Application of Machine Learning Models in Social Sciences: Managing Nonlinear Relationships

Subjects: Social Sciences, Mathematical Methods

Contributor: Theodoros Kyriazos , Mary Poga

The increasing complexity of social science data and phenomena necessitates using advanced analytical techniques to capture nonlinear relationships that traditional linear models often overlook. This chapter explores the application of machine learning (ML) models in social science research, focusing on their ability to manage nonlinear interactions in multidimensional datasets. Nonlinear relationships are central to understanding social behaviors, socioeconomic factors, and psychological processes. Machine learning models, including decision trees, neural networks, random forests, and support vector machines, provide a flexible framework for capturing these intricate patterns. The chapter begins by examining the limitations of linear models and introduces essential machine learning techniques suited for nonlinear modeling. A discussion follows on how these models automatically detect interactions and threshold effects, offering superior predictive power and robustness against noise compared to traditional methods. The chapter also covers the practical challenges of model evaluation, validation, and handling imbalanced data, emphasizing cross-validation and performance metrics tailored to the nuances of social science datasets. Practical recommendations are offered to researchers, highlighting the balance between predictive accuracy and model interpretability, ethical considerations, and best practices for communicating results to diverse stakeholders. This chapter demonstrates that while machine learning models provide robust solutions for modeling nonlinear relationships, their successful application in social sciences requires careful attention to data quality, model selection, validation, and ethical considerations. Machine learning holds transformative potential for understanding complex social phenomena and informing data-driven psychology, sociology, and political science policy-making.

machine learning in social sciences nonlinear relationships model interpretability predictive analytics imbalanced data handling

Overview of Nonlinear Relationships in Social Sciences

Nonlinear relationships are fundamental to understanding the complexity of social phenomena [1]. In much social science research, variables are traditionally assumed to interact in simple, proportional ways. However, this assumption often overlooks the reality that many relationships are inherently nonlinear [2][3]. In nonlinear relationships, changes in one variable do not consistently lead to proportional changes in another. Instead, the effect of a variable may vary based on other factors, leading to curvilinear, threshold, or even chaotic patterns.

These patterns are particularly prevalent in psychology, sociology, and demography, where human behavior and social systems exhibit dynamic, context-dependent interactions [4][5].

Linear models assume a direct, proportional relationship between independent variables and a dependent outcome [6]. For example, in a typical linear regression, each additional year of education is expected to result in a uniform increase in income, regardless of baseline education levels [7]. However, this assumption of uniform effects across all values fails to capture complexities. Nonlinear models, by contrast, allow for more flexible relationships, where the impact of a predictor may grow, shrink, or change direction depending on its value or the values of other variables [8]. For example, the effect of education on income might be modest up to a certain point, such as completing high school, but becomes significantly larger with further higher education.

Nonlinear models have emerged as essential tools for analyzing the complex interactions that linear models frequently overlook. For example, the effect of education on voting behavior can vary significantly across different socioeconomic groups and regions. Research indicates that higher educational attainment generally correlates with increased political engagement, but this relationship is not uniform. Individuals without a high school diploma exhibit minimal political engagement, but those with a college degree show marked increases in participation [9][10][11]. Similarly, cognitive performance follows a curvilinear trajectory: it tends to improve in adolescence and young adulthood, peaks in midlife, and declines in later years. This pattern exemplifies the limitations of linear models, which assume a constant rate of change and fail to account for varying improvements and declines across the lifespan [12][13].

Nonlinear relationships are also evident in the context of income and health outcomes. While higher income typically leads to better access to healthcare and improved health outcomes, the positive effects diminish beyond a certain income threshold. Once essential healthcare needs are met, further increases in income provide little additional health benefit [14][15]. These examples underscore the necessity of nonlinear models for capturing the intricate and multifaceted nature of social phenomena, providing a more accurate representation of dynamic relationships than linear models. This has important implications for policy-making and interventions addressing complex social issues [10][11].

