"IMPROVING" PREDICTION OF HUMAN BEHAVIOR USING BEHAVIOR MODIFICATION

Galit Shmueli
Institute of Service Science
National Tsing Hua University
Hsinchu, Taiwan
galit.shmueli@iss.nthu.edu.tw

Abstract

The fields of statistics and machine learning design algorithms, models, and approaches to improve prediction. Larger and richer behavioral data increase predictive power, as evident from recent advances in behavioral prediction technology. Large internet platforms that collect behavioral big data predict user behavior for internal purposes and for third parties (advertisers, insurers, security forces, political consulting firms) who utilize the predictions for personalization, targeting and other decision-making. While standard data collection and modeling efforts are directed at improving predicted values, internet platforms can minimize prediction error by "pushing" users' actions towards their predicted values using behavior modification (BM) techniques. The better the platform can make users conform to their predicted outcomes, the more it can boast its predictive accuracy and ability to induce behavior change. Hence, platforms are strongly incentivized to "make predictions true". This strategy is absent from the ML and statistics literature. Investigating its properties requires incorporating causal notation into the correlation-based predictive environment—an integration currently missing. To tackle this void, we integrate Pearl's causal $do(.)$ operator into the predictive framework. We then decompose the expected prediction error given BM, and identify the components impacting predictive power. Our derivation elucidates the implications of such BM to data scientists, platforms, their clients, and the humans whose behavior is manipulated. BM can make users' behavior more predictable and even more homogeneous; yet this apparent predictability might not generalize when clients use predictions in practice. Outcomes pushed towards their predictions can be at odds with clients' intentions, and harmful to manipulated users.

***K*eywords** behavior modification · behavioral big data · machine learning · prediction error · causal intervention · internet platforms

1 Introduction

Recent years have seen an incredible growth in predictive modeling of user behavior using behavioral big data in both industry and in academia. Behavioral big data (BBD) are large and highly detailed datasets on human and social actions and interactions (Shmueli, 2017). BBD-based predictions now shape almost every aspect of modern life, both online and on ground (Agrawal et al., 2018). In contrast to how statistics and machine learning have approached the task of reducing prediction error by improving predictions, a surprising new approach relies on behavior modification techniques, now popularly used in industry. Such behavior modification can be aimed at pushing user actions towards their predicted values, thereby making predictions more certain.

In her enlightening and alarming book, Zuboff (2019) describes the processes used by several large internet platforms that collect BBD to package the raw material of users' actively shared data and passively generated data (e.g. location data, video usage, friendship ties) into "prediction products" that are then sold to business customers–insurance companies, marketers, advertisers, security forces, political consulting firms, etc.–in "behavioral futures markets". The predictions are used to modify users' behaviour, shaping it toward desired commercial or other outcomes.[1] Often, the BBD platform[2] delivers the interventions on its client's behalf.

One example is the recently launched Google Analytics "predictive audiences" service that "automatically enriches your data by bringing Google machine-learning expertise to bear on your dataset to predict the future behavior of your users".[3] Another example is Facebook's "loyalty prediction" service which offers advertisers the ability

[1] www.theguardian.com/technology/2019/jan/20/shoshana-zuboff-age-of-surveillance-capitalism-google-facebook

[2] We use the term "BBD platform" for internet platforms that collect users' BBD

[3] "Purchase Probability, which predicts the likelihood that users who have visited your app or site will purchase in the next seven days…Churn Probability, predicts how likely it is that recently active users will not visit your app or site in the next seven days." https://blog.google/products/marketingplatform/analytics/new-predictive-capabilities-google-analytics/

to target users based on how they will behave, what they will buy, and what they will think. Commenting on this service,[4] Frank Pasquale, a law professor at the University of Maryland and scholar at Yale's Information Society Project said he

> worried how the company could turn algorithmic predictions into "self-fulfilling prophecies," since "once they've made this prediction, they have a financial interest in making it true." That is, once Facebook tells an advertising partner you're going to do some thing or other next month, the onus is on Facebook to either make that event come to pass, or show that they were able to help effectively prevent it (how Facebook can verify to a marketer that it was indeed able to change the future is unclear).

In other words, the more accurate these prediction products, the higher value they provide their customers, and in turn the higher the revenues for the BBD platform. Moreover, the better the BBD platform is able to make users conform to their algorithmically determined destiny, the more it can boast both its predictive accuracy and its ability to induce behavior change (Rushkoff, 2019). Hence, BBD platforms have a strong incentive to improve prediction accuracy, that is, to reduce prediction error.

While Zuboff (2019) uncovered the dangers of a company using behavior predictions to manipulate its customers' behavior for its own commercial gain, we take this one step further: BBD platforms have the technical ability and incentive to manipulate their users' behaviors not only in directions of increasing their clients' gains, but also in a direction that showcases its prediction capabilities, thereby misleading its clients and manipulating humans in possibly dangerous directions. An extreme example is predicting mental health risk for a healthcare stress reduction app. While the app maker aims to lower stress of high risk users, the BBD platform can demonstrate high prediction accuracy by turning high risk predictions into high risk realities.

The goal of this work is to introduce a technical vocabulary which enables investigating this new behavior modification approach to minimizing prediction error. Technical terminology and notation is needed in order to identify the properties and implications of the behavior modification approach to resulting predictive power. Various questions arise: Can behavior modification mask poor predictive algorithms? Can one infer from the manipulated predictive power the counterfactual of non-manipulated predictive power? Can platform clients running routine A/B testing detect this scheme? What are the roles of personalized predictions and of personalized behavior modifications within the error minimization strategy?

Using the $do(.)$ operator by Pearl (2009), we aim to enable the analysis and evaluation of the effect of behavior manipulation on predictive power. While do calculus is well developed for causal effects identification (Pearl, 2009), the challenge here lies in combining the causal $do(.)$ operator into the existing correlation-based predictive framework. Our goal is to make transparent the effects of behavior modification on predictive power, thereby enable the study of its impact on business, social, and humanistic aspects, and its potential implications.

2 The statistical and machine learning approach to reducing prediction error: improve predictions

The fields of statistics and machine learning have been introducing new and improved models, algorithms, approaches, and even data, aimed at improving predictive power. Approaches such as regularization, boosting, and ensembles have proven highly useful in generating more precise predictions. From transparent regression models and tree-based algorithms, to more blackbox support vector machines, k-nearest neighbors, neural nets and especially deep learning algorithms, their justification and adoption lies in their ability to capture intricate signals linking inputs and a to-be-predicted output.

Predictive performance is typically measured by out-of-sample prediction errors, which compare predicted values with actual values for new observations. More formally, the prediction error e_i for record i is defined as the difference between the actual outcome value y_i and its prediction $\hat{y}_i$, that is $e_i = y_i - \hat{y}_i$. For a sample of n records, we have a set of actual outcome values $\vec{y} = [y_1, y_2, \ldots, y_n]$, a set of predicted outcome values $\vec{\hat{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n]$, and a set of prediction errors $\vec{e} = [e_1, e_2, \ldots, e_n]$. For each record i we also have predictor information in the form of p measurements $\vec{x}_i = [x_{i,1}, \ldots, x_{i,p}]$. The predictor information for n records is contained in the matrix X. Predicted values are obtained from $\hat{f}$, the algorithm trained on (or model estimated from) data on inputs X and actual outcomes $\vec{y}$, so that $\vec{\hat{y}} = \hat{f}(X)$.

2.1 Targets of statistical and machine learning efforts

Predictive algorithms and methods are designed and tuned to minimize some aggregation of the error values ($\vec{e}$) by operating on the predicted values ($\vec{\hat{y}}$). Improving predicted values is typically achieved by improving three components:

[4] `https://theintercept.com/2018/04/13/facebook-advertising-data-artificial-intelligence-ai/`

1. structure of the algorithm/model $\hat{f}$ that relates the predictor information X to the outcome (e.g., new algorithms and methods),
2. estimation/computation of $\hat{f}$,
3. quality and quantity of predictor information X. Larger, richer behavioral datasets have been shown to improve predictive accuracy (Martens et al., 2016).

In all these approaches, the actual outcome values $\vec{y}$ are considered fixed. The top panel of Figure 1 illustrates the statistical and machine learning approach of improving the above three components in order to minimize prediction error.

2.2 Components affecting predictive power: dissecting the expected predicted error

When predicting an outcome y that is not expected to be manipulated between training and deployment, we anticipate prediction error due to the inability of the model $\hat{f}$ to (1) correctly capture the underlying f even with unlimited training data (bias), (2) correctly estimate f due to insufficient data (variance), (3) capture the errors for individual observations $\vec{\epsilon}$ (noise). For predicting a numerical outcome or probability for a new observation, these three sources are formalized through a bias-variance decomposition of the expected prediction error (EPE), using squared-error loss[5](Geman et al., 1992):

$$\begin{aligned} EPE(\vec{x}) &= E\left((y|\vec{x}) - \hat{f}(\vec{x})\right)^2 \\ &= E(\epsilon^2) + \left(f(\vec{x}) - E(\hat{f}(\vec{x})\right)^2 + E\left(\hat{f}(\vec{x}) - E(\hat{f}(\vec{x})\right)^2 \\ &= \sigma^2 + Bias^2(\hat{f}(\vec{x})) + Var(\hat{f}(\vec{x})). \end{aligned} \tag{1}$$

In statistics and machine learning, prediction is based on an assumption of continuity, where the predicted observations come from the same underlying processes and environment as the data used for training the predictive algorithm and testing its predictive performance. The deterministic underlying function f and the random noise distribution are both assumed to remain unchanged between the time of model training and evaluation and the time of deployment. This assumption underlies the practice of randomly partitioning the data into separate training and test sets (or into multiple "folds" in cross validation), where the model is trained on the training data and evaluated on the separate test data. Of course, the continuity assumption is often violated to some degree depending on the distance (temporal, geographical, etc.) between the training/test data and the to-be-predicted data and how fast or abruptly the environment changes between these two contexts. These challenges increase prediction errors beyond the disparity observed between training and test prediction errors. Hence, predictive power based on the test data might provide an overly optimistic estimate compared to actual performance at deployment.

2.3 Statistical and machine learning efforts at BBD platforms

Companies such as Google, Facebook, Uber, Netflix, and Amazon have been investing in improving prediction algorithms through collecting, buying, storing and processing unprecedented amounts and types of data. They have also hired top statistics and machine learning talent, purchased AI companies, and developed in-house predictive algorithms and platforms. These are aimed at improving predictions along the three strategies described earlier.

3 A new industry approach to reducing prediction error: manipulating user actions (outcome values)

BBD platforms now have the incentive and technology[6] to minimize prediction errors in a direction that is absent from academic prediction research: by manipulating *actual* outcomes ($\vec{y}$). When the outcome of interest is a human behavior online or offline (clicking an ad, purchasing an item, posting sensitive information, visiting a doctor, voting, etc.), this action can be indirectly manipulated by using *behavior modification* techniques.[7] The most popular technique is the *nudge*, defined as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (Thaler and Sunstein, 2009, p. 6). Zuboff (2019) identifies two more types of behavior modification: *herding*, which is controlling key elements in a person's immediate context in order to guide their behavior towards a predictable one; and *operant conditioning*,

[5]Assuming underlying model $E(y|\vec{x}) = f(\vec{x}) + \epsilon$, where ϵ has zero mean and variance σ^2.

[6]For example, Facebook's "AI backbone" FBLearner Flow combines machine learning and experimentation capabilities that can be applied to the entire Facebook userbase `https://engineering.fb.com/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/`

[7]*Behavior modification*, or *behavior change techniques* are "an observable, replicable and irreducible component of an intervention designed to alter or redirect causal processes that regulate behavior." (Michie et al., 2013)

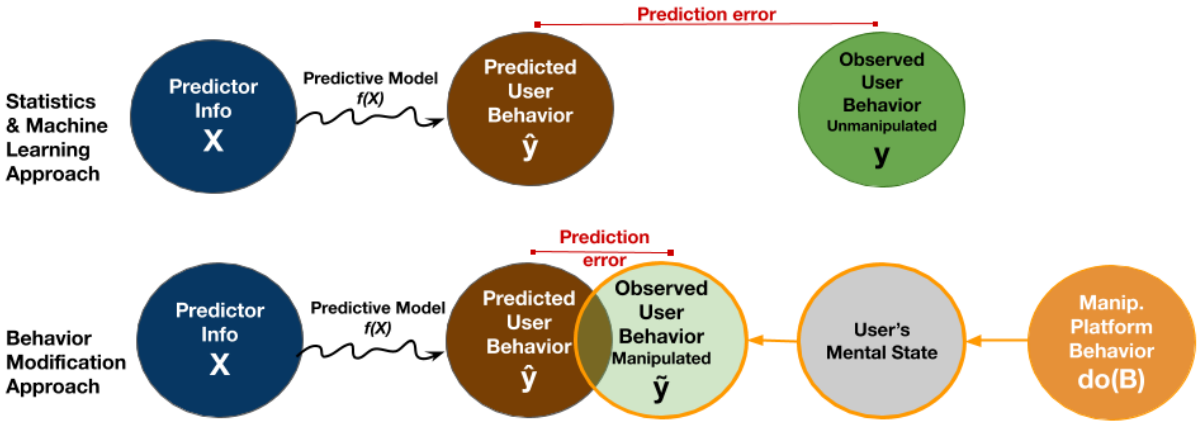

Figure 1: Prediction error with no behavior modification (top) vs. with behavior modification (bottom). Manipulating platform behavior $do(B)$ pushes the observed user behavior towards its predicted value. Note that only orange arrows denote a causal effect. The squiggly black arrows denote a correlation-based predictive relationship.

