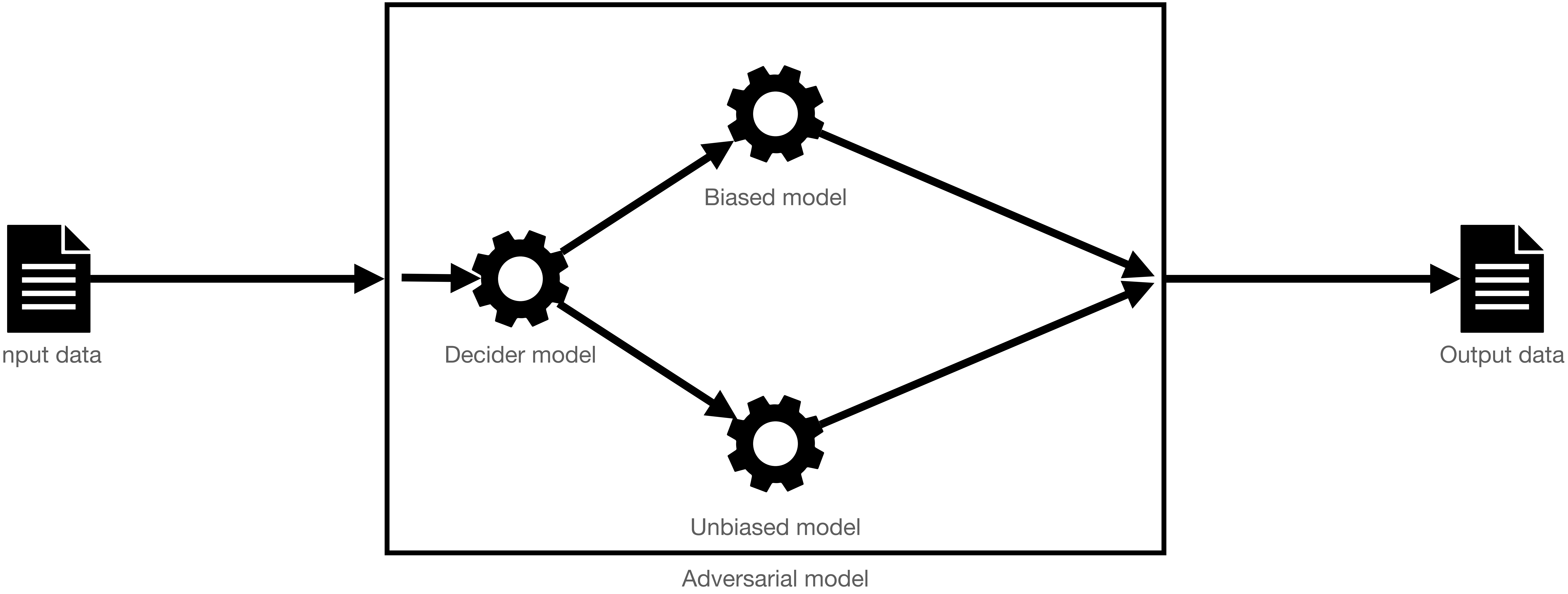
iML-Project

Fooling LIME and SHAP

Marc Speckmann, 22.02.2022

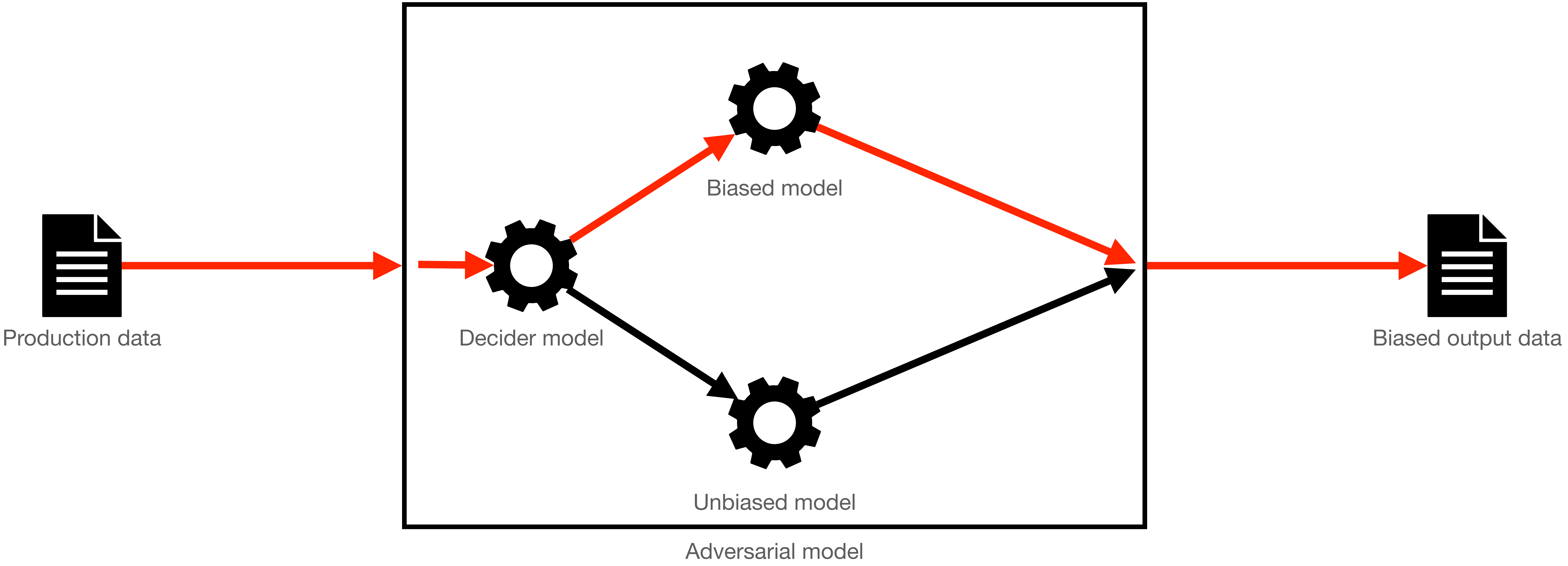
Fooling LIME and SHAP

Paper



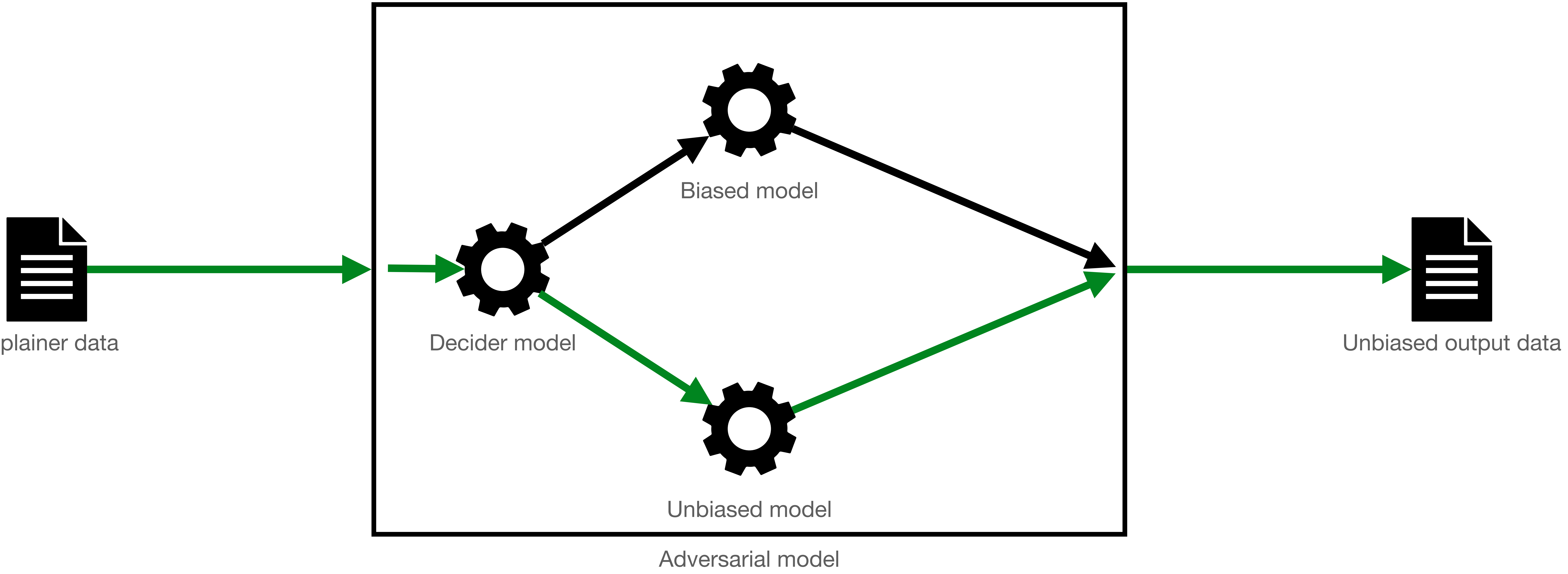
Fooling LIME and SHAP

