Solving the Long-Range Forecasting Problem in Machine Learning

Eugenia Leonova
Prescient Models LLC
eugenia.leonova@prescientmodels.com

Joseph L Breeden
Prescient Models LLC
breeden@prescientmodels.com

Lenders around the world are generating massive databases of borrower behavior. This often includes unstructured or highly nonlinear and multidimensional alternative data. Correspondingly, machine learning is at the forefront of finding patterns in such data sets. Although machine learning plus big data is a highly successful formula in the right context, this article highlights some serious limitations in long-range forecasting and proposes remedies through careful hybridization with established techniques.

In a world where big data is everywhere, no one has big data relative to the economic cycle. Data volume needs to be thought of along two dimensions. (1) How many accounts / transactions / data fields do we have? (2) How much time history do we have? Few, if any, big data sets include history covering one economic cycle (back to 2005) or two economic cycles (back to 1998). Therefore, unstructured learning algorithms will be unable to distinguish between long-term macroeconomic drivers and point-in-time variations across accounts or transactions. This is the multicolinearity problem that is well known in consumer lending.

This paper presents a solution to the multicolinearity problem in the context of applying neural networks and gradient tree boosting to modeling consumer behavior. An initial model is built using methods that are specifically tuned to capture long-term drivers of performance. That model is treated as given information to a machine learning algorithm that then learns potentially highly nonlinear dynamics relative to the given knowledge. In our examples here, the long-range dynamics are modeled with an Age-Period-Cohort algorithm. Neural networks and stochastic gradient boosting are chosen as the machine learning examples, because they represent both probabilistic and discriminant methods, respectively.

This approach of incorporating given knowledge into a machine learning algorithm solves a deep but rarely recognized problem. In short, almost all models created via machine learning will not give correct long-range forecasts when long-term drivers are present without corresponding data covering multiple cycles for those drivers. Thus, the need for such a solution is great and the solution provided here is sufficiently general that it can apply to a

broad range of applications where both high frequency and low frequency drivers are present in the data.