## **Representing Model Ensembles in PMML**

Tridivesh Jena<sup>1,\*</sup>, Alex Guazzelli<sup>1,</sup>, Wen Ching Lin<sup>1</sup>, Michael Zeller<sup>1</sup>

1. Zementis, Inc.

\*Contact author: tridivesh.jena@zementis.com

Keywords: Predictive Analytics, PMML, Random Forest, R, Standards

PMML, the Predictive Model Markup Language, is the de facto standard to represent predictive analytics models [1,2,3]. It is currently supported by many of the leading commercial and open-source data mining systems, including R. With PMML, it is extremely easy to build a predictive model in one system (PMML producer) and exchange with another solution (PMML consumer) avoiding incompatibility problems and custom coding. The R **pmml** package was designed to export many popular predictive algorithms into the PMML standard [4,5]. Given the recent interest in ensemble models and their applicability to large datasets, it was only natural to add functionality to the R **pmml** package to convert these models into PMML.

The R **pmml** package is now able to export PMML for ensemble models via the ada and randomForest functions. In this presentation, we describe all the steps necessary to export random forest and stochastic boosting models from R into PMML and show how the PMML standard is capable of representing not only model ensembles but also any R specified treatments for missing and invalid values as well as outliers. Additional functions available to the data scientist through the R **pmml** package include the ability to perform data pre- and post-processing. By using the R **pmml** and the **pmmlTransformations** package [6,7], a scientist can read in input data in R, perform transformations on the input data, build the ensemble model and finally output the entire predictive workflow containing model and any pre or post-processing steps in PMML format. Once operationally deployed, the resulting PMML then generates predictions directly from raw input data.

Being able to export the entire model ensemble in PMML format, together with any data validation and transformation steps is remarkable, since it allows for these models to be moved to the operational environment without the need for any recoding. Once in PMML, models can be deployed in minutes and executed in a variety of Big Data platforms, including Hadoop, in-database or cloud computing.

## References

- [1] The Data Mining Group (DMG) website: www.dmg.org
- [2] A. Guazzelli, W. Lin, T. Jena (2010). *PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics (2nd Edition)*. CreateSpace (available on <u>Amazon.com</u>).
- [3] A. Guazzelli (2010). What is PMML? Explore the power of predictive analytics and open standards. IBM developerWorks website.
- [4] A. Guazzelli, M. Zeller, W. Lin, G. Williams (2009). **PMML: An Open Standard for Sharing Models**. *The R Journal*, Volume 1/1.
- [5] The R pmml package: <a href="http://cran.r-project.org/web/packages/pmml/index.html">http://cran.r-project.org/web/packages/pmml/index.html</a>
- [6] T. Jena, A. Guazzelli, W. Lin, M. Zeller (2013). The R pmmlTransformations Package. In Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
- [7] The R pmmlTransformations package: <a href="http://cran.r-project.org/web/packages/pmmlTransformations/index.html">http://cran.r-project.org/web/packages/pmmlTransformations/index.html</a>