Predictive analytics of insurance claims using multivariate decision trees

Zhiyu Quan [*1] and Emiliano A. Valdez [†1]

[1]Department of Mathematics, 341 Mansfield Road, University of Connecticut, Storrs, Connecticut 06269-1009

Abstract

Because of its many advantages, the use of decision trees has become an increasingly popular alternative predictive tool for building classification and regression models. Its origins date back to about five decades where the algorithm can be broadly described by repeatedly partitioning the regions of the explanatory variables and thereby creating a tree-based model for predicting the response [3]. The use of decision trees for predictive analytics has many advantages. First, a decision tree model is considered to be a nonparametric and thereby does not require distribution assumptions. Second, apart from the ability to handle missing data, it can detect non-linear effects and possible interactions between the explanatory variables. Third, it can be considered as a variable selection process by assessing the relative importance of the explanatory variables. Finally, decision trees, especially smaller-sized trees, are straightforward to interpret by a visualization of the tree structure in the plot. These advantages are particularly useful for insurance and actuarial data.

Innovations and extensions to the original methods, such as random forest and gradient boosting, have further improved the capabilities of using decision trees as a predictive model. Random forest refers to ensembles of trees whereby a set of unpruned fully-grown trees are generated based on a bootstrap sampling of the original data using a subsample of the explanatory variables. The boosting algorithm builds trees sequentially so that for each new iteration, a tree is grown using the residuals from previously grown trees.

In addition, the extension of using decision trees with multivariate response variables started to develop and it is the purpose of this paper to apply multivariate tree models to insurance claims data with correlated responses. Multivariate trees extends the univariate trees by identifying important explanatory variables across a set of responses that covary thereby allowing for a more logical grouping of the responses. For the purpose of this paper, we examined two methods of multivariate tree models: multivariate regression tree [1] and multivariate tree boosting [4]. The extension to multivariate response variables inherits several advantages of the univariate tree models such as its distribution-free feature, ability to rank essential explanatory variables, and high predictive accuracy.

[*]E-mail address: zhiyu.quan@uconn.edu
[†]E-mail address: emiliano.valdez@uconn.edu

To illustrate the approach, we analyze a data drawn from the Wisconsin Local Government Property Insurance Fund (LGPIF) which offers multi-line insurance coverage of property, motor vehicle, and contractors' equipments. This dataset has been used in [2] where the multivariate analysis of the data is based on the use of copulas which is a departure from our method. With multivariate tree models, we were able to capture the inherent relationship among the response variables. We compared marginal results from multivariate tree models to some univariate tree models and we find that the marginal predictive model based on multivariate trees is an improvement from that based on simply the univariate trees.

Keywords: univariate decision trees, multivariate regression trees, multivariate tree boosting, predictive model of insurance claims.

Acknowledgements: We thank Professor Jed Frees and Gee Lee for providing us the data used in this paper.

References

[1] Glenn De'ath (2002), "Multivariate regression trees: a new technique for modeling species-environment relationships. " *Ecology*, vol. **83**(4), pp. 1105-1117.

[2] Frees, Edward W., Lee, Gee, and Yang, Lu (2016), "Multivariate Frequency-Severity Regression Models in Insurance." *Risks*, vol. **4**(1), pp. 1-36.

[3] Loh, Wei-Yin (2014), "Fifty Years of Classification and Regression Trees." *International Statistical Review*, vol. **82**(3), pp. 329-348.

[4] Miller, Patrick J., Lubke, Gitta H., McArtor, Daniel B., and Bergeman, C. S. (2016), "Finding structure in data using multivariate tree boosting." *Psychological Methods*, vol. **21**(4), pp. 583-602.