Predictive versus Generative Modelling: 
a Challenge for (Social) Sciences

G. Saporta
CEDRIC, Conservatoire National des Arts et Métiers, Paris
gilbert.saporta@cnam.fr

Key Words: black-box models, data science, inference, machine learning

We carry on the reflexion initiated in [1]. Donoho [2], reinterpreting [3], notes that most developments in academic statistics are oriented towards inference: ie fitting, estimating and validating generative models, preferably parcimonious, rather than for prediction where “unfortunately, accuracy and simplicity (interpretability) are in conflict”. Classical inference corresponds to the role of statistics as an auxiliary of sciences. However in most sciences, a good model should also provides accurate predictions, which becomes the sole criterium in decision sciences like pattern recognition, customer behaviour, etc. The most efficient predictive models are rather black-box algorithms like random forests or deep learning.

Some consider that “statistics is the least important part of data science” [4] while others claim that even science is obsolete [5]! Meanwhile, renowned scientists [6] are calling practitioners of social sciences to study Machine Learning. The use of black-box models fitted for massive data is probably the main challenge for social sciences due to their lack of interpretability. Getting better predictions, thanks to a better understanding of the real world, needs to combine statistics and machine learning with causal inference.

References: