ICIS 2022 Proceedings

AI in Business and Society

Dec 12th, 12:00 AM

A Strategic Analysis of Algorithm Manipulation

Jiali ZHOU
HKUST, jzhoubf@connect.ust.hk

Jiexin ZHENG
hkust, jzhengas@connect.ust.hk

A Strategic Analysis of Algorithm Manipulation: a Lending Game perspective

Short Paper

Jiali ZHOU
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
jiali.zhou@connect.ust.hk

Jiexin ZHENG
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
jzhengas@connect.ust.hk

Abstract

Machine learning models are widely used in many business contexts, but there is a growing concern that strategic individuals may manipulate their features to obtain desirable outcomes from the machine learning models. This paper offers a theoretical analysis of the impact of feature manipulation on the performance of the machine learning models and the payoffs of firms in an online lending context. Contrary to the common belief, our interesting finding is that manipulation may not be harmful to a firm under some circumstances. Instead, it could increase the classification model's performance and raise a firm's payoff and the social welfare when high-quality individuals manipulate more. Overall, our findings suggest that manipulation can bring strategic value to machine learning models instead of just being a harmful activity. Our findings provide useful insights for feature engineering and lay a foundation for future research about optimal strategies to cope with manipulation activities.

Keywords: Algorithm manipulation, strategic classification, machine learning

Introduction

Machine learning models are widely used for decision-making in many business contexts. For example, Banks leverage risk assessment models to make lending decisions; E-commerce platforms use consumers' social connections and purchase history to predict what products to advertise. In general, machine learning models do not have access to individuals' true outcomes (e.g., willingness to pay, preference), and need to rely on observable characteristics (features) to predict the outcomes. One important assumption is that individuals always behave honestly, and their features are reliable for making inferences. However, that's not always the case. Individuals may strategically behave or manipulate their characteristics to obtain the desired inference from machine learning models. For example, in the online lending context, some borrowers may intentionally sign up for more credit cards to get better ratings from the bank's risk assessment model to obtain a loan; Similarly, consumers may adjust their browsing behavior to act as if they have low-willingness to pay in order to get coupons from the platform. Increasingly, many worry that the performance of the machine learning model would drastically decay when facing individuals' strategic manipulation (Dalvi et al., 2002).

Given the increasing adoption of machine learning in business applications and the potential backlash of feature manipulation, it is important and timely to analyze how such manipulation activities affect machine learning models and the payoffs of firms who apply these models. The results of such analyses would be relevant for both policy decisions (e.g., whether manipulation is socially harmful? Should manipulation be punished?) and machine model designs (e.g., should machine learning models abandon the use of manipulable features? How manipulable features affect the use of other unmanipulable features?).

This paper seeks to shed light on these issues by analyzing how feature manipulation in an online lending context affects the classification models a lender deploys and its payoff. Specifically, in our setting, one lender decides whether to approve loan requests from strategic borrowers. Borrowers can be high-quality or low-quality, but it is only profitable for the lender to lend to high-quality borrowers. The quality of borrowers is unobservable, so the lender relies on a classification model to make loan decisions based on borrowers' features. However, in order to improve the classification outcomes, some borrowers may manipulate their features (e.g., manipulate the bank asset proof in order to increase the chance of obtaining the loan), which affects the performance of the classification model and the lender's payoff. In this setting, we ask how the lender's optimal classification strategy should adjust and how its payoff would be affected by borrowers' manipulation? In future work, we would also like to answer the question of what strategies the lender can adopt in order to cope with borrowers' feature manipulation.

Our interesting finding is that manipulation may not be harmful to the lender under certain circumstances. Instead, it could increase the classification model's performance and raises the lender's payoff. This happens when high-quality borrowers have more advantages in manipulation than low-quality borrowers, and so manipulation mainly allows high-quality borrowers who otherwise cannot get the loan (i.e., false negative) to obtain the loan through manipulation. This increases the classification's performance, the lender's payoff, and social welfare. On the other hand, if both types of borrowers are equally likely to manipulate, or low-quality borrowers have more advantages in manipulation, then manipulation decreases the classification model's performance, the lender's payoff, and social welfare. We further consider several model extensions, such as incorporating the lender’s algorithm transparency choice and considering deep learning algorithms, and obtain similar findings.

Overall, our theoretical analyses suggest manipulation can bring strategic value to machine learning models, instead of just being a harmful activity as many have suggested (Dalvi et al., 2002; Tu et al., 2019). This finding offers a plausible explanation for firms’ incentive to continue applying manipulable features in their classification models despite the growing concerns from both industry and academia. Our model also sets a foundation for further analyses of optimal strategies to leverage the benefit or mitigate the cost due to feature manipulation.

