Profitability prediction model for dairy farms using the random forest method

Maria Yli-Heikkilä[1], Jukka Tauriainen[2]

[1] Agrifood Research Finland MTT, Economic Research, Animale, Tietotie, 31600 Jokioinen, Finland, email: maria.yli-heikkila@mtt.fi

[2] Agrifood Research Finland MTT, Economic Research, Kampusranta 9 C, 60320 Seinäjoki, Finland, email: jukka.tauriainen@mtt.fi

Poster paper prepared for presentation at the EAAE 2014 Congress
'Agri-Food and Rural Innovations for Healthier Societies'

August 26 to 29, 2014
Ljubljana, Slovenia

Abstract

We applied an ensemble learning method known as random forests, which is widely used in supervised machine learning, to predict the profitability ratio of dairy farms based on financial and production related variables. The predictive model was implemented as a web service to enable farmers to calculate the profitability of their business. Hereby, farmers can better assess the sustainability of their business over time, or in comparison to other farms in the sector.
Keywords: Predictive modelling, random forests, learning algorithms, dairy farms, profitability.

1. Introduction

The European Union's (EU) agricultural policy aims at viable food production, sustainable management of natural resources and balanced development across all of 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, member states collect accountancy data from a sample of agricultural holdings. Based on the The Finnish accountancy data, which is anonymously available to academic researchers, we have been able to use an extensive set of variables and observations for modelling. Our aim was to develop a tool enabling 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. In addition, estimating the financial performance of a business requires the determination the monetary value of a farm's assets, liabilities, and capital. Due to the limited amount of accounting data available, in most cases farmers are unable to calculate financial indicators, such as profitability ratios. This prevents them from assessing the sustainability of their 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 based on variables available to all the farmers.

The profitability ratio indicates how operative costs, including family factors – i.e. the wage and interest claims – are covered by family farm income. A profitability ratio of 1.0 indicates that all production costs, including the costs of family factors (opportunity costs), have been covered and the entrepreneur's profit is zero. As a relative concept, the profitability ratio is ideal for making comparisons between years, as well as for farms from various size classes and production sectors.

2. Method

In this study, we applied an ensemble learning method called random forests (RF), which is widely used in supervised machine learning. RF are frequently applied due to their high prediction accuracy and ability to identify informative variables. RF are equally applicable to regression and classification problems. As in a regression case, a random forest is a predictor consisting of a collection of randomized base regression decision trees which are combined to form the aggregated regression estimate. Breiman (2001) first demonstrated that substantial gains can be achieved in classification and regression accuracy by using ensembles of decision trees, where each tree in the ensemble is grown in accordance with a random parameter.

We extracted unbalanced panel data from the FADN database. We combined data from years 2000-2012, resulting in 4370 observations of 300 variables of Finnish dairy farms. We split the data into a training set and test set, on a 2/3 basis respectively. The training set was used for variable selection, model fitting and performance evaluation.

In variable selection, we applied a backwards selection algorithm, as described in Kuhn (2014). The algorithm produces several orderings of variables, by computing importance measures for each training set in a 10-fold cross-validation. The procedure was repeated five times in order to smooth out variability. Based on the mean difference of the prediction accuracies observed for each tree in terms of the mean-squared error (MSE) before and after random permutation of a predictor variable, RF identified 35 variables as the most informative predictor set. The 35 predictors included in the 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, amount of silage), indebtedness (interest costs), and level of investment (tax deductions on production facilities and support payment entitlements). The final model was then validated using the test set data. The predictive performance was measured in terms of root-mean-squared error RMSE $= 0.26$ and adjusted $R^2 = 0.66$ on the test set.

As the aim of the study was to develop an online tool based on the predictive model, we also had to consider the usability of the application. Could the model have fewer predictors with a tolerable decline in predictive performance? We revised the variable selection results. Figure 1.a shows that the RMSE improves only slightly when the number of predictor variables increases. We used a tolerance function from Kuhn (2014) in order to select the least complex model within 1% loss of performance of the best RMSE value. The new set of predictors included 14 variables, as shown in Figure 1.b.

