Comparing Out-of-Sample Performance of Machine Learning Methods to Forecast U.S. GDP Growth

Ba Chu, Shafiullah Qureshi

Abstract

We run a 'horse race' among popular forecasting methods, including machine learning (ML) and deep learning (DL) methods, that are employed to forecast U.S. GDP growth. Given the unstable nature of GDP growth data, we implement a recursive forecasting strategy to calculate the out-of-sample performance metrics of forecasts for multiple subperiods. We use three sets of predictors: a large set of 224 predictors [of U.S. GDP growth] taken from a large quarterly macroeconomic database (namely, FRED-QD), a small set of nine strong predictors selected from the large set, and another small set including these nine strong predictors together with a high-frequency business condition index. We then obtain the following three main findings: (1) when forecasting with a large number of predictors with mixed predictive power, density-based ML methods (such as bagging, boosting, or neural networks) can somewhat outperform sparsity-based methods (such as Lasso) for short-horizon forecast, but it is not easy to distinguish the performance of these two types of methods for long-horizon forecast; (2) density-based ML methods tend to perform better with a large set of predictors than with a small subset of strong predictors, especially when it comes to shorter horizon forecast; and (3) parsimonious models using a strong high-frequency predictor can outperform other sophisticated ML and DL models using a large number of low-frequency predictors at least for long-horizon forecast, highlighting the important role of predictors in economic forecasting. We also find that ensemble ML methods (which are the special cases of density-based ML methods) can outperform popular DL methods.