Learning to Incentivize Improvements from   
Strategic Agents

Course: Strategic and Societal Aspects of Machine Learning.

By Cfir Hadar.

July 2023.

Abstract

TODO

Summary

# What is the Problem and Why is it Interesting?

Machine learning systems are often used in settings where individuals adapt their features to obtain a desired outcome. Individuals, subject to a classifier’s predictions, may act strategically to influence their predictions (‘strategic manipulation’). In such settings, strategic behavior leads to a major loss in model performance in deployment.

In this work, the researchers present a training method that both maintain the accuracy of existing approaches, while inducing higher levels of improvement and less manipulation.

# Main Contributions

In this research paper, the authors make the following key contributions:

Introducing a New Approach for Strategic Adaptation

The paper presents a novel approach to address strategic adaptation in machine learning. This approach centers around the concept of constructive adaptation risk, which involves training classifiers that incentivize decision subjects to adapt their features in ways that improve true outcomes. The authors provide formal evidence that this risk captures both the strategic and constructive dimensions of decision subjects' behavior. They assume a feature taxonomy that distinguishes improvable features, which, if changed, lead to changes in the true qualification, from non-causal features, which do not affect the true qualification.

## Characterizing the Dynamics of Strategic Decision Subjects

To understand the dynamics between strategic decision subjects and the model designer in a classification setting, the paper formulates a two-player sequential game. Building upon previous works on strategic classification, the authors generalize cost functions to utilize the Mahalanobis distance, allowing for the capture of correlations between changes in different features. By doing so, they derive closed-form expressions for decision subjects' optimal strategies, shedding light on their behavior when the model designer employs non-causal features as predictors.

## Formulating the Problem as Risk Minimization

The paper formulates the problem of training the desired classifier as a risk minimization problem. They propose a method to train the classifier that addresses the objectives of incentivizing improvement or discouraging adversarial manipulation. To validate the effectiveness of this method, the authors conduct evaluations on simulated and real-world datasets. The empirical results demonstrate the superior performance of their approach compared to existing methods, even in cases where some feature types are misspecified. Additionally, the paper suggests a potential extension of the main result into a non-linear setting using LIME

# How did they do it?

## Recap

Standard prediction settings . This classifier performs poorly in a setting with strategic adaptation since the model is deployed on a different population.

Existing approaches in strategic classification tackle these issues by training a classifier that is robust to *all* adaptation (i.e., all adaptations are undesirable) , where is the best adaption. This design choice misses the opportunity to encourage a profile to truly improve to change their .

Ideal solution is , that is when the true label changes with the adapted features (impossible to solve since is unknown).

## Proposed Approach

To authors divide features to three separate groups - . – immutable, cannot be altered (race, age). – Improvable, can be altered in a way that will change the true outcome. – manipulable, can be altered without changing the true outcome.

The mechanism designer is to incentivize helpful changes in , and no change in .

,

whereas is the manipulation risk, which penalizes pure manipulation. is the improvement risk, which rewards decision subjects for playing their improving best response.

Note that is the manipulating (improving) best response, which involves an adaptation that only alters manipulable (improvable) features.

## Best Response

Suppose a linear classifier . Define the cost of altering using Mahalanobis norm , whereas subscript ‘A’ is the actionable part of the vector (excluding the immutable features), and is positive definite. This allows for more complex cost regimes.

Followed up is the obvious utility definition .

The researchers found a close form solution for that allow for solving using the BFGS algorithm.

# Key Take-Aways

The main take-away of this paper is the new concept of utilizing strategic behavior of players to improve positive rate in classification. As mentioned before, this may be helpful for companies making a profit out of people that are classified positive and correctly.

Implementation

Can be found in my github TODO add link.

Extensions

As seen in the paper, the proposed method is better than just dropping (ignoring) the manipulative features in term of incentivizing improvement and preventing deterioration (true 1s becoming 0s). For example, in the ‘credit’ dataset, the ManipulationProof method achieves 36% improvement and 23% deterioration, while the proposed approach achieves 54% improvement while only 13% deterioration. In what way is this great result reflecting in the learnt classifier?

