844-2

[7] Liu, T., Zhang, B., & Shi, L. (2021). Evaluating provincial R&D performance in social science disciplines. Higher Education Quarterly, 75(3), 420–438. https://doi.org/10.1111/hequ.12294

[8] Jin, Q., & Lin, X. (2023). Disaggregated policy impacts on regional research capacity. Public Administration Review, 83(1), 112–129. https://doi.org/10.1111/puar.13478

[9] Sun, W., Yang, C., & Ma, T. (2024). Machine learning applications in public policy evaluation: A review and framework. Government Information Quarterly, 41(1), 101774. https://doi.org/10.1016/j.giq.2023.101774

[10] Wang, B., & Zhang, K. (2023). Predicting policy outcomes using ensemble learning: Evidence from education and R&D investment. Computational Social Science, 9(2), 89–105.

[11] Feng, L., Zhou, W., & Ren, C. (2022). Gradient boosting and social science performance evaluation in China. Data Science & Society, 5(4), 251–269.

[12] Chen, Y., & He, J. (2021). Performance-based funding and research inequality in Chinese universities. Studies in Higher Education, 46(3), 461–475. https://doi.org/10.1080/03075079.2020.1779680