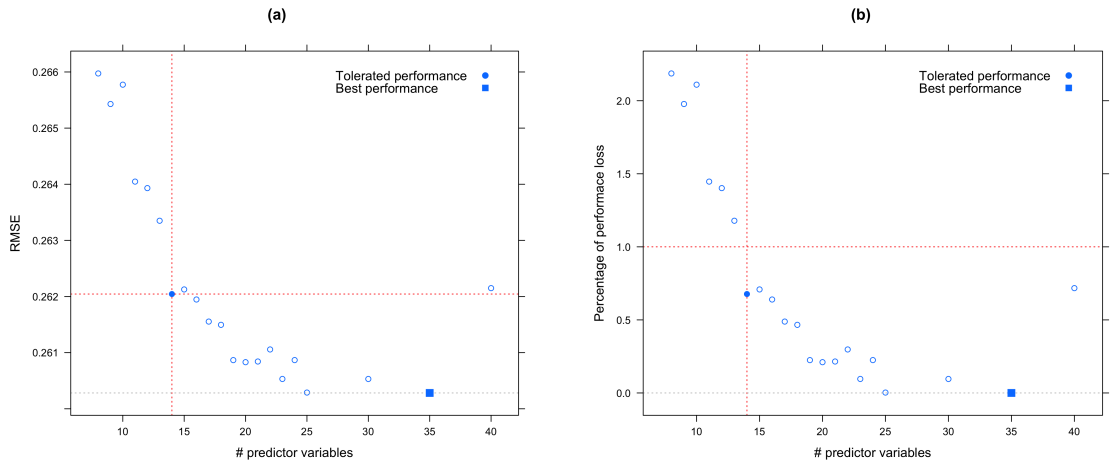


Figure 1. Variable selection. The highest accuracy in terms of the RMSE is achieved with 35 predictor variables. However, if 1% loss of performance is tolerated, a model with 14 predictors achieves only slightly higher RMSE of 0.262.

Finally, we also considered the most evident question a user might pose in relation to our model: how reliable is a prediction? With a standard prediction, a single point estimate (a conditional mean) is returned for each new instance. Such a point estimate does not contain information on the dispersion of observations around the predicted value. We must therefore calculate the prediction intervals of a new observation. Meinshausen (2006) showed that RF provide information on the full conditional distribution of the response variable, not only on the conditional mean. Conditional quantiles can be inferred with quantile regression forests, which is a generalisation of RF (see Meinshausen, 2006). Quantile regression forest (QRF) algorithm is computationally heavier than Breiman's random forests and performs slightly worse by having RMSE $= 0.26$ and adjusted $R^2 = 0.64$ on test set.

The new set of predictor variables was also tested on other prominent predictive model

methods, namely, a generalized linear model (GLM, Nelder and Wedderburn, 1972), a generalized linear model via penalized maximum likelihood (GLM+, Friedman et al., 2010) and a support vector machine (SVM, Cortes and Vapnik, 1995) with a Gaussian radial basis function (Karatzoglou et al., 2004). The RF model moderately outperformed these in terms of accuracy, with the other models producing RMSE metrics on the test set: 0.29 (GLM), 0.29 (GLM+), 0.32 (SVM). The corresponding adjusted R^2 metrics were as follows: 0.63 (GLM), 0.56 (SVM), 0.63 (GLM+). Figure 4 combines the predictive performance of all the models on validation and test set data.