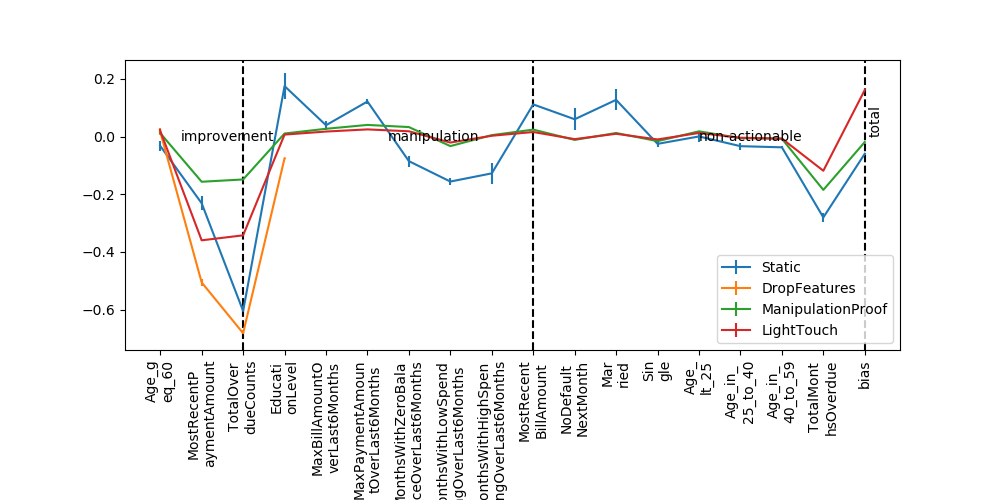


Figure 1 ‘credit’ dataset weights (per feature) mean +- std plot (over folds) for the different methods. Features are divided into 3 groups. Improvement -> manipulative -> non-actionable.

As we can see in Figure 1, while the standard method fell for the manipulations, both the proposed approach, and the ManipulationProof method learnt close-to-zero weights for the manipulative features. But while the ManipulationProof method didn’t incentivize high reward for the improvement features, the proposed approach did learn to reward those more.

(Note, the feature values are centered by subtracting the mean and dividing by the std, per feature).

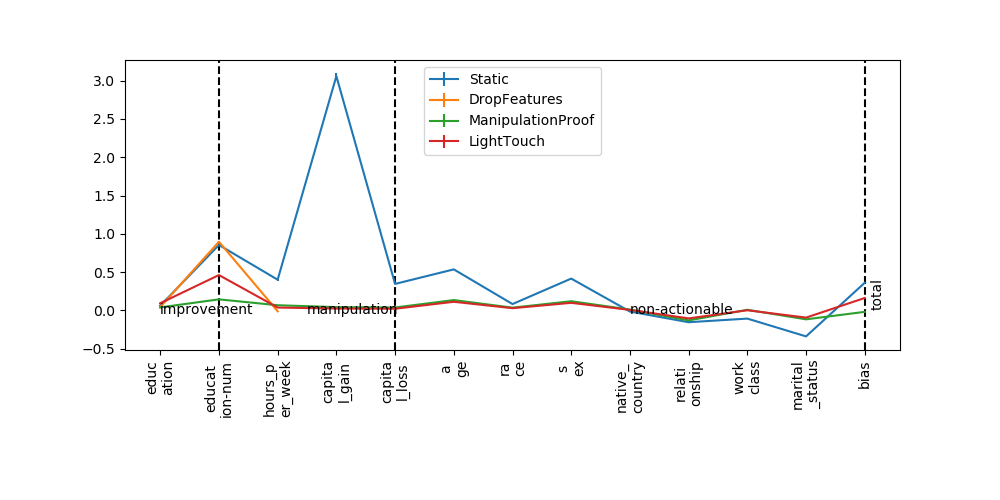


Figure 2 ‘adult’ dataset.

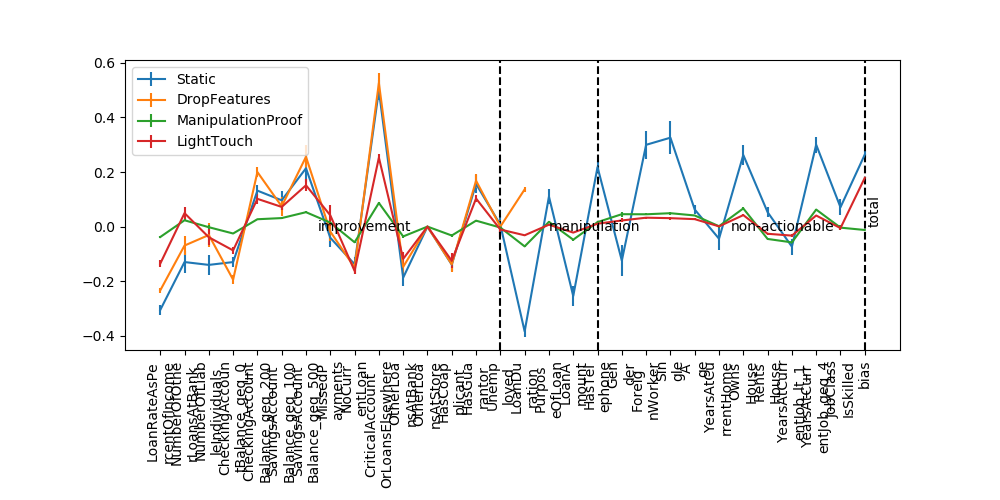


Figure 3 'germen' dataset.