The paper contributes to the growing literature on the impact of algorithm manipulation (e.g., Dalvi et al., 2002; Tu et al., 2019; Frankel and Kartik, 2019; Jann and Schottmüller, 2020; Wang et al., 2020). Most existing studies are interested in ways to minimize manipulation's impact based on the assumption that manipulation is harmful to the classification models (e.g., Dalvi et al., 2002; Tu et al., 2019; Frankel and Kartik, 2019; Jann and Schottmüller, 2020). This paper adds to the literature by helping clarify the fundamental question of whether manipulation is harmful to the classification models and society. Although some recent studies have shown some potential benefits of manipulation based on the assumptions that manipulating a feature has productivity benefit (e.g., Wang et al. (2020) shows manipulation incentivizes job candidates to seek education, which leads to an increase in their productivity), our model shows a potential benefit can arise just by strategic interaction between the classification model and the individuals being classified. Second, we also provide new results regarding how feature manipulation affects the optimal use of unmanipulable features in classification models. This result provides useful insights into feature engineering, i.e., select features to be included in the machine learning models, with limited budgets. Last, our analysis framework also provides a foundation for analyzing and comparing different policies that the firm can apply to either leverage the benefit or mitigate the cost of feature manipulation. We believe this is a promising avenue for future research.

Model Setting

We use online lending as an example scenario while try to abstract from details in order to increase generality. A lender is faced with a pool of borrowers who wish to get loan from the lender. The size of borrower pool is N, which we normalize to be 1. Among the pool of borrowers, there are two types: a θ proportion high-quality borrowers, and $1-\theta$ proportion of low-quality borrowers.

The lender incurs a cost of W by approving a loan, but only lending to high-quality borrowers generates sufficient benefit to cover the cost. Therefore, from the lender’s perspective, only lending to high-quality borrowers is desirable. Normalize the benefit from lending to low-quality borrowers to be zero and the benefit from lending to high-quality borrowers to be B_L where $B_L > W$.

As is often the case, the quality of a borrower is unobservable to the lender, so the lender asks borrowers to upload their features and relies on a classification algorithm to make predictions about their quality.

We categorize features as manipulable or unmanipulable. A manipulable feature means the borrowers can manipulate the value of the feature by incurring some cost. For example, a borrower can manipulate the bank asset proof by borrowing money from families or friends and making a temporary deposit to the bank account. On the other hand, by unmanipulable feature, we mean the value of the feature is not susceptible to manipulation, at least in the short term (or equivalently, the cost of manipulation is too high). Features such as gender, race, and highest diploma obtained may belong to this case.

Without loss of generality, we assume there are two features, one manipulable and one unmanipulable feature, that serve as input for the classification model. Further, we assume each feature can only take two values, H and L, and having H on a feature is more likely to be associated with high-quality borrowers than having L on the feature.

We follow the literature (Wang et al., 2020) to assume the unmanipulable feature predicts the correct type with a probability of γ_u. $\gamma_u \in [0.5, 1]$ so that a borrower with value H in unmanipulable feature is more likely to be high-quality. Similarly, we define γ_m: $\gamma_m \in [0.5, 1]$ for the manipulable feature.

With this assumption, the distribution of borrowers with no manipulation along the two features is presented in Table 1.

		Manipulable Feature (m)			
		H		L	
Unmanipulable Feature (u)	H	$\boldsymbol{HH}$:	$\boldsymbol{HH, h}$: $\theta\gamma_u\gamma_m$ $\boldsymbol{HH, l}$: $(1-\theta)(1-\gamma_u)(1-\gamma_m)$	$\boldsymbol{HL}$:	$\boldsymbol{HL, h}$: $\theta\gamma_u(1-\gamma_m)$ $\boldsymbol{HL, l}$: $(1-\theta)(1-\gamma_u)\gamma_m$
	L	$\boldsymbol{LH}$:	$\boldsymbol{LH, h}$: $\theta(1-\gamma_u)\gamma_m$ $\boldsymbol{LH, l}$: $(1-\theta)\gamma_u(1-\gamma_m)$	$\boldsymbol{LL}$:	$\boldsymbol{HL, h}$: $\theta(1-\gamma_u)(1-\gamma_m)$ $\boldsymbol{HL, l}$: $(1-\theta)\gamma_u\gamma_m$

Table 1. The distribution of borrowers with no manipulation

In our setting, we allow for the possibility that borrowers may manipulate the manipulable feature in order to obtain better classification outcomes, i.e., to be predicted as high-quality. A high(low) quality borrower incurs a cost of $C_H(C_L)$ to manipulate the feature. Further, we assume the cost can be different for different types of borrowers. For example, consider the bank asset proof as one manipulable feature: a high-quality borrower may already have a relatively decent bank asset proof or can easily borrow money from his wealthy friends or families to create a good one. While a low-quality borrower may need to incur a considerate cost (e.g. borrow more money to fake the proof) to create a decent bank asset proof. When borrowers pass the classification and obtain the loan, they obtain a benefit of B_B.

The problem for the lender is to design a classification strategy to decide which borrowers to lend given their features. The lender knows W, B_L and if manipulation is prevalent in each state. The borrower knows whether he/she is high or low-quality, C_H or C_L and B_B. They can also infer the classification outcome if they choose to manipulate feature, for example, by communicating with others who might have comparable features. Throughout the paper, we are interested in the Nash equilibrium wherein both the lender and borrowers would not deviate from their selected strategy even given others' strategy known.

Payoff

The lender's decision given the borrowers pool is to design a classification strategy to decide which borrowers to lend. Because borrowers have four states in our setting, so its strategy can be represented by four binary decision variables $\{D_{HH} \in \{0,1\}, D_{HL} \in \{0,1\}, D_{LH} \in \{0,1\}, D_{LL} \in \{0,1\}\}$, where $D_i = 1$ means the lender would lend to borrowers in state i, and the subscript represents the state (e.g. HL refers to borrowers with H in unmanipulable feature and L in manipulable feature).

