4th International Conference on Computational Finance and Business Analytics

Beyond Linear Assumptions: Decoding the Innovation Performance of High-Tech Enterprises with Machine Learning—Evidence from Guangxi's Top 100 Firms

CFBA-2610

Cao Liang and Yin RenZhan

School of Education and Humanities/ASEAN Vocational Research Center, Guangxi Polytechnic of Construction, Guangxi 530000, China; Scientific Research Center, Scientific research Center, Guangxi Polytechnic of Construction, Guangxi 530000, China

Introduction or abstract

This study challenges the monotonic assumptions of traditional linear models by leveraging machine learning to unravel the complex innovation-performance relationship in high-tech firms. Using 2022 data from Guangxi's top 100 enterprises, we first apply K-Means clustering to identify three archetypes: "Innovation Elites," "Capital Giants," and "Market Followers." An XGBoost model achieves high accuracy (R^2=0.893) in predicting per capita operating income, significantly outperforming linear regression. SHAP-based interpretation unveils key non-linear dynamics. A core finding is an inverted U-shaped relationship for R&D personnel proportion, where performance peaks at ~15% before diminishing returns emerge due to coordination costs. Sectoral heterogeneity is quantified: high-tech services rely on human capital (SHAP=+0.32), while advanced manufacturing hinges on fixed assets (SHAP=+0.28). Theoretically, this work refutes monotonicity in innovation studies. Methodologically, it demonstrates a powerful explainable AI framework. Practically, findings support differentiated strategies and evidence-based policies, offering a robust tool for assessing innovation and guiding resource allocation in regional ecosystems.

Objectives

This study's primary objective is to transcend the limitations of traditional linear models by unraveling the complex, non-linear relationships between innovation inputs and firm performance. Specifically, it aims to: 1) identify and quantify non-linear patterns, such as the hypothesized inverted U-shaped effect of R&D personnel proportion and the saturation points for patent density; 2) investigate the moderating role of sectoral heterogeneity, demonstrating how the drivers of innovation vary systematically between knowledge-intensive and capital-intensive industries; and 3) pioneer a robust, explainable AI (XAI) framework that combines the high predictive accuracy of XGBoost with the deep interpretability of SHAP analysis. Ultimately, the research seeks to provide a more nuanced, data-driven understanding that can guide both firm-level strategy and evidence-based industrial policy.

Materials & Methods

This study employs a three-stage analytical framework using data from the 2022 report on Guangxi's Top 100 High-Tech Enterprises for the 2021 fiscal year. The dependent variable, per capita operating income, is modeled against key independent variables including R&D personnel ratio, R&D intensity, patent density, and per capita fixed assets. **First**, K-Means clustering is applied to standardized data to identify distinct enterprise archetypes. **Second**, an XGBoost regression model, optimized through grid search and 5-fold cross-validation, is trained to predict innovation performance with high accuracy. **Third**, the "black box" nature of the XGBoost model is interpreted using SHAP (SHapley Additive exPlanations) to quantify each feature's non-linear contribution and interaction effects through summary plots, dependence plots, and force plots. The methodology's robustness was confirmed through alternative algorithms, cross-validation, and sensitivity analyses.

Results

The empirical results reveal several key findings. First, K-Means clustering successfully identified three distinct enterprise archetypes: "Innovation Elites" (39%), characterized by high R&D and patent density; "Capital Giants" (21%), defined by massive fixed assets; and "Market Followers" (40%), which lag across most metrics. Second, the XGBoost model demonstrated superior predictive power (R^2 = 0.893) compared to linear regression, identifying R&D personnel proportion as the most critical driver of performance. Most importantly, SHAP analysis uncovered significant non-linear relationships, validating the core hypotheses: the proportion of R&D personnel exhibits a clear inverted U-shaped effect on performance, with an optimal threshold around 15%, beyond which returns diminish. Furthermore, the analysis confirmed significant sectoral heterogeneity, showing that knowledge-intensive industries like biotechnology are driven by human capital, while capital-intensive sectors like advanced manufacturing rely more on fixed asset efficiency.

Conlcusion

This study concludes that the relationship between innovation inputs and performance in high-tech enterprises is fundamentally non-linear, challenging traditional linear assumptions. The research empirically validates an inverted U-shaped relationship for R&D personnel proportion, identifying an optimal threshold around 15%, and confirms that the key drivers of performance are highly heterogeneous across industries. Methodologically, the study demonstrates the value of a SHAP-XGBoost framework, which successfully balances high predictive accuracy with robust interpretability, overcoming the limitations of both traditional econometrics and "black-box" machine learning. These findings yield critical practical implications, advising firms to optimize R&D team size and tailor strategies to their specific industrial context, and urging policymakers to shift from one-size-fits-all incentives to precise, data-driven industrial policies. While limited by cross-sectional data, future research should leverage panel data and mixed-methods approaches to further explore causal dynamics and enhance the generalizability of these findings.

References

1.Guangxi Institute of Scientific and Technical Information. Report on the Top 100 High-tech Enterprises, Top 10 Innovation Capacity Enterprises, Top 10 Innovation Vitality Enterprises, and Top 10 Gazelle Enterprises in Guangxi (2022) [R]. Nanning: Guangxi Zhuang Autonomous Region Institute of Scientific and Technical Information (2022).
2.Nohria, N., Gulati, R.: Is slack good or bad for innovation? Academy of management Journal 39, 1245-1264 (1996).
3.Cohen, W.M., Levinthal, D.A.: Absorptive capacity: A new perspective on learning and innovation. Administrative science quarterly 35, 128-152 (1990).
4.Griliches, Z.: Issues in assessing the contribution of research and development to productivity growth. The bell journal of economics 92-116 (1979).
5.Hall, B.H., Jaffe, A., Trajtenberg, M.: Market value and patent citations: A first look. University Library of Munich, Germany (2002).
6.Hannan, M.T., Freeman, J.: Structural inertia and organizational change. American sociological review 149-164 (1984).
7.Tan, J., Peng, M.W.: Organizational slack and firm performance during economic transitions: Two studies from an emerging economy. Strategic management journal 24, 1249-1263 (2003).
8.Laursen, K., Salter, A.: Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. Strategic management journal 27, 131-150 (2006).
9.Pavitt, K.: Sectoral patterns of technical change: towards a taxonomy and a theory. Research policy 13, 343-373 (1984).
10.Malerba, F.: Sectoral systems of innovation and production. Research policy 31, 247-264 (2002).
11.Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems 30, (2017).

Acknowlegements or Contact

The authors would like to express their sincere gratitude to all those who supported this research. We also extend our appreciation to the relevant science and technology administrative departments in Guangxi for their valuable assistance in the data collection process regarding the top 100 high-tech enterprises, which provided the essential empirical foundation for this study. Finally, we wish to thank our colleagues, as well as the anonymous reviewers and editors, for their insightful comments and constructive feedback, which significantly contributed to refining our methodological framework and enhancing the overall quality of this manuscript.