ROBUST TEXT CLASSIFICATION IN THE PRESENCE OF CONFOUNDING BIAS

Virgile Landeiro & Aron Culotta
Illinois Institute of Technology
Chicago

INTRODUCTION

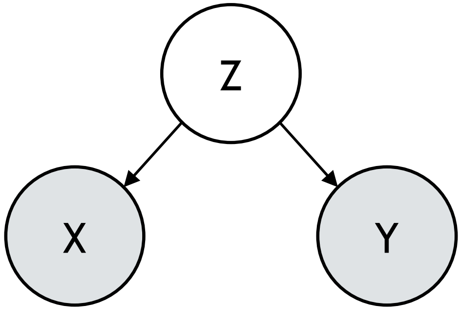
- Development of text classification over more than 50 years;
- Mostly centered around categorization of documents into topics;
- New areas of research:
  - Public health surveillance;
  - Political science;
  - Marketing;
  - …
- But algorithms stay the same: standard supervised classification algorithms.
- To ensure validity of study ➔ need classifiers robust to confounding variables.

nyc	angeles	ny	york	california
los	la	brooklyn	snow	disneyland
jersey	city	san	ca	hollywood
monica	santa	nj	manhattan	losangeles
earthquake	team	dodgers	hills	cute
heart	vegas	chill	state	happiness
makeup	pacific	cali	father	brother
also	guess	socal	field	job
cant	venice	tacos	boo	wonderful
laugh	train	single	wanna	brothers

50 TOP FEATURES FOR LOGISTIC REGRESSION

Male (resp. Female) and New York (resp. Los Angeles) are highly correlated.

WHAT IS A CONFOUNDING VARIABLE?



Graphical model: a confounding variable Z correlated with both X and Y.

- Prediction vs. causal inference.
- Assume same impact in training and testing sets.

- Small training datasets;
- Confounder shifts over time.

RELATED WORK

- Matching
- Stratification
- Features removal
- **J. Pearl developed the back-door adjustment**

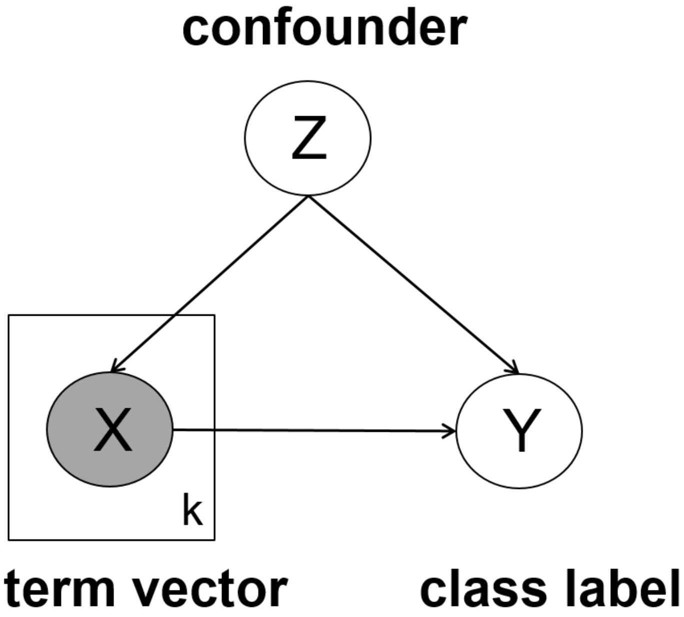
- $P_{train}(X) \neq P_{test}(X)$
- $P_{train}(Y) \neq P_{test}(Y)$

- We focus on:

$$P_{train}(Y|Z) \neq P_{test}(Y|Z)$$

BACK-DOOR ADJUSTMENT FOR TEXT CLASSIFICATION

- $D = \{(\boldsymbol{x_i}, y_i, z_i)\}_{i=1}^{n}$
- The back-door criterion requires that:
  - No node in Z is a descendant of X;
  - Z blocks every path between X and Y that contains an arrow pointing to X.
- The back-door criterion is met:



$$p(y|do(\boldsymbol{x})) = \sum_{z \in Z} p(y|\boldsymbol{x}, z) \times p(z)$$

BACK-DOOR ADJUSTMENT FOR TEXT CLASSIFICATION

$$p(y|do(\boldsymbol{x})) = \sum_{z \in Z} p(y|\boldsymbol{x}, z) \times p(z)$$

- Restrict to binary variables.
- Fit a logistic regression model on $p(y|\boldsymbol{x}, z)$ at training time by appending two features $c_{i,0}$ and $c_{i,1}$ to every $\boldsymbol{x_i}$.
- Z is not observed at testing time.

