Training Machine Learning to Anticipate Manipulation

Joshua Blumenstock

Abstract :
An increasing number of decisions are guided by machine learning algorithms. But when consequential decisions are encoded in algorithms, individuals may strategically alter their behavior to achieve desired outcomes. This paper develops an empirical approach to adjust decision algorithms to anticipate manipulation. By explicitly modeling incentives to manipulate, our approach produces decision rules that are stable under manipulation, even when the rules are fully transparent. We stress test this approach through a large field experiment with smartphone users in Kenya, designed to mimic the "digital loan" services that have recently proliferated in East Africa. When implemented, decision rules estimated with this strategy-robust approach outperform those based on standard machine learning approaches.