Off to the Races: A Comparison of Machine Learning and Alternative Data for Nowcasting of Economic Indicators

Jeffrey Chen, Abe Dunn, Kyle Hood, Andrea Batch

Question. In recent years, timely, highly granular alternative data sources have increased in availability, while increasingly sophisticated machine learning (ML) techniques have also proliferated. ML methods, such as Random Forests and Neural Networks, encompass non-parametric, non-linear, and computationally intensive techniques that are optimized for predictive accuracy. Alternative data, such as credit card transactions and search queries, serve as proxies of market-moving activity. Investment firms have already shown that a combination of these techniques can be used to produce accurate nowcasts of market behavior. We ask whether such an approach can extend to official national statistics. As an initial approach, we focus on improving estimates for personal consumption of services, an important component of quarterly GDP. We select this sector because a primary data source, the Quarterly Services Survey (QSS) produced by the U.S. Census Bureau, is not available for the initial "advance" estimate of GDP, thus necessitating some type of imputation.

We pose the following questions: *Do algorithms or data drive prediction accuracy? Is the accuracy gain of a pure prediction-based approach worth the departure from theory-based approaches? (I.e., what are the risks and benefits of alternative methods?) What is the implication of ML and alternative data for current estimates of GDP? How do we apply best practices to ML methods and alternative data?*

Method(s). We conduct a "prediction horse race" using a standard model validation paradigm to evaluate the predictive accuracy associated with an array of algorithms and data. In this paradigm, the out-of-sample predictive accuracy of a given model is measured using a one-step-ahead forecast in which each successive $y_{t=T}$ is predicted using models calibrated for $t < T$. Our target is always the final QSS estimates for each NAICS code, predicted using data that are available within a month after the end of a reference quarter. Each industry model is defined as a combination of two dimensions:

- *Algorithm*: We test a battery of ML methods that are effective in the context of high-dimensional data sets: Principal Components Regression, LASSO, Ridge Regression, Gradient Boosting, Adaptive Boosting, and Random Forests. As OLS is only feasible in cases where n < k, we also test regressions by constructing parsimonious specifications (e.g., top most correlated variables).
- *Data Selection*: We test a variety of alternative data sources and data inclusion criteria (e.g., only conceptually related predictors, all available predictors, etc.).

The result is an $m \times n \times k$ array containing the out-of-sample Mean Absolute Error for each algorithm (m), data selection (n), and QSS industry (k) combination. This results matrix serves as the basis of an analysis that answers the questions posed above.

Data. We draw on data from both conventional and non-conventional sources. Our target is the QSS final estimate available in time for BEA's third estimate of GDP. Initial input data sources include (1) BLS's Current Employment Survey, (2) BLS's Consumer Price Indexes, (3) Credit Card Transactions (timely private economic data source), and (4) Google Trends Search Query Data (timely socially-related public data set).