$\boldsymbol{x_0}$	$\boldsymbol{x_1}$	$\boldsymbol{c_0}$	$\boldsymbol{c_1}$	$\boldsymbol{z}$
0	0	0	1	1
0	1	0	1	1
1	0	1	0	0
1	1	0	1	1

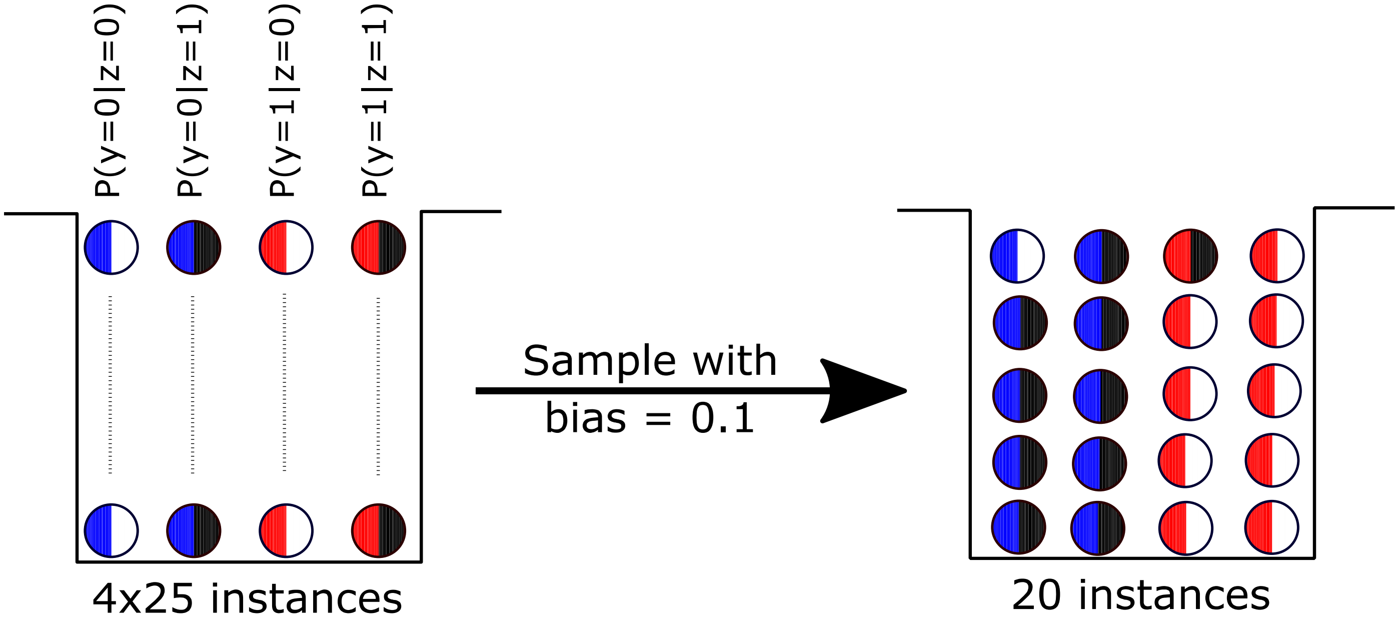
Dataset	Target variable	Confounder
Twitter	Location of a user: New York City or Los Angeles	Gender of the user: Male or Female
IMDb	Sentiment of the review: Positive or Negative	Genre of the film: Horror or Other
Canadian Parliament	Political affiliation: Liberal or Conservative	Political position: Government or Opposition

DATASETS

3 differents datasets to experiment with back-door adjustment.