Traditional linear models have been widely utilized in social science due to their simplicity and ease of interpretation. However, they often fail to capture the complexities of social phenomena, leading to significant limitations. One key drawback is their tendency to oversimplify relationships, such as assuming that social support always has a linear positive effect on mental health. This neglects the possibility of diminishing returns or adverse effects, such as dependency or stress from excessive support [16][17]. Moreover, linear models struggle with threshold effects, where a predictor’s influence only becomes significant after crossing a critical point. For example, the relationship between years of schooling and job satisfaction may only emerge after an individual obtains a formal degree or certification [18][19].

Linear models also face challenges in adequately representing interactions between variables. For instance, the effect of parental involvement on student achievement may vary depending on the school’s quality or the student’s socioeconomic background. While linear models can incorporate interaction terms, this can lead to issues like multicollinearity and over-specification, particularly in high-dimensional datasets [20][21]. Furthermore, the manual specification of interactions increases the likelihood of overlooking essential patterns in the data [22].

Finally, linear models can lead to misleading inferences when ignoring nonlinear relationships. For example, studies on the relationship between income and happiness often reveal that happiness levels off after a certain income threshold, contradicting the linear assumption that happiness increases indefinitely with income [23]. Neglecting this nonlinearity can result in policies that overemphasize income to enhance well-being while neglecting other factors like social relationships and personal fulfillment [24].

In summary, while traditional linear models offer a straightforward approach, they often fail to capture the complexity of social phenomena. Their limitations in simplifying relationships, missing threshold effects, and inadequately representing interactions highlight the need for more flexible models. Nonlinear models, which can better accommodate the intricate dynamics of social behavior and interactions, are crucial for generating meaningful insights in social science research [25]. As social science increasingly relies on large and complex datasets, adopting nonlinear modeling techniques becomes critical to avoid the pitfalls of overly simplistic assumptions.

Introduction to Machine Learning

Linear models have historically been the cornerstone of social science research due to their simplicity and interpretability. However, as social phenomena are increasingly recognized as complex, the limitations of linear models hinder their ability to capture the intricacies of human behavior, social interactions, and psychological processes. This has led to the adoption of more flexible approaches, such as machine learning (ML), which can handle the complexity of real-world social data [26][27][28][29][30].

Machine learning models differ from linear models in both their assumptions and goals. While linear models focus on estimating specific parameters to describe relationships between variables, ML models prioritize prediction and pattern recognition. This distinction is particularly relevant for social science, where the true relationships between variables are often unknown or highly complex. Machine learning algorithms can learn these relationships directly from the data, making them better suited to modeling nonlinear dynamics that traditional approaches might overlook [31][32][33].

A key strength of ML models is their ability to capture nonlinear relationships. Algorithms such as decision trees, random forests, and neural networks are designed to manage nonlinear interactions. Decision trees, for example, split data into branches based on decision rules, effectively capturing sudden changes or threshold effects. Neural networks can model complex nonlinear interactions through their layered architecture by adjusting connection weights between neurons, offering a more nuanced understanding of variable relationships [34][35][36].

ML models also excel in automatically detecting interactions and threshold effects, which would require manual specification in linear models. Random forests, for example, consist of multiple decision trees, each potentially uncovering different combinations of interacting variables contributing to predicting outcomes. This automatic detection is particularly beneficial in high-dimensional datasets, where the number of possible interactions makes manual specification impractical [37][38][39].

Another advantage of ML models is their robustness to noise and outliers. Ensemble methods like random forests mitigate the influence of outliers by averaging predictions across multiple trees, producing more stable results. Similarly, neural networks employ regularization techniques like dropout to reduce overfitting and increase resilience to noisy data, making them suitable for complex, real-world social science data [40][41][42].

The scalability of ML models for high-dimensional data is another significant advantage. Algorithms like support vector machines (SVMs) and gradient boosting machines (GBMs) are designed to handle large predictor sets and can automatically select the most relevant features. This capability is invaluable in sociology, psychology, and demography, where datasets are increasingly large and complex [43][44].

