TowerDebias

A Novel Debiasing Method Based on the Tower Property

By: Aditya Mittal and Norm Matloff

Department of Statistics & Computer Science

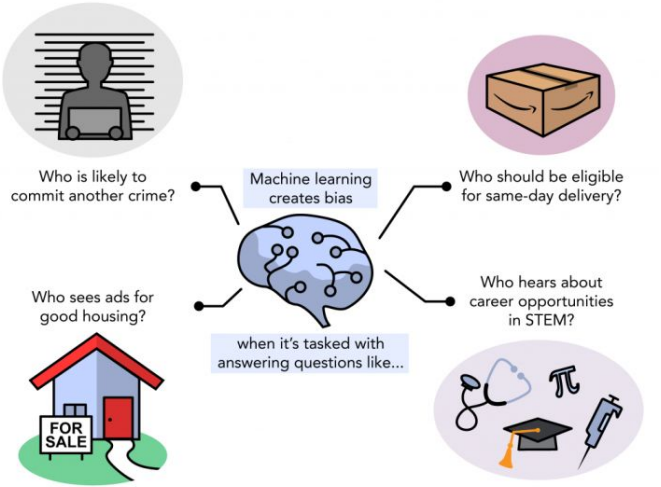
Agenda

1. Background Information
2. Example of Related Work
3. Conceptual Theory behind TowerDebias
4. Empirical Analysis
5. Discussion

Introduction

TowerDebias: An application to enhancing *Fairness in Machine Learning Algorithms*

- Machine learning has become more relevant in business, healthcare, legal systems, etc.
  - ➢ Example of *algorithmic bias*: **COMPAS**
- Eliminate the influence of **sensitive variables [S]** on predictions produced by black-box machine learning models.
  - ➢ Examples of [S]: Race, Gender, Age
- **Fairness-Utility Tradeoff:** balance between fairness and predictive accuracy.



Case Study: COMPAS

ProPublica vs Northpointe

- **Context**: Using machine learning algorithms to predict the probability that a criminal will re-commit a crime in the future.
  - ‘Black-Box’ algorithms aid judges in decision-making processes.
  - Extremely impactful on the lives of defendants.

- Northpointe’s development of the **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) algorithm.

- Faced scrutiny under Propublica's analysis which alleged **algorithmic bias** against black defendants relative to comparable whites.
  - Northpointe refutes Pro-Publica’s assertion; ProPublica maintains its position using statistical analysis.

Measuring Fairness

Avenues to enhance fairness:

- **Pre-processing:** Process the data itself *before* training an algorithm to reduce the bias associated with sensitive features.
- **In-processing:** Implementing fairness to the design and algorithm *during* training of models to induce fairness.
- **Post-processing:** Modification of the model's predictions *after* it has been trained.

Proposed Fairness Metrics:

- **Statistical Parity:** Requires an equal likelihood for individuals in both marginalized and non-marginalized groups to be assigned to the positive class.
- **Equalized Odds:** Requires that the protected and unprotected groups have equal true and equal false positive rates.
- **Correlation Coefficient:** Correlation between predicted response [Y] and sensitive variable [S].

= Corr(P(Defendant re-commits crime| X), Race)

Related Work

fairml: A Statistician's Take on Fair Machine Learning Modelling

fairml: A Statistician's Take on Fair Machine Learning Modelling (by Marco Scutari)

Statistical approach to developing fair machine learning models via the **FairML** package.

- Penalizing the weights of sensitive variables to reduce their predictive power and create interpretable, fair models.
- Incorporates an unfairness parameter, where 0 signifies perfect fairness and 1 indicates no fairness constraints.

The package includes several functions for regression and classification settings. It makes sense to apply towerDebias to these algorithms! *(We set the unfairness parameter to 0.2)*

Tower Property

Tower Property of Conditional Expectation: **$E(Y \mid X) = E[E(Y \mid X, S) \mid X]$**

The conditional expectation of Y given X is equivalent to the conditional expectation of Y given S and X both, conditioned solely on X.