a term coined by the famous behavioral psychologist B.F. Skinner, which uses positive and negative reinforcement to encourage certain behaviors and extinguish others. BJ Fogg, Standford university's Behavior Design Lab director, lists seven types of "persuasive" technology tools (Fogg, 2002). While the field of marketing has used behavior modification even prior to the advent of the internet (Nord and Peter, 1980), today's technologies and big data enable more covert, pervasive, and powerful manipulation due to their networked, continuously updated, dynamic and pervasive nature (Yeung, 2017). Zuboff (2019) explains,

> These interventions are designed to enhance certainty by doing things: they nudge, tune, herd, manipulate, and modify behavior in specific directions by executing actions as subtle as inserting a specific phrase into your Facebook news feed, timing the appearance of a BUY button on your phone, or shutting down your car engine when an insurance payment is late.

While these examples do not necessarily involve prediction, prediction-based behavior modification is common in recommendation systems, targeted advertising, precision marketing, and other "personalized" interventions that intend to cause human users to change their behavior in a specific direction that is beneficial to the intervention initiator: towards longer online engagement, higher purchase propensity, increased information sharing, or, in our case, *towards the platform's predicted values.*

The two key points are that (1) BBD platforms have a plethora of powerful and tested behavior modification tools, and (2) behavior modification techniques are designed to modify behavior *in a predictable way* –here "pushing" outcome values towards their predicted values – in order "to shape individual, group, and population behavior in ways that continuously improve their approximation to guaranteed outcomes." (Zuboff, 2019, p. 339)

3.1 Two hypothetical scenarios

Consider an insurance company interested in acquiring new customers, but trying to avoid high-risk customers. Now consider an internet platform that is interested in selling to the insurance company the risk scores of their users. To showcase their predictive power, the platform can generate risk scores for a set of users, then use behavior modification to "push" users' behaviors towards their predicted scores (e.g. encouraging/discouraging engagement with the app during driving; showing/not-showing ads for alcoholic beverages during work hours). Such a strategy would turn high-risk predictions into high-risk realities. Note that in this scenario, the gap between the platform's goal (showcase accurate predictions) and the insurance company's objective (avoiding high risk customers) leads to pushing high-risk users' behaviors in a direction that is not only ethically dubious but also at odds with the insurance company's interests.

Another example is a political consulting firm who is interested in reaching likely "Vote for T" individuals. An internet platform wants to sell predicted "Vote for T" scores of their users might prove their predictive power by generating "Vote for T" probability predictions, and then using behavior modification to push users' behaviors towards voting for T. In this scenario, the platform's strategy would turn voting predictions into voting realities. Note that in this scenario, on top of serious ethical and legal implications, the platform's strategy might be in line with the political consulting firm's goal if the firm is trying to promote voting for T, but at odds if the consulting firm is non-partisan, or trying to promote a different candidate.

3.2 Why would a platform follow this strategy?

Given the financial incentives and technical capabilities of internet platforms to showcase predictive power for their prediction products, using behavior modification for "improving" prediction might be used intentionally by a platform's management or by a data scientist under pressure to showcase performance.

Even without intention, such a strategy might be taking place on a platform due to the now-poplar use of algorithms such as reinforcement learning, which employ behavior modification (and user feedback) in order to optimize a predetermined objective function. The common objective of machine learning algorithms to minimize prediction error would lead to this new outcome of "improved" prediction.

3.3 Prediction error under behavior modification: Reducing uncertainty by manipulating the deployment scenario

In the case of no behavior modification, differences between the training and deployment environments introduce uncertainty, typically by increasing bias and/or changing the noise distribution, and are therefore likely to cause larger prediction errors at deployment. Behavior modification implies that, by design, the contexts of training and deployment of the predictive model are made different, albeit in a way that *reduces* uncertainty. While differences between the deployment and training contexts arising from uncontrollable and unforeseeable conditions increase uncertainty, behavior modification intends to shift actual outcome values y in a specific direction, which is by design closer to $\hat{y}$. For example, when predicting that a user is likely to become depressed, displaying depressing news, friends' posts, and depression-related ads increases that user's chance of depression (Facebook's emotional contagion experiment by Kramer et al. (2014) displays such capability). When predicting the arrival time of a delivery, incentivizing faster/slower driving can increase the accuracy of the predicted arrival time. Displaying donation amounts by "friends" with amounts similar to the user's predicted amount can increase the chance the user donates the predicted amount.

3.4 Notation for intentional manipulation in predictive scenarios

To study the components affecting prediction error under behavior modification, it is useful to decompose the new form of expected prediction error (under behavior modification) into separate meaningful sources. This can help us identify components such as bias, variance, and noise. However, we need technical vocabulary that can encode both correlation-based and causal-based terminology.

The challenge is that standard notation and terminology used in statistics and machine learning for predictive modeling is insufficient for formalizing the problem of minimizing prediction error by intentionally manipulating the actual outcome values by way of behavior modification. The bottom panel of Figure 1 illustrates this new scenario. Specifically, predictive terminology lacks notation for denoting an intentional manipulation, as distinguished from correlation-based relationships. At the same time, while causal notation does exist in the world of causal effects and causal inference (Pearl, 2009; Rubin, 1974), in that world correlation-based prediction is excluded. Figure 1, which includes both causal arrows (orange) and a correlational connector (depicted as a squiggly black arrow, but with no causal interpretation[8]) is incoherent in the world of causal diagrams, as well as in the world of prediction. We therefore propose to integrate causal notation into the existing predictive terminology and context in a parsimonious way. We do this by adopting the $do(.)$ operator by Pearl (2009), where $do(B)$ denotes that variable B is not simply observed but rather manipulated.[9] This allows us to incorporate intentional behavioral modification into the predictive modeling context. We then use this notation to decompose the expected prediction error, in order to identify the different components that affect predictive power.

