Can Machine Learning discover the determining factors in participation in insurance schemes?

A comparative analysis

1 Background

External shocks, such as extreme events in weather conditions, markets or policy, significantly impact agriculture. Farmers use various risk management tools to deal with these risks, where insurance takes the lion's share (Finger *et al.*, 2022).

The agricultural insurance literature (f.e., Meuwissen, Mey and van Asseldonk (2018) has analysed several aspects that affect the relationship between farmers and insurance. In particular, El Benni, Finger and Meuwissen (2016) emphasised the character of the variables selection and accuracy of prediction. Furthermore, the complementary effects of farm-specific characteristics and risk management strategies regarding both farm income and household income risk are analysed in El Benni, Finger and Mann (2012a) and Trestini *et al.* (2018), while the role of subsidies and the farm size in the stabilisation of farm income is evidenced in Aleksandrova, Zhmykhova and Viira, (2022). Moreover, Zubor-Nemes et al. (2018) highlighted the correlation between the economic performances of crop-producing farms with agricultural insurance contracts, finding that these farms outperform those who do not employ this risk-management instrument.

2 Topic and objectives of the analysis

We explore the elements that affect farmers' participation in an insurance scheme. This assessment is required to build or modify the structure of the insurance contract or to meet special insurance requirements based on the unique features of individual farms. Hence, this can support insurance companies and policymakers in creating contracts that satisfy farmers' needs.

This study analyses the (many) characteristics that potentially affect farmers' behaviour when considering participating in an insurance scheme using different Machin Learning tools. The number of characteristics influencing participation choice is usually large, making the task challenging. The additional problem is that these factors are interrelated and can mask the influence in the prediction of adhesions by misleading the forecast. Performing an accurate prediction and recognising the factors that affect farmers' participation are the main objectives of this analysis. Unfortunately, traditional methodologies (GLM) cannot satisfactorily use this large set of variables because of problems such as multicollinearity and overfitting. Problems that ML tools could overcome.

3 Methodology and data

The analysis uses individual data from the Italian FADN from 2016 to 2019. To have a homogeneous group of farms, we focus on field crop farms (type of farming 1), yielding 10,926 observations.Focusing only on one type of farm, we analyse the homogeneous class of farm insurances that covers crop production risk. To evaluate the participation in insurance schemes, we have focused on insurance subsidised by the Rural Development Program (RDP). Then, we identify the dichotomous dependent variable, taking the value of 1 when the farm buys subsidised insurance and zero otherwise.

We use 66 characteristics that the literature commonly considers to affect insurance participation choices. In particular, we consider economic, technical, financial, topographic and climatic characteristics (see f.e., (Mishra and El-Osta, 2001; Yee, Ahearn and Huffman, 2004; El Benni, Finger and Mann, 2012b; El Benni, Finger and Meuwissen, 2016; Severini, Tantari and Di Tommaso, 2016).

Because the number of participants is small (i.e., around 4% of the observations), we recur to simultaneous over- and under-sampling to create a valuable dataset for the estimation (Menardi and Torelli, 2014).

We use three Machine Learning (ML) approaches that are: LASSO, Boosting and Random Forest (Hastie, Tibshirani and Friedman, 2009; Storm, Baylis and Heckelei, 2020) to explore the issue and compare the results from these with those derived from a GLM model that has been traditionally used in insurance assessment. The considered ML approaches use a large set of variables by selecting the variable.

The ML approaches are analysed considering the following aspects: goodness-of-fit, ability to perform variable selection, and performing in variables setting. To compare the goodness-of-fit, we use Confusion Matrix analysis and metrics to compare predicted and observed values (MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error)). Moreover, one analyses the performance in variables selections, collinearity treatment and the ease-of-use (requirement of tuning).

Variable selection can be explored in two ways. First, consider the cases in which the single variables are selected. Second, consider the relative importance of each variable in affecting participation. All these results could be helpful in practice because focusing only on the subset of information that is more valuable to explain insurance participation could

reduce the cost of gathering and processing information and related costs.

