ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Predicting the profitability of agricultural enterprises in dairy farming

Maria Yli-Heikkilä[1], Jukka Tauriainen[2] and Mika Sulkava[3]

1- Natural Resources Institute Finland (Luke) - Economics and Social Sciences
Tietotie 2 C, FI-31600 Jokioinen - Finland

2- Natural Resources Institute Finland (Luke) - Economics and Social Sciences
Kampusranta 9 C, FI-60320 Seinäjoki - Finland

3- Natural Resources Institute Finland (Luke) - Economics and Social Sciences
Latokartanonkaari 9, FI-00790 Helsinki - Finland

Abstract. Profitability and other economic aspects of agriculture can be analyzed using various machine learning methods. In this paper, we compare linear, additive and recursive partitioning -based models for predicting the profitability of farms using information easily available to a dairy farmer. We find that an ensemble of recursive partitioning methods provides the best prediction accuracy. We also analyze the importance of the predictor variables. These findings may turn out to be useful in increasing our understanding of the factors affecting farm profitability and developing a web-service for farmers to predict the performance of their own farm enterprise.

1 Introduction

The profitability of farm enterprises is very important, as it makes it possible for farms to stay in business in the long-term and, thus, be part of a stable food supply chain. Farm profitability in Finland has fluctuated strongly in recent years [1], which can complicate farmers' planning for the future.

In this paper, various machine learning methods are compared in the task of predicting the profitability of agricultural enterprises. The data were collected from a sample of bookkeeping farms that are a source of data for characterizing Finnish agriculture in the EconomyDoctor service of Natural Resources Institute Finland (Luke) [2].

EconomyDoctor is a source of various information concerning the economy and production process of different types of farms and horticultural enterprises as well as reindeer farming. Annual data is availabe in the service since 2000, and forecasts of structural development are made until 2020.

The aim is to build a new web-service in EconomyDoctor for farmers that would predict the profitability of their enterprises. Therefore, choosing a method with good prediction accuracy is important. Another aim is to understand better the factors affecting the profitability of an agricultural enterprise. Interpretable and understandable models are useful in realizing this second aim.

Earlier, neural networks have been used in predicting the sufficiency of internal financing of farms [3] as well as for analyzing profiles of farm profitability [4]. Farm size change has been predicted using machine learning techniques [5].

2 Profitability bookkeeping data

Annual profitability data for Finnish agricultural and horticultural enterprises show the average results of over 50 000 enterprises and are calculated from the profitability bookkeeping maintained by Luke. The profitability of Finnish farms is monitored annually using a sample of approximately 1 000 farms. Data from the period 2000–2011 were used in this study. We focused only on dairy farms in this work, since the factors affecting profitability are not necessarily the same in different types of production. The number of bookkeeping dairy farms varied between 291 and 387 in the study period.

The form of the bookkeeping data is similar to data in the Farm Accountancy Data Network (FADN) [6]. There are thousands of variables in the bookkeeping data bank. The aim was to select variables that are easily available to farmers. To address this issue, a selection of variables related to single-entry bookkeeping accounts, production, work-load and taxation were chosen. Euro-based variables were deflated to match the consumer price index. Variables having nearly zero variance were removed. Highly correlated variables were removed when the pairwise absolute value of Pearson's correlation coefficient was over 0.90. The final data set included 220 variables, with 4 228 observations.

3 Methods for prediction

In this study, several predictive modeling methods incorporating a variable selection algorithm were chosen. The following methods and their R [7] applications were used in the variable selection and prediction.

- Linear least squares models:

  Linear Regression with Backwards Selection [8], **leapBackward** [9]; Linear Regression with Forward Selection [8], **leapForward** [9]; Linear Regression with Stepwise Selection [8], **leapSeq** [9];

- Penalized linear models:

  Elastic Net [10], **enet** [11]; Elastic Net [12], **glmnet** [13]; Ridge Regression with Variable Selection [14], **foba** [15]; Least Angle Regression [16], **lars** [17]; Sparse Regression [18], **lasso** [17]; Relaxed Lasso [19], **relaxo** [20]

- Additive model:

  Gradient Boosting with Smooth Components [21], **gamboost** [22]

- Recursive partitioning models:

  Implementation of the CART algorithm [23], **rpart** [24]; Bagged CART [25], **treebag** [26]; Implementation of M5 rule-based model tree [27] with additional corrections based on nearest neighbors [28], **cubist** [29]; Conditional Inference Tree, **ctree** [30]; Random Forest [31], **rf** [32]; Quantile Random Forest [33], **qrf** [34]; Interpretable tree-like estimator [35], **nodeHarvest** [36]; Stochastic Gradient Boosting [37], **gbm** [38]; Multivariate

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Adaptive Regression Spline [39], **earth** [40]; Multivariate Adaptive Regression Spline [39] with Generalized Cross Validation (GCV) penalty per knot, **gcvEarth** [40];

- Ensemble selection model:

  Forward stepwise selection of models into the ensemble when maximizing the performance to the Root Mean Square Error (RMSE) on a training set [41], **ensemble** [42]

All the predictive models were trained through the interface of the `train` function provided by the **caret** package [43]. The `train` function develops the parameter tuning, selecting the values that maximize accuracy in the RMSE.

