Predicting voting behavior using digital trace data

Online Supplement

Appendix A: Descriptive Statistics

Table S1. Summary Statistics for Outcome Variables

	n Yes	% Yes	n No	% No	N
Low income[1]	556	0.313	1220	0.687	1776
High income[2]	147	0.083	1629	0.917	1776
Under 25 y.	304	0.153	1687	0.847	1991
Over 60 y.	282	0.142	1709	0.858	1991
Male	901	0.453	1090	0.547	1991
East	377	0.19	1604	0.81	1981
Married	704	0.354	1287	0.646	1991
No partner	811	0.407	1180	0.593	1991
No children	1349	0.678	642	0.322	1991
Unemployed	112	0.056	1879	0.944	1991
Full-time emp.	927	0.466	1064	0.534	1991
Undecided	344	19.1	1455	80.9	1799
Voted	1589	92.7	126	7.3	1715
AfD	224	14.6	1307	85.4	1531

[1] Personal net income <= 1000 Euro.

[2] Personal net income >= 3000 Euro.

Table S1. Summary Statistics for Outcome Variables

	n Yes	% Yes	n No	% No	N
GREEN	121	7.9	1410	92.1	1531
CDU	365	23.8	1166	76.2	1531
SPD	329	21.5	1202	78.5	1531
Left	245	16	1286	84	1531
FDP	145	9.5	1386	90.5	1531

Appendix B: Overview of Feature Blocks

Table S2. Types of Digital Trace Data

General Use	News Media	Domains/Apps Commonly Visited
1. Device type (e.g., tablet) 2. Type of connection (e.g., wifi) 3. Device (brand & model) 4. Operating system and version (e.g., iOS 10.2.1) 5. Length of time, overall and by each of the above	1. Total time spent on news media, public-service broadcast and alternative news media websites 2. Proportion of time spent on news media, public-service broadcast and alternative news media websites 3. Total time spent using news media, public-service broadcast and apps 4. Proportion of time spent on news media, public-service broadcast and apps	1. Total time spent on each website and app (visited at least 80 times for a total time of 1 minute or more by at least 1% of the sample)

Appendix C: List of news domains or blogs labelled populist, propagandistic or “alternative/fake” news

anonymousnews.ru

rt.com

sputniknews.com

breitbart.com

compact-online.de

compact-shop.de

der-kleine-akif.de

tichyseinblick.de

jungefreiheit.de

pi-news.de

contra-magazin.com

freie-presse.net

unzensuriert.de

epochtimes.de

journalistenwatch.com

neopresse.com

freiewelt.net

deutschland-kurier.org

quotenqueen.wordpress.com

deutsch.rt.com

philosophia-perennis.com

unzensuriert.at

schluesselkindblog.com

blauenarzisse.de

sezession.de

morgengagazin.com

antaios.de

uncut-news.ch

propagandaschau.wordpress.com

anderweltonline.com

andreas-unterberger.at

bazonline.ch

bayernistfrei.com

conservo.wordpress.com

zuercherin.com

eike-klima-energie.eu

einprozent.de

eva-herman.net

einwanderungskritik.de

ef-magazin.de

frankjordanblog.wordpress.com

fdogblog.wordpress.com

freitum.de

de.gatestoneinstitute.org

geolitico.de

haunebu7.wordpress.com

identitaere-bewegung.de

kpkrause.de

kopp-report.de

korrektheiten.com

michael-mannheimer.net

mmnews.de

oliverjanich.de

sciencefiles.org

de.sputniknews.com

wissensmanufaktur.net

refcrime.info

rapefugees.net

kontrapunkt.social

terminegegenmerkel.wordpress.com

zuerst.de

bayernweit.blogspot.de

dushanwegner.com

dieunbestechlichen.com

de.europenews.dk

freiezeiten.net

guidograndt.de

danisch.de

indexexpurgatorius.wordpress.com

info-direkt.eu

inge09.blog

michaelgrandt.de

michael-klonovsky.de

moshpitscorner.wordpress.com

peymani.de

luegenpresse2.wordpress.com

lepenseur-lepenseur.blogspot.de

nachtgespraechblog.wordpress.com

nixgut.wordpress.com

nation24.de

pboehringer.de

politikstube.com

politikversagen.net

prabelsblog.de

preussische-allgemeine.de

publicomag.com

rundertischdgf.wordpress.com

signal-online.de

de.sott.net

Appendix D: XGBoost Tuning

The (lower-level) tree growing process and the (higher-level) ensemble building sequence in XGBoost is controlled by a number of tuning parameters. On the tree level, *max_depth* controls the size of the trees, i.e., the maximum number a given tree is allowed to be split. Tree size is further controlled by *gamma*, the minimum loss reduction required per split and by *min_child_weight*, the degree of impurity that precludes a node from been subject to further splitting. *lambda* (L2 regularization) and *alpha* (L1 regularization) control the amount of shrinkage that is applied to the scores of a given node. The next tuning parameter, *colsample_by_tree,* represents the ratio of columns that are sampled prior to building a tree. s*ubsample* controls sampling on the case level, i.e. the subsampling ratio for sampling training observations prior to tree building. The tree ensemble is built by growing *nrounds* of trees, whereas the learning rate *eta* determines the shrinkage by which the prediction scores are scaled after growing an individual tree in the boosting sequence.

In order to tune these parameters efficiently, we implemented a two-step approach by first considering a set of try-out values for the tree level parameters while fixing *nrounds* and *eta* such that only a small ensemble is built. In the second step, we scaled up *nrounds* and *eta* while focusing on the respective *max_depth*, *gamma*, *min_child_weight* and *alpha* values that performed best in the first step. The tuning grids of both steps are summarized in Table S1.

Note that since XGBoost models are tedious to tune, we also modeled our outcomes using random forests as a robustness check (Breiman 2001, Wright & Ziegler 2017). The results were similar. Due to space constraints, we only discuss the XGBoost results in the paper.

Table S1. XGBoost Tuning Grids

Parameter	First grid	Second grid
nrounds	100	250, 500, 750, 1000
eta	0.05	0.025, 0.01
gamma	0, 0.5, 1	best
max_depth	1, 3, 5, 7, 9, 11	best-1, best, best+1
min_child_weight	0, 1, 2, 3, 4, 5	best
subsample	1	0.7, 1
colsample_by_tree	1	0.7, 1
lambda	1	1
alpha	0, 0.5	best

References

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5–32.

Wright, M., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software, 77,* 1–17.