4 Results

The main preliminary results are summarised qualitatively in the following table that allows to compare the ML approaches and the GLM models.

			GLM	LASSO	Boosting	Random Forest
Goodness-of-fit	Confusion Matrix Analysis	AUC	0.694	0.710	0.886	0.945
		Accuracy	0.854	0.864	0.952	0.894
		Sensitivity	0.428	0.454	0.776	0.491
		Specificity	0.959	0.965	0.995	0.994
		Posit. Prediction Value	0.722	0.763	0.977	0.950
		Negat. Prediction Value	0.872	0.878	0.947	0.888
		Detection Rate	0.117	0.118	0.157	0.102
		Balanced Accuracy	0.694	0.710	0.886	0.742
	Metrics	MAE	0.146	0.136	0.048	0.149
		MSE	0.146	0.136	0.048	0.079
		RMSE	0.382	0.369	0.219	0.281
Selection of Variables			66	64	41	66
Treatment of Collinearity			↓	↑	●	●
Automatic (requires little tuning)			↑	↑	●	●

Legend	↑ Good	● Fair	↓ Poor

Regarding the goodness-of-fit, one found Boosting overcame the performance of Random Forest and that, in turn, outperforms Lasso and GLM, which perform very poorly in predicting the number of farmers joining the subsidised insurance scheme. Despite the over-under sample techniques, the number of positive values is not detected in the same way by Boosting and Random Forest, with the latter resulting in poor performance (Sensitivity, Negative Prediction Value, Detection Rate and Balanced Accuracy). MAE, MSE and RMSE confirm the best Boosting performance, the poor outcomes reached by GLM and LASSO, and finally, Random Forest shows mixed results. Boosting also prevails in selecting variables: this model can reach high performance using only 41 variables on 66. Other models present low capacity in selection variables. These models present different capacities to fight collinearity, with LASSO as the best performers, followed by Boosting and Random Forest. Moreover, the powerlessness of GLM to select variables makes this model off the comparison. Finally, we must draw attention to various difficulties encountered while setting up these instruments: Contrary to Boosting and Random Forest, where it is essential to pay attention to specific non-automatic processes, GLM and LASSO do not provide the need for tuning.

Two additional aspects are under investigation, and extensive results will be provided in the full version of the paper: i) Which variables are selected the most? ii) Which variables are the most important? Preliminary results show that the most important factors that affect insurance participation (in order of importance) are: farm economic size, presence of other gainful activities, amount of utilised agricultural area, kW of available machinery, production diversification (Herfindahl index), degree of intensification (as total revenue per unit of utilised agricultural area), fixed capital on total capital, and mechanical expenses.

5 Discussion

Although participation in an insurance scheme is a complex decision, ML ensures relatively good prediction for sure better than GLM models. Within the considered ML approaches, Boosting offers better performances in this regard than the other two considered ML tools. Furthermore, it also uses a smaller set of variables as regressors. Conversely, the setting of Boosting can be challenging, and the evaluation of trade-offs with performance must be necessary to consider the different variables utilised in the estimation. The proposed ML tools allow identifying the essential variables in explaining participation choice. The general conclusion is that ML is a helpful tool for exploring the factors that explain farmers' participation in insurance schemes. Furthermore, results obtained using these approaches can be useful to better design insurance schemes and, hopefully, boost farmers' participation. Therefore, the ML approach is a key step that should be done carefully considering the characteristics of the empirical case study.

References

Aleksandrova, O., Zhmykhova, T. and Viira, A. (2022) 'The role of subsidies in stabilising farm income: Evidence from Estonia', *Agricultural and Food Science*, 31(1), pp. 24–36. doi: 10.23986/afsci.112241.

El Benni, N., Finger, R. and Mann, S. (2012a) 'Effects of agricultural policy reforms and farm characteristics on income risk in Swiss agriculture', *Agricultural Finance Review*, 72(3), pp. 301–324.

El Benni, N., Finger, R. and Mann, S. (2012b) 'The effect of agricultural policy change on income risk in Swiss agriculture', *123rd EAAE Seminar - Price Volatility and Farm Income Stabilisation: Modelling Outcomes and Assessing Market and Policy Based Responses*, p. 16. doi: 10.1108/00021461211277204.