Example: We are predicting probability of recidivism (re-commit a crime) based on 5 prior crimes and race.

Avg (Recidivism| Prior Crimes = 5) = Avg [Avg (Recidivism|Race, Prior Crimes = 5) | Prior Crimes = 5]

- Compute average conditional probability of recidivism based on race and number of prior crimes within inner expectation
- Condition probability of recidivism again with prior crimes = 5 on the outer expectation
- Updated prediction solely conditioned on prior crimes, **eliminating the effect of race**
- Represents average probability of recidivism of individuals at population level with prior crimes = 5

Relation to towerDebias: Motivating Example

Defendant Profile

Name: John Doe

Age: 18

Number of Prior Crimes: 5

Race: White

We want to predict the probability that John Doe will recommit a crime, while making sure we don’t factor in his race into our decision making.

Relation to TowerDebias (Example Continued)

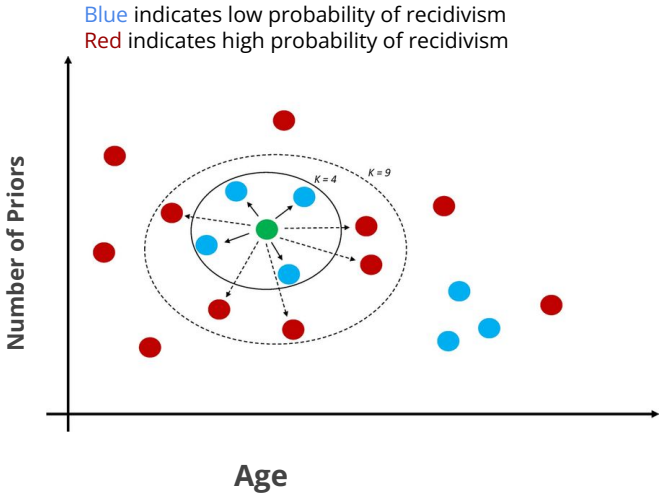
towerDebias: In a sample, we may not any have individuals of exactly age 18 and 5 crimes. So, we average the probabilities of recidivism for individuals whose committed number of crimes is NEAR 5.

The parameter **k** is the number of nearest neighboring *rows* to compute the average probability of recidivism with.

1. Small **k:** May lead to an overly narrow selection that might not cause significant reduction in the correlation.

2. Larger **k:** May include rows that are too distant and not representative of the new data point being debiased.

This choice of k is quite crucial to *minimizing* the extent of influence of the sensitive variable.



Empirical Study

Application of towerDebias on *traditional* ML algorithms: Linear/Logistic Regression, K-Nearest Neighbors, XGBoost, Neural Network.

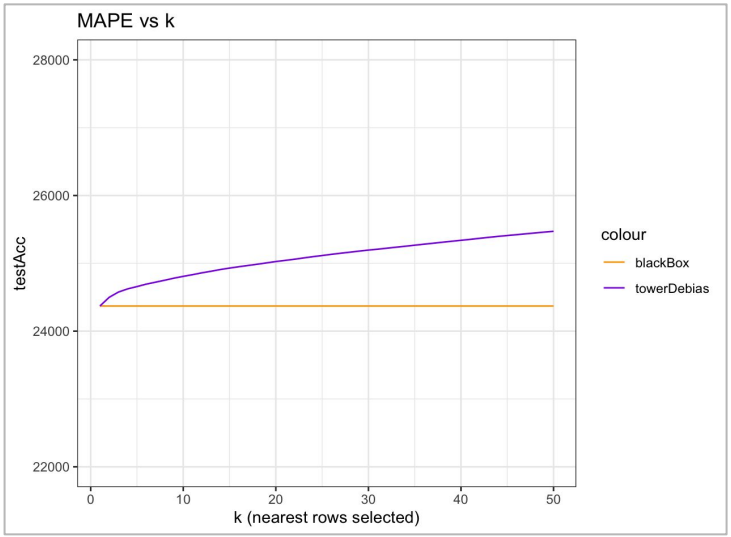
Application of towerDebias on *fairML* algorithms: FRRM, FGRRM, ZLM, ZLRM