While linear models often focus on inference, machine learning models emphasize prediction accuracy. This shift is particularly relevant in domains like behavioral psychology and public health, where the primary goal is to predict outcomes—such as mental health disorders or voting behavior—rather than test specific hypotheses [45][46]. Machine learning's predictive power makes it a valuable tool for understanding and forecasting social phenomena [47][48].

However, despite their strengths in prediction and nonlinear modeling, ML models often sacrifice interpretability, which remains essential in social science research. Understanding the mechanisms driving observed relationships is as crucial as making accurate predictions [49][50]. Hybrid approaches are emerging to address this challenge. These involve using ML to explore nonlinear relationships and identify patterns, followed by applying more interpretable models like generalized additive models or decision trees to understand the nature of these relationships. New tools, such as SHAP values and LIME, enhance the ability to extract interpretable insights from even the most complex ML models [51][52].

Ultimately, while linear models have historically been foundational in social science research, the growing complexity of social data necessitates more flexible, data-driven approaches. With their ability to manage nonlinear dynamics, detect interactions, and scale to high-dimensional data, machine learning models offer a powerful alternative to traditional models. As social science evolves, machine learning will play a central role in uncovering the intricate, nuanced relationships that shape human behavior and societal outcomes [53][54].

References

1. Room, G. The Empirical Investigation of Nonlinear Dynamics in the Social World. Ontology, Methodology and Data. Sociologica 2020, 14, 163–193.

2. Kravchenko, S. The birth of “normal trauma”: The effect of nonlinear development. Econ. Sociol. 2020, 13, 150–159.
3. Strydom, G.; Ewing, M.T.; Heggen, C. Time lags, nonlinearity and asymmetric effects in an extended service-profit chain. Eur. J. Mark. 2020, 54, 2343–2363.
4. Girme, Y.U. Step out of line: Modeling nonlinear effects and dynamics in close-relationships research. Curr. Dir. Psychol. Sci. 2020, 29, 351–357.
5. Sanclemente Ibáñez, F.J.; Gamero Vázquez, N.; Arenas Moreno, A.; Medina Díaz, F.J. Linear and nonlinear relationships between job demands-resources and psychological and physical symptoms of service sector employees. When is the midpoint a good choice? Front. Psychol. 2022, 1329, 950908.
6. Hope, T.M. Linear regression. In Machine Learning; Academic Press: Cambridge, MA, USA, 2020; pp. 67–81.
7. Okoye, K.; Hosseini, S. Regression Analysis in R: Linear Regression and Logistic Regression. In R Programming: Statistical Data Analysis in Research; Springer Nature Singapore: Singapore, 2024; pp. 131–158.
8. Munir, K.; Kanwal, A. Impact of educational and gender inequality on income and income inequality in South Asian countries. Int. J. Soc. Econ. 2020, 47, 1043–1062.
9. Caffrey-Maffei, L. Education, Self-Importance, and the Propensity for Political Participation. Perceptions 2019, 5.
10. Oser, J.; Hooghe, M. Democratic ideals and levels of political participation: The role of political and social conceptualisations of democracy. Br. J. Politics Int. Relat. 2018, 20, 711–730.
11. Pellicer, M.; Assaad, R.; Krafft, C.; Salemi, C. Grievances or skills? The effect of education on youth political participation in Egypt and Tunisia. Int. Political Sci. Rev. 2022, 43, 191–208.
12. Dim, E.E.; Schafer, M.H. Age, Political Participation, and Political Context in Africa. J. Gerontol. Ser. B Psychol. Sci. Soc. Sci. 2024, 79, gbae035.
13. Pickering, D. Political activation and social movements: Addressing non-participation in Aotearoa New Zealand. Sociol. Compass 2023, 17, e13022.
14. Džunić, M.; Golubović, N. Civic and Political Participation in Transition Countries: The Case of Serbia. Facta Univ. Ser. Econ. Organ. 2018, 15, 001–013.
15. Kutuk, Y.; Usturali, A. The nonlinear relationship between political trust and nonelectoral political participation in democratic and nondemocratic regimes. Soc. Sci. Q. 2023, 104, 478–504.
16. Nickels, S.; Steinhauer, K. Prosody–syntax integration in a second language: Contrasting event-related potentials from German and Chinese learners of English using linear mixed effect models.