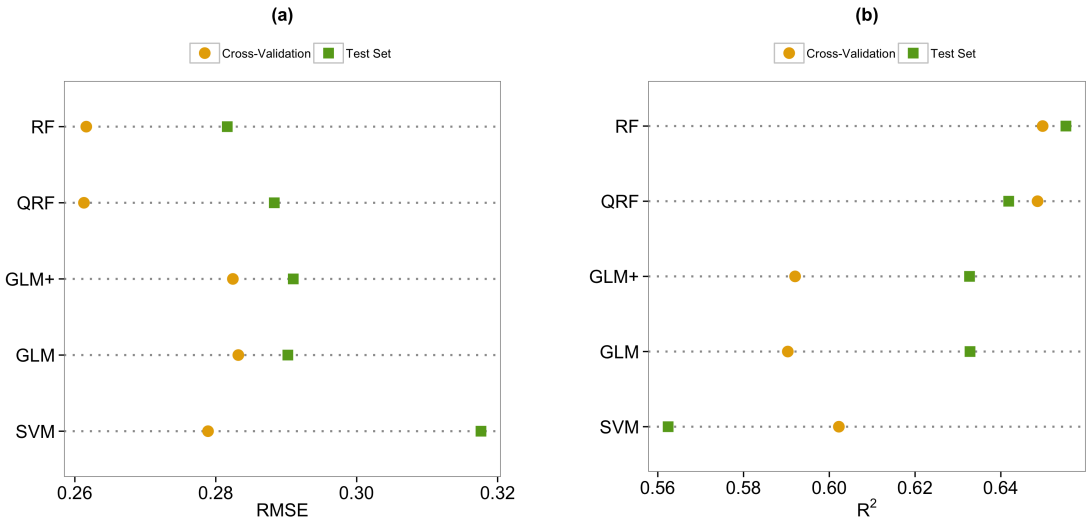


Figure 2. Predictive performance of the models. RF shows the best performance both in terms of the RMSE and the adjusted R^2 metrics.

3. Results

The predictors in the final model are presented in Figure 3 in the form of a word cloud. The predictors are ranked with the importance score, which is the total decrease in node impurities, measured by the Gini index from splitting on the variable, averaged over all trees. The font size indicates the relative importance of the variable in the predictive model. Clearly, the net farm result has the highest score. The net farm result is the interest to the equity invested in the enterprise (taxes not subtracted). The gross return has the second highest score followed by the annual wage claim and the ratio of total expenses and revenues. Three predictors are related to depreciations on support payment entitlements, which indicates growing scale of investment, i.e. entitlements have been transferred (sale, lease) to another farmer or a newcomer.

Note that the importance score tells the ranking of variables for model refining purposes. Although RF outperform the other approaches in this study in accuracy, RF provide little understanding of the data beyond predictions. Therefore, the importance score should not be used to for inferring the interplay of the predictors. In fact, it is a subject of further research to analyze the mechanism of the model with simulations. However, the predictor set indicates that profitability is related to the productivity, the scale of operations, indebtedness, and the level of investments.

QRF method was applied to build a model to predict the profitability ratio of a dairy farm. In addition to superior accuracy over other prominent predictive model approaches, QRF the

provides prediction intervals for a point prediction, and thus provides information about the reliability of a future prediction. Figure 4 shows how the prediction intervals are narrower within the lower and upper quartiles of the response variable, i.e. profitability ratio. The development of profitability ratio of Finnish dairy farms has been quite stable in average (Figure 4.a). The median has fluctuated between 0.5 and 0.7. However, the interquartile range has widened in this period. Figure 4.b shows the 90% prediction intervals on the validation data.