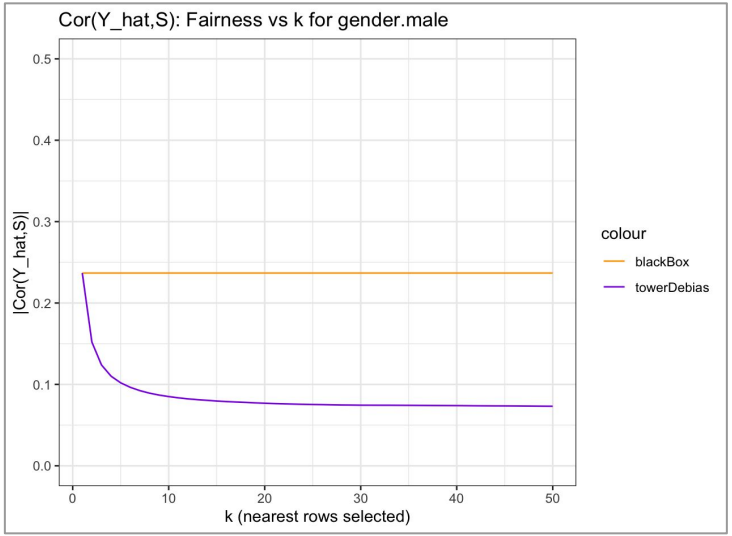
Data Name	**Response Variable**	**Sensitive Variable**	**Type**
Svcensus	Wage Income	Gender	*Regression*
Law schools admissions	LSAT Score	Race	*Regression*
Compas	Two-year Recidivism	Race	*Classification*
Iranian Churn	Exited	Gender, Age	*Classification*
Dutch Census	Occupation	Gender	*Classification*

Svcensus data

Subset of US census data from back in early 2000, focusing on six different engineering occupations. The goal is predict the **income** [Y] of a person, with **gender** as the sensitive variable [S].

Neural Network vs. towerDebias

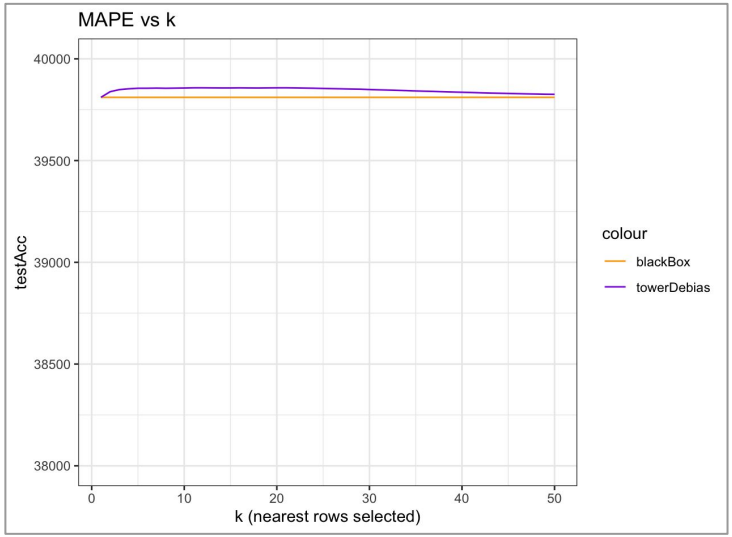


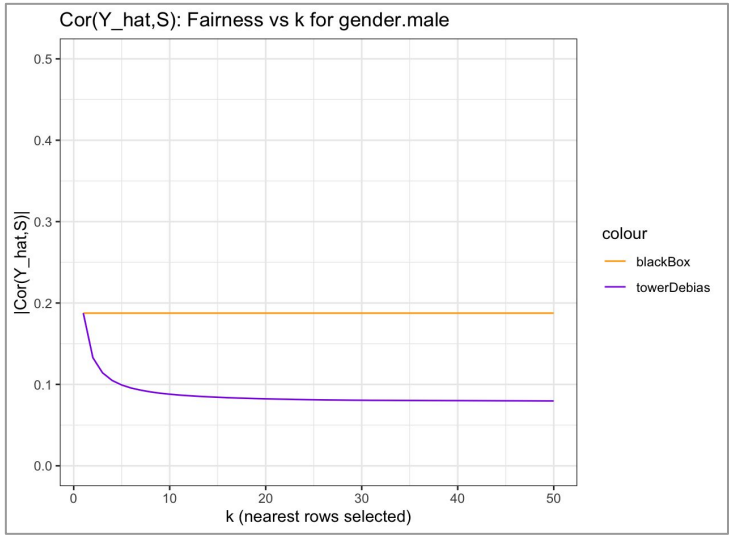


Svcensus data

Subset of US census data from back in early 2000, focusing on six different engineering occupations. The goal is predict the **income** [Y] of a person, with **gender** as the sensitive variable [S].

Fair Ridge Regression vs towerDebias

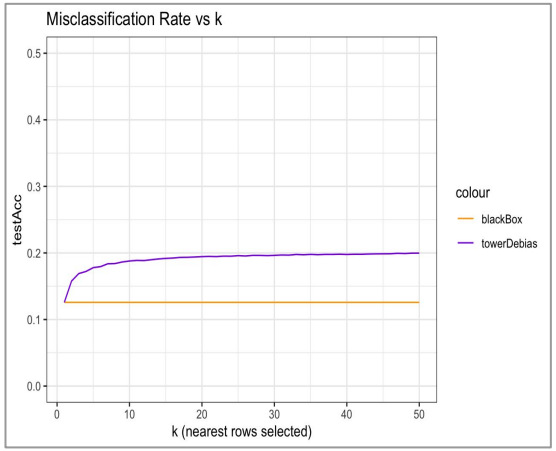


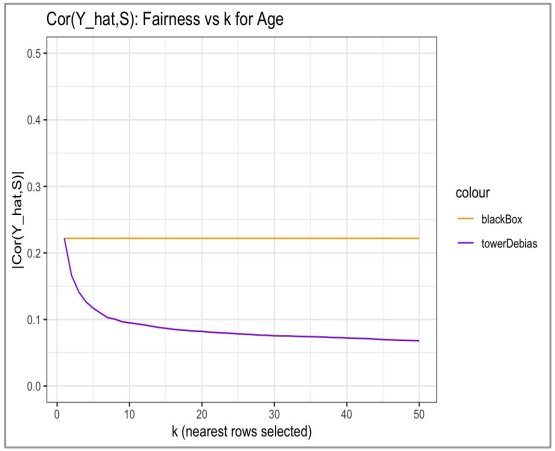


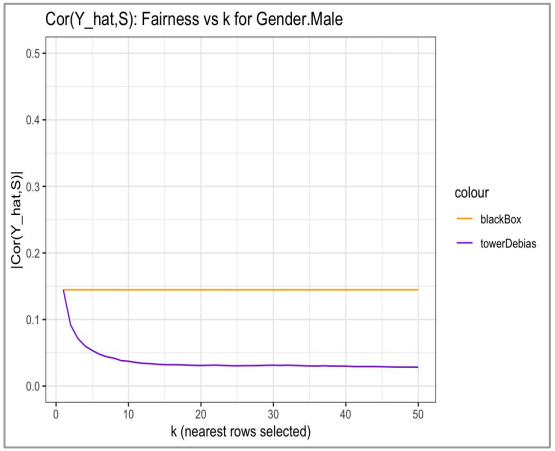
Iranian Churn data

The Iranian Churn dataset is used to predict the discontinuation of a customer's relationship with a company using **Exited** as the response variable [Y] with respect to **Gender** and **Age** as the sensitive variables [S].

Neural Network vs. towerDebias



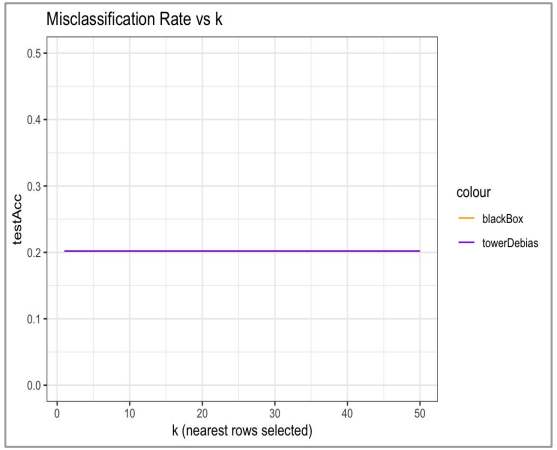


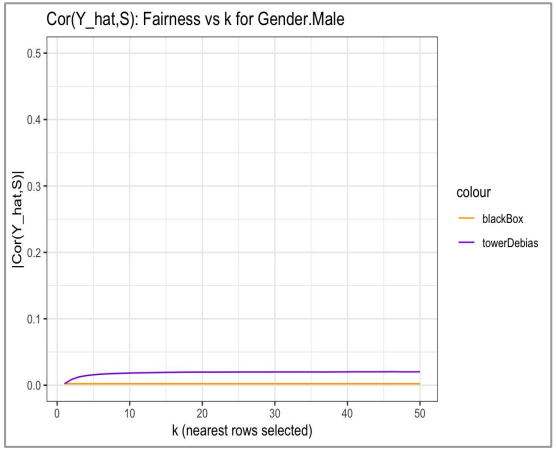


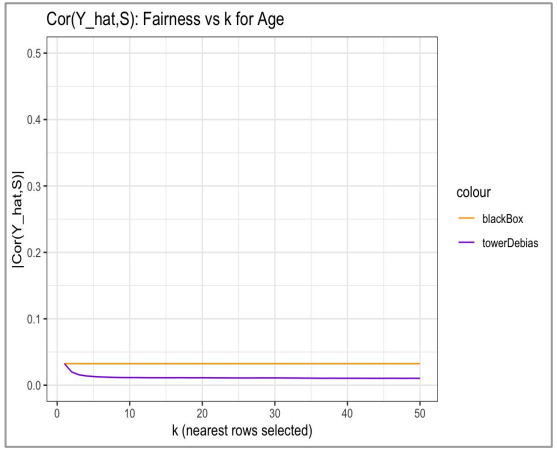
Iranian Churn data

The Iranian Churn dataset is used to predict the discontinuation of a customer's relationship with a company using **Exited** as the response variable [Y] with respect to **Gender** and **Age** as the sensitive variables [S].

FairML vs towerDebias



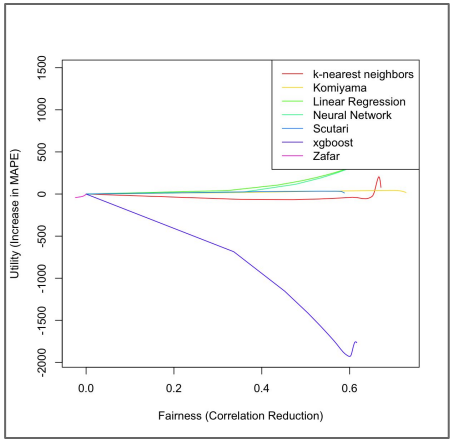


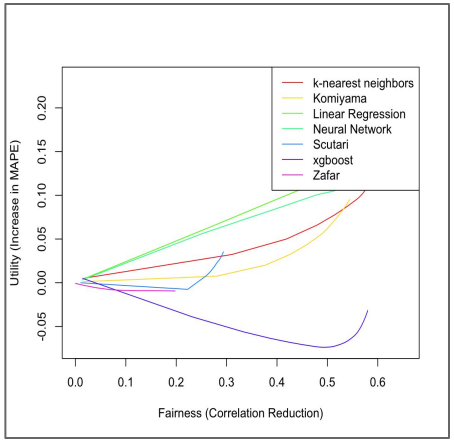


Fairness vs. Utility Graphs

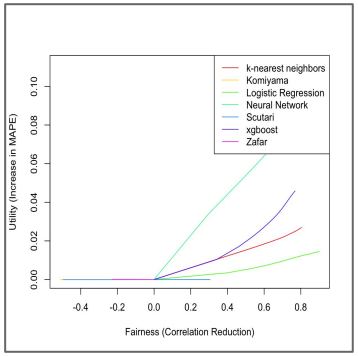
Another perspective into our results...

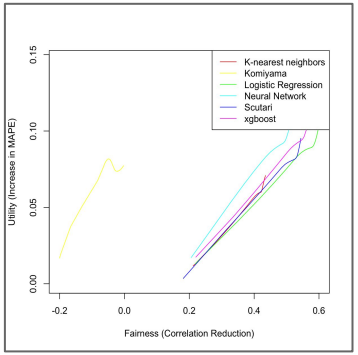
towerDebias on **Regression** datasets

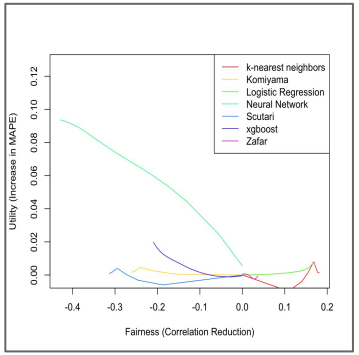




towerDebias on **Classification** datasets







Discussion

- Fairness in Machine Learning is an increasingly growing and important topic, especially with the application of extremely complex AI algorithms throughout different sectors.
- **towerDebias**: Utilize the Tower Property to enhance fairness during post-processing in Machine Learning.
- Provides a convenient framework to improve fairness across various different applications.
- Important to weigh trade-offs between fairness and utility in decision-making processes.

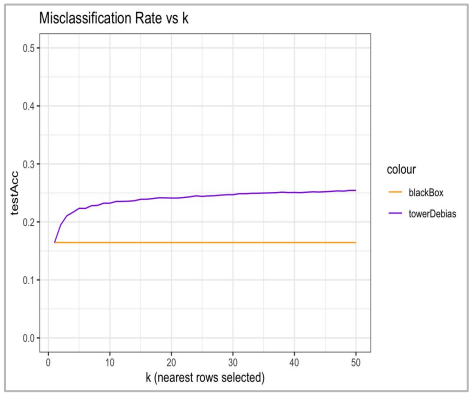
Thank You!

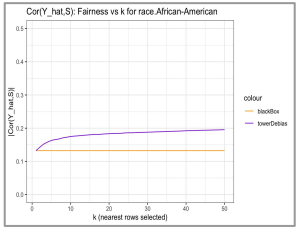
Questions?

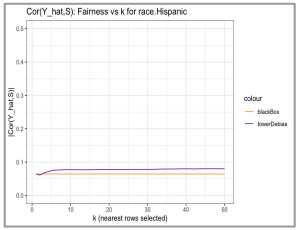
Compas data

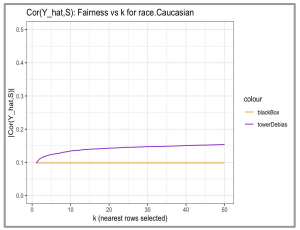
A collection of criminal offenders screened in Florida (US) during 2013-14. The goal is predict whether a defendant is a **recidivist** [Y], with **race** as the sensitive variable [S].

Neural Network vs. towerDebias









FairML vs towerDebias