Second Lang. Res. 2018, 34, 9–37.

17. Weng, S.F.; Reps, J.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 2017, 12, e0174944.
18. Bone, A.E.; Gomes, B.; Etkind, S.N.; Verne, J.; Murtagh, F.E.; Evans, C.J.; Higginson, I.J. What is the impact of population ageing on the future provision of end-of-life care? Population-based projections of place of death. Palliat. Med. 2018, 32, 329–336.
19. Guimarães, M.H.; Sousa, C.; Garcia, T.; Dentinho, T.; Boski, T. The value of improved water quality in Guadiana estuary—A transborder application of contingent valuation methodology. Lett. Spat. Resour. Sci. 2011, 4, 31–48.
20. Laparra, V.; Malo, J. Visual aftereffects and sensory nonlinearities from a single statistical framework. Front. Hum. Neurosci. 2015, 9, 557.
21. Simpson, A.H.; Richardson, S.J.; Laughlin, D.C. Soil–climate interactions explain variation in foliar, stem, root and reproductive traits across temperate forests. Glob. Ecol. Biogeogr. 2016, 25, 964–978.
22. Wouters, A.; Pauwels, B.; Lambrechts, H.A.; Pattyn, G.G.; Ides, J.; Baay, M.; Meijnders, P.; Lardon, F.; Vermorken, J.B. Counting clonogenic assays from normoxic and anoxic irradiation experiments manually or by using densitometric software. Phys. Med. Biol. 2010, 55, N167.
23. Parkes, L.; Kim, J.Z.; Stiso, J.; Calkins, M.E.; Cieslak, M.; Gur, R.E.; Gur, R.C.; Moore, T.M.; Ouellet, M.; Roalf, D.R.; et al. Asymmetric signaling across the hierarchy of cytoarchitecture within the human connectome. Sci. Adv. 2022, 8, eadd2185.
24. Rørvik, E.; Fjæra, L.F.; Dahle, T.J.; Dale, J.E.; Engeseth, G.M.; Stokkevåg, C.H.; Thörnqvist, S.; Ytre-Hauge, K.S. Exploration and application of phenomenological RBE models for proton therapy. Phys. Med. Biol. 2018, 63, 185013.
25. Bonnebaigt, R.; Caulfield, C.P.; Linden, P.F. Detrainment of plumes from vertically distributed sources. Environ. Fluid Mech. 2018, 18, 3–25.
26. Alpaydin, E. Machine Learning; MIT Press: Cambridge, MA, USA, 2021.
27. El Naqa, I.; Murphy, M.J. What Is Machine Learning? Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 3–11.
28. Sammut, C.; Webb, G.I. (Eds.) Encyclopedia of Machine Learning; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
29. Wang, H.; Lei, Z.; Zhang, X.; Zhou, B.; Peng, J. Machine Learning Basics . 2016. Available online: http://whdeng.cn/Teaching/PPT_01_Machine%20learning%20Basics.pdf (accessed on 20 November 2024).