The methods were trained and compared using 10 fold cross-validation with 5 repeats. The data were divided into training (60%), validation (6%) and test sets (33%). First, each method was used for recursive feature selection incorporating resampling using the `rfe` function in the **caret** package. The `rfe` function finds the optimum subset of predictors for the most accurate model. We used the ranked lists of predictors from each method to train the final models. As the aim of the study was to build a web application, for usability reasons we considered having a predictive model based on a maximum of 20 predictors. The preliminary results from the 5 best models in the variable selection (**cubist**, **gbm**, **earth**, **gcvEarth**, **rf**) indicated that within the variable subsets from 10 to 20, the subsets with 18 variables were the most accurate. Thus, we decided to use a subset size of 18 predictors for the final model tuning.

In the final model tuning, the `train` function was used to optimize the parameters for each model. Finally, an **ensemble** was built by selecting the subset of models that yielded the best performance on RMSE. Models are selected for inclusion in the ensemble using greedy forward stepwise model selection with 1000 iterations. Models added multiple times obtain more weight in the ensemble average. The indexes of the training set for each fold are the same for all the models in the final model training.

4 Results

The methods were compared as described in the previous section. The results on accuracy are presented in Table 1. The **ensemble** model that included the **cubist** (weight = 0.396), **gbm** (weight = 0.395) and **earth** (weight = 0.209) models yielded the best performance on the training set. The single model with the highest prediction accuracy was the **cubist** model. The small differences between the training and test set results indicate no major overfitting.

The ensemble model's weights were further utilized to rank the predictors. Each predictor was given points according to its placing in the ranked predictor list of 18 variables in each model. The first placing got 18 points, the last placing 1 point. The points were then penalized according to the corresponding weight in the ensemble model and summed. There were altogether 30 variables in the ensemble model. Table 2 shows the 15 most important predictors.

Model	*Training*	*Test*
ensemble	0.241	0.252
cubist	0.251	0.256
gbm	0.252	0.266
earth	0.262	0.275
gcvEarth	0.264	0.275
rf	0.264	0.265
gamboost	0.278	0.292
qrf	0.289	0.295
nodeHarvest	0.293	0.305
treebag	0.294	0.300
lars	0.300	0.306
foba	0.301	0.306
lasso	0.301	0.306
leapForward	0.303	0.308
leapBackward	0.303	0.307
leapSeq	0.303	0.307
glmnet	0.314	0.337
ctree	0.319	0.329
rpart	0.326	0.331
relaxo	0.683	0.690

Table 1: The prediction performance of the models in RMSE for training set and test set data.

Predictors	*Points*
Net result	18.0
Wage claim	14.0
Total depreciation	12.2
Depreciation max. 10% e.g. on support payment entitlements	10.3
Total expenses per revenues	9.0
Milk produced per milk quota	6.9
Undepreciated balance, year end	6.3
Investment and improvement cost depreciation max. 25%	5.3
Total work load	4.2
Support and recompense excl. VAT	4.0
Feed unit yield	4.0
Interest costs	4.0
Depreciation max. 25% e.g. on machinery	3.6
Work load in cattle husbandry per milk produced	3.5
Workload involved in investments	3.2

Table 2: The 15 most important predictors for profitability in dairy farms. Predictors in a cursive font are taxation-related bookkeeping variables. Others are production and workload related.

5 Conclusions

The prediction capabilities of the methods varied in regard to this problem. The best methods provided useful results, which were considered in the implementation of a web-based prediction system. We found that an ensemble of recursive partitioning methods provides the best prediction accuracy. The most accurate single prediction method was the **cubist** model, a rule-based tree model with additional corrections based on nearest neighbors. The RMSE value of 0.25 is acceptable for web application use. Typically, the profitability ratio of a dairy farm is between -0.3 and 1.4. Hence, there is useful predictive power in the best models. There is, however, still room for improvement and it is probable that not all factors affecting profitability were present in the data.

The feature selection procedure resulted in scores for the importance of predictors and provides a standpoint for the further study of factors affecting the profitability of dairy farms. The ranked predictor list indicates that profitability is related to the productivity, the scale of operations, indebtedness, and the level of investments. These findings are plausible, and suggest that real economic relatioships may have been captured by the models. Similar analysis of other types of agricultural production is a subject of future research.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

References

[1] Olli Rantala and Jukka Tauriainen. Development of results and profitability of agriculture and horticulture. In Jyrki Niemi and Jaana Ahlstedt, editors, *Finnish Agriculture and Rural Industries 2011*, volume 111a of *Publications*, chapter 4.1, pages 52–56. MTT Economic Research, Agrifood Research Finland, 2011.