Denote the manipulated outcome as $\tilde{y}$. Using the $do(.)$ operator,[10] we note that it is incorrect to write $\tilde{y} \doteq do(y)$ because the user's outcome y is not directly manipulated. Instead, the modified outcome is fully mediated: the platform tailors its behavior B ($do(B)$ or personalized $do(B_i)$) to manipulate the user's instinct or mental state (e.g. emotion, thought), which in turn leads to the modified outcome $\tilde{y}_i$. This manipulation is specifically aimed at pushing the outcome towards its prediction. We therefore write

$$\tilde{y_i} \doteq y_i | do(B_i). \tag{2}$$

[8] We chose to use a single-headed squiggly arrow rather than a bi-directional straight arrow for the correlation-based predictive relationship to convey the asymmetric input-output roles of X and y. In causal diagrams, bi-directional arrows convey an unobservable variable affecting the two variables at the arrowheads, and there is no way to represent an asymmetric correlation-based predictive relationship.

[9] "The $do(x)$ operator is a mathematical device that helps us specify explicitly and formally what is held constant, and what is free to vary" (Pearl, 2009, p. 358)

[10] It is possible to use Rubin's potential outcomes notation intended for estimating treatment effects (e.g. Imbens and Rubin, 2015, p. 33). This requires defining $B = \{0, 1, 2, \ldots\}$ as the intervention assignment, and denoting by $y_i(B)$ the outcome, where $y_i(0)$ is the un-manipulated outcome. The quantity $y_i|do(B), \vec{x}$ is written as $y_i(B)|B, \vec{x}$. We prefer the $do(.)$ operator since it conveys the causal nature of the manipulation B and clearly differentiates it from the correlation-based prediction components $\vec{x}$.

Table 1: Short and full notation

Short notation	Full notation/definition	Description
y_i	$y_i\|\vec{x}_i = f(\vec{x}_i) + \epsilon_i$	Outcome under no manipulation
f	$f(\vec{x})$	True function under no manipulation
$\hat{f}$	$\hat{y}\|\vec{x} = \hat{f}(\vec{x})$	Predicted outcome under no manipulation
σ^2	$Var(\epsilon) = E(\epsilon^2)$	Noise variance under no manipulation
f_{do}	$g(do(B), \vec{x})$	True function under $do(B)$
$\tilde{y}_i$	$y_i\|do(B_i), \vec{x}_i = g(do(B_i), \vec{x}_i) + \tilde{\epsilon}_i$	Manipulated outcome
$\tilde{\sigma}^2$	$Var(\tilde{\epsilon}) = Var(\tilde{y}) = E(\tilde{\epsilon}^2)$	Noise variance under $do(B)$

To allow heterogeneous effects of the behavioral modification, by the user's specific predictor information $X_i = x_i$ (e.g. user i's browsing history, demographics, location), we can write:

$$\tilde{y}_i \doteq y_i|do(B_i), \vec{x}_i. \tag{3}$$

Second, to denote the predictive (correlation-based) relationship between the outcome and the predictors, we continue using the standard predictive notation:

$$y_i = f(\vec{x}_i) + \epsilon_i. \tag{4}$$

Third, for the manipulated outcome, we use f_{do} to denote the underlying function, which can be a completely different function from f:

$$\tilde{y}_i = f_{do}(do(B_i), \vec{x}_i) + \tilde{\epsilon}_i = g(do(B_i), \vec{x}_i) + \tilde{\epsilon}_i. \tag{5}$$

We use $\sim$ on top of terms affected by $do(B)$. We note that the quantity $E(\tilde{y}_i|\vec{x}_i) - E(y_i|\vec{x}_i) = E(\tilde{y}_i - y_i|\vec{x}_i)$, is called the (population) *Conditional Average Treatment Effect* (CATE) (Athey and Imbens, 2016; Imbens and Rubin, 2015) or *Individual Treatment Effect* (ITE) (Shalit et al., 2017) and is of key interest in treatment effect estimation and testing.[11]

In the predictive modeling phase, we estimate f using $\hat{f}$. We then compare the predicted value $\hat{y}$ to the manipulated outcome $\tilde{y}$. Table 1 provides the short notation, full notation and description for each of the above terms. Together with equations 2-5, we now have a sufficient vocabulary for examining the prediction error under behavior modification.

3.5 Behavior modification for improving predictive power vs. estimating or predicting the effect of behavior modification

Behavior modification techniques are used by BBD platforms for two purposes other than reducing prediction error: for estimating the overall effect of a behavior modification intervention (A/B testing), and for predicting personalized user reactions to a behavior intervention for precision targeting (uplift modeling). These two purposes differ from the focus in this paper, and are also different from each other. Using the terminology and $do(.)$ operator, we briefly describe these approaches in Appendix A, summarizing the key differences in Table 2.

3.6 Expected prediction error of manipulated outcomes ($\widetilde{EPE}$)

When outcome values are intentionally "pushed" towards their predictive values, it is intuitive that the resulting expected prediction error will be lower than the no-manipulation outcome values.[12] We can now formalize the following questions: Given a specific prediction algorithm $\hat{f}$, trained on data with no behavior modification $(X, \vec{y})$, when will the expected prediction error for a manipulated user with predictors $\vec{x}$ and manipulation $do(B_i) = b$ be lower than if the user was not manipulated? That is, for a p-norm loss function L_p, when will we get

$$E[L_p(\tilde{y}, \hat{f}(b, \vec{x}))] < E[L_p(y, \hat{f}(\vec{x}))]? \tag{6}$$

When might the manipulation lead to worse predictive power?

To answer these questions, we proceed to break down the EPE into several non-overlapping components. Using the standard L_2 loss function, we can obtain the EPE under behavior modification as follows (the full derivation is

[11] Note that $y_i|\vec{x}_i$ assumes no manipulation. Pearl (2009, p. 70-72) offers an alternative formulation to encode manipulation vs. no manipulation, by adding a binary intervention indicator I_B that obtains values in $\{do(B_i), idle\}$. In our case $I_B = idle$ for $y_i|\vec{x}_i$.

[12] In the prediction minimization process *all* subjects are initially not B-manipulated and later a sample is B-manipulated using personalized modifications.