Given the classification strategy, the lender's payoff would be:

$$\Pi^k = \sum_{i \in \{HH, HL, LH, LL\}} D_i^k \left[N_{i,h}^k (B_L - W) + N_{i,l}^k(-W)\right],$$

where $k =$ "u" corresponds the case with no user manipulation while $k =$ "m" corresponds the case with user manipulation. The lowercase h or l refers the high or low quality of borrowers.

The lender optimal strategy is to choose $D_i^k = 1$ if and only if the following condition holds:

$$N_{i,h}^{k}(B_L - W) + N_{i,l}^{k}(-W) \geq 0 \; \Rightarrow \frac{N_{i,h}^{k}}{N_{i,h}^{k}+N_{i,l}^{k}} \geq \frac{B_L}{W} \;, \qquad (1)$$

which basically says the lender is willing to lend borrowers in state i if and only if it expects the proportion of high-quality borrowers in that state is sufficiently high. This is intuitive, because high-quality borrowers bring benefit to the lender while low-quality borrowers bring only cost by assumption.

A borrower's strategy is whether to manipulate its manipulable feature. Given a borrower's true type $i \in \{h, l\}$and her original state $j \in \{HH, HL, LH, LL\}$ and new state after manipulation $j' \in \{HH, HL, LH, LL\}$, her payoff can be represented by:

$$U_{i,j} = \begin{cases} D_j B_B, if\ not\ manipulate \\ D_{j'}(B_B - C_i), if\ manipulate \end{cases} . \qquad (2)$$

Clearly, a borrower is willing to manipulate only if $D_{j'} > D_j$ and $B_B \geq C_i$, which says a borrower is willing to manipulate if and only if such manipulation can improve the prediction outcome and the manipulation is not too costly. Because we assume being H in one feature is more likely to be associated with a high-quality borrower, so it is straightforward to see a borrower only has incentive to manipulate the feature from L to H but not the other way. In addition, to avoid the uninteresting case that no borrower has incentive to manipulate, we assume $B_B > \min\{C_H, C_L\}$.

Strategies with no feature manipulation

We first analyze the baseline case when borrowers do not manipulate features. Given the four states that borrowers belong to, i.e., $\{HH, HL, LH, LL\}$ and the two potential choices in each state, i.e., $D_i \in \{HH, HL, LH, LL\} \in \{0,1\}$, the lender has 16 possible strategies in total. However, we can easily rule out most strategies as the optimal one by making use of the following lemma. Assume a sequence $H > L$, and use i, i' to represent the unmanipulable feature and j, j' to represent the manipulable feature, then:

Lemma 1 $D_{i,j} \geq D_{i'j'}\ if\ i \geq i'\ and\ j \geq j'$

Throughout the paper, formal proofs of Lemmas and Propositions are presented in the Online Appendix to save space. We focus on explaining the intuitions here. Referring to Figure 1, this Lemma basically says whenever the lender chooses to lend to borrowers in a particular cell A, then the lender would strictly prefer to lend to borrowers in the cell left to or above cell A. This is because our assumption that a borrower having an H in one feature is more likely to be high-quality than if he had an L in one feature, so the proportion of high-quality borrowers would be higher on the right or at the top cells, which increases the lender's incentive to lend. According to condition (1), the lender would strictly prefer to lend to borrowers on the right or at the top cells.

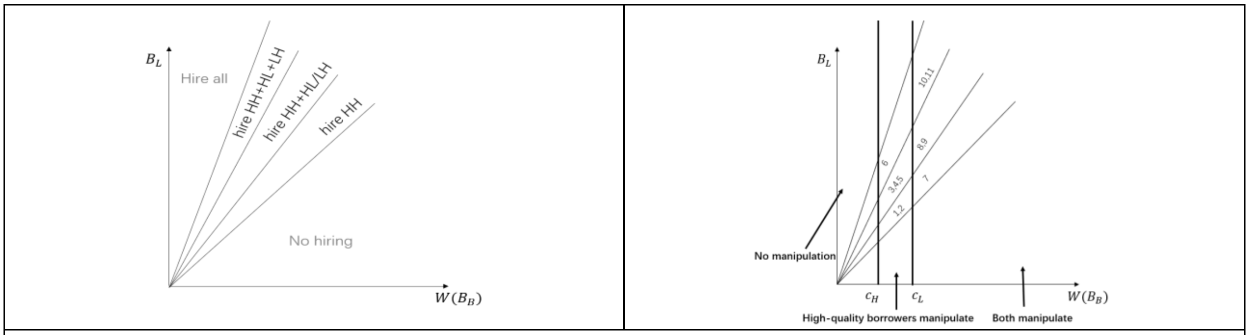

Figure 1. Lender's strategy with no manipulation(left) and with manipulation (right)

According to this lemma, a strategy like $D_{HH} = 0, D_{HL} = 1$ can not be optimal. Further, the case becomes trivial if the lender choose to lend all borrowers (i.e., $D_{HH} = D_{HL} = D_{LH} = D_{LL} = 1$) or the lender does not want to lend any (i.e., $D_{HH} = D_{HL} = D_{LH} = D_{LL} = 0$), so we shall focus only on cases that at least two D_i have different values. We also know $D_{HH} = 1$ and $D_{LL} = 0$ must hold, otherwise, Lemma 1 would suggest we will return to the two trivial cases (i.e., lend to all (none) borrowers).

