Strategic Feature Selection

Algorithmic predictions are increasingly used to inform decision-making about allocation of resources. Decision-makers rely on individuals' *features* to determine eligibility and set allocation amounts, with the aim of implementing normative priorities. For example, eligibility for social welfare programs is determined using poverty-targeting scores [1], and government payments to health providers and insurers is based on patient risk scores [2]. Such algorithmic decision-making systems incentivize organizations that serve individuals to respond strategically and "game" the prediction rule.

We consider the U.S. Medicare Advantage (MA) program as a running example, where the government determines payments to private insurers using a public risk-adjustment model that is trained to predict patient costs given health data from the previous year [3]. The goal of risk adjustment is to ensure that insurers receive higher payments for higher-risk enrollees who are expected to need more services. This payment rule inadvertently introduces incentives for private insurers to overreport diagnosis codes, thereby inflating risk-adjusted payments, a practice known as "upcoding." In 2024, higher MA risk scores were estimated to translate into $50 billion in overpayments, as a result of upcoding [4].

To counteract the effect of upcoding, Centers for Medicare & Medicaid Services (CMS) excludes diagnoses that are at risk of inappropriate coding by health plans and providers [5]. In 2024, CMS removed the conditions corresponding to Protein-Calorie Malnutrition and Angina Pectoris from the payment model to limit the sensitivity of the model to higher coding intensity in MA and maintain the ability to accurately predict costs [5]. Despite the use of feature selection as a policy lever to combat manipulation, it remains difficult to reason about which features a decision-maker should exclude in response to strategic behavior, since dropping features comes at the cost of predictive accuracy.

To address this gap, we develop a formal framework to reason about feature selection under strategic behavior. We build on existing frameworks of strategic learning [6], but with a focus on policy levers commonly used in practice that are perhaps more coarse and simple, but as a result more widely applied. In addition, while general strategic learning requires detailed information about costs to manipulation, we focus on realistic limited information settings.

We present a theoretical model of a decision-maker's choice to drop or retain features in a prediction model when such features can be strategically manipulated. We focus on a regression setting, which aligns with the risk-adjustment models used by CMS. We give sufficient conditions for the decision-maker to be better off dropping or retaining features, which we also pair with simulations and examples. Finally, we discuss future directions towards practical policy recommendations.

[1] A. Camacho and E. Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 2011.

[2] M. Geruso and T. Layton. Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, 2020.

[3] G. C. Pope, J. Kautter, R. P. Ellis, A. S. Ash, J. Z. Ayanian, L. I. Lezzoni, M. J. Ingber, J. M. Levy, and J. Robst. Risk adjustment of medicare capitation payments using the CMS-HCC model. *Health care financing review*, 2004.

[4] Medicare Payment Advisory Commission. *The Medicare Advantage program: Status report*. Report to the Congress: Medicare Payment Policy Chapter 12, Medicare Payment Advisory Commission (MedPAC), 2024.

[5] Centers for Medicare & Medicaid Services. *Advance notice of methodological changes for calendar year (CY) 2024 for medicare advantage (MA) capitation rates and part C and part D payment policies*. Technical report, Centers for Medicare & Medicaid Services, 2024.

[6] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, 2016.