Notes on Statistical Modeling: The Two Cultures by Leo Breiman

Jiaqi Bi, University of Toronto

July 12, 2021

1 General Ideas

The article argues about two cultures of data analysis:

1. The Data Modeling Culture
2. The Algorithm Modeling Culture

The data modeling culture generates responses from predictors using specific functions, such as linear regression, logistic regression. The algorithmic modeling culture has machine learning techniques, that the response variables can be predicted by using algorithms by using decision trees and neural networks. The author deems that data modeling has some imperfections:

- Focusing on data modeling could lead to irrelevant theory and questionable scientific conclusions
- Focusing only on data modeling could lead researchers not to use better algorithmic models
- Focusing only on data modeling could prevent statisticians from working on new problems.

2 Example of The Ozone Project

During the time that the author worked for the government that he needed to set up the alert for ozone levels. By using 450 meteorological variables for a period of 7 years, and generating linear regressions as well as using first five years as training data sets and last two years as testing sets, the author found the project was a failure since the false alarm rate of the final predictor was too high.