PRIN: RESEARCH PROJECTS OF RELEVANT NATIONAL INTEREST – Call 2020

Prot. 20205L79R8

"Towards a holistic approach to Sustainable Risk management in agriculture" Sus-Risk

Report

Deliverable D3.1 - Report assessing the performances of machine learning in comparison with traditional econometric analyses.

Document Title	Report	Author	Simone Severini, Biagini Luigi
Document type	Deliverable	Due date	01/12/2023
First issue		Ref.	
Dissemination level	Internal	Revised	

PROJECT	Towards a holistic approach to sustainable risk management in agriculture	Prot.	……
Call identifier	PRIN: RESEARCH PROJECTS OF RELEVANT NATIONAL INTEREST – 2020 Call for proposals Prot. 20205L79R8		
Work Package	3		
Deliverable n°	D3.1	Lead beneficiary:	
Coordinator			
Project Manager	…		
Project Start date	…		
Project Duration	36 month		

Report assessing the performances of machine learning in comparison with traditional econometric analyses

1 Introduction

The rapid expansion of agricultural data and computational advancements has sparked a methodological evolution in how risks are assessed, premiums are set, and outcomes are forecasted in the agricultural insurance sector. Two central approaches now contend for prominence: traditional econometric analyses, exemplified by Generalized Linear Models (GLM), and machine learning (ML) methodologies such as LASSO, Elastic Net, and Boosting. An in-depth appraisal of their performances reveals not only technical distinctions but also varied implications for policy, practice, and long-term sustainability of risk management tools in agriculture.

2 Traditional Econometric Approaches: Stability and Transparency

Traditional econometric analyses, particularly Generalised Linear Models (GLM), have long formed the backbone of insurance ratemaking and risk prediction models in agriculture. Strong theoretical underpinnings and clear interpretability characterise these methods. GLMs operate by positing a functional relationship—usually linear or log-linear—between explanatory variables and the predicted response (such as expected indemnity or premium). The specification of these models is informed by theory and existing literature, which provides transparency and makes the results comprehensible for a wide audience, including policymakers and practitioners. However, this interpretability comes at the cost of flexibility. GLMs require a priori identification of relevant variables and functional forms, making them sensitive to overfitting if too many predictors are included, or vulnerable to underfitting if key variables are omitted (Breiman, 2001; Efron and Hastie, 2016).

Moreover, in complex risk environments like agriculture, where the occurrence of adverse events (drought, flood) is rare but catastrophic and data distributions are often zero-inflated and fat-tailed, traditional models struggle to capture the full complexity of the data (Yang, Qian and Zou, 2018; Saha *et al.*, 2020). This often results in poor out-of-sample predictive performance and instability in the estimated parameters across different years, as seen in empirical results where GLM models exhibited higher RMSE, indicating greater errors in predicting indemnities. These weaknesses undermine the credibility of risk assessments and

threaten the financial sustainability of insurance pools when estimated premiums diverge from real risk, resulting in multi-annual deficits for insurers (Hastie, Tibshirani and Friedman, 2009; James *et al.*, 2013).

3 Machine Learning Approaches: Flexibility, Predictive Power, and Parsimony

Machine learning constitutes a fundamentally data-driven paradigm, emphasising predictive accuracy and adaptability rather than an explicit specification of theoretical relationships. Methods such as LASSO and Elastic Net are designed not only to fit complex data structures but also to select variables efficiently, limiting overfitting and reducing multi-collinearity even when hundreds of potential predictors are available (Tibshirani, 1996; Tay, Narasimhan and Hastie, 2021). Boosting, as an ensemble method, iteratively improves predictions by focusing on the hardest-to-predict cases, offering particularly impressive results in terms of minimising prediction error (Bühlmann and van de Geer, 2011).

Empirical studies in the context of agricultural insurance reveal the strengths of these ML approaches (Severini *et al.*, 2020). Not only did ML models such as LASSO and Boosting achieve much lower RMSE values than GLM, indicating superior goodness-of-fit, but they also managed to do so while selecting far fewer explanatory variables on average. This means insurers can potentially reduce information-gathering costs without sacrificing the accuracy or stability of premium estimates. The stability of ML methods, particularly Boosting, is further demonstrated in their out-of-sample performance—ML-based ratemaking consistently produced premiums closely aligned with realised indemnity levels, maintaining actuarial balance and financial sustainability over multiple years. The greater stability and predictive accuracy of ML approaches translated directly into more equitable and sustainable insurance offerings for farmers, addressing both the economic and social objectives of agricultural insurance policy.

The application of ML is particularly effective when paired with suitable probability distributions such as the Tweedie, which is adept at modeling the highly skewed and zero-inflated nature of agricultural indemnities—both areas where GLMs relying on standard distributions falter. Furthermore, the capacity of ML methods to uncover complex, non-linear relationships and adapt to variable-rich environments significantly expands their applicability, especially as the volume and richness of agricultural data grow through advances in digital technology and remote sensing

4 Trade-offs: Interpretability and Transparency

While the predictive superiority of ML is substantial, these methods do present challenges. Most notably, the "black box" quality of many ML algorithms, especially Boosting, complicates interpretability—a critical feature when stakeholders require clear explanations for premium rates or policy impacts. Traditional econometric models, though less powerful in prediction, remain easier to interpret and justify in regulatory or academic contexts, making them preferable in settings where transparency and communicability are paramount (Lv and Fan, 2009; Fan and Lv, 2010; Erasmus, Brunet and Fisher, 2020).