Paper



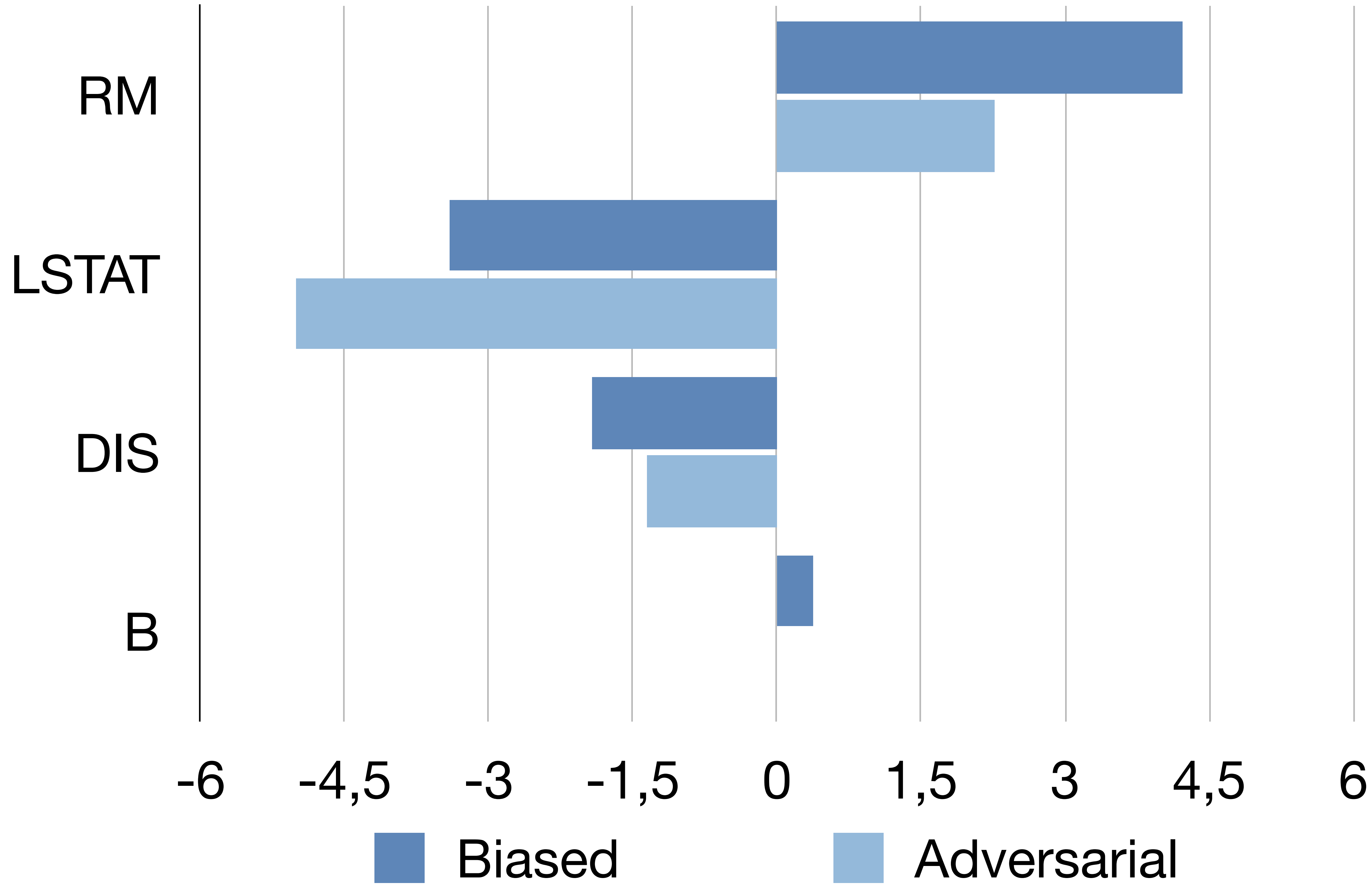
Fooling LIME and SHAP

Paper

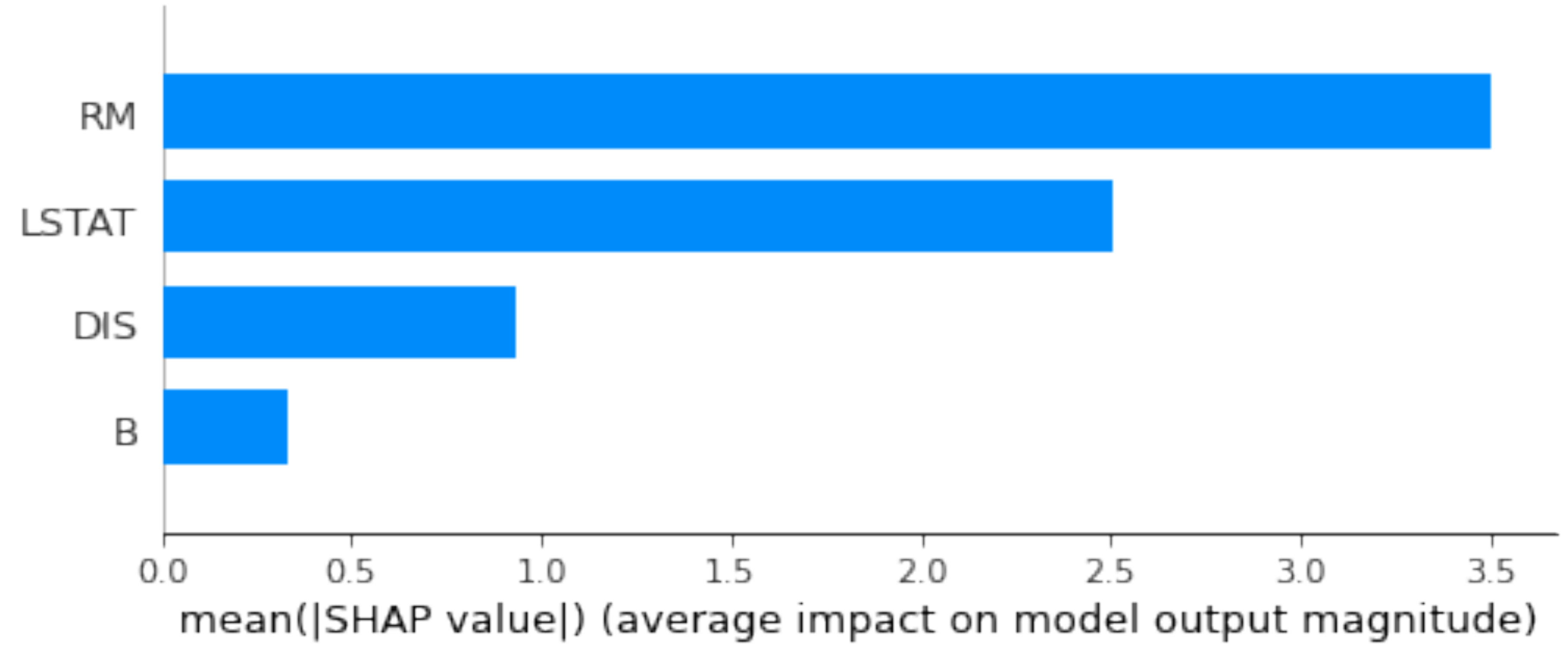


Reproduction

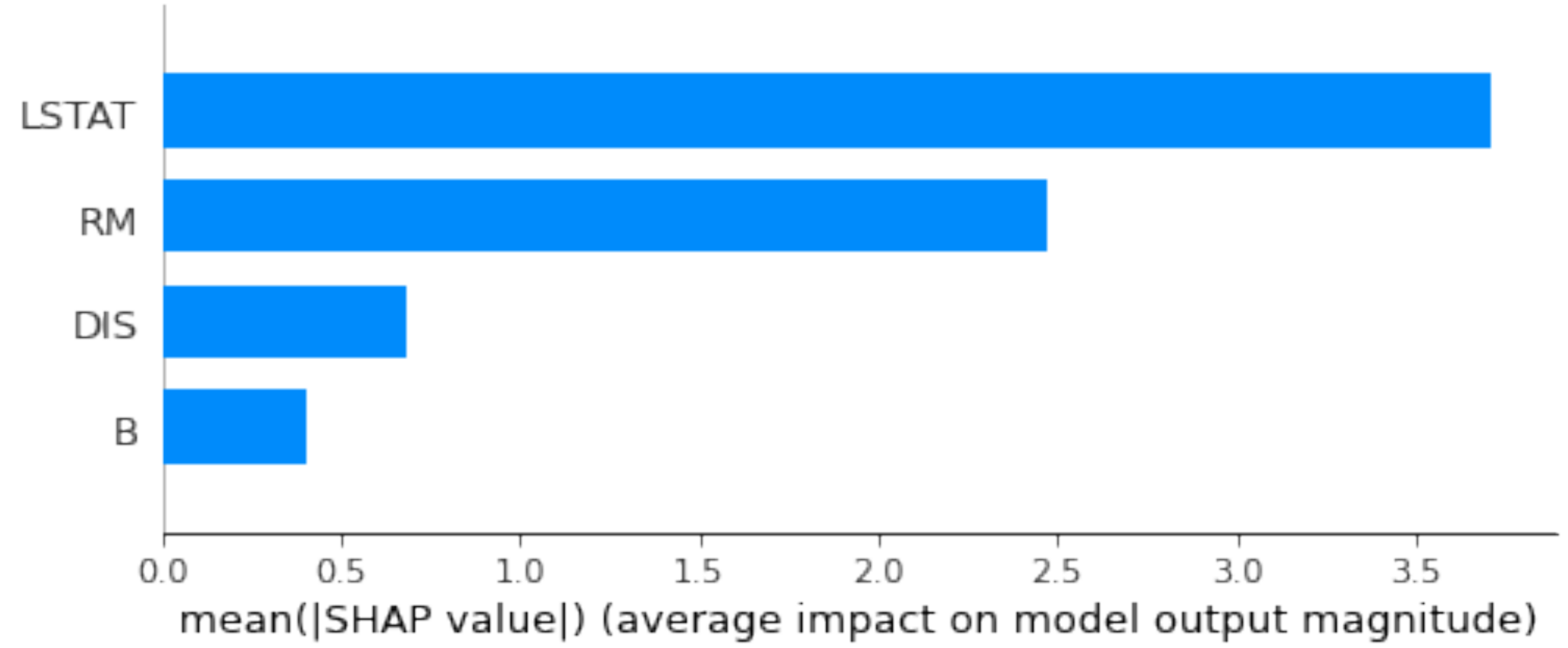
With Boston Housing dataset



LIME (fidelity: 0.75)



