The 21st PACIOLI workshop

Variable selection method for FADN data in predicting profitability

Maria Yli-Heikkilä & Jukka Tauriainen

MTT Agrifood Research Finland
Economic Research

September 25, 2013

Outline

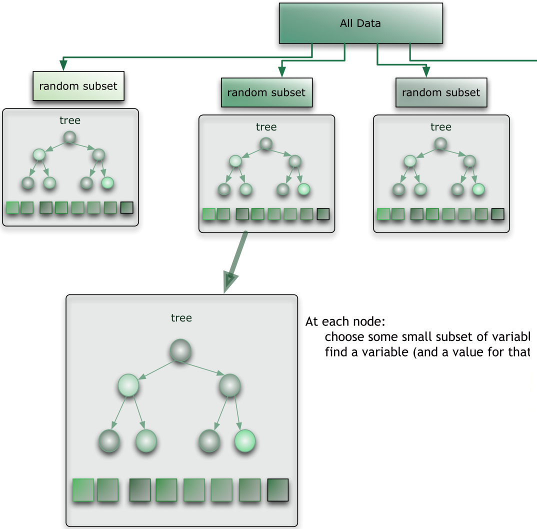
1. Motivation
2. Random forest method
3. Results
4. Discussion

Motivation

- Find important variables for interpretation of profitability of dairy farms
- Design a good prediction model of profitability
- Develop a web based tool

Random forest

- Statistical method for classification and regression problems (Breiman, 2001)
- Ensemble approach, a single decision tree vs. a forest of decision trees



At each node:

choose some small subset of variabl

find a variable (and a value for that

Source: http://citizennet.com/blog/wp-content/uploads/2012/11/RF.jpg

Random forest: function

- Combines many binary decision trees built using several bootstrap samples coming from the learning sample L and choosing randomly at each node a subset of explanatory variables X (with replacement).
- In classification a new object from an input vector is given to each of the trees in the forest. Each tree votes, i.e. gives a classification for the object.
- The forest chooses the classification having the most votes over all the trees in the forest (majority or average voting).

Random forest: features

- No pruning, all the trees of the forest are maximal trees.
- No overfitting as the number of trees increases.
- Fast algorithm
- Can also handle missing values.
- Works on continuous and categorical responses
- Gives variable importance based on permutation

Random forest: Out-of-bag error estimate

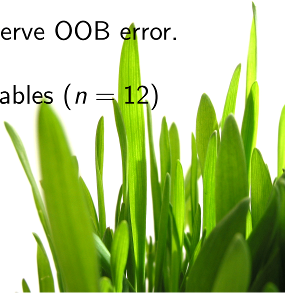
- No test set validation step needed
- Similar to leave-one-out cross-validation, but almost without any additional computational burden.
- OOB error is a random number, since based on random resamples of the data

Variable importance

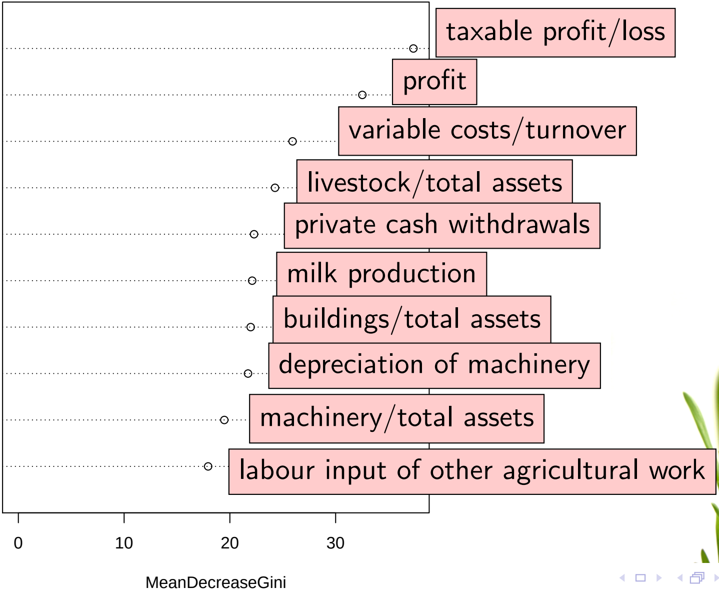
- Gini index accumulates the Gini gain over all splits and trees to evaluate the discriminatory power of a variable.

