Predicting profitability of dairy farms with a learning ensemble model

Maria Yli-Heikkilä, Jukka Tauriainen, Mika Sulkava

MTT Agrifood Research Finland

European Union's (EU) agricultural policy aims at viable food production, sustainable management of natural resources and balanced development across all Europe's rural areas. The Farm Accountancy Data Network (FADN) is an instrument for evaluating the income of agricultural holdings and the impact of the EU's agricultural policy. Every year, the member states collect accountancy data from a sample of the agricultural holdings. The Finnish accountancy data is also anonymously available for academic researchers. Thus, we have been able to use an extensive set of variables and observations for modeling. Our aim has been to develop a tool for individual farmers in the Finnish dairy sector to estimate the profitability of their business.

Finnish family farms are not subject to an accounting obligation. Their income statement is typically based on cash-based single-entry bookkeeping and is prepared for tax purposes only. Estimating the financial performance of a business requires, in addition, determining the monetary value of farms assets, liabilities, and capital. In a typical case, with a limited amount of accounting data, the farmer cannot calculate financial indicators, such as profitability ratios. Therefore, the farmer cannot assess the sustainability of the business over time or in comparison to other farms in the sector. To address this issue, we have built a model that predicts the profitability ratio from such variables that are at hand for a farmer.

The response variable profitability ratio indicates how operative costs including family factors, meaning the wage claim and the interest claim of agriculture, are covered by family farm income. As a relative concept, profitability ratio is well suited for comparisons between years, as well as for farms representing different size classes and production sectors.

In this study, we applied a learning ensemble method (e.g. Friedman et al., 2008), which combines several prediction models based on different learning algorithms. Ensemble prediction is then a linear combination of the predictions of each of the ensemble members.

We extracted unbalanced panel data from the FADN database. We combined data from the years 2000-2011, resulting in 4228 observations of 334 variables on Finnish dairy farms. We split the data into a training set and a test set by 2/3. The training set was first used for variable selection.

Random forest (RF), first introduced by Breiman (2001), is a supervised learning algorithm based on decision trees. RF combines tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Internal esti-

mates of the model are also used to measure variable importance. (Breiman, 2001.)

In the variable selection, we applied the backwards selection wrapper algorithm, as described in Kuhn (2014). The algorithm produces several orderings of variables by RF's computation of importance measures on each training set of a 10-fold cross-validation. The procedure was repeated five times to smooth out the variability. Based on the mean difference of prediction accuracies observed for each tree in terms of mean squared error before and after random permutation of a predictor variable, RF identified 20 variables as the most informative predictor set. The final model was then validated with the test set data. The predictive performance was measured in terms of root mean squared error $RMSE = 0.25$ and adjusted $R^2 = 0.68$. The predictors included in the final model indicated productivity (annual cattle care workload per produced milk), scale of operations (total income in relation to expenses, profit/loss, advance payment of tax), indebtedness (interest costs), and level of investment (tax deductions on production facilities and support payment entitlements).

In the second phase, five other prominent predictive model methods were tested. The models were the generalized linear model (GLM), generalized boosted regression model (GBM), multivariate adaptive regression splines (MARS), projection pursuit regression (PPR), and support vector machine (SVM) with a Gaussian radial basis function. The models produced RMSE metrics on the test set as follows: 0.26 (GLM), 0.23 (GBM), 0.22 (MARS), 0.23 (PPR), and 0.26 (SVM). Corresponding adjusted R^2 metrics were: 0.67 (GLM), 0.73 (GBM), 0.75 (MARS), 0.73 (PPR), and 0.67 (SVM). RF, GBM, MARS, PPR, and SVM were then combined into a linear regression ensemble. The prediction accuracy outperformed the individual models by having $RMSE = 0.22$ and adjusted $R^2 = 0.76$.

We have previously used a profitability prediction tool based on random forest, in a web application serving farmers. As the ensemble prediction model outperforms our earlier model based on random forest, we will replace our existing tool with the new model.

Keywords: Predictive modelling, learning ensembles, dairy farms, profitability.

References:

Breiman, L. (2001). Random Forests *Machine Learning, Volume 45(1)*, 5-32.

Friedman, J. H., Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics 2 (2008), no. 3*, 916–954.

Kuhn, M. (2014). Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer and the R Core Team. *caret: Classification and Regression Training.* R package version 6.0-24.