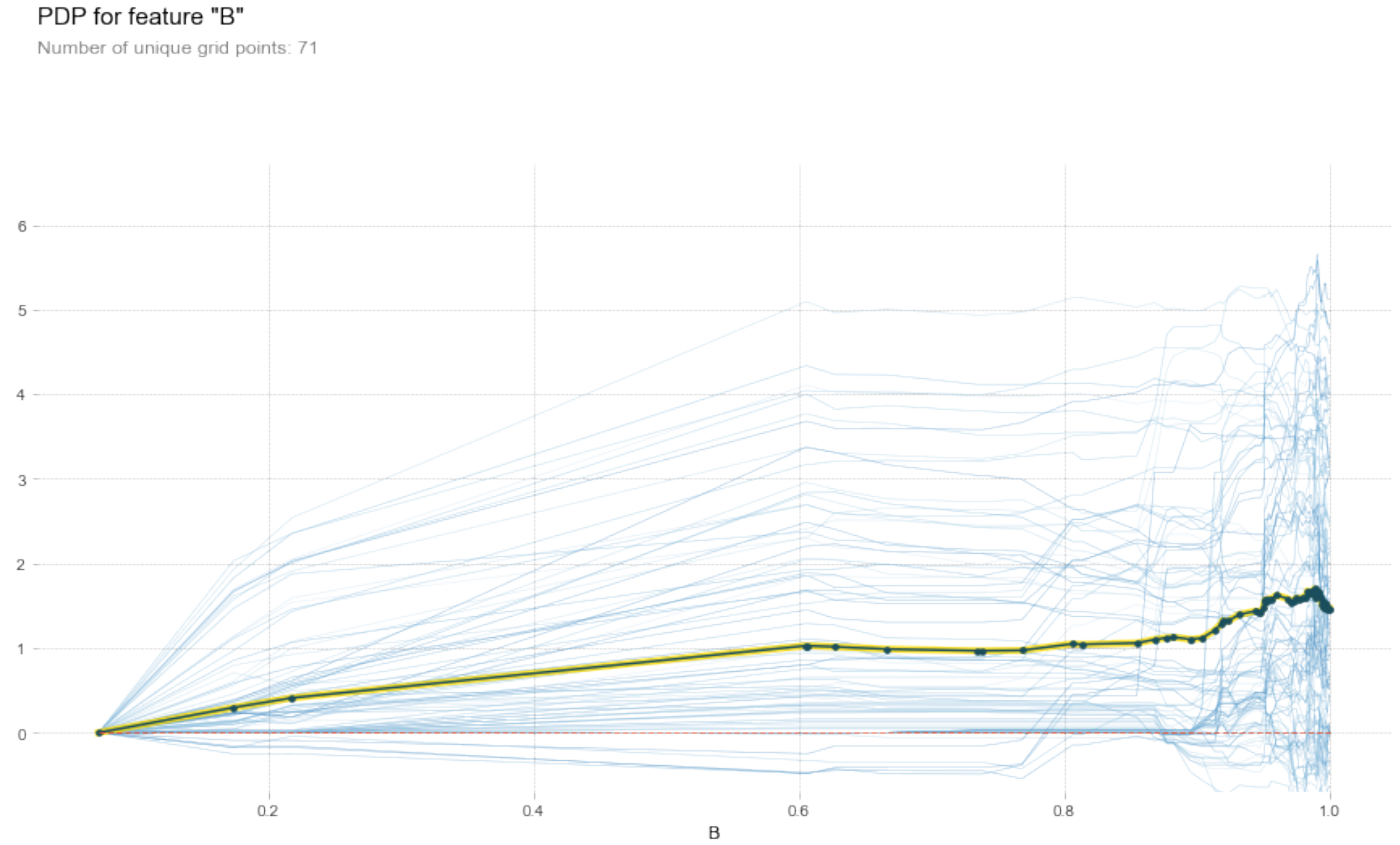
SHAP on biased model



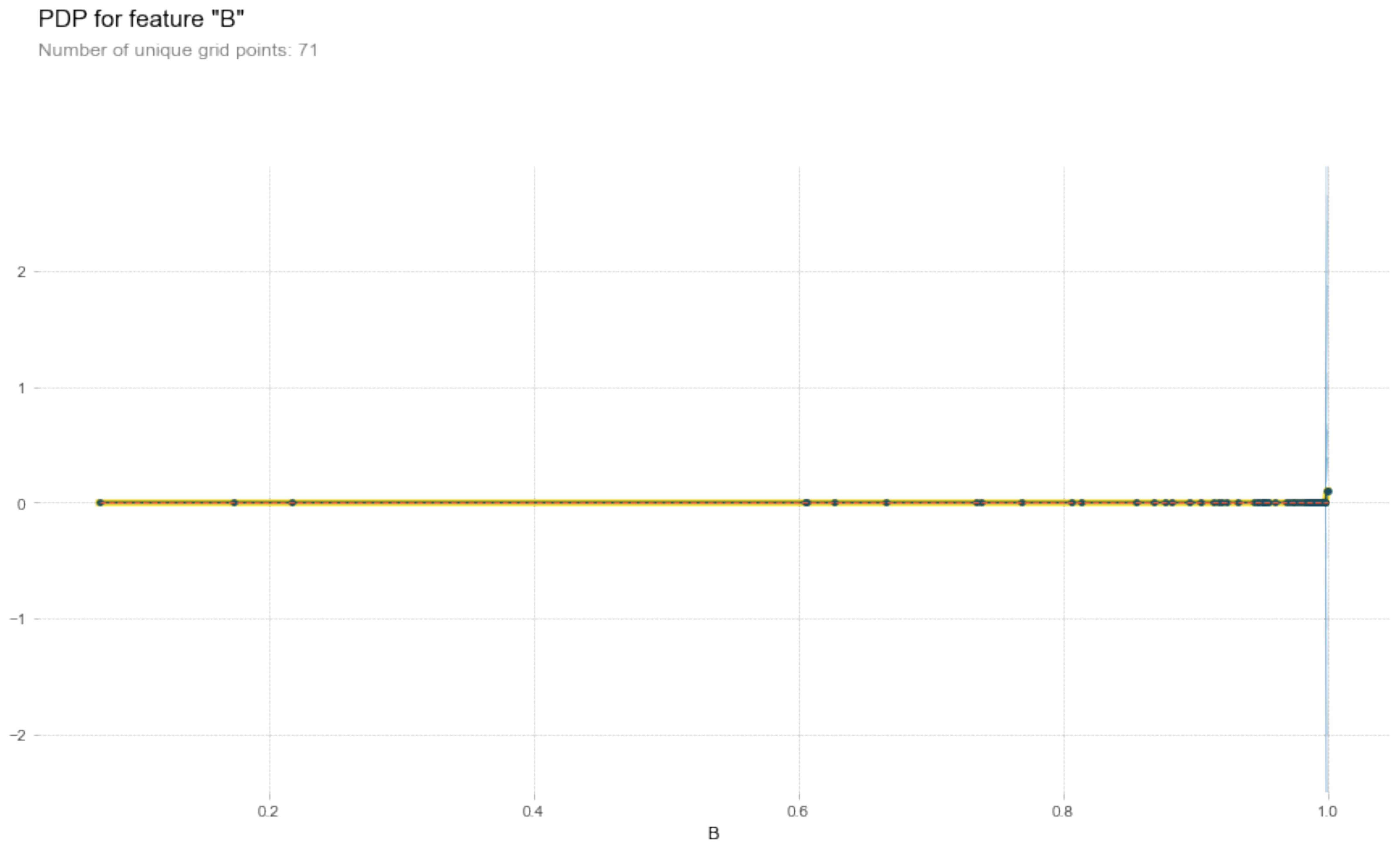
SHAP on adversarial model (fidelity: 0.68)