Through this elimination, the lender is left to vary only D_{HL} and D_{LH}, and this gives four different strategies at the lender's side: (1) $D_{HH} = 1, D_{HL} = D_{LH} = D_{LL} = 0$, (2) $D_{HH} = D_{HL} = 1, D_{LH} = D_{LL} = 0$, (3) $D_{HH} = D_{LH} = 1, D_{HL} = D_{LL} = 0$, (4) $D_{HH} = D_{HL} = D_{LH} = 1, D_{LL} = 0$. Depending on the parameter ranges, the lender may choose one of the four strategies. Below proposition summarizes the detail.

Proposition 1 Under no manipulation and $C(HH > 0)$ (meaning profitable to lend HH) and $C(LL < 0)$ (meaning unprofitable to lend LL), the lender's strategy and payoff would be:

$$\Pi^{\mathrm{u}} = \begin{cases} (B_L - W)\theta\gamma_u\gamma_m - W(1-\theta)(1-\gamma_u)(1-\gamma_m)\ lend\ to\ HH\ if\ not\ C(HL > 0) and\ not\ C(LH > 0) \\ (B_L - W)\theta\gamma_u - W(1-\theta)(1-\gamma_u)\ lend\ to\ HH + HL\ if\ C(HL > 0) and\ not\ C(LH > 0) \\ (B_L - W)\theta\gamma_m - W(1-\theta)(1-\gamma_m)\ lend\ to\ HH + LH\ if\ not\ C(HL > 0)\ and\ C(LH > 0) \\ (B_L - W)\theta(\gamma_u + \gamma_m - \gamma_u\gamma_m) - W(1-\theta)(1-\gamma_u\gamma_m)\ lend\ to\ HH + HL + LH\ if\ C(HL > 0)\ and\ C(LH > 0) \end{cases}$$

Where $C(HL > 0)$ (i.e., profitable to lend to HL), $C(LH > 0)$ (i.e., profitable to lend to LH), $C(HH > 0)$ and $C(LL < 0)$ are parameter ranges. $C(HH > 0)$: $\frac{B_L}{W} \geq 1 + \frac{(1-\theta)(1-\gamma_u)(1-\gamma_m)}{\theta\gamma_u\gamma_m}$, $C(HL > 0)$: $\frac{B_L}{W} \geq 1 + \frac{(1-\theta)(1-\gamma_u)}{\theta\gamma_m}$, $C(LH > 0)$: $\frac{B_L}{W} \geq \frac{(1-\theta)(1-\gamma_m)}{\theta\gamma_m}$, $C(LL < 0)$: $\frac{B_L}{W} < 1 + \frac{(1-\theta)\gamma_u\gamma_m}{\theta(1-\gamma_u)(1-\gamma_m)}$

The equilibrium strategies under different parameter ranges can be illustrated in Figure 1. As $\frac{B_L}{W}$ grows, the firm will first find it optimal to only lend to borrowers in state HH, because, by Lemma 1, the proportion of high-quality borrowers is the highest in the HH state. When B_L becomes larger relative to W, meaning the lender enjoys higher benefit from high-quality borrowers, it would tolerate lower proportion of high-quality borrowers among approved loans, so it would consider HL or LH. As $\frac{B_L}{W}$ continue to grow, both HL and LH would be considered.

Because there is no borrower manipulation in this situation, a borrower in state i will simply accept the lending when he is predicted to be a high-quality type, i.e., $D_i = 1$.

Strategies with feature manipulation

When borrowers manipulate features, the proportion of borrowers in each state will change. Expecting this, the lender can also change its classification strategy. Naturally, we are interested in the equilibrium.

Again, the lender would have 16 possible strategies to consider. We rely on the following result to rule out impossible strategies. We say a classification strategy is more stringent on one feature when the optimal classification requires higher or equal value on this feature, then

Lemma 2 The lender will be more stringent on the manipulable feature under feature manipulation.

Referring to Table 1, suppose if there were no manipulation, the classification model predicts the states with color to approve the loan, then the color area can only shrink at the horizontal dimension with manipulation. For example, if the lender does not want to lend to borrowers in HL under the no manipulation case, then it will not lend to borrowers in HL with manipulation. This is straightforward–Otherwise, if the lender is willing to lend to borrowers in state HL with manipulation but unwilling to do so with no manipulation, then no borrowers in HL will manipulate as this only incurs manipulation cost but does not affect the lending outcome, hence the lender's benefit of lending HL borrowers would be the same as if there were no manipulation. But by assumption, the lender is unwilling to lend to borrowers in HL with no manipulation but willing to lend to borrowers in HL with manipulation. This leads to a contradiction.

With lemma 2, the lender's possible strategies under each scenario can be summarized as follows:

Without manipulation	**With manipulation**
HH	HH or $HH + LH$
$HH + HL$	HH or $HH + LH$ or $HH + HL + LH$
$HH + LH$	HH or $HH + LH$
$HH + HL + LH$	HH or $HH + HL$ or $HH + HL + LH$

Table 2. Lender's possible strategies under different conditions.