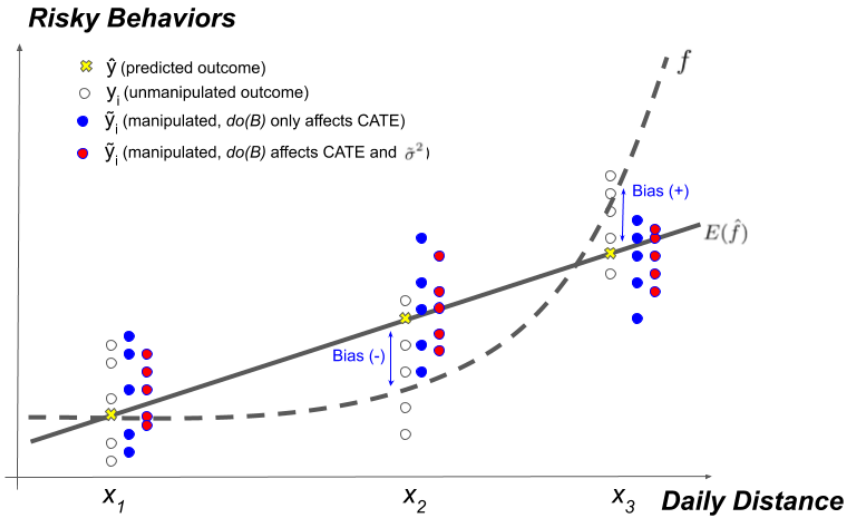

Figure 2: Hypothetical prediction of risky driving behaviors given distance, by ride-sharing platform. Illustrates the effect of behavior modification on shifting the average outcome by $CATE = -Bias$ (blue circles) or on both shifting the average outcome and shrinking the variance $\tilde{\sigma}^2$ (red circles). Yellow X marks are predicted values $\hat{f}(x_i)$. (The schematic assumes a very large training sample, and thus $\hat{f} \approx E(\hat{f})$.)

given in Appendix B):

$$\begin{aligned} \widetilde{EPE}(\vec{x}) &= E\left(y|do(B), \vec{x} - \hat{f}(\vec{x})\right)^2 \\ &= \tilde{\sigma}^2 + \left[CATE(\vec{x}) + Bias(\hat{f}(\vec{x}))\right]^2 + Var(\hat{f}(\vec{x})). \end{aligned} \tag{7}$$

Each of the terms in eq. 7 has an interesting meaning and different implications on the effect of behavior modification on EPE. The additive nature of this formulation provides insights on the roles of data size, predictive algorithm properties, and behavior modification qualities. By comparing $\widetilde{EPE}(\vec{x})$ to $EPE(\vec{x})$ (the manipulated and non-manipulated scenarios), we can see the following:

Data size: Whether manipulating or not, data size affects $\widetilde{EPE}$ via the variance of the predictive algorithm,[13] indicating that larger training samples can improve not only predictions, but also the average manipulated prediction error. Pushing the outcome towards a more stable prediction leads to smaller errors.

Magnitude of behavior modification effect: The second term shows the role of the average behavior modification magnitude (CATE) in countering the bias of the predictive algorithm. This term is minimized when $CATE = -Bias(\hat{f})$, that is, when, on average, $do(B)$ pushes the user's behavior in a direction and magnitude that exactly counters the prediction algorithm's bias. Thus, an effective behavior modification can improve predictive power by combating the predictive algorithm's bias, as long as $0 < CATE < -2Bias$ or $-2Bias > CATE > 0$.

Noise (homogeneity of prediction errors): Compared to σ^2 in the no-manipulation EPE, the first term in $\widetilde{EPE}$ is $\tilde{\sigma}^2$, the noise variance *under behavior modification*. This means behavior modification can decrease/increase also the *variability* of prediction errors (or equivalently, of $\tilde{y}$ relative to y) across different users.

To modify or not to modify? Three Scenarios

To better understand the trade-offs and implications of the four $\widetilde{EPE}$ sources ($\tilde{\sigma}, CATE, Bias(\hat{f}), Var(\hat{f})$) on the expected prediction error, we consider three scenarios. Figure 2 is a simplistic illustration of the roles of CATE and $\tilde{\sigma}$. Suppose predictions are drivers' risk scores in terms of risky driving behaviors and the goal is to minimize the (squared) differences between the predicted and actual values. A ride-sharing or social media platform can modify the driver's behavior by manipulating the driver's engagement with their app while driving. Suppose the x-axis is the daily distance traveled, so that risk is a quadratic function of distance. Yet the predictive model estimates a linear relationship. We consider three specific distances: x_1, x_2, and x_3. While $\hat{f}$ is a biased algorithm, for x_1 there

[13]The machine learning "bias" is asymptotic in sample size: an algorithm is biased "if no matter how much training data we give it, the learning curve will never reach perfect accuracy" (Provost and Fawcett, 2013).

is no bias, for x_2 the bias is negative, and for x_3 bias is positive (for simplicity, the schematic assumes a very large training sample and thus $\hat{f} \approx E(\hat{f})$).

3.6.1 Scenario 1: Low-bias $\hat{f}$ trained on a very large sample

This scenario would be akin to deep learning algorithms applied to massive training data. The very large sample means $Var(\hat{f}) \approx 0$ and $Bias(\hat{f})$ is very small. The strategy of setting $CATE = -Bias(\hat{f})$ is optimal if the behavior modification also decreases error heterogeneity so that $\tilde{\sigma}^2 \leq \sigma^2$. Because the bias is small, the optimal behavior modification should have a small effect. In Figure 2 $\hat{f}(x_1)$ has no bias, and therefore applying behavior modification to drivers with distance x_1 will introduce bias, and is only useful if it can sufficiently shrink the variability of the resulting risky behaviors (red points).

3.6.2 Scenario 2: High-bias $\hat{f}$ trained on a very large sample

High-bias algorithms include models with relatively few parameters (e.g. naive Bayes, linear regression, shallow trees, and k-NN with large k). As in scenario 1, here too $Var(\hat{f}) \approx 0$. The strategy of setting $CATE = -Bias(\hat{f})$ (e.g. in Figure 2 increasing average risky behaviors for x_2 by $|Bias(\hat{f}(x_2))|$ and decreasing it for x_3 by $|Bias(\hat{f}(x_3))|$) is optimal if the behavior modification does not increase error heterogeneity, so that $\tilde{\sigma}^2 \leq \sigma^2$. While a small modification effect (in the right direction) can help counter the bias, the ideal modification effect must be as large as the bias. Note that EPE is computed for a specific $\vec{x}$, and therefore generalizing the above rule to any $\vec{x}$ requires either assuming homoskedastic errors $\tilde{\epsilon}$, or that the inequality holds for all $\vec{x}$ ($\forall \vec{x}\ \tilde{\sigma}^2_{\vec{x}} \leq \sigma^2_{\vec{x}}$).