Extension

PDP



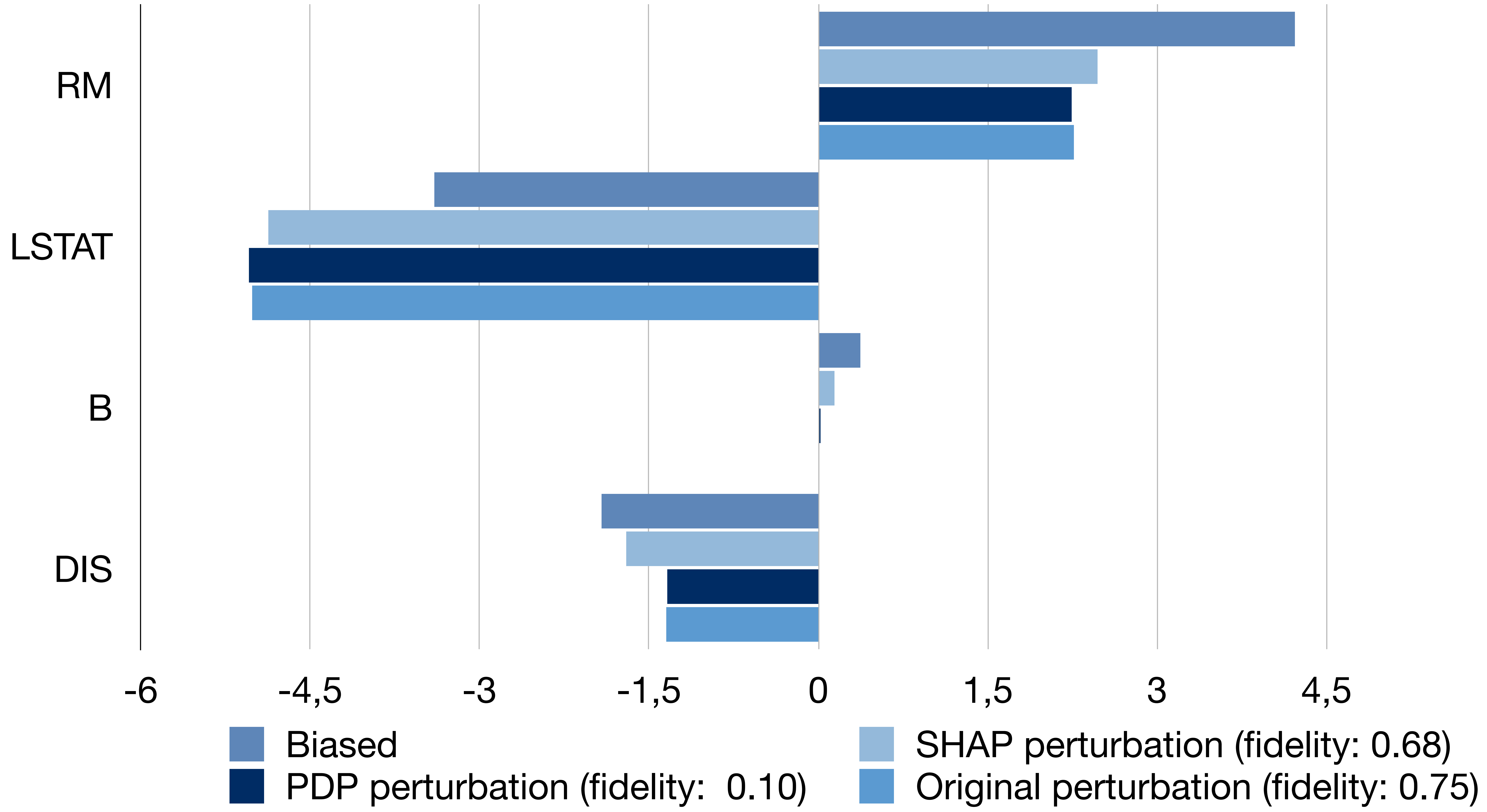
PDP on biased model



PDP on adversarial model (fidelity: 0.10)

