

Conflict of Interest