INJECTING CONFOUNDING BIAS

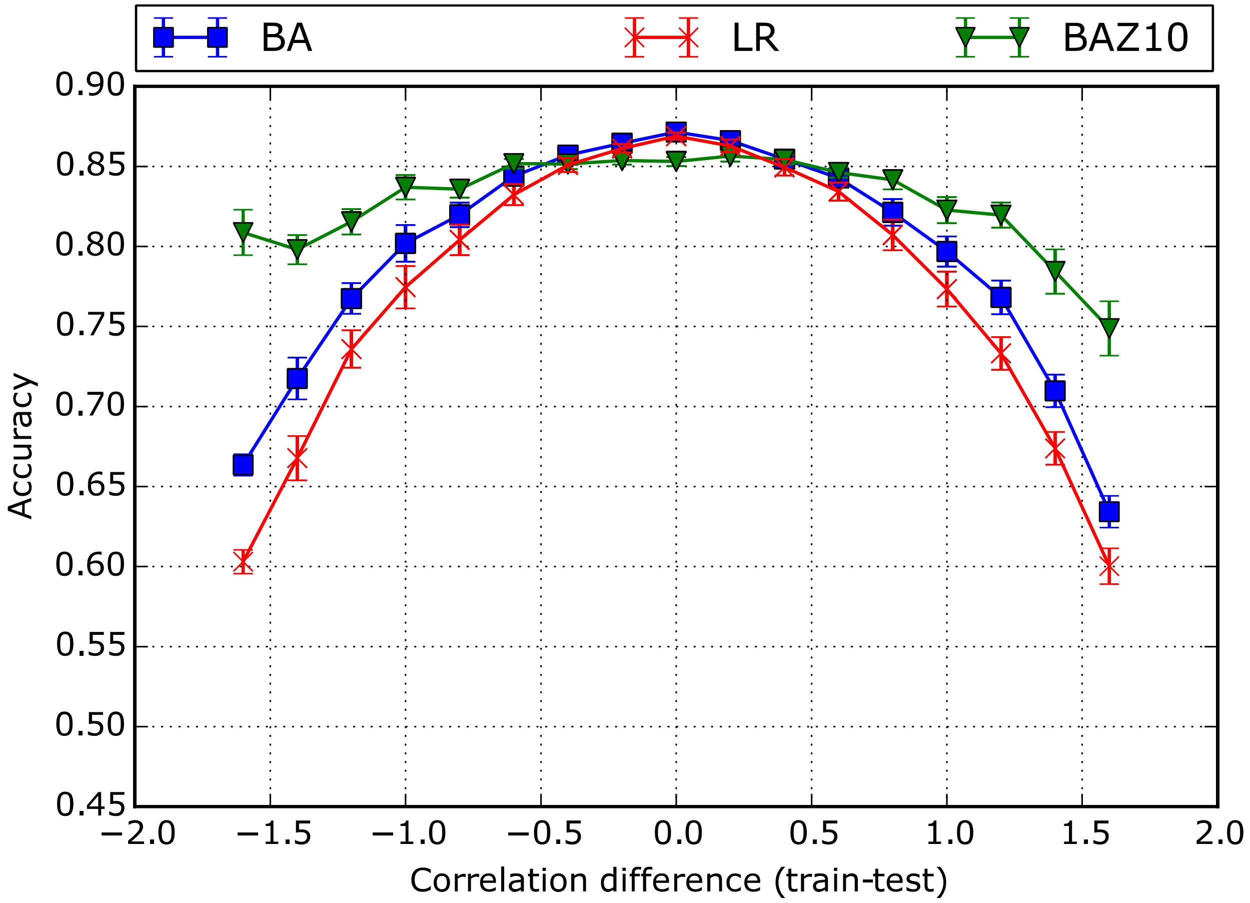
- Introduce confounding bias according to the following constraints:
  - $P_{train}(y = 1|z = 1) = b_{train}$
  - $P_{test}(y = 1\ |z = 1) = b_{test}$
  - $P_{train}(Y) = P_{test}(Y)$
  - $P_{train}(Z) = P_{test}(Z)$