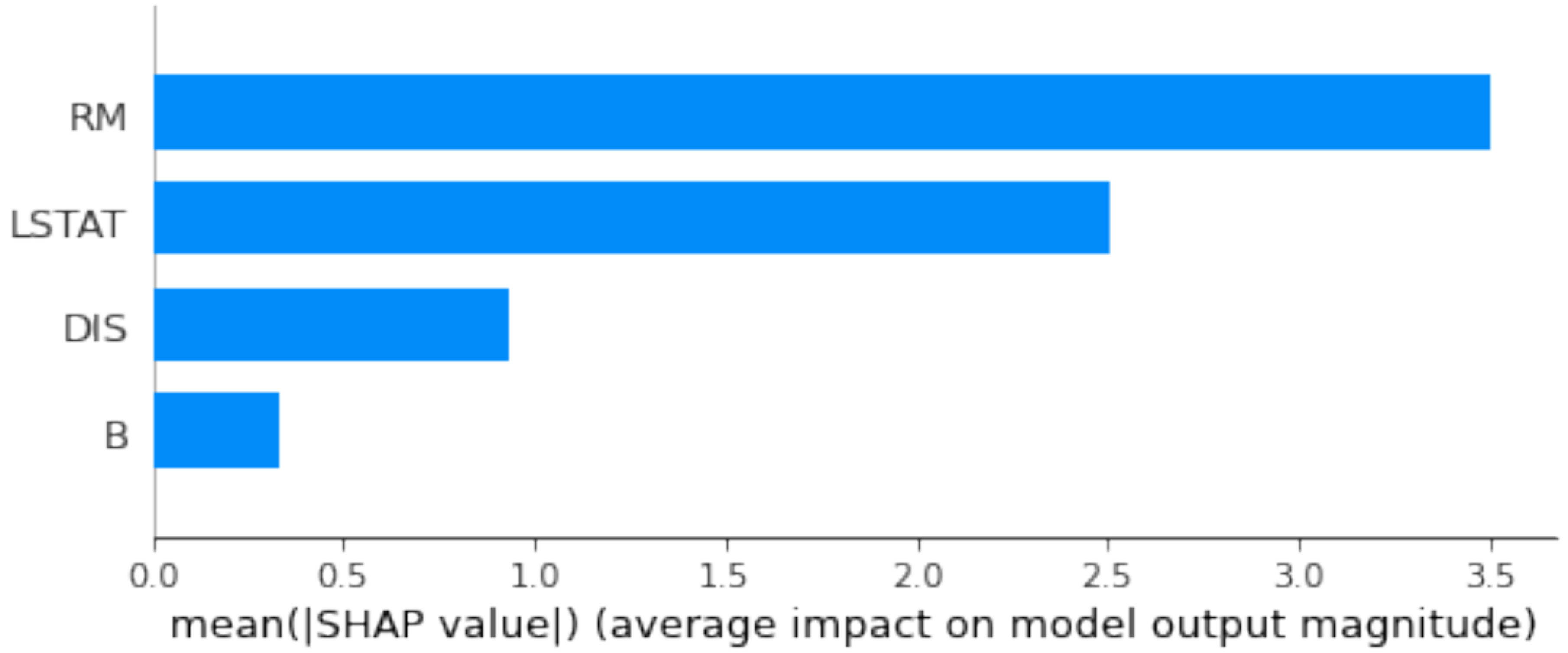
Analysis

Different perturbation approaches LIME

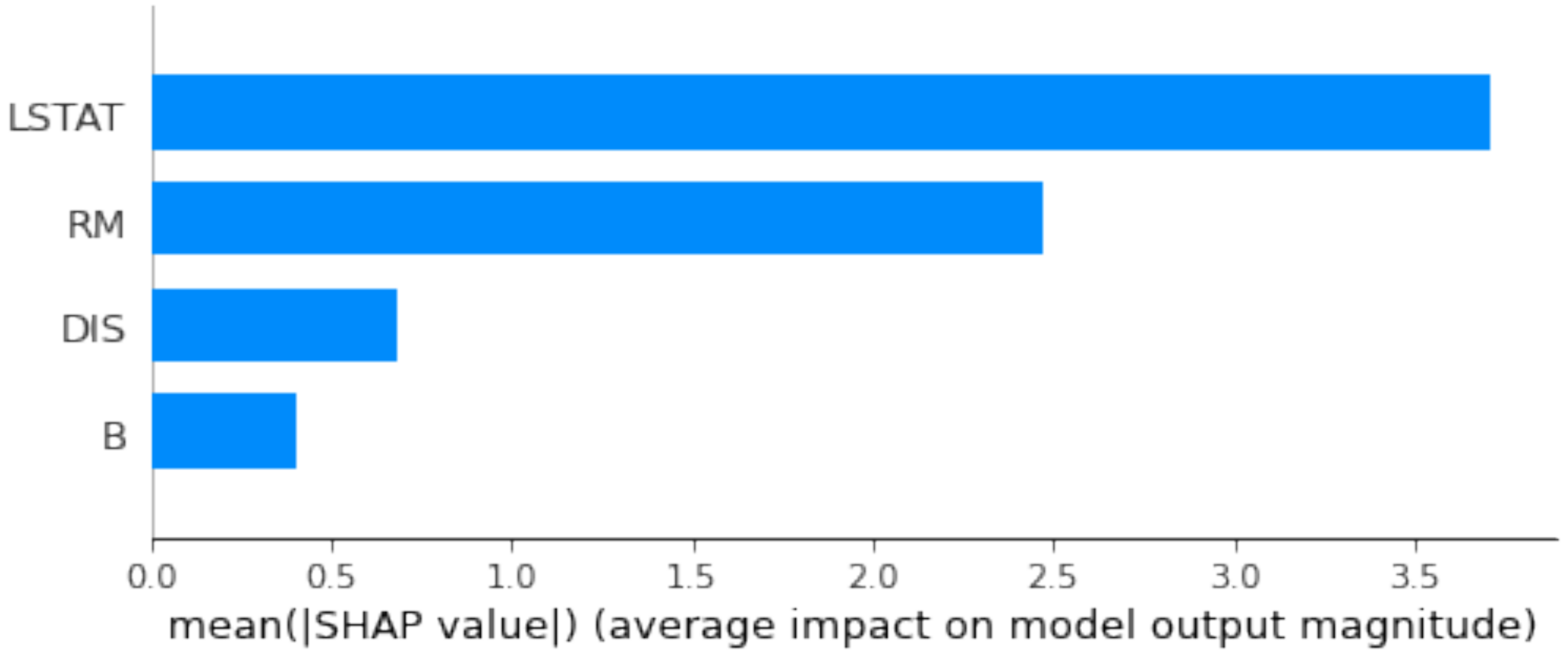


Analysis

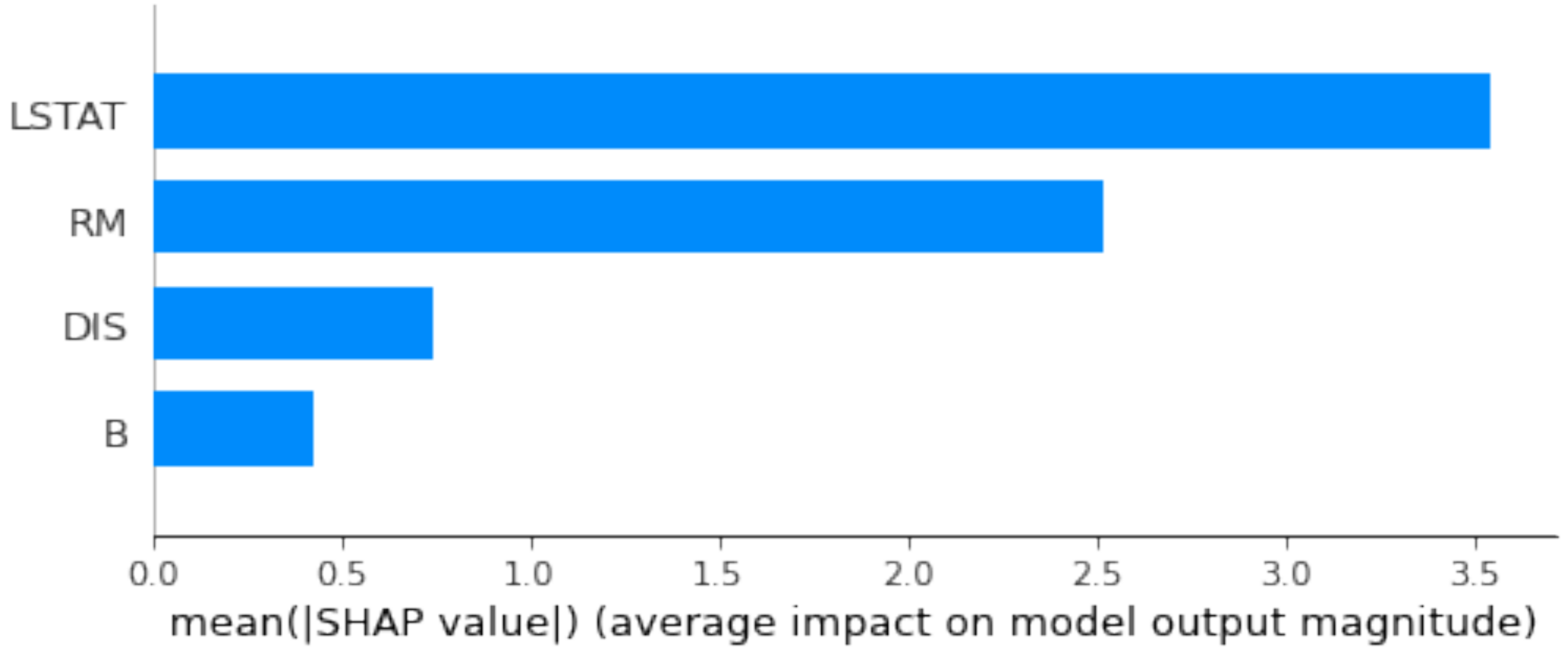