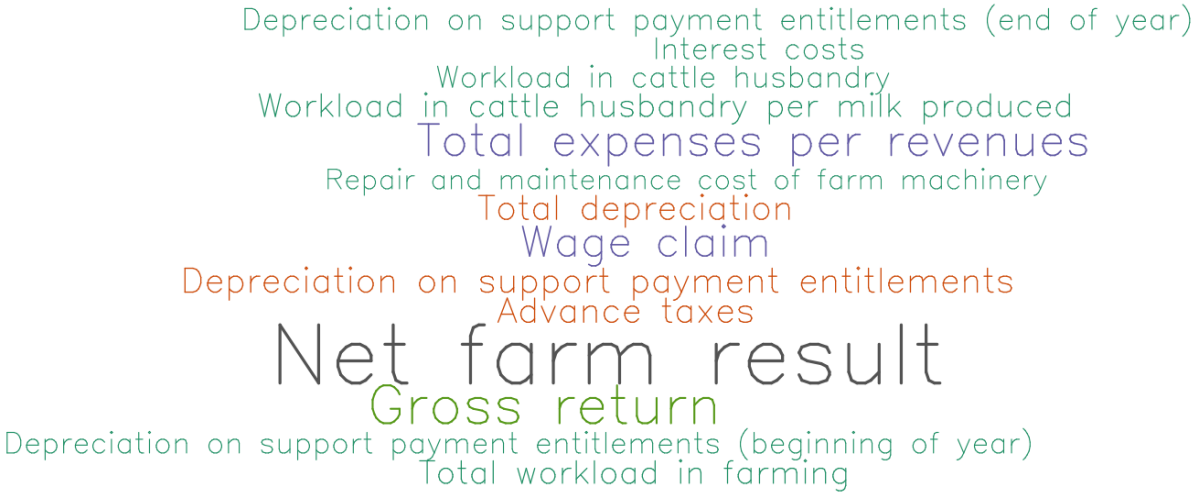
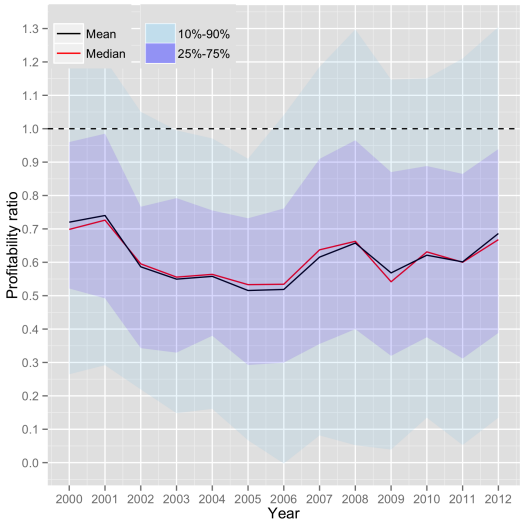
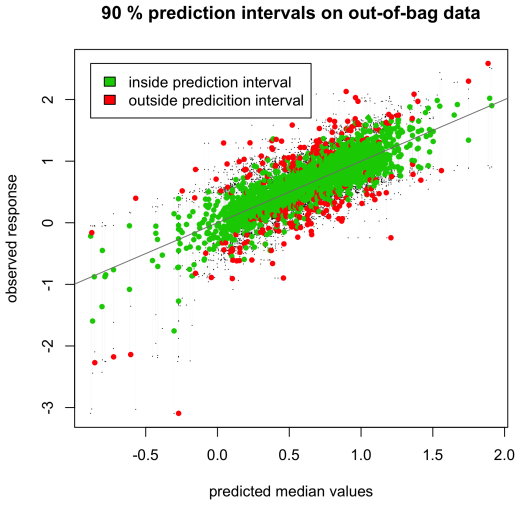


Figure 3. The predictor variables presented as a word cloud. The font size indicates the relative importance of the variable in the predictive model. The ranking of the variables is based on the importance score produced by the random forest algorithm.



(a) The profitability of Finnish dairy farms in 2000-2012 (n ≈ 336 farms annually).



(b) Predicted vs. observed values with 90% prediction intervals on validation data.

Figure 4. The development of the profitability of the Finnish dairy farms is presented in the Figure a. QRF model provides prediction intervals as shown in Figure b. The prediction intervals are narrower within the lower and upper quartiles of the response variable profitability ratio.

The predictive model was implemented as a web service for farmers who wish to calculate the profitability of their business. The profitability calculator will be integrated with

the MTT Agrifood Research Finland's Economy Doctor web service site (`www.mtt.fi/economydoctor`), which provides a wide range of EU and national level agricultural sector information and benchmarking tools for farmers.

References

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

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning* 20(3): 273-297.

Friedman, J., Hastie, T,. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1): 1-22.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9): 1-20.

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.

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research* 7: 983-999. Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*, Series A, General, 135:370-384.