EPPS 6323 “Knowledge Mining” Spring 2023, Dr. Ho
Assignment 1: Review of Breiman (2001) and Sheuli (2010)
Tyler Vintila

The rapid development of algorithmic modeling methods along with large sets of digitized data has spurred monumental gains for predictive models in industry and select fields of study. In his 2001 article “Statistical Modeling: The Two Cultures”, Leo Breiman gave an impassioned argument that the field of statistics at large was mistaken for remaining wholly dedicated to stochastic data modeling when advances in algorithmic modeling were proven to have unmatched predictive power. Breiman argued that a reliance on data modeling alone was leading to irrelevant theory, questionable scientific practice, and the prevention of novel progress. Breiman’s consulting experience provided examples of how adopting a mindset favoring accurate predictions- rather than testing parametric representations of theory- is ultimately more beneficial. The virtues of algorithmic modeling were further compared to data modeling along the lines of multiplicity of functions to match up to true nature and tradeoffs between accuracy and parsimony. Breiman also contradicted established assumptions about the curse of dimensionality and our ability to convene information from algorithmic methods. Ultimately, Breiman conceded that data modeling approaches do have a time and place, but algorithmic modeling is generally better for its predictive power and potential to discover useful information.

Galit Shmueli picked up Breiman’s dichotomous approach with his 2010 article “To Explain or Predict?” in *Statistical Science.* Nearly 10 years later, Shmueli argued that predictive modeling (comparable to Breiman’s algorithmic modeling) continued to be underutilized in the field of statistics. To a larger point, Shmueli was admonishing the passive ease with which many studies incorrectly attribute explanatory value with predictive strength. Where Breiman provided personal experience and examples to highlight his arguments, Shmueli primarily abstracted explanatory modeling and predictive modeling as distinctive statistical methods. The two (not incompatible) approaches were described in terms of causation-association representations, theory-data functional bases, retrospective-prospective applications, and bias-variance mitigation along successive steps of the statistical modeling process inherent to typical studies. For each step, Shmueli highlighted the dependence of explanatory models on theory and sampling and the complimentary virtues of predictive models for superior predictive power and additional theory development. They followed with comparative examples of the 2006 Netflix challenge and online auctions for how each approach might alternatively develop and arrive at varied conclusions. Finally, Shmueli described how explanation and prediction might be synthesized by particular attention to study purpose and comprehensive reporting of both explanatory and predictive qualities.

Both articles establish a competitive stance that I don’t see so warranted, or helpful for that matter. I am decidedly not a statistician, so I don’t know the prevailing attitudes in the field; but pitting parametric/explanatory and algorithmic/predictive modeling approaches against each other seems “apples to oranges”. Daoud and Dubhashi (2020) were more in the right presentation by explicitly acknowledging the foundations of mainstream data modeling/explanatory/causal approaches being exercises for deductive knowledge. The prototypical hypothetico-deductive scientific method is the gold standard for establishing faith in theory. We have established standards for the iterative process of accepting and refuting assumptions made during these types of studies through subsequent testing and publication. On the other hand, algorithmic methods follow inductive logic to an extreme. Both authors make successful appeals to the predictive power of these methods, but they miss the point of purpose in maintaining testable hypotheses with randomized controls.

We may be better served to establish similar standards for acceptance and dissemination of these useful inductive methodologies than to call for a replacement of a “better” kind of model. Shmueli, in fact, provides clear characterization for how predictive modeling can aid theory development and evaluation. Additionally, focus may be better attributed to improved study design before considering “best” statistical practice; surely an optimized input space with quality, relevant measures is of principal importance for either approach. To extend Shmueli’s and Daoud and Dubhashi’s application of prediction models in the more traditional process, I might suggest a focus on standards for identifying necessary and/or sufficient measures to given theoretical constructs, seeking standard practice for deriving relevant inferences from predictive models, and reinforcing evaluative benchmarks for working theories across fields. Neither argument can hold such an extreme stance as to say predictive modeling is “better” just by predictive power alone- such would be “throwing out the baby with the bathwater” or “missing the forest for the trees” or some other idiom; however, they are completely right in that predictive methods are extremely valuable to explore and inform, especially with complex or ethically challenging problems.