Different perturbation approaches SHAP



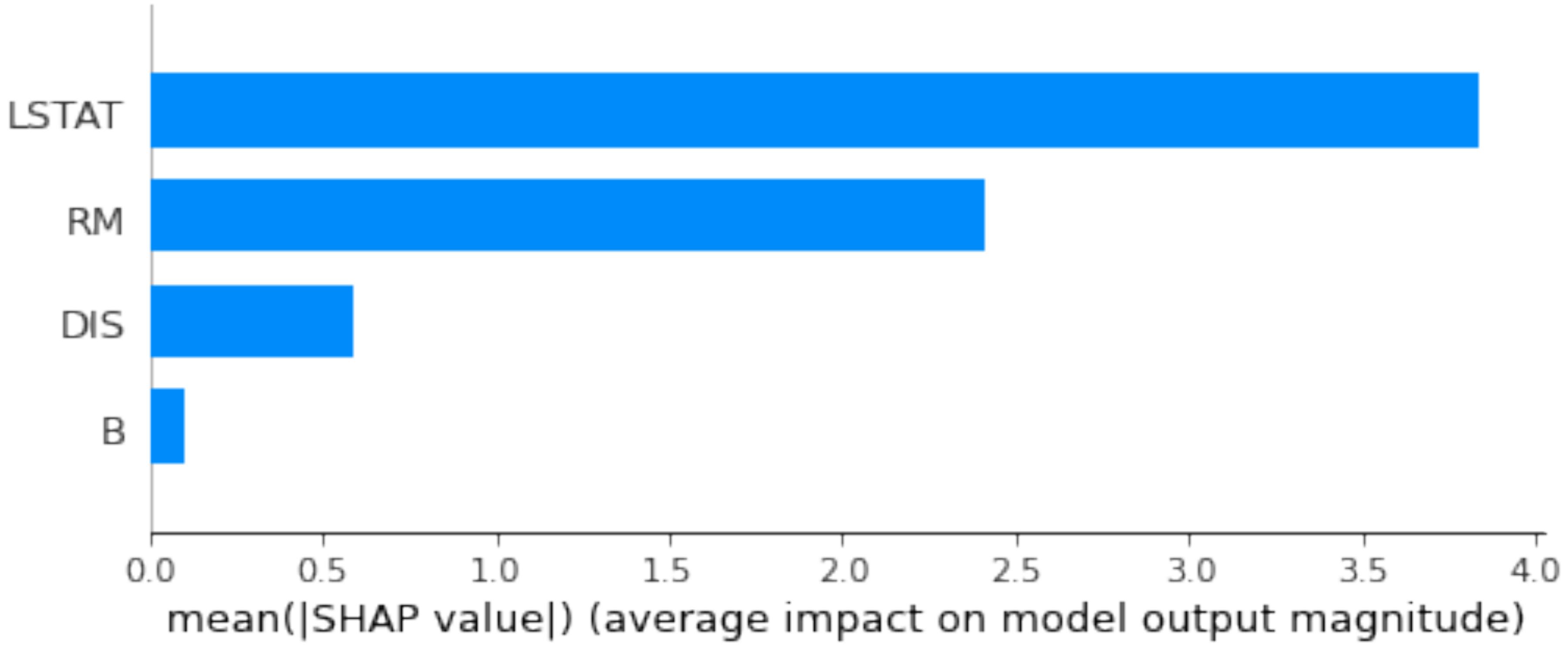
Biased



Original perturbation (fidelity: 0.68)



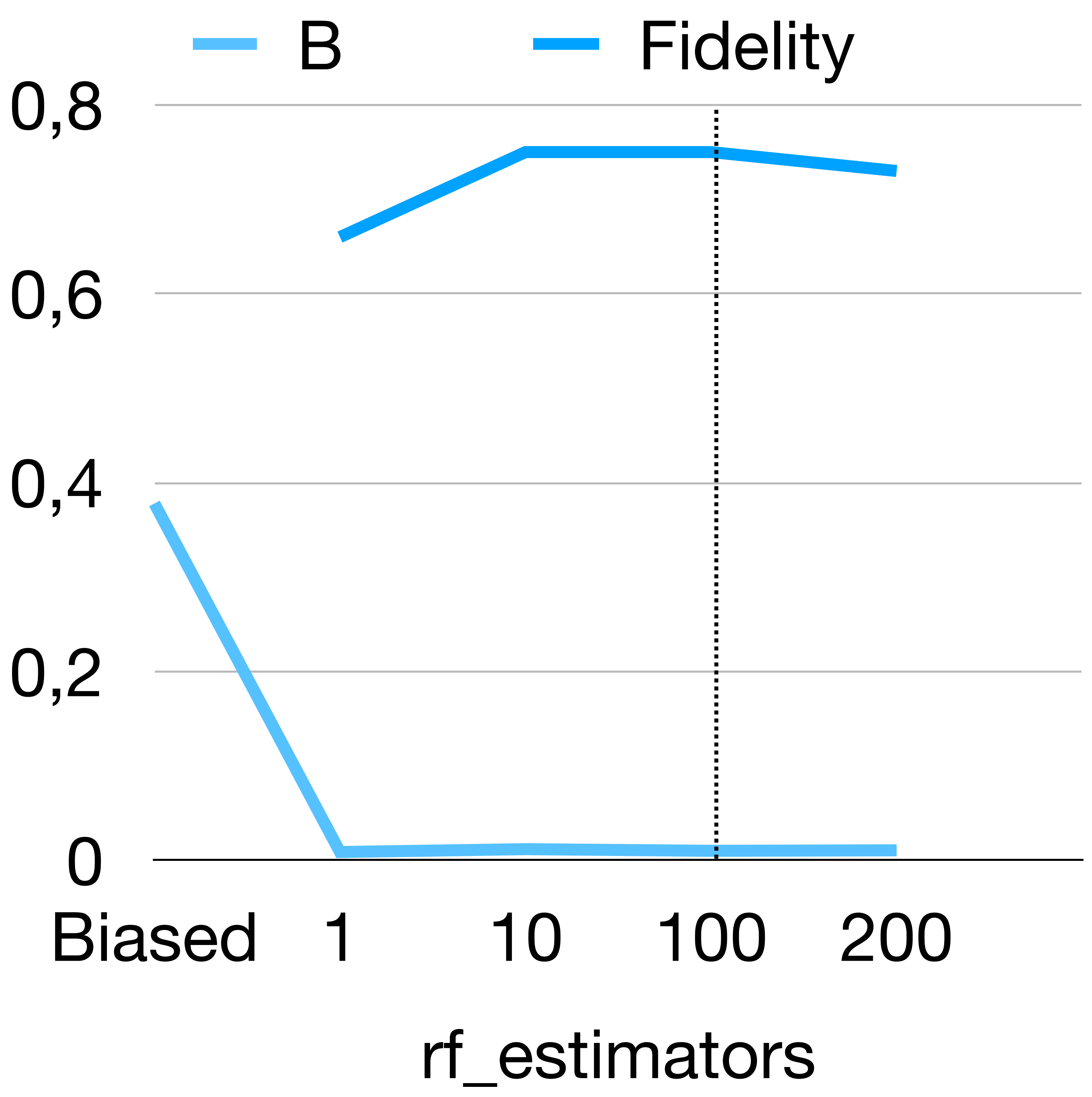
LIME perturbation (fidelity: 0.75)