El Benni, N., Finger, R. and Meuwissen, M. P. M. (2016) 'Potential effects of the income stabilisation tool (IST) in Swiss agriculture', *European Review of Agricultural Economics*, 43(3), pp. 475–502.

Cai, J., de Janvry, A. and Sadoulet, E. (2020) 'Subsidy Policies and Insurance Demand', *American Economic Review*, 110(8), pp. 2422–2453.

Diaz-Caneja, M. B. *et al.* (2008) *Agricultural Insurance Schemes*. Edited by Office for Official Publications of the European Communities. Luxembourg: Office for Official Publications of the European Union.

Enjolras, G., Capitanio, F. and Adinolfi, F. (2012) 'The demand for crop insurance: Combined approaches for France and Italy', *Agricultural Economics Review*. 2012, 13(1), pp. 5–22. doi: 10.2139/ssrn.1836798.

Feng, H., Du, X. and Hennessy, D. A. (2020) 'Depressed demand for crop insurance contracts, and a rationale based on third generation Prospect Theory', *Agricultural Economics*, 51(1), pp. 59–73. doi: 10.1111/agec.12541.

Finger, R. *et al.* (2022) 'The Importance of Improving and Enlarging the Scope of Risk Management to Enhance Resilience in European Agriculture', *Resilient and Sustainable Farming Systems in Europe*, pp. 18–37. doi: 10.1017/9781009093569.003.

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

Knapp, L. *et al.* (2021) 'Revisiting the diversification and insurance relationship: Differences between on- and off-farm strategies', *Climate Risk Management*. The Author(s), p. 100315. doi: 10.1016/j.crm.2021.100315.

Menardi, G. and Torelli, N. (2014) *Training and assessing classification rules with imbalanced data, Data Mining and Knowledge Discovery*. doi: 10.1007/s10618-012-0295-5.

Meuwissen, M. P. M., van Asseldonk, M. A. P. M. and Huirne, R. B. M. (2008) *Income stabilisation in European agriculture: Design and economic impact of risk management tools, Income Stabilisation in European Agriculture: Design and Economic Impact of Risk Management Tools*. Edited by M. P. M. Meuwissen, M. A. P. M. van Asseldonk, and R. B. M. Huirne. The Netherlands: Wageningen Academic Publishers. doi: 10.3920/978-90-8686-650-2.

Meuwissen, M. P. M., Mey, Y. De and van Asseldonk, M. (2018) 'Prospects for agricultural insurance in Europe', *Agricultural Finance Review*, 78(2), pp. 174–182. doi: 10.1108/AFR-04-2018-093.

Mishra, A. K. and El-Osta, H. S. (2001) 'A temporal comparison of sources of variability in farm household income', *Agricultural Finance Review*. Emerald, 61(2), pp. 181–198. doi: 10.1108/00214820180001123.

Severini, S., Tantari, A. and Di Tommaso, G. (2016) 'The instability of farm income. Empirical evidences on aggregation bias and heterogeneity among farm groups', *Bio-based and Applied Economics*, 5(1), pp. 63–81. doi: 10.13128/BAE-16367.

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.

Trestini, S. *et al.* (2018) 'Assessing the risk profile of dairy farms: application of the Income Stabilisation Tool in Italy', *Agricultural Finance Review*, 78(2), pp. 195–208. doi: 10.1108/AFR-06-2017-0044.

Yee, J., Ahearn, M. C. and Huffman, W. (2004) 'Links among Farm Productivity, Off-Farm Work, and Farm Size in the Southeast', *Journal of Agricultural and Applied Economics*. Cambridge University Press ({CUP}), 36(3), pp. 591–603. doi: 10.1017/S1074070800026882.

Zubor-Nemes, A. *et al.* (2018) 'Farmers' responses to the changes in Hungarian agricultural insurance system', *Agricultural Finance Review*, 78(2), pp. 275–288. doi: 10.1108/AFR-06-2017-0048.