Model Selection in Data Analysis Competitions

The use of data analysis competitions for selecting the most appropriate model for a problem is a recent innovation in the field of predictive machine learning. Two of the most well-known examples of this trend were the Netflix Competition and recently the competitions hosted on the online platform Kaggle. In this paper, we will state and try to verify a set of qualitative hypotheses about predictive modelling, both in general and in the scope of data analysis competitions. To verify our hypotheses we will look at previous competitions and their outcomes, use qualitative interviews with top performers from Kaggle and use previous personal experiences from competing in Kaggle competitions. The stated hypotheses about feature engineering, ensembling, overfitting, model complexity and evaluation metrics give indications and guidelines on how to select a proper model for performing well in a competition on Kaggle.

General information
State: Published
Organisations: Department of Applied Mathematics and Computer Science, Cognitive Systems
Contributors: Wind, D. K., Winther, O.
Pages: 55-60
Publication date: 2014

Host publication information
Title of host publication: Meta-learning and Algorithm Selection: Proceedings of the International Workshop on Meta-learning and Algorithm Selection co-located with 21st European Conference on Artificial Intelligence (ECAI 2014)
Editors: Vanschoren, J., Brazdil, P., Soares, C., Kotthoff, L.
Electronic versions:
paper_14.pdf
URLs:
http://ceur-ws.org/Vol-1201/
Source: PublicationPreSubmission
Source-ID: 100924956
Research output: Research - peer-review › Article in proceedings – Annual report year: 2014