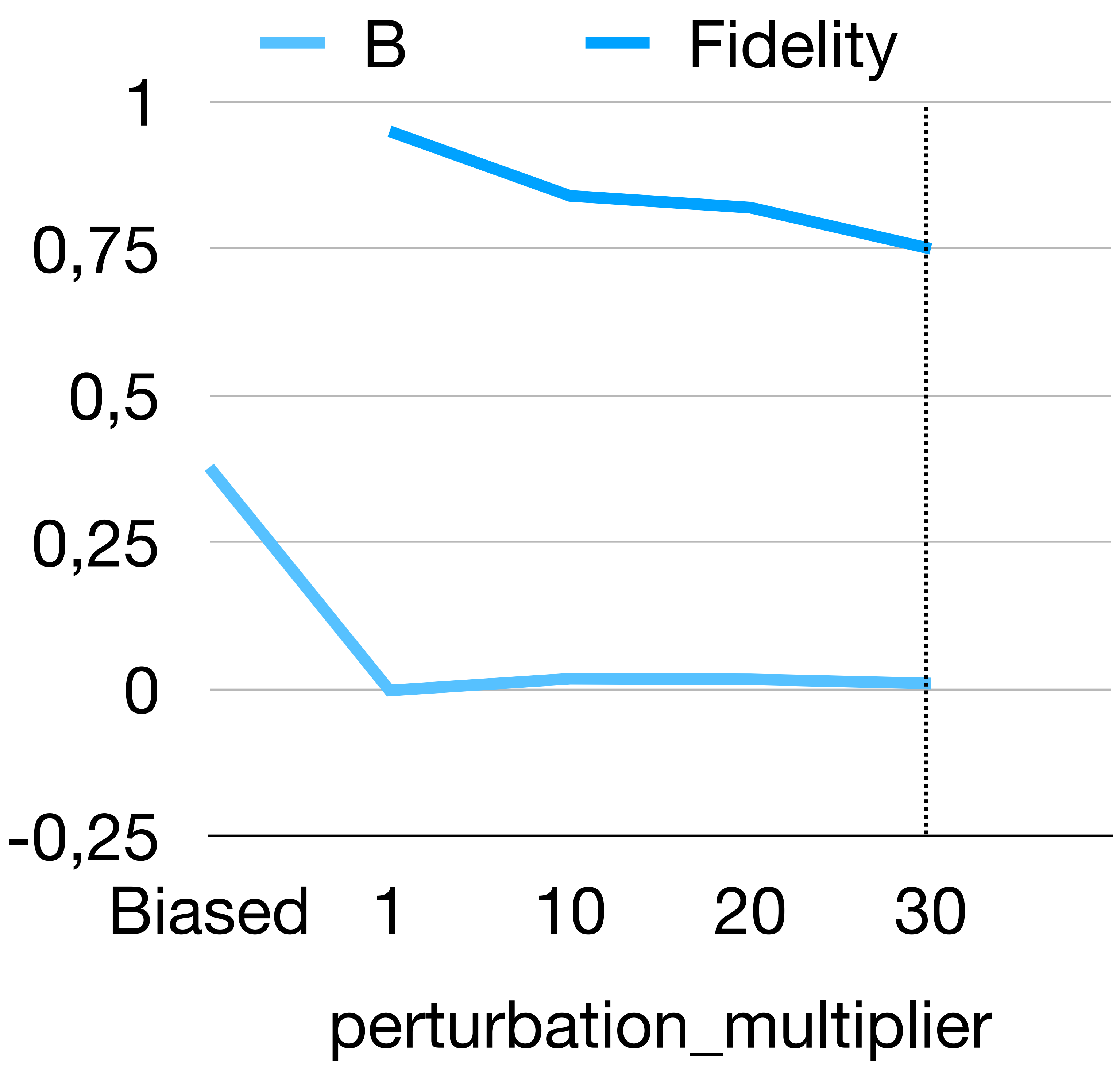


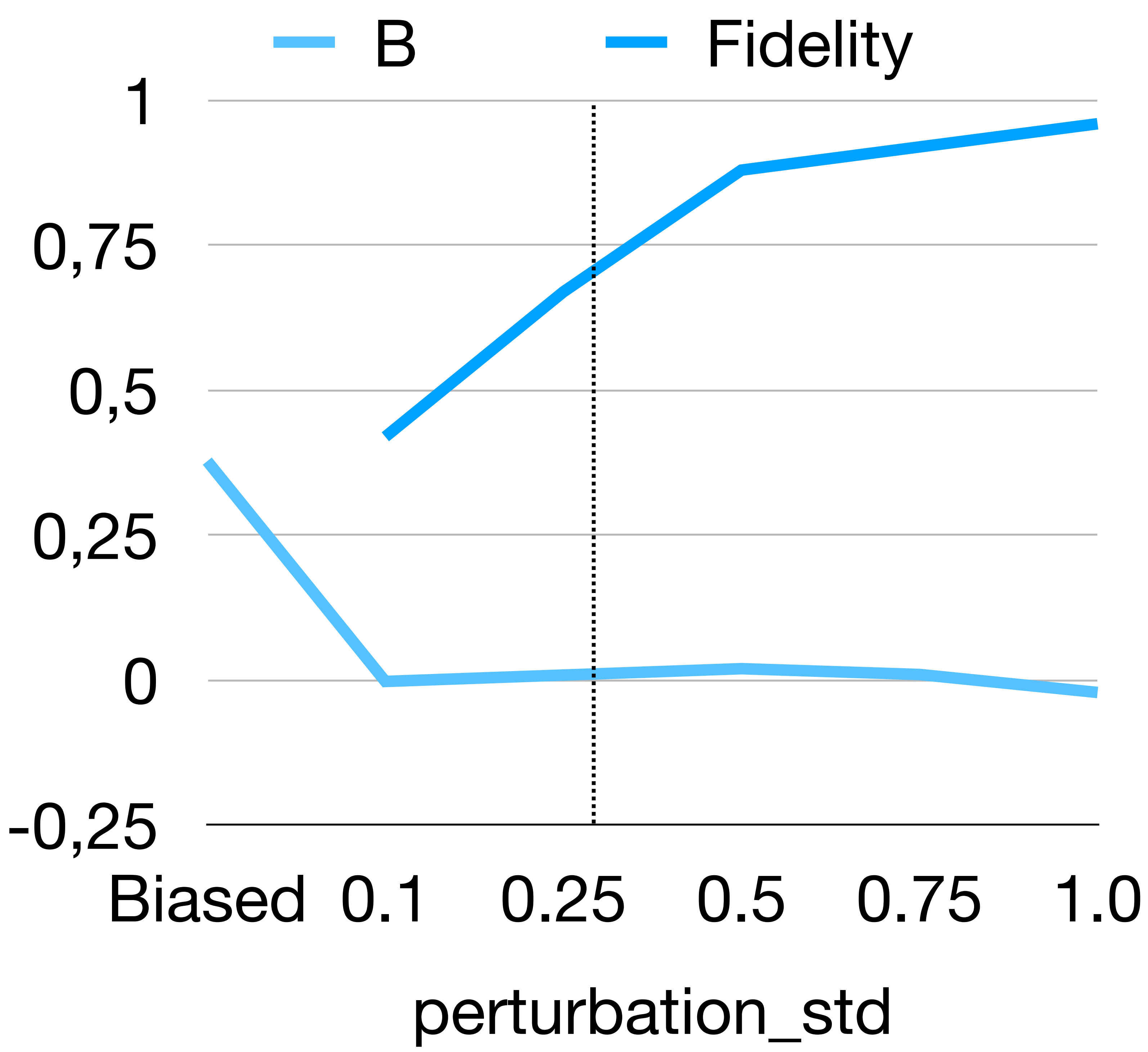
PDP perturbation (fidelity: 0.10)