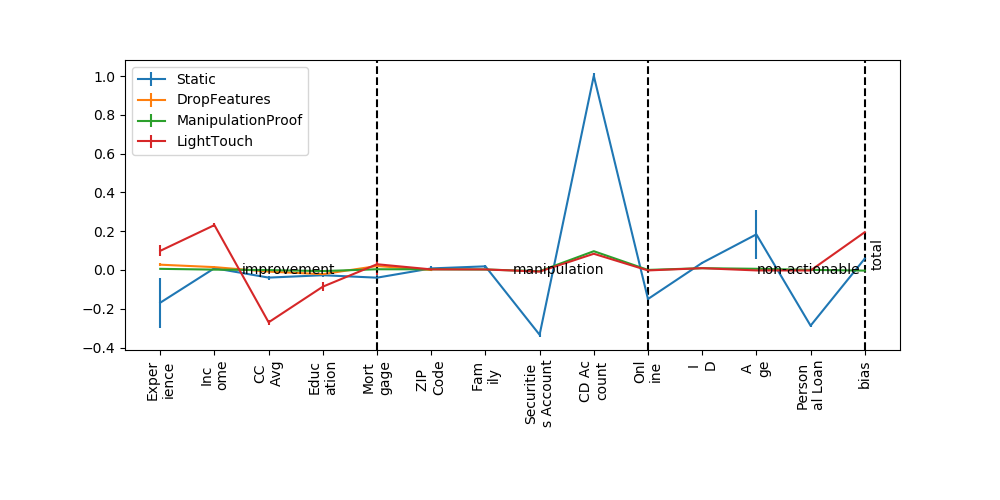


Figure 4 'bank' dataset.

As we can see the analysis is consistent across datasets.

The scientific community is well-aware of the ‘cherry picking’ phenomena, where researchers show only results that support their claims.

[*THERA BANK - Personal Loans Conversion Analysis*](https://www.kaggle.com/code/debajyotipodder/personal-loans-conversion-binary-classification/notebook)*,* the goal is too predict the likelihood of a liability customer buying personal loans (while retaining them as depositors).

The proposed approach yielded brilliant results with as far of 88% user improvement, and not more than 1% deterioration, while preserving approximately the same accuracy.

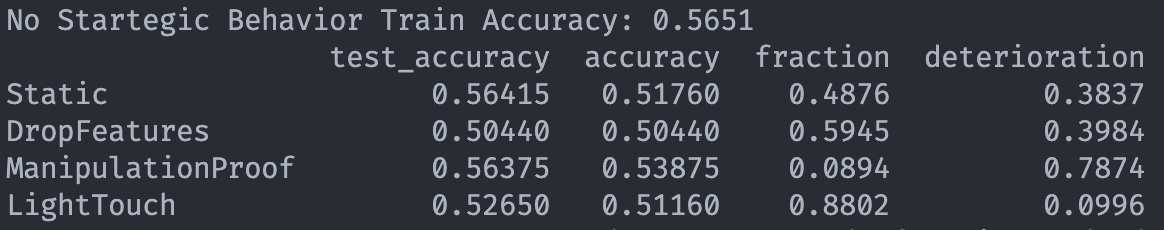
**

Figure 5 notes: (1) test\_accuracy: accuracy of the classifier (without strategic behavior) on the test set. (2) accuracy: (%) of examples classified correctly even after manipulation. (3) fraction: (%) changes from 0 (true label) to 1 (predicted) after improvement. (4) deterioration: (%) changes from 1 (true label) to 0 (predicted) after improvement.

Conclusions

In addition to the fascinating insights gained from this project, there are several valuable lessons I have learned along the way. Firstly, this project has highlighted the importance of innovative thinking and approaching problems from different angles. The paper's novel approach to incentivizing improvement, rather than solely focusing on combating manipulations, has broadened my understanding of mechanism design and its practical applications.

Future work may include analyzing the non-linear model extension suggested in the paper powered by LIME. I think LIME might not be the best-fit approach for this task, as it presents only local interpretation of the model, which as we already know deep neural networks may be too non-smooth for that.

Reference

1. Dataset - <https://www.kaggle.com/code/debajyotipodder/personal-loans-conversion-binary-classification/notebook>.
2. Paper - <https://openreview.net/forum?id=W98AEKQ38Y>.