3.6.3 Scenario 3: High-variance $\hat{f}$

If the predictive model has high variance and is estimated on a relatively small sample, then potential minimization of $\tilde{\sigma}^2$ and/or $\left[CATE + Bias(\hat{f})\right]^2$ by way of behavior modification might be negligible relative to $Var(\hat{f})$. Because behavior modification is based on "pushing" behavior towards $\hat{f}(\vec{x}_i)$, a highly volatile $\hat{f}$ might result in erratic $do(B_i)$ modifications in terms of magnitude or even direction. Hence the choice of predictive model or algorithm can be detrimental to the effectiveness of behavior modification.

4 Discussion

We described a new strategy that might be used by BBD platforms for reducing prediction error which is completely different from approaches taken by the fields of statistics and machine learning. This strategy involves behavior modification, and therefore formalizing it into technical language requires supplementing predictive notation with causal terminology. While our $\widetilde{EPE}$ formula also applies to behavior modification for commercial benefit (e.g. advertising), we have focused on the more extreme case of a potentially rogue BBD platform aiming to minimize prediction errors or unintentionally doing so by using automated personalization techniques such as reinforcement learning. These two efforts can be misaligned, as in risk prediction applications where the client aims to reduce risk, while the platform pushes risky users towards the risky action. Using the $do(\cdot)$ operator, we are able to describe the entire system that includes the training dataset, the predictive algorithm, and the behavior modification. We are also able to distinguish between this strategy and two related, but different, behavior modification usages commonly employed by companies: A/B testing and uplift modeling. The same can be applied to other related approaches such as reinforcement learning, which is especially relevant due to its combined use of prediction and behavior modification for personalization.

4.1 Technical and business implications

Contrasting the bias-variance decomposition of the manipulated and non-manipulated scenarios highlighted two key sources of the manipulated prediction error: the CATE-bias relationship, and its tradeoff with the manipulated noise variance. We now use these insights to return to the questions we posed earlier.

4.1.1 Can behavior modification mask poor predictive algorithm performance?

BBD has been shown to be very noisy, sparse, and high-dimensional (De Cnudde et al., 2020). Behavior modification can improve $\widetilde{EPE}$ by countering the predictive algorithm's bias as well as by reducing the noise variance. This means that poor performance of a predictive algorithm, due to the algorithm's bias and variance, and/or due to the data noisiness, can be masked by $do(B)$. Therefore, customers of BBD platforms wanting to achieve the (manipulated) prediction accuracy level demonstrated by the platform, must purchase both the predictions *and* the ability to apply behavior modification similar to the one performed by the BBD platform. Purchasing the

predictions alone might uncover a much weaker predictive performance when deployed to non-manipulated users (or by applying a less effective modification).

4.1.2 Can one infer the counterfactual EPE from the manipulated $\widetilde{EPE}$?

The difference between the two quantities of no-manipulation EPE and behavior-modified $\widetilde{EPE}$ involves $CATE$, $bias(\hat{f})$, σ, and $\tilde{\sigma}$.[14] Some of these quantities can be estimated (e.g., CATE), while others are more difficult, if impossible. Hence, it is unlikely the no-manipulation predictive power can be ascertained from the manipulated $\widetilde{EPE}$. This means customers of BBD platforms who want to evaluate the no-manipulation predictive power will need to obtain (purchase) information about the estimated EPE at the time of algorithm testing.

4.1.3 Can clients detect the manipulation via A/B tests?

Platform clients who regularly run A/B tests on the platform are not likely to detect the error minimization strategy, because of the random allocation of users in an A/B test. This randomization spreads B-"affected" users across the A and B conditions, and therefore the difference between the group averages will cancel out the B effect. The A/B test statistic and its statistical significance are therefore not impacted by B.

4.1.4 What are the roles of personalized prediction and personalized behavior modification in error minimization?

An ideal behavior modification reduces not only the average magnitude of the errors, but also their variability so that errors are more consistent across users. This highlights the role of *personalized prediction* that companies now invest in: a user's personal $\vec{x}_i$ data is used to select the best modification $do(B_i) = h(B_i, \vec{x}_i)$, where the range of B_i choices includes not only different stimuli[15] (e.g. different content display), but also different types of reinforcement (e.g. positive reinforcement such as rewards, recognition, praise, vs. negative reinforcement such as time pressure, social pressure). For example, Kosinski et al. (2013) showed how Facebook users' Likes can predict their psychological attributes, ranging from sexual orientation to intelligence, and suggested that including such attributes can improve personalized interventions.[16] Personalized interventions are also becoming more powerful with the introduction of reinforcement learning, which personalizes the system's behavior by using users' traits ($\vec{x}_i$) combined with their implicit feedback (den Hengst et al., 2020).

Personalized predictions have the potential to minimize $\widetilde{EPE}$ more equally both within a certain user profile $\vec{x}$ and across different user profiles, by lowering conditional bias via manipulating CATE, and by shrinking the (manipulated) outcomes' variance.

Finally, the bias-variance decomposition highlights the important role of a large training dataset in minimizing the predictive algorithm's variance $Var(\hat{f})$. Since predictive models trade off bias and variance, in the behavior modification context low-bias algorithms are advantageous in terms of requiring a smaller manipulation effect to minimize EPE. One avenue for further research is the effect of behavior modification on classification accuracy (for binary outcomes), where the effects of bias and variance on EPE are multiplicative rather than additive, and the literature reports conflicting results on their roles (e.g. Friedman, 1997; Domingos, 2000).

4.2 Humanistic and societal implications

Behavior modification, now pervasively applied by BBD platforms to their "data subjects", is geared towards optimizing the platform's commercial interest, often at the cost of users' well-being and agency. "Persuasive technology", a design philosophy now implemented on platforms from e-commerce sites and social networks to smartphones and fitness wristbands, aim at generating "behavioral change" and "habit formation", most often without the user's knowledge or consent (Rushkoff, 2019). This application of behavior modification to platform *users* diverges from application to employees for increasing the organization's productivity and workers' job satisfaction. And, clearly, such use diverges from the original intention of behavior modification procedures "to change socially significant behaviors, with the goal of improving some aspect of a person's life" (Miltenberger, 2015).

