SUPPLEMENTARY MATERIAL TO: Verster et al. S Afr J Sci. 2021;117(1/2), Art. #7607.

HOW TO CITE:
Verster T, Harcharan S, Bezuidenhout L, Baesens B. Predicting take-up of home loan offers using tree-based ensemble models: A South African case study [supplementary material]. S Afr J Sci. 2021;117(1/2), Art. #7607. https://doi.org/10.17159/sajs.2021/7607/suppl

Appendix 1: Additional techniques used to predict take-up rate

Supplementary table 1 includes the results of all the techniques that were investigated and Supplementary table 2 provides a brief explanation of each of these techniques. The principal aim of the predictive models built in this paper is generalisation. Generalisation means the ability to predict the outcome on novel cases.[1] One method to test how well a model generalises is to compare the fit statistics of a model between the training and validation data sets.[1-3] In Supplementary table 1, the percentage difference between the Gini coefficient of the training and validation data sets is given. Only two of the models built had a very large difference (indicated by an asterisk in the table): memory-based reasoning and random forest. Large differences between the performance on the training and validation data sets usually indicate overfitting.[1] Of the remaining techniques, the simplest model with the highest validation Gini coefficient was considered the 'best'.[3] Boosting was the 'best' technique based on the above reasoning, with bagging a close second. Noteworthy in third place was the neural network and in fourth place the logistic regression with interaction terms. All six variables as well as all two-way interaction terms were considered in the logistic regression with interaction terms.

In the paper, logistic regression was used as a baseline model to compare these results. Logistic regression is a common technique used in most financial industry applications.[2] The three techniques used in the main paper are highlighted in bold.

Supplementary table 1: Gini results of all modelling techniques considered

Modelling technique	Training Gini	Validation Gini	% Difference between training and validation Gini
Neural network	0.452	0.445	1.55%
Rule induction	0.406	0.398	1.97%
Bagging	**0.472**	**0.467**	**1.06%**
Data mining regression	0.410	0.405	1.22%
Data mining neural network	0.411	0.403	1.95%
Logistic regression	**0.410**	**0.403**	**1.71%**
Partial least squares	0.405	0.398	1.73%
Least angle regression	0.405	0.398	1.73%
Memory-based reasoning	0.498	0.354	28.92%*
Boosting	**0.477**	**0.469**	**1.68%**
Random forest	0.728	0.490	32.69%*
Logistic regression with interactions	0.436	0.428	1.83%
Support vector machines	0.169	0.163	3.55%

**indicates a large difference (>15%) between training and validation Gini*

Supplementary table 2: Brief description of modelling techniques

Modelling technique	Brief explanation
Neural network	Neural networks are inspired by the human brain, specifically how a biological neuron works. A neural network attempts to learn, by means of repeated trials, how to organise itself to achieve optimal prediction.[4]
Rule induction	Rule induction combines decision-tree and neural network models to predict nominal targets. It is intended to be used when one of the nominal target levels is rare.[3]
Bagging	Bagging applies random sampling with replacement to create several samples.[5] Each observation has the same chance to be drawn for each new sample.[6] A decision tree is built for each sample.[7]
Data mining regression	Data mining regression first bins all variables and then performs forward selection from all binned and original inputs.[3]
Data mining neural network	Data mining neural network starts by transforming the original inputs into principal components, which are orthogonal linear transformations of the original variables. The three principal components with the highest target correlation are selected for the next step. Next, one of eight possible continuous transformations is applied to the three principal component inputs. The target is predicted by a regression model using the selected principal components and transformation. The process is repeated on the residuals.[3]
Logistic regression	Logistic regression is a generalised linear regression using the logit transformation on the dependent variable.[1,2,8]
Partial least squares	The partial standard regression model identifies input combinations, called factors, which are correlated with both input and target distributions.[27]
Least angle regression	The least angle regression is a generalisation of forward regression using a penalised best-fit criterion.[3]
Memory-based reasoning	Memory-based reasoning is based on k-nearest-neighbour prediction[9], i.e. decisions are made based on the prevalence of each target level in the nearest kcases[3].
Boosting	Boosting performs weighted resampling to boost the accuracy of the model by focusing on observations that are more difficult to classify or to predict.[6] At the end of each iteration, the sampling weight is adjusted for each observation in relation to the accuracy of the model result.[5] Correctly classified observations receive a lower sampling weight, and incorrectly classified observations receive a higher weight. A decision tree is built for each sample.[7]
Random forest	A forest model is an ensemble, or combination, of decision-tree models in which only a subset of selected variables is considered for each decision tree built.[3]
Logistic regression with interactions	Logistic regression is a generalised linear regression using the logit transformation on the dependent variable.[1,2,8] In this specific implementation of logistic regression, all two-way interactions were also considered as possible input variables.
Support vector machines	A support vector machine is a supervised machine learning method[10] that is used to perform classification and regression analyses[3].

References

1. SAS Institute Inc. Predictive modelling using logistic regression (SAS Institute course notes). Cary, NC: SAS Institute Inc.; 2010.
2. Siddiqi N. Credit risk scorecards. Hoboken, NJ: John Wiley & Sons; 2006.
3. SAS Institute Inc. Applied analytics using SAS Enterprise Miner (SAS Institute course notes). Cary, NC: SAS Institute Inc.; 2015.
4. Caudill M. Neural network primer: Part I. San Diego: Al Expert; 1989.
5. Schubert S. The power of the group processing facility in SAS Enterprise Miner. Paper SAS123-2010. Cary, NC: SAS Institute Inc.; 2010. Availabe from: https://support.sas.com/resources/papers/proceedings10/123-2010.pdf
6. Maldonado M, Dean J, Czika W, Haller S. Leveraging ensemble models in SAS Enterprise Miner. Paper SAS1332014. Cary, NC: SAS Institute Inc.; 2014. Available from: https://support.sas.com/resources/papers/proceedings14/SAS133-2014.pdf
7. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–140.
8. Anderson R. The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation. New York: Oxford University Press; 2007.
9. Breed DG, Verster T. The benefits of segmentation: Evidence from a South African bank and other studies. S Afr J Sci. 2017;113(9/10), Art. #2016-0345. http://dx.doi.org/10.17159/sajs.2017/20160345
10. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.