We can also determine borrowers' manipulation strategy according to (2) and the lender's potential classification strategies, which affects the composition of high-quality and low-quality borrowers in each state. Consider the number of high-quality borrowers in the state HH, $N^m_{HH,h}, N^m_{HH,l}$, there are three possibilities depending on the lender's classification strategy and the relationship among B_B, C_H, C_L: (1) $N^m_{HH,h} = N^u_{HH,h} + N^u_{HL,h}$ and $N^m_{HH,l} = N^u_{HH,l}$ (The superscript $m(u)$ denote the state with manipulation (no manipulation)) if $C_H \leq B_B < C_L$. This happens when the manipulation is only profitable for the high-quality borrower but too costly for a low-quality borrower. Because the number of high-quality borrowers is increased but the low-quality borrower is decreased so the proportion of high-quality borrowers in HH is increased. (2) $N^m_{HH,h} = N^u_{HH,h} + N^u_{HL,h}$ and $N^m_{HH,l} = N^u_{HH,l} + N^u_{HL,l}$ if $\max\{C_H, C_L\} \leq B_B$. This happens when manipulation is profitable for both borrowers. Now because all borrowers originally in HL are mixed with borrowers in HH, so the proportion of high-quality borrowers in HH is decreased. (3) $N^m_{HH,h}$=$N^u_{HH,h}$ and $N^m_{HH,l}$=$N^u_{HH,l} + N^u_{HL,l}$ if $C_L \leq B_B < C_H$. This is just a reversed case (1). Manipulation is only profitable for low-quality borrowers and the proportion of high-quality borrowers in HH would decrease.

We can follow the same logic to obtain the composition of borrowers corresponding to each possible classification strategy. Given the lender's possible classification strategies and the borrowers' manipulation, we can obtain the equilibrium as follows:

Proposition 2 With feature manipulation, the lender's strategy and payoff can be summarized in Table 3[1]:

Initial Distribution	$\mathbf{\Pi}^{\mathrm{u}}$	$\mathbf{\Pi}^m$ **when** $C_H \leq B_B < C_L$	$\mathbf{\Pi}^m$ **when** $\max(C_H, C_L) < B_B$	$\mathbf{\Pi}^m$ **when** $C_L \leq B_B < C_H$
$not\ C_{HL>0}\ and\ not\ C$	Lend to HH: $\mathbf{\Pi}^{\mathrm{u}}_{HH}$	Lend to HH: $\mathbf{\Pi}^{m(1)}_{HH} > \mathbf{\Pi}^{\mathrm{u}}_{HH}$ Lend to $HH + LH$: $\mathbf{\Pi}^{m(2)}_{HH+LH} > \mathbf{\Pi}^{\mathrm{u}}_{HH}$	Lend to HH: $\mathbf{\Pi}^{m(7)}_{HH} < \mathbf{\Pi}^{\mathrm{u}}_{HH}$	Lend to HH : $\mathbf{\Pi}^{m(12)}_{HH} < \mathbf{\Pi}^{\mathrm{u}}_{HH}$
$C_{HL>0}\ and\ not\ C_{LH>0}$	Lend to $HH + HL$: $\mathbf{\Pi}^{\mathrm{u}}_{HH+HL}$	Lend to HH: $\mathbf{\Pi}^{m(3)}_{HH} > \mathbf{\Pi}^{\mathrm{u}}_{HH+HL}$ Lend to $HH + LH$: $\mathbf{\Pi}^{m(4)}_{HH} > \mathbf{\Pi}^{\mathrm{u}}_{HH+HL}$	Lend to HH: $\mathbf{\Pi}^{m(8)}_{HH} = \mathbf{\Pi}^{\mathrm{u}}_{HH+HL}$	Lend to $HH + HL$: $\mathbf{\Pi}^{m(13)}_{HH+HL} = \mathbf{\Pi}^{\mathrm{u}}_{HH+HL}$
$not\ C_{HL>0}\ and\ C_{LH>0}$	Lend to $HH + LH$: $\mathbf{\Pi}^{\mathrm{u}}_{HH+LH}$	Lend to $HH + LH$: $\mathbf{\Pi}^{m(5)}_{HH+LH} > \mathbf{\Pi}^{\mathrm{u}}_{HH+LH}$	Lend to HH: $\mathbf{\Pi}^{m(9)}_{HH} < \mathbf{\Pi}^{\mathrm{u}}_{HH+LH}$ Lend to $HH + LH$: $\mathbf{\Pi}^{m(10)}_{HH+LH} < \mathbf{\Pi}^{\mathrm{u}}_{HH+LH}$	Lend to HH: $\mathbf{\Pi}^{m(14)}_{HH} < \mathbf{\Pi}^{\mathrm{u}}_{HH+LH}$
$C_{HL>0}\ and\ C_{LH>0}$	Lend to $HH + HL + LH$: $\mathbf{\Pi}^{\mathrm{u}}_{HH+HL+LH}$	Lend to $HH + LH$: $\mathbf{\Pi}^{m(6)}_{HH+LH} > \mathbf{\Pi}^{\mathrm{u}}_{HH+HL+LH}$	Lend to HH: $\mathbf{\Pi}^{m(10)}_{HH} < \mathbf{\Pi}^{\mathrm{u}}_{HH+HL+LH}$ Lend to $HH + LH$: $\mathbf{\Pi}^{m(11)}_{HH+LH} < \mathbf{\Pi}^{\mathrm{u}}_{HH+HL+LH}$	Lend to $HH + HL$: $\mathbf{\Pi}^{m(15)}_{HH+HL} < \mathbf{\Pi}^{\mathrm{u}}_{HH+HL+LH}$

Table 3. Lender's payoff under manipulation and their comparisons with the initial ones.

