Editorial　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　Open Access

A Useful Role for Data Mining in Economics

Sunil Sapra[*]

Department of Economics and Statistics, California State University, Los Angeles, CA 90032, USA

[*]**Corresponding author:** Sunil Sapra, Department of Economics and Statistics, California State University, Los Angeles, CA 90032, USA, Tel: 323-343-2941; Fax: 323-343-5462; E-mail: ssapra@exchange.calstatela.edu

Editorial

Many economists view data mining as simply extracting information from data that is not supported by the data. Nevertheless, in recent years, data mining has become a well-defined inter-disciplinary field at the intersection of computer science and statistics, which employs methods from artificial intelligence, machine learning, statistics, and database systems to discover interesting patterns in large datasets. It is a relatively young field, which has shown great promise in business and economics. Despite recognizing merits of data mining, economists remain skeptical of data mining. Schonlau (2005) explains the differences between economists and analysts in the data-mining community with respect to their approach to regression analysis as follows. While economists traditionally build a model from theory and then use the data to estimate parameters of their model, data miners tend to favor the approach in which all or most available regressors are used in the regression model. Since their econometric model is justified by economic theory, economists are sometimes less interested in testing how well the model fits the data. The standard approach used by data miners for validating their models is to split the data into training and a test dataset. The model is fitted to the training dataset. The fitted model is then used to make predictions on the test data.

To understand the attitudes of economists towards data mining, it is also helpful to note that statisticians have long used two approaches to statistical modeling to derive conclusions [1]. One of these approaches assumes that the data are generated by a known data generating process based on which researchers fit known statistical distributions to data. The second approach employs algorithmic models and treats the data generating process as unknown. While statisticians have begun to embrace the second approach in recent years following advances in statistical computing and availability of high-speed computing, economists have been reluctant to embrace it due to strict adherence to theories about individual and aggregate economic behavior.

It is our view that economists could benefit from a positive attitude towards data mining. Among economic methodologies, the pro-data mining attitude is apparent in the London School of Economics (LSE) methodology [2]. The proponents of this methodology view data mining as an essential empirical activity, which is exploratory in nature. The LSE methodology takes a general-to-specific approach. Employing the LSE framework for model specification involves seeking out models that are valid parsimonious restrictions of the completely general model based on the belief that a sufficiently complicated model can, in principle, describe the economic world. Any more parsimonious model is considered an improvement on such a complicated model if it conveys all of the same information in a simpler, more compact form.

Data mining is not without limitations. Overzealous use of data mining can result in over-fitting and false discovery. Over-fitting risks including too many parameters that need to be estimated leading to a high level of uncertainty. It relates to the case where some residual variation is included as if it were structural. By assessing the model on test data rather than on the training data, data mining ensures that the model is not overfitted and is generalizable [3]. Economists can benefit from competent use of data mining as an exploratory tool to arrive at models, which not only fit the data well but also generate good forecasts.

References

1. Breiman, Leo (2001) Statistical Modeling: The Two Cultures. Stat Sci 16: 199-231.
2. Hoover, Kevin D, Perez, Stephen J (2000) Three attitudes towards data mining, J Econ Methodology 7: 195-210.
3. Schonlau, Matthias (2005) Boosted regression (boosting): An introductory tutorial and a Stata plugin. The Stata Jl 5: 330-354.