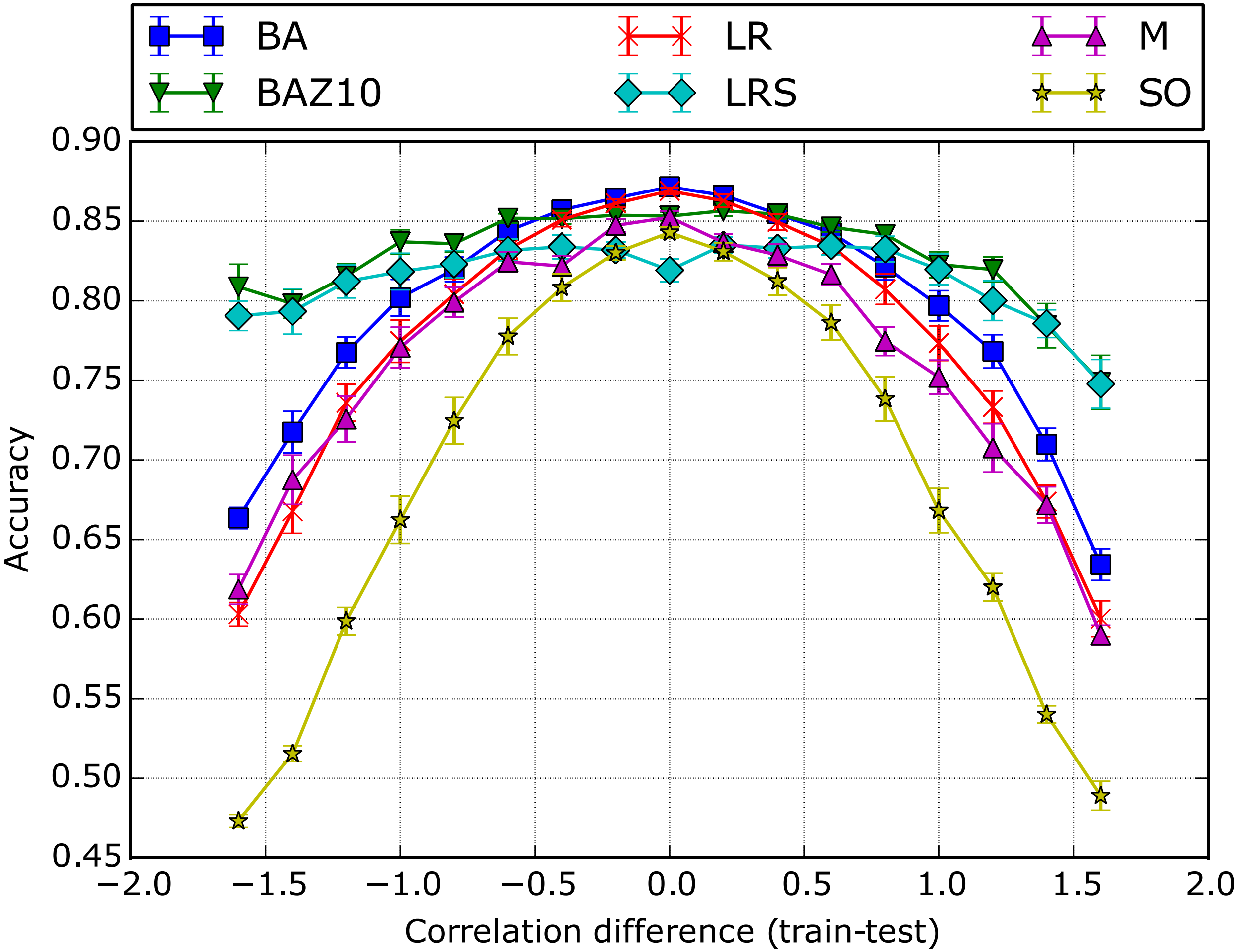
BASELINES

- Logistic Regression (LR)
- Back-door Adjustment (BA and BAZ10)
- Subsampling (S)
- Matching (M)
- Sum Out (S)