[2] EconomyDoctor. `http://www.mtt.fi/economydoctor`, February 2015.

[3] Arto Latukka. *Predicting Financial Distress of Farms using Neural Network Application*. Lic.Sc. thesis, University of Helsinki, Department of Economics and Management No. 22, Production Economics and Farm Management, Helsinki, Finland, December 1998. In Finnish.

[4] Mika Sulkava, Anne-Mari Sepponen, Maria Yli-Heikkilä, and Arto Latukka. Clustering of the self-organizing map reveals profiles of farm profitability and upscaling weights. *Neurocomputing*, 147(5):197–206, 2015.

[5] Diti Oudendag, Zoltán Szlávik, and Hennie van der Veen. The use of machine learning techniques to predict farm size change – an implementation in the dutch dairy sector. *American Academic & Scholarly Research Journal*, 4(5), 2012.

[6] Farm accounting data network. `http://ec.europa.eu/agriculture/rica/index.cfm`, February 2015.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. R version 3.0.2.

[8] Alan J. Miller. *Subset selection in regression*. Monographs on statistics and applied probability. Chapman & Hall/CRC, 2002.

[9] Thomas Lumley. *leaps: regression subset selection*, 2009. R package version 2.9, Fortran code by Alan Miller.

[10] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[11] Hui Zou and Trevor Hastie. *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*, 2012. R package version 1.1.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[13] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *glmnet: Lasso and elastic-net regularized generalized linear models*, 2014. R package version 1.9-8.

[14] Tong Zhang. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. Technical report, Statistics Department, Rutgers University, 2008.

[15] Tong Zhang. *foba: greedy variable selection*, 2008. R package version 0.1.

[16] Brad Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression - with discussion. *The Annals of Statistics*, 32(2):407–499, 2004.

[17] Trevor Hastie and Brad Efron. *lars: Least Angle Regression, Lasso and Forward Stagewise*, 2013. R package version 1.2.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.

[19] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, pages 374–393, 2007.

[20] Nicolai Meinshausen. *relaxo: Relaxed Lasso*, 2012. R package version 0.1-2.

[21] Peter Buehlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting – with discussion. *Statistical Science*, 22(4):477–505, 2007.

[22] Torsten Hothorn, Peter Buehlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. *mboost: Model-Based Boosting*, 2014. R package version 2.3-0.

[23] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.* Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[24] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. R package version 4.1-8.

[25] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[26] Andrea Peters and Torsten Hothorn. *ipred: Improved Predictors*, 2013. R package version 0.9-3.

[27] John R. Quinlan. Learning with Continuous Classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348. World Scientific, 1992.

[28] John R. Quinlan. Combining instance-based and model-based learning. In *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*, pages 236–243. Morgan Kaufmann, 1993.

[29] Max Kuhn, Steve Weston, Chris Keefer, and Nathan Coulter. C code for Cubist by Ross Quinlan. *Cubist: Rule- and Instance-Based Regression Modeling*, 2014. R package version 0.0.18.

[30] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

[31] Leo Breiman. Random forests. In *Machine Learning*, volume 45, pages 5–32, 2001.

[32] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[33] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

[34] Nicolai Meinshausen. *quantregForest: Quantile Regression Forests*, 2012. R package version 0.2-3.

[35] Nicolai Meinshausen. *The Annals of Applied Statistics*, 4(4):2049–2072, 2010.

[36] Nicolai Meinshausen. *nodeHarvest: Node Harvest for regression and classification*, 2013. R package version 0.6.

[37] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 1999.

[38] Greg Ridgeway. *gbm: Generalized Boosted Regression Models*, 2013. With contributions from others. R package version 2.1.

[39] Jerome H. Friedman. Multivariate adaptive regression splines – with discussion. *The Annals of Statistics*, 19(1):1–141, 1991.

[40] Stephen Milborrow. *earth: Multivariate Adaptive Regression Spline Models*, 2014. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. R package version 3.2-7.

[41] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning*, pages 137–144. ACM Press, 2004.

[42] Zach Mayer and Jared E. Knowles. *caretEnsemble: Framework for combining caret models into ensembles*, 2013. R package version 1.0.

[43] Max Kuhn. *caret: Classification and Regression Training*, 2014. R package version 6.0-35. Contributions from Jed Wing and Steve Weston and Andre Williams and Chris Keefer and Allan Engelhardt and Tony Cooper and Zachary Mayer and the R Core Team.