While new methods are emerging to address these gaps, the lack of robust, theory-based inference procedures can be a hindrance in settings that demand formal hypothesis testing or where decisions hinge on quantifying uncertainty.

5 Practical Implications: Cost, Robustness, and Policy Suitability

A central advantage of ML approaches is their efficiency in variable selection, which directly reduces the logistical and financial burdens of data collection. Especially for new or untested insurance products, ML frameworks allow for the design of fair, actuarially sound premiums without requiring an exhaustive and costly list of variables for every policyholder. This directly benefits both insurers, who see improved portfolio solvency, and farmers, who are offered more affordable and fair coverage aligned to their actual risk profiles (Varian, 2014; Storm, Baylis and Heckelei, 2020).

ML's robustness to changing data environments—such as year-to-year disturbances or shifts in underlying production conditions—further enables more resilient insurance design. In contrast, the traditional econometric reliance on static and often oversimplified specifications leaves insurers exposed to volatility and unforeseen shifts in claim experiences, ultimately threatening the financial sustainability of insurance pools.

However, the deployment of ML in real-world policy and insurance contexts must also contend with issues such as model governance, regulatory acceptance, and the need to maintain trust among policyholders, many of whom may be sceptical of opaque or seemingly inscrutable algorithms. Hybrid approaches that leverage machine learning (ML) for predictive tasks while retaining interpretable, generalised econometric frameworks for policy explanation and compliance may prove most effective, particularly as regulatory environments evolve in response to technological advancements.

6 Conclusion

In conclusion, the comparative assessment of machine learning and traditional

econometric analyses in agricultural insurance reveals a clear performance distinction. ML approaches—especially when equipped to handle distributional complexities and high-dimensional data—consistently provide greater predictive power, stability, and economic efficiency than traditional econometric models. This translates into more accurate, fair, and sustainable ratemaking, improving both the insurer's risk management and the insured's access to equitable coverage. Nonetheless, trade-offs remain, particularly in interpretability and the capacity for statistical inference, underscoring the continuing value of traditional econometrics in applications where transparency and regulatory scrutiny are required.

The future of risk assessment and insurance ratemaking in agriculture thus lies not in the wholesale replacement of traditional methods by machine learning, but in a strategic integration of both, harnessing the strengths of each to build more resilient, fair, and efficient insurance systems for a landscape characterized by rapid change and pervasive uncertainty

References

Breiman, L. (2001) 'Statistical modeling: The two cultures', *Statistical Science*, 16(3), pp. 199–215. doi: 10.1214/ss/1009213726.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data*, *Springer Series in Statistics*. Berlin, Heidelberg: Springer Berlin Heidelberg (Springer Series in Statistics). doi: 10.1007/978-3-642-20192-9.

Efron, B. and Hastie, T. (2016) *Computer Age Statistical Inference*. Cambridge: Cambridge University Press (Institute of Mathematical Statistics Monographs). doi: 10.1017/CBO9781316576533.

Erasmus, A., Brunet, T. D. P. and Fisher, E. (2020) 'What is Interpretability?', *Philosophy & Technology*. doi: 10.1007/s13347-020-00435-2.

Fan, J. and Lv, J. (2010) 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica*, 20(1), p. 101.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. Springer.

James, G. *et al.* (2013) *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer New York (Springer Texts in Statistics). doi: 10.1007/978-1-4614-7138-7.

Lv, J. and Fan, Y. (2009) 'A unified approach to model selection and sparse recovery

using regularized least squares', *Annals of Statistics*, 37(6 A), pp. 3498–3528. doi: 10.1214/09-AOS683.

Saha, D. *et al.* (2020) 'Application of the Poisson-Tweedie distribution in analyzing crash frequency data', *Accident Analysis and Prevention*, 137(October 2019), p. 105456. doi: 10.1016/j.aap.2020.105456.

Severini, S. *et al.* (2020) 'Applications of Machine Learning for the Ratemaking of Agricultural Insurances', (2005), pp. 1–42.

Storm, H., Baylis, K. and Heckelei, T. (2020) 'Machine learning in agricultural and applied economics', *European Review of Agricultural Economics*, 47(3), pp. 849–892. doi: 10.1093/erae/jbz033.

Tay, J. K., Narasimhan, B. and Hastie, T. (2021) 'Elastic Net Regularization Paths for All Generalized Linear Models'. Available at: http://arxiv.org/abs/2103.03475.

Tibshirani, R. (1996) 'Regression shrinkage and selection via the LASSO', *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Varian, H. R. (2014) 'Big Data: New Tricks for Econometrics', *Journal of Economic Perspectives*, 28(2), pp. 3–28. doi: 10.1257/jep.28.2.3.

Yang, Y., Qian, W. and Zou, H. (2018) 'Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models', *Journal of Business & Economic Statistics*, 36(3), pp. 456–470. doi: 10.1080/07350015.2016.1200981.