In Table 3, the first character superscript of payoffs represents the no manipulation case ("u") and manipulation case ("m"), the second number superscript corresponds to the area in Figure 1. Whether borrowers would manipulate is decided by the cost of manipulation (C_H or C_L) relative to the benefit of improving the classification outcome (B_B). We represent the detailed payoff expressions in Online Appendix.

The impact of manipulation on the payoff and classification strategy

This section examines how manipulation affects the lender's classification strategy and payoff.

By comparing the lender's classification strategy under the no manipulation case and the feature manipulation case, we can analyze how manipulation affects the optimal classification strategy. We are interested in how manipulation affects the prediction power of manipulable and unmanipulable features, measured by the probability that having H at a feature predicts high-quality borrowers, and the criterion the lender would set on each feature.

Proposition 3 Compared with no manipulation case, with manipulation, the prediction power of the manipulable feature is increased when $C_H < B_B < C_L$ and decrease otherwise. The prediction power of unmanipulable feature is unchanged, but the lender's criterion on unmanipulable feature tends to be softened, i.e., $D^m_{i,j} \geq D^u_{i,j}$, when $C_H \leq B_B < C_L$ and tightened, i.e., $D^m_{i,j} \leq D^u_{i,j}$, otherwise.

Although manipulation is commonly believed to hurt classification performance, our analysis shows this may not be the case under some circumstances. Particularly, when high-quality borrowers manipulate more than lower quality borrowers (i.e., $C_H \leq B_B < C_L$), the impact of manipulation is to make high-quality borrowers more likely to have H at the manipulable feature, which makes it easier for the lender to

[1] Cases that cannot achieve equilibrium are omitted.

differentiate high-quality borrowers from low-quality ones. Refer to Table 1, with no manipulation, the prediction power of manipulable feature is $\frac{(\theta\gamma_m)}{\theta\gamma_m+(1-\theta)(1-\gamma_m)}$ (i.e., $\frac{\text{high}-\text{quality borrowers having } H \text{ at the manipulable feature}}{\text{all borrowers having } H \text{ at the manipulable feature}}$). With manipulation and if it is high-quality borrowers have more advantages in manipulation (i.e., if $C_H \le B_B < C_L$), then according to (2), high-quality borrowers at L would have incentive to manipulate to obtain H, making the prediction power of manipulable feature becomes $\frac{\theta}{1-(1-\theta)\gamma_m}$ (i.e., $\frac{\#HH,h+\#LH,h+\#HL,h+\#LL,h}{\#HH+\#LH+\#HL,h+\#LL,h}$), which is larger than $\frac{(\theta\gamma_m)}{\theta\gamma_m+(1-\theta)(1-\gamma_m)}$, thus the H at manipulable feature can predict high-quality borrowers with a higher probability.

However, if it is mainly low-quality borrowers have more incentive to manipulate ($C_L \le B_B < C_H$), then the above logic is reversed. Manipulation would allow more low-quality borrowers to obtain the higher feature value and thus hurt the performance of classification, i.e. the prediction power of manipulable feature becomes $\frac{\theta\gamma_m}{1-\theta(1-\gamma_m)}$ (i.e., $\frac{\#HH,h+\#HL,h+\#LH,l+\#LL,l}{\#HH+\#LH+\#HL,l+\#LL,l}$), which is smaller than $\frac{\theta\gamma_m}{\theta\gamma_m+(1-\theta)(1-\gamma_m)}$. The result is similar in the case both borrowers manipulate.

On the other hand, because borrowers cannot manipulate the unmanipulable feature, the distribution of borrowers on the unmanipulable feature dimension does not change hence there is no change in the prediction power of unmanipulable feature.

Although the prediction power of unmanipulable feature is unchanged, manipulation can pose externality on the lender's criterion on the unmanipulable feature, and different from what Lemma 2 suggests about the manipulable feature, the lender can either soften or tighten its criterion on unmanipulable features.

Consider a concrete example in Table 4, when with $\theta = 0.5, \gamma_u = 0.7, \gamma_m = 0.8, B_L = 1.5, W = 1$. With no manipulation, the distribution of borrowers indicates the lender only has incentive to lend to borrowers in HH state. However, when with manipulation and high-quality borrowers have more advantages in manipulation, the lender's incentive to consider LH would be increased, because a softened criterion on unmanipulable feature would induce high-quality borrowers in state LL to manipulate to obtain LH, which makes the proportion of high-quality borrowers higher than that if there were no manipulation (i.e., the proportion of high-quality borrowers increase from $\theta(1-\gamma_u)\gamma_m = 63\%$ to $\theta(1-\gamma_u) = 68\%$). Since the increase in high-quality borrowers in state LH is substantial enough (making the profit of lending to LM turn from -0.01 to 0.005), the lender would be incentivized to lower its criterion on the unmanipulable feature by lending to borrowers in LH state.