Given the often conflicting goals of data subjects and the platforms that collect and use their data as well as manipulate their behavior, it is important to introduce these causal mechanisms into the predictive environment,

[14] $\widetilde{EPE} - EPE = \tilde{\sigma}^2 - \sigma^2 + CATE^2 + 2 \times CATE \times Bias(\hat{f})$

[15] "People who do a lot of research on products may see an ad that features positive product reviews, whereas those who have signed up for regular deliveries of other products in the past might see an ad offering a discount for those who "Subscribe & Save."" `www.nytimes.com/2019/01/20/technology/amazon-ads-advertising.html`

[16] "online insurance advertisements might emphasize security when facing emotionally unstable (neurotic) users but stress potential threats when dealing with emotionally stable ones." Kosinski et al. (2013)

so that our statistics, machine learning, and computational social science communities can study their technical properties and implications. By introducing and integrating causal notation into the predictive terminology, we can start studying how behavior modification can create “better" predictions. This allows examining the effects of different behavior modification types, magnitudes, variation, and directions on anticipated outcomes.

Conclusion

Behavior modification can make users’ behavior not only more predictable but also more homogeneous. However, this apparent “predictability" is not guaranteed to generalize when the predictions are used by platform clients outside of the platform environment, or within the platform with a different or no behavior modification strategy. Outcomes pushed towards their predictions can also be at odds with the client’s intention, and harmful to the manipulated users. While platforms have the incentive and capabilities to minimize prediction errors, such minimization might even occur without the platform’s knowledge, due to automated personalization techniques that combine users’ data and their implicit feedback. It is therefore critical to have a useful technical vocabulary that integrates intentional behavior modification into the correlation-based predictive framework to enable studying such contemporary strategies.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

De Cnudde, S., Martens, D., Evgeniou, T., and Provost, F. (2020). A benchmarking study of classification techniques for behavioral data. *International Journal of Data Science and Analytics*, 9(2):131–173.

den Hengst, F., Grua, E. M., el Hassouni, A., and Hoogendoorn, M. (2020). Reinforcement learning for personalization: A systematic literature review. *Data Science*, pages 1–41.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*, pages 231–238.

Fogg, B. J. (2002). *Persuasive technology: using computers to change what we think and do*. Morgan Kaufmann.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kohavi, R. and Longbotham, R. (2017). Online controlled experiments and a/b testing. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning and Data Mining*. Springer.

Kohavi, R. and Thomke, S. (2017). The surprising power of online experiments. *Harvard Business Review*.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Lo, V. S. (2002). The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86.

Martens, D., Provost, F., Clark, J., and de Fortuny, E. J. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly*, 40(4):869–888.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., and Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, 46(1):81–95.

Miltenberger, R. G. (2015). *Behavior modification: Principles and procedures*. Cengage Learning, 6 edition.

Nord, W. R. and Peter, J. P. (1980). A behavior modification perspective on marketing. *Journal of Marketing*, 44(2):36–47.

Pearl, J. (2009). *Causality*. Cambridge university press, 2 edition.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Radcliffe, N. J. and Surry, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, pages 1–33.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rushkoff, D. (2019). *Team Human*. WW Norton & Company.

Rzepakowski, P. and Jaroszewicz, S. (2012). Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, pages 43–50.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR.

Shmueli, G. (2017). Research dilemmas with behavioral big data. *Big data*, 5(2):98–119.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Yeung, K. (2017). 'hypernudge': Big data as a mode of regulation by design. *Information, Communication & Society*, 20(1):118–136.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

Appendix

A Differences between A/B testing, uplift modeling, and error minimization

As mentioned earlier, behavior modification techniques are used by BBD platforms for two purposes other than reducing prediction error: for estimating the overall effect of a behavior modification intervention (A/B testing), and for predicting personalized user reactions to a behavior intervention for precision targeting (uplift modeling). We now show how these two purposes and error minimization are different from each other. Table 2 summarizes the key differences.

Table 2: Differences between A/B testing, uplift modeling, and error minimization

	A/B testing	**Uplift modeling**	**Error minimization**
Business goal	Test effectiveness of new design/feature	Effective user targeting	Increasing value of predictive products
Analysis goal	Test average effect of new feature: $ATE = \frac{1}{n_B}\sum_{i=1}^{n_B}\tilde{y}_i - \frac{1}{n_A}\sum_{i=1}^{n_A}\tilde{y}_i$	Predict uplift for each user i: $\text{uplift}_i = \hat{\tilde{y}}_{i,do(B=1)} - \hat{\tilde{y}}_{i,do(B=0)}$	Minimize overall prediction error: e.g. MSE= $\frac{1}{n}\sum_i (\tilde{y}_i - \hat{y}_i)^2$
Sequence of events	intervention $\rightarrow$ estimation	intervention $\rightarrow$ prediction	prediction $\rightarrow$ intervention
Intervention levels	$do(B=0)$, $do(B=1)$	$do(B=0)$, $do(B=1)$	$do(B_i)$ (personalized)
X used for	not used, or for CATE	training predictive model(s)	training $\hat{f}$ and personalizing $do(B_i)$

A.1 Estimating the overall effect of a behavior modification intervention for product improvement (A/B testing)

BBD platforms try to improve their products and users' experience on an ongoing basis, from updating website design to introducing new features, new ad layouts, marketing emails, and more. To do this, they employ A/B tests, which compare the impact a new test design/feature (version B) vs. an existing one (version A). Amazon, Microsoft, Facebook, Google and similar companies conduct thousands to tens of thousands of A/B tests each year, on millions of users, testing user interface changes, enhancements to algorithms (search, ads, personalization, recommendation, etc.), changes to apps, content management system, and more (Kohavi and Longbotham, 2017; Kohavi and Thomke, 2017). The A/B test is a simple randomized experiment comparing the average outcome of a treatment group to that of a control group. Suppose that version A is the current website functionality ($B = 0$), and version B is a new feature ($B = 1$). The A/B testing process is as follows:

1. randomly assign n_A users to condition A ($do(B = 0)$) and n_B users to condition B ($do(B = 1)$).
2. measure the outcome for users in condition A ($\tilde{y}_i$, $i = 1, \ldots, n_A$) and in condition B ($\tilde{y}_i$, $i = 1, \ldots, n_B$).
3. estimate the Average Treatment Effect $ATE = \frac{1}{n_B}\sum_{i=1}^{n_B}\tilde{y}_i - \frac{1}{n_A}\sum_{i=1}^{n_A}\tilde{y}_i$.
4. use statistical inference and effect magnitude estimates to determine whether the new feature adds value.

When there is interest in the ATE for certain subgroups, such as by the user's language, steps 3-4 can be supplemented with CATE. In summary: A/B tests are used to determine the effectiveness of a new behavior modification feature compared to an existing one. This is done by comparing the *average outcome* of the two randomly assigned $do(B)$ groups, and estimating and testing ATE or CATE.