30. Zhou, Z.H. Machine Learning; Springer Nature: Berlin/Heidelberg, Germany, 2021.
31. Elhanashi, A.; Saponara, S.; Dini, P.; Zheng, Q.; Morita, D.; Raytchev, B. An integrated and real-time social distancing, mask detection, and facial temperature video measurement system for pandemic monitoring. J. Real-Time Image Process. 2023, 20, 95.
32. Levy, J.; Mussack, D.; Brunner, M.; Keller, U.; Cardoso-Leite, P.; Fischbach, A. Contrasting classical and machine learning approaches in the estimation of value-added scores in large-scale educational data. Front. Psychol. 2020, 11, 2190.
33. Yılmaz, K.; Turanlı, M. A multi-disciplinary investigation of linearization deviations in different regression models. Asian J. Probab. Stat. 2023, 22, 15–19.
34. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. Science 2015, 349, 255–260.
35. Hainmueller, J.; Mummolo, J.; Xu, Y. How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. Political Anal. 2019, 27, 163–192.
36. Wu, J.; Chen, S.; Zhou, W.; Wang, N.; Fan, Z. Evaluation of feature selection methods using bagging and boosting ensemble techniques on high throughput biological data. In Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology, Tokyo, Japan, 15–18 September 2020; pp. 170–175.
37. Mitchell, T.M.; Mitchell, T.M. Machine Learning; McGraw-hill: New York, NY, USA, 1997; Volume 1.
38. Morris, C.; Raman, S.; Seymour, S. Openness to social science knowledges? The politics of disciplinary collaboration within the field of UK food security research. Sociol. Rural. 2019, 59, 23–43.
39. Ray, L. Explaining Violence-Towards a Critical Friendship with Neuroscience? J. Theory Soc. Behav. 2016, 46, 335–356.
40. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. 2022, 23, 40–55.
41. Neuman, Y.; Cohen, Y. AI for identifying social norm violation. Sci. Rep. 2023, 13, 8103.
42. van Putten, I.; Kelly, R.; Cavanagh, R.D.; Murphy, E.J.; Breckwoldt, A.; Brodie, S.; Cvitanovic, C.; Dickey-Collas, M.; Dickey-Collas, M.; Melbourne-Thomas, J.; et al. A decade of incorporating social sciences in the integrated marine biosphere research project (IMBeR): Much done, much to do? Front. Mar. Sci. 2021, 8, 662350.
43. Lebaron, F.; Castro, T.A.F. Some contributions from Geometry to linear models’ construction in Social Sciences. Bull. Sociol. Methodol./Bull. Méthodol. Sociol. 2018, 140, 90–109.

44. Yuan, Y.; Zhu, W. Artificial Intelligence-Enabled Social Science: A Bibliometric Analysis. In Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Education (IC-ICAIE 2022), Chengdu, China, 24–26 June 2022; Atlantis Press: Dordrecht, The Netherlands, 2022; pp. 1602–1608.
45. Leach, M.; Scoones, I. The social and political lives of zoonotic disease models: Narratives, science and policy. Soc. Sci. Med. 2013, 88, 10–17.
46. Veltri, G.A. Big data is not only about data: The two cultures of modelling. Big Data Soc. 2017, 4, 2053951717703997.
47. Sarker, I.H. Machine learning: Algorithms, real-world applications and research directions. SN Comput. Sci. 2021, 2, 160.
48. Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. Electron. Mark. 2021, 31, 685–695.
49. Edelmann, A.; Wolff, T.; Montagne, D.; Bail, C.A. Computational social science and sociology. Annu. Rev. Sociol. 2020, 46, 61–81.
50. Li, Y.; Wang, S.; Song PX, K.; Wang, N.; Zhou, L.; Zhu, J. Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. Stat. Its Interface 2018, 11, 721.
51. Ahearn, C.; Brand, J.E. Predicting layoff among fragile families. Socius Sociol. Res. Dyn. World 2019, 5, 237802311880975.
52. Nakagawa, S.; Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol. Evol. 2013, 4, 133–142.
53. Kong, D.; Zhu, J.; Duan, C.; Lu, L.; Chen, D. Bayesian linear regression for surface roughness prediction. Mech. Syst. Signal Process. 2020, 142, 106770.
54. Playford, C.J.; Gayle, V.; Connelly, R.; Gray, A.J. Administrative Social Science Data: The Challenge of Reproducible Research. Big Data Soc. 2016, 3, 2053951716684143.

Retrieved from https://encyclopedia.pub/entry/history/show/128900