When without manipulation			**When high-quality borrowers manipulate**	
$\boldsymbol{HH},\boldsymbol{h}$: 0.28 $\boldsymbol{HH},\boldsymbol{l}$: 0.03	$\boldsymbol{HL},\boldsymbol{h}$: 0.07 $\boldsymbol{HL},\boldsymbol{l}$: 0.12		$\boldsymbol{HH},\boldsymbol{h}$: 0.35 $\boldsymbol{HH},\boldsymbol{l}$: 0.03	$\boldsymbol{HL},\boldsymbol{h}$: 0 $\boldsymbol{HL},\boldsymbol{l}$: 0.12
$\boldsymbol{LH},\boldsymbol{h}$: 0.12 $\boldsymbol{LH},\boldsymbol{l}$: 0.07	$\boldsymbol{HL},\boldsymbol{h}$: 0.03 $\boldsymbol{HL},\boldsymbol{l}$: 0.28		$\boldsymbol{LH},\boldsymbol{h}$: 0.15 $\boldsymbol{LH},\boldsymbol{l}$: 0.07	$\boldsymbol{HL},\boldsymbol{h}$: 0 $\boldsymbol{HL},\boldsymbol{l}$: 0.28

Table 4. Illustrative example about lender's strategies (colored cell) with and without the manipulation of high-quality borrowers.

This result has an interesting implication for practice. With no manipulation, some high-quality borrowers having a disadvantageous unmanipulable feature, such as being black or female, may be rejected by the lender because they are mixed with too many low-quality borrowers. Yet, with manipulation and if high-quality borrowers have more advantages in manipulating (e.g., it could be easier for a high-quality borrower who can obtain decent bank asset proof by manipulation), then the lender's incentive to include these borrowers with disadvantageous unmanipulable features could be increased by softening its criterion on unmanipulable features such as race or gender but tighten the criterion on the manipulable feature (refer to Lemma 2), because such change in criterion would mainly admit more high-quality borrowers.

On the other hand, if it is mainly the low-quality borrowers having more advantages in manipulation, manipulation tends to induce the lender to tighten its criterion on the unmanipulable feature. The logic is just a reverse of the previous discussion—when it is mainly the low-quality borrowers having more advantages in manipulation, manipulation could decrease the proportion of high-quality borrowers in some states that the lender originally finds profitable to lend. Thus, the lender would become even stricter in the criterion set at the unmanipulable feature.

In sum, contrary to some common beliefs, strategic manipulation can either increase or decrease classification performance. Whether manipulation increases or decreases classification performance depends on whether it is the high-quality borrowers or low-quality borrowers manipulate more.

We next examine how manipulation affects the lender's payoff.

Proposition 4 Compared with no manipulation case, the lender's payoff under manipulation would increase when $C_H \leq B_B < C_L$ and decrease otherwise.

If high-quality borrowers manipulate more, then the lender's classification algorithm can predict high-quality borrowers better, and Proposition 4 shows this better performance can translate to higher profit for the lender. Note the lender's payoff is primarily decided by two factors: the number of borrowers it can reach and the proportion of high-quality borrowers within these borrowers. If it is high-quality maintainers who manipulate more, previous analyses suggest manipulation can benefit both factors: On the one hand, manipulation allows borrowers who originally cannot pass the lender's classification criterion to meet the criterion by manipulating its feature, thus increasing the number of borrowers the lender can reach. On the other hand, because it is high-quality borrowers who manipulate more, the increased borrowers would mainly be high-quality borrowers, thus the proportion of high-quality borrowers also tends to increase.

Additionally, it is straightforward to see that lending to high-quality borrowers would improve social welfare. As manipulation leads more high-quality borrowers to receive the loan with $C_H \leq B_B < C_L$, so the social welfare can also be improved with manipulation in this case. However, manipulation is bad for the lender if low-quality borrowers manipulate more. In that case, the strategic manipulation would give more advantages to low-quality borrowers (i.e., in terms of being easier to be classified into the desired category) and so harms the lender. Although there are recent studies suggesting that manipulation could be beneficial, those settings are based on the assumption that manipulation itself could increase individuals' performance (e.g., Wang et al. (2020) suggests manipulation allows job candidates to seek education which leads to an increase in their productivity). Our result is different because the manipulable feature does not have productivity impacts. The source of benefit just comes from the strategic interaction itself.

Extension

We consider two extensions of our main model. First, we incorporate algorithm transparency. It is possible that the lender can hide the features used for classification to avoid borrowers' manipulation. For example, the lender can either ask borrowers to directly provide the bank asset proof or work with banks to obtain borrowers' bank asset condition without borrowers knowing it. In the second case, because borrowers do not know whether their bank assets will be used for classification, they are less likely to manipulate this feature. We can endogenize the lender's algorithm transparency decision. Suppose the lender can choose either to be "opaque" or "transparent" on a manipulable feature. In an opaque case, the lender does not disclose whether a feature will be used for classification, so borrowers will not manipulate the feature. While in a transparent case, borrowers can continue to manipulate the feature. Apparently, the "opaque" case is the same as the no manipulation case, so the same analysis can be applied to analyze the lender's algorithm transparency decisions. Proposition 3 and 4 suggest that the lender has greater incentives to be transparent on manipulable features that high-quality borrowers have advantages in manipulation, while to be opaque on manipulable features that low-quality borrowers have advantages in manipulation.