A.2 Predicting personalized behavior modification effects for maximizing conversion/revenue (uplift modeling)

Uplift modeling, also known as *differential response analysis*, or *true lift modeling* (Rzepakowski and Jaroszewicz, 2012; Radcliffe and Surry, 2011), is used in precision marketing and in political persuasion for identifying people who will modify their behavior (e.g. purchase a product, or vote for a candidate) conditional on being given a particular treatment (e.g. receiving a coupon, or a phone call), assuming the treatment can also cause a negative outcome for some people. In uplift modeling, an intervention is applied to a randomly sampled group of people. The resulting data, along with data from a control group, is used to build predictive model(s) for predicting a person's *change in response*, or *uplift* due to the intervention. The predictive model(s)[17] produce predicted outcomes under

[17]Uplift modeling includes a two-model approach, where models of the form $\tilde{y} = g(X) + \tilde{\epsilon}$ are trained separately on the treatment and control groups, and a single-model approach that trains a single model on the combined dataset ($\tilde{y} = h(X, B) + \tilde{\delta}$).

$do(B=0)$ and $do(B=1)$, which are then combined to obtain uplift$_i = \hat{\tilde{y}}_{i,do(B=1)} - \hat{\tilde{y}}_{i,do(B=0)}$. Finally, the uplift values are used to determine which users to treat ($do(B=1)$) and for which to avoid treatment ($do(B=0)$). This process is as follows:

1. randomly assign n_A users to condition A ($do(B=0)$) and n_B users to condition B ($do(B=1)$.
2. measure the outcome and predictors for users in conditions A ($\{\tilde{y}_i, \vec{x}_i\}$, $i = 1, \ldots, n_A$) and B ($\{\tilde{y}_i, \vec{x}_i\}$, $i = 1, \ldots, n_B$).
3. train a predictive model of $\tilde{y}$ on X, B (e.g., Lo, 2002).
4. use the model to predict $\tilde{y}_{i,do(B=1)}$ and $\tilde{y}_{i,do(B=0)}$, and compute uplift$_i = \hat{\tilde{y}}_{i,do(B=1)} - \hat{\tilde{y}}_{i,do(B=0)}$.
5. Based on uplift, determine who should be treated.

A.3 Behavior modification for minimizing prediction error

The process of minimizing prediction error can be summarized as follows:

1. collect observational predictors X, B and outcome $\vec{y}$ for a sample of n users.
2. build a predictive model $\hat{y} = \hat{f}(X, B)$. Compute personalized predicted scores $\hat{y}_i = \hat{f}(\vec{x}_i, B_i)$ for each user.
3. apply behavior modification $do(B_i)$ to each user to push their outcome towards their predicted value $\hat{y}_i$.
4. compute manipulated prediction errors $\tilde{y}_i - \hat{y}_i$, and summarize them to demonstrate impressive predictive power.

B Derivation of $\widetilde{EPE}$ bias-variance decomposition (Equation 7)

The derivation for Equation 7 (bias-variance decomposition under $do(B)$) is as follows. For convenience, we use f and $\hat{f}$ to denote the no-manipulation true function and its estimated model. For the manipulated scenario, we use f_{do} to denote the true function under behavior modification $do(B)$.

For a new observation with inputs $\vec{x}$ and manipulated outcome $y|do(B), \vec{x}$, we can decompose the expected prediction error as follows (for convenience, we drop subscript i):

$$\begin{aligned}
\widetilde{EPE}(\vec{x}) &= E\left(y|do(B), \vec{x} - \hat{f}(\vec{x})\right)^2 \\
&= E(\tilde{y} - \hat{f})^2 \\
&= E(\tilde{y} - f_{do} + f_{do} - \hat{f})^2 \\
&= E(\tilde{y} - f_{do})^2 + E(f_{do} - \hat{f})^2 + 2E(\tilde{y} - f_{do})(f_{do} - \hat{f}).
\end{aligned} \tag{8}$$

These three terms can be further simplified. The first term can be simplified by noting that $\tilde{y} = f_{do} + \tilde{\epsilon}$:

$$E(\tilde{y} - f_{do})^2 = E(\tilde{\epsilon}^2) = \tilde{\sigma}^2. \tag{9}$$

The second term can be written as:

$$\begin{aligned}
E(f_{do} - \hat{f})^2 &= E\left(f_{do} - E(\hat{f}) + E(\hat{f}) - \hat{f}\right)^2 \\
&= \left(f_{do} - E(\hat{f})\right)^2 + E\left(\hat{f} - E(\hat{f})\right)^2 \\
&= \left(f_{do} - E(\hat{f})\right)^2 + Var(\hat{f})
\end{aligned} \tag{10}$$

because the cross product is zero:

$$2E\left(f_{do} - E(\hat{f})\right)\left(E(\hat{f}) - \hat{f}\right) = 2\left(f_{do} - E(\hat{f})\right)\left(E(\hat{f}) - E(\hat{f})\right) = 0. \tag{11}$$

We can further write Equation 10 as a function of the bias and variance of $\hat{f}$:

$$\begin{aligned}
\left[f_{do} - E(\hat{f})\right]^2 + Var(\hat{f}) &= E\left[f_{do} - f + f - E(\hat{f})\right]^2 + Var(\hat{f}) \\
&= \left[CATE + Bias(\hat{f})\right]^2 + Var(\hat{f}).
\end{aligned} \tag{12}$$

Finally, using the independence of the new observation's prediction error $\tilde{\epsilon}$ from the prediction $\hat{f}$ based on the training data ($E(\tilde{\epsilon}\hat{f}) = 0$), the third term can be shown to be zero:

$$2E(\tilde{y} - f_{do})(f_{do} - \hat{f}) = 2E(\tilde{\epsilon})(f_{do} - \hat{f}) = 2f_{do}E(\tilde{\epsilon}) - 2E(\tilde{\epsilon}\hat{f}) = 0. \tag{13}$$

Therefore, we can write $\widetilde{EPE}$ from Equation (8) as:

$$\begin{aligned} \widetilde{EPE}(\vec{x}) &= E\left(y|\vec{x}, do(B) - \hat{f}(\vec{x})\right)^2 \\ &= \tilde{\sigma}^2 + \left[CATE + Bias(\hat{f})\right]^2 + Var(\hat{f}). \end{aligned} \tag{14}$$

Acknowledgements

I thank Foster Provost, Rob Hyndman, Ali Tafti, Soumya Ray, Sam Ransbothan, Travis Greene, Boaz Shmueli, Patricia Kuo, Raquelle Azran, and participants of SCECR 2020 symposium for invaluable feedback, and Noa Shmueli for Fig. 1 artwork.