Hyperparameter Sensitivity

LIME

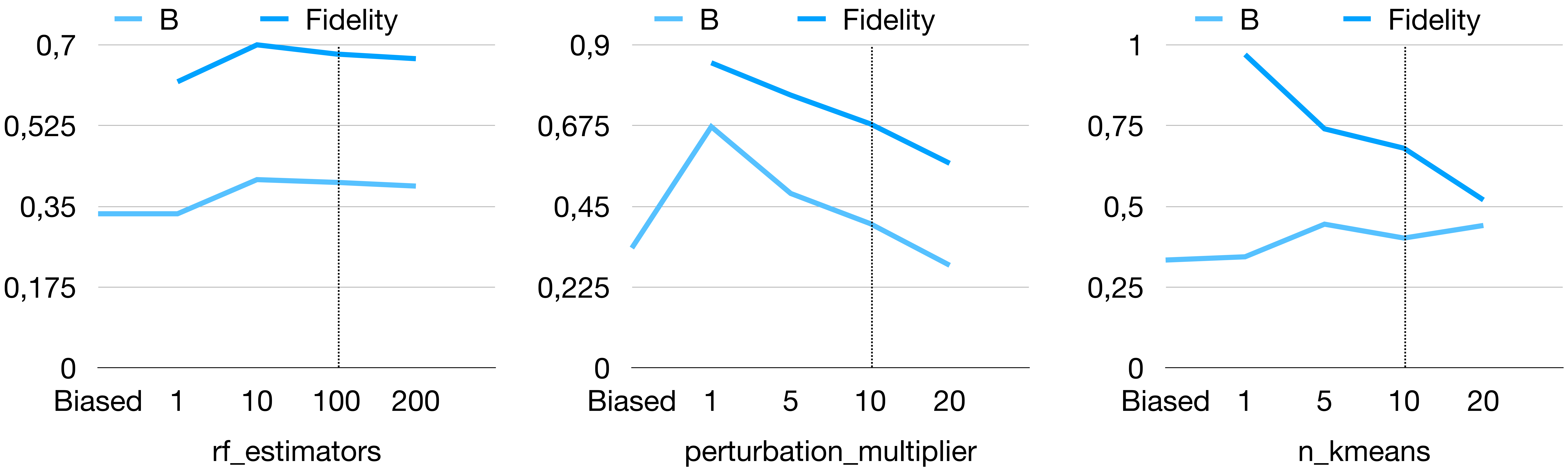






Hyperparameter Sensitivity

SHAP



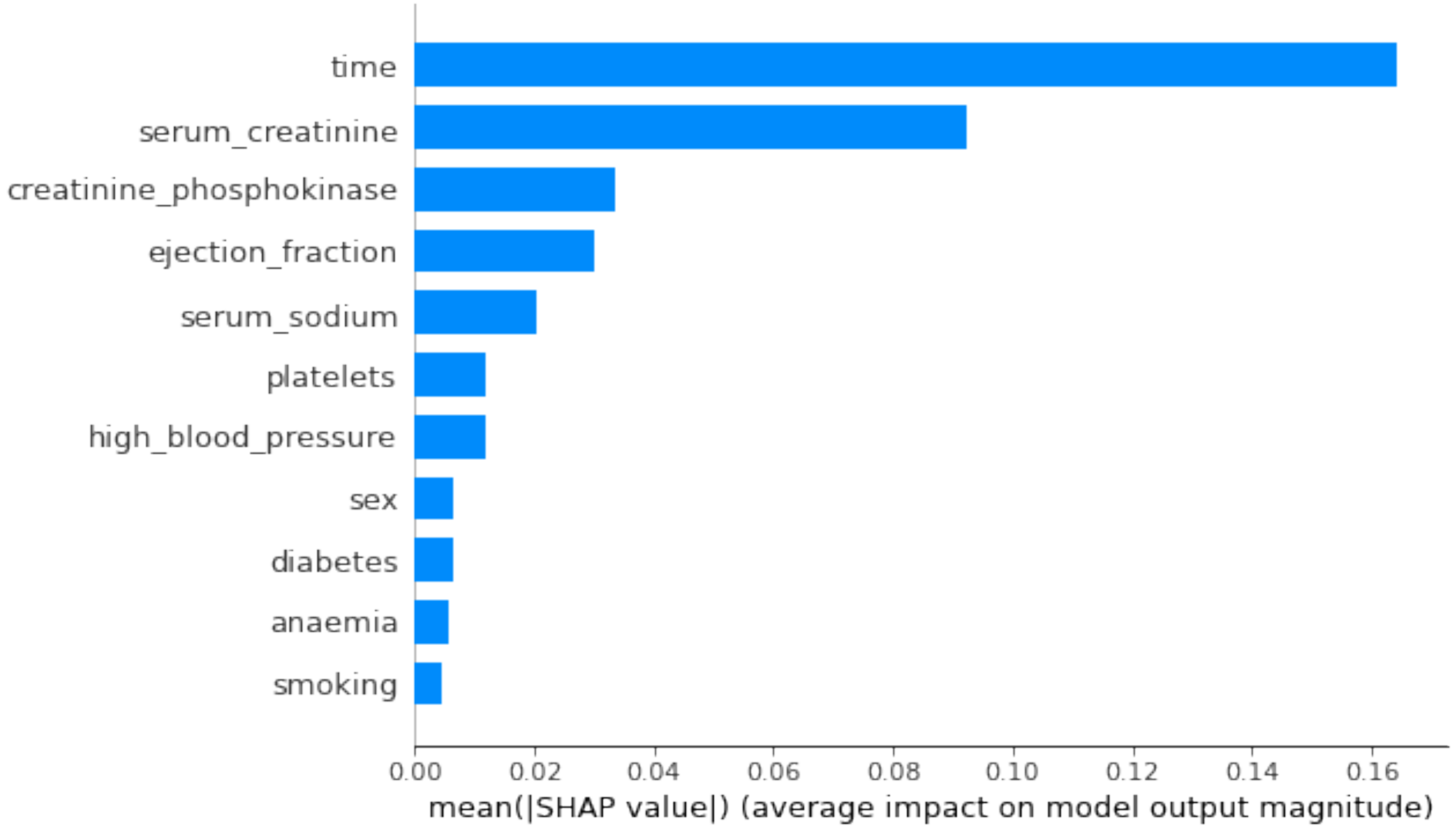
New Datasets

- Gender discrimination dataset
  - All female doctors at Houston College of Medicine who claimed that the College has engaged in a pattern and practice of discrimination against women in giving promotions and setting salaries.
- Heart failure prediction dataset
  - Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide.

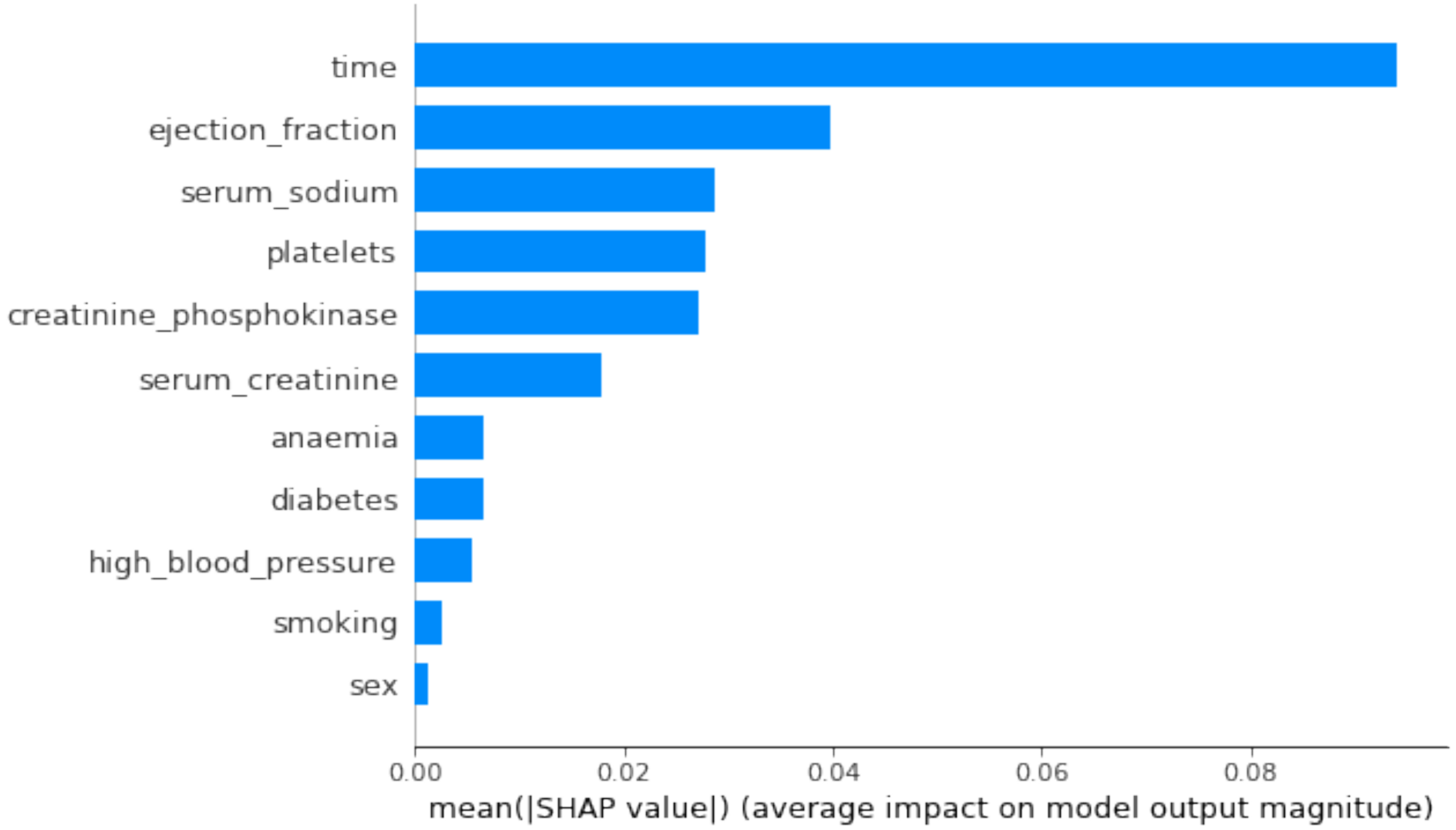
Gender discrimination dataset

Exper
Cert=1
Dept=1
Rank=1
Clin=1
Gender=1
Rank=2
Gender=0
Clin=0
Cert=0
Rank=0
Dept=0
Prate
-40000
-30000
-20000
-10000
0
10000
20000
30000
Biased
Adversarial (fidelity: 0.58)

Heart failure prediction dataset



Biased model



Adversarial model (fidelity: 0.92)

Thanks for the attention

GitHub