Second, we can incorporate classification algorithms that automatically extract features, as in the case of deep learning models. To model this, assume the lender applies a deep learning model to classify borrowers. Instead of directly using the user features for classification, the deep learning model will take user features as input and automatically generates "deep features" that are correlated with users' features for classification. Typically, these deep features are not interpretable to humans, so borrowers will face uncertainty regarding whether their manipulation can lead to the desired outcome. Suppose that, based on these deep features, the algorithm classifies borrowers into two states, $deepH$ and $deepL$. Assume that the lender will lend to borrowers having $deepH$ state, so the lender's payoff can be represented as $\Pi = \mathrm{N}_{\mathrm{deepH,h}}(B_L - W) + \mathrm{N}_{\mathrm{deepH,l}}(-W)$, where $\mathrm{N}_{\mathrm{deepH,h}}$ ($\mathrm{N}_{\mathrm{deepH,l}}$) represents the number of high (low) quality borrowers that have $deepH$ states. We model the payoff of a borrowers with type $i \in \{h, l\}$ and state j (or j')$\in \{deepH, deepL\}$ as $U_{i,j} = \begin{cases} D_j B_B & \text{if not manipulate} \\ D_{j'} p(B_B - C_i) + D_j(1-p)(-C_i) & \text{if manipulate} \end{cases}$, where j' represents the state after manipulation. This specification models that borrowers will face uncertainties, measured by $p \in [0,1]$, about whether their manipulation would result in the desired outcome, i.e., being classified as $deepH$ by the deep learning model. p can be thought of as the complexity of the deep learning model. For a complex deep learning model, because it is difficult for borrowers to understand the model structure and parameters, so p tends to be small, i.e., there is great uncertainty in manipulation. In contrast, for a simple deep learning model, borrowers may gauge the decision mechanisms of the deep learning model by using XAI techniques (e.g., Arrieta et al. 2020), so p tends to be large. We find our main insights continue to hold with this extension, that is, manipulation can increase the firm's payoff when high type borrowers manipulate more, i.e., when $C_H < pB_B < C_L$. Further, we find that (a) a deep learning model reduces borrowers' incentive to manipulate. (b) a more complex model (measured by smaller p) can increase (decrease) the firm's payoff when low (high)-quality borrowers who manipulate more.

Concluding Remark

This paper develops a classification game in the online lending context, in which borrowers can manipulate their features to get lending from institutions. Our model creates a basis for discussing the implication of algorithm manipulation and provides useful insights for organizations to design and apply algorithms in settings wherein feature manipulation is possible. Our findings also offer rationales for why firms practically deploy manipulable features in their machine learning models despite the widespread concerns.

The next step of this study is to analyze and compare different policies that firms can use to either mitigate the cost or leverage the benefit due to manipulation. The firm may reduce manipulation activities by increasing the cost for users to manipulate, for example, by asking users to certificate the features or by including more features in the machine learning models; An alternative way is that the firm can incur costs to make their machine learning models more robust to manipulation (e.g., Hardt 2016). Which type of approach is more appropriate? In addition, given that manipulation sometimes can benefit the classification, the firm may also want to encourage manipulation under some circumstances. How should the firm implement such incentives? We believe these are interesting questions to explore in the future.

Last, it is worth noting that we do not try to make a legal or moral judgment about manipulation. Actually, in some cases, manipulation may not cause severe moral or legal issues. For example, consider a university admission scenario, and one feature for an applicant could be the essay writing sample. Applicants may ask professional institutions to help polish the essay, which could also be regarded as one type of manipulation. This kind of manipulation seems to be acceptable.

References

Arrieta AB, D´ıaz-Rodr´ıguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garc´ıa S, Gil-L´opez S, Molina D, Benjamins R, et al. 2020. "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion* 58:82–115.

Dalvi, N., Domingos, P., Sanghai, S. and Verma, D. 2004. "Adversarial classification," In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 99-108).

Dekel, O., Fischer, F. and Procaccia, A.D. 2010. "Incentive compatible regression learning," *Journal of Computer and System Sciences*, *76*(8), pp.759-777.

Frankel, A. and Kartik, N. 2019. "Muddled information," *Journal of Political Economy*, *127*(4), pp.1739-1776.

Goodhart, C.A. 1984. "Problems of monetary management: the UK experience," In *Monetary theory and practice* (pp. 91-121). Palgrave, London.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. 2016. "Strategic classification," In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science* (pp. 111-122).

Hu, L., Immorlica, N. and Vaughan, J.W. 2019. "The disparate effects of strategic manipulation," In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 259-268).

Jann, O. and Schottmüller, C. 2020. "An informational theory of privacy," *The Economic Journal*, *130*(625), pp.93-124.

Milli, S., Miller, J., Dragan, A.D. and Hardt, M. 2019. "The social cost of strategic classification," In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 230-239).

Tu, Z., Zhang, J. and Tao, D., 2019. "Theoretical analysis of adversarial learning: A minimax approach," *Advances in Neural Information Processing Systems*, *32*.

Wang, Q., Huang, Y., Jasin, S., & Singh, P. V. 2022. "Algorithmic transparency with strategic users". *Management Science*. *Forthcoming*