Results: Variable selection

- Dairy data contains 334 observations and 744 variables.
- Minimum # of variables in the model – reasonable tool.
- Procedure:
    - {Run model for all vars, select best n vars, give each selected var one point} repeat 10 times
    - Take vars having 10 points. Run with these, observe OOB error.
    - Run with $n = 5, 7, 9, 10, 15, 20, 30, 40, 50, 60$
    - Best model with lowest OOB found with 10 variables ($n = 12$)



Dairy model with 10 variables



Results: Prediction model

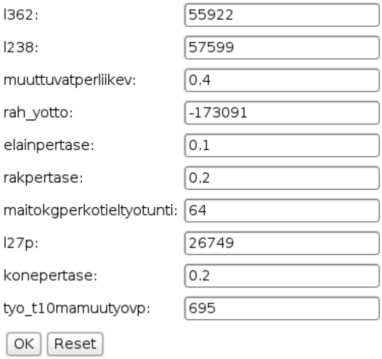
```
Call:
 randomForest(formula = response ~ ., data = datata,
                         ntree = n,      mtry = 2)
               Type of random forest: classification
                     Number of trees: 8000
No. of variables tried at each split: 2

        OOB estimate of  error rate: 38.62%
Confusion matrix:
       poor low modest ok class.error
poor     34  14      4  0   0.3461538
low       8  53     20  6   0.3908046
modest    3  25     48 20   0.5000000
ok        1   7     21 70   0.2929293
```

Results: Test case

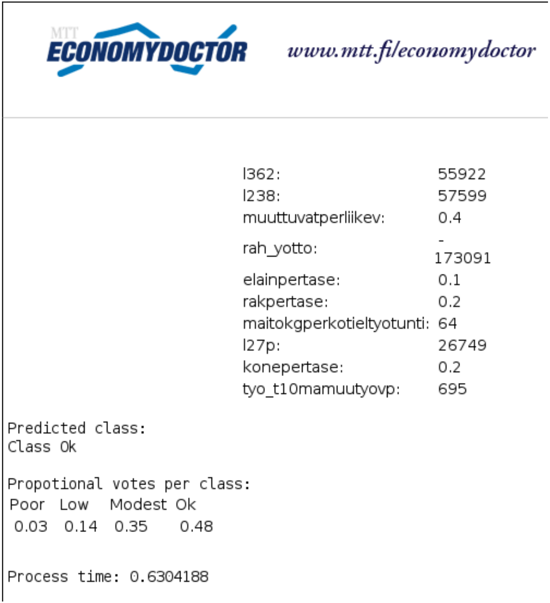
One case left out of the data. Case150 is of class Ok.
The model predicts Case150 as follows:

	poor	low	modest	ok
Case150	0.07	0.27	0.27	0.39

Random Forest model to predict profitability



l362:	55922
l238:	57599
muuttuvatperliikev:	0.4
rah_yotto:	-173091
elainpertase:	0.1
rakpertase:	0.2
maitokgperkotieltyotunti:	64
l27p:	26749
konepertase:	0.2
tyo_t10mamuutyovp:	695



MTT ECONOMYDOCTOR www.mtt.fi/economydoctor

l362:	55922
l238:	57599
muuttuvatperliikev:	0.4
rah_yotto:	- 173091
elainpertase:	0.1
rakpertase:	0.2
maitokgperkotieltyotunti:	64
l27p:	26749
konepertase:	0.2
tyo_t10mamuutyovp:	695

```
Predicted class:
Class Ok

Propotional votes per class:
 Poor  Low   Modest Ok
 0.03  0.14  0.35   0.48


Process time: 0.6304188
```

Discussion

- Accuracy rather low (38% of cases missclassified). Refine the results with other methods?
- What kind of cases get well/poorly predicted by the model?
- How to utilize the information of the prototypes of each class?
- Usability study of the Tool
- Interpretation of the output

Thank you!

MTT
ECONOMYDOCTOR