RESULTS FOR THE TWITTER DATASET

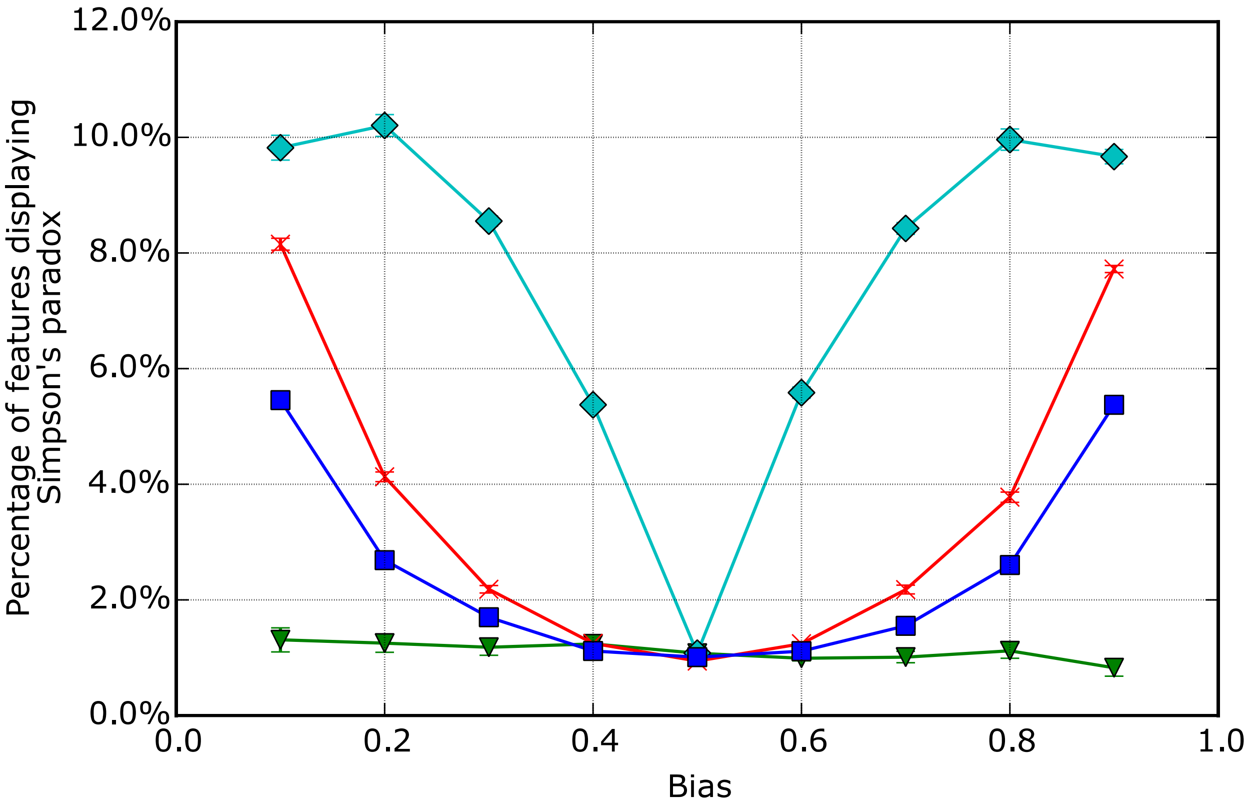


RESULTS FOR THE TWITTER DATASET



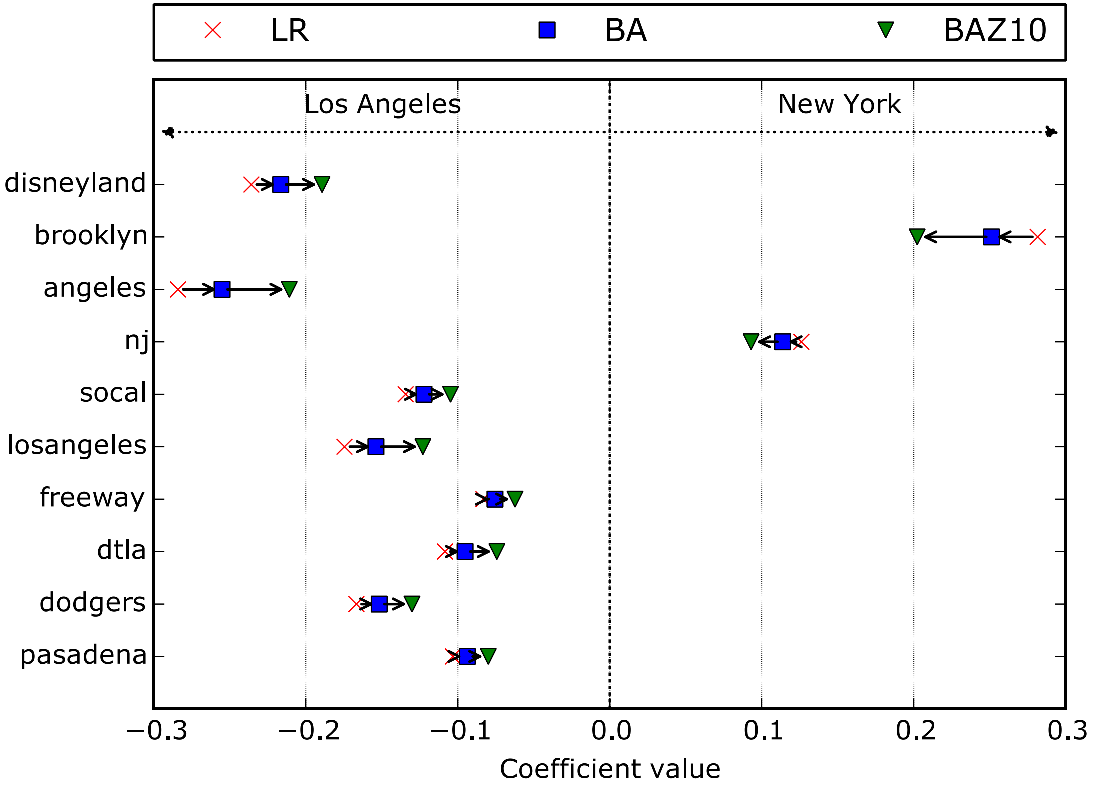
EFFECTS OF BACK-DOOR ADJUSTMENT

- Simpson's Paradox



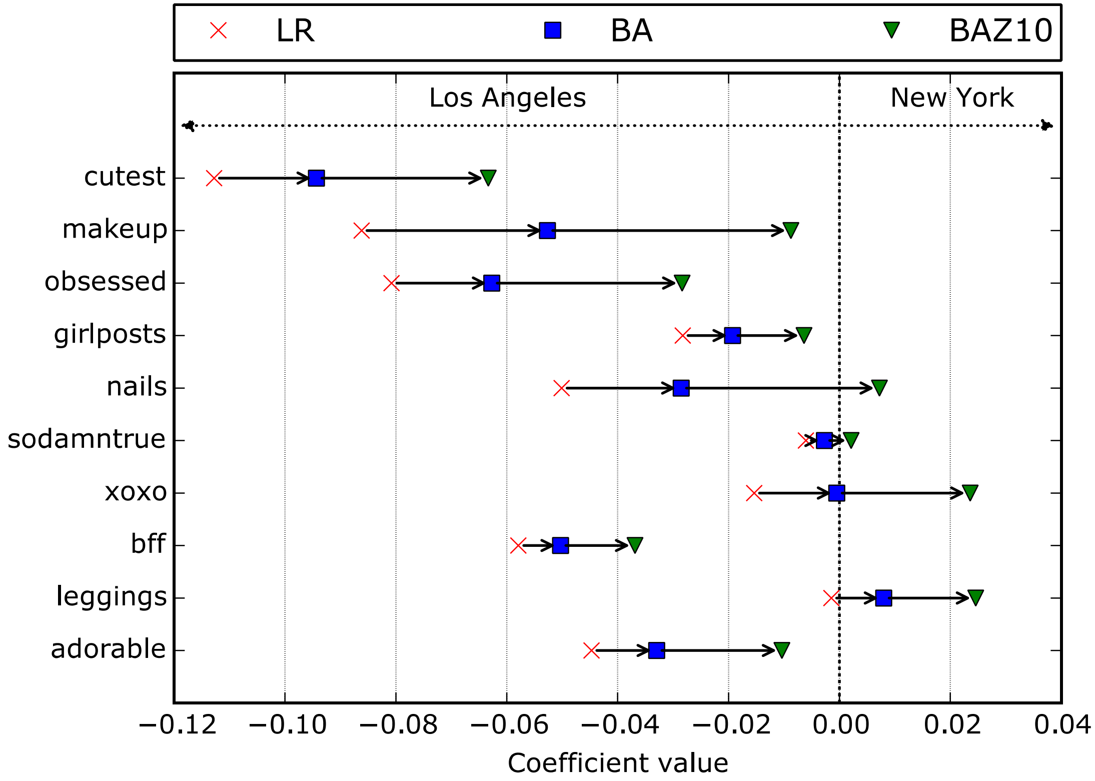
EFFECTS OF BACK-DOOR ADJUSTMENT

- Coefficients of features correlated with the classes



EFFECTS OF BACK-DOOR ADJUSTMENT

- Coefficients of features correlated with the confounders


× LR
■ BA
▼ BAZ10
Los Angeles
New York
cutest
makeup
obsessed
girlposts
nails
sodamntrue
xoxo
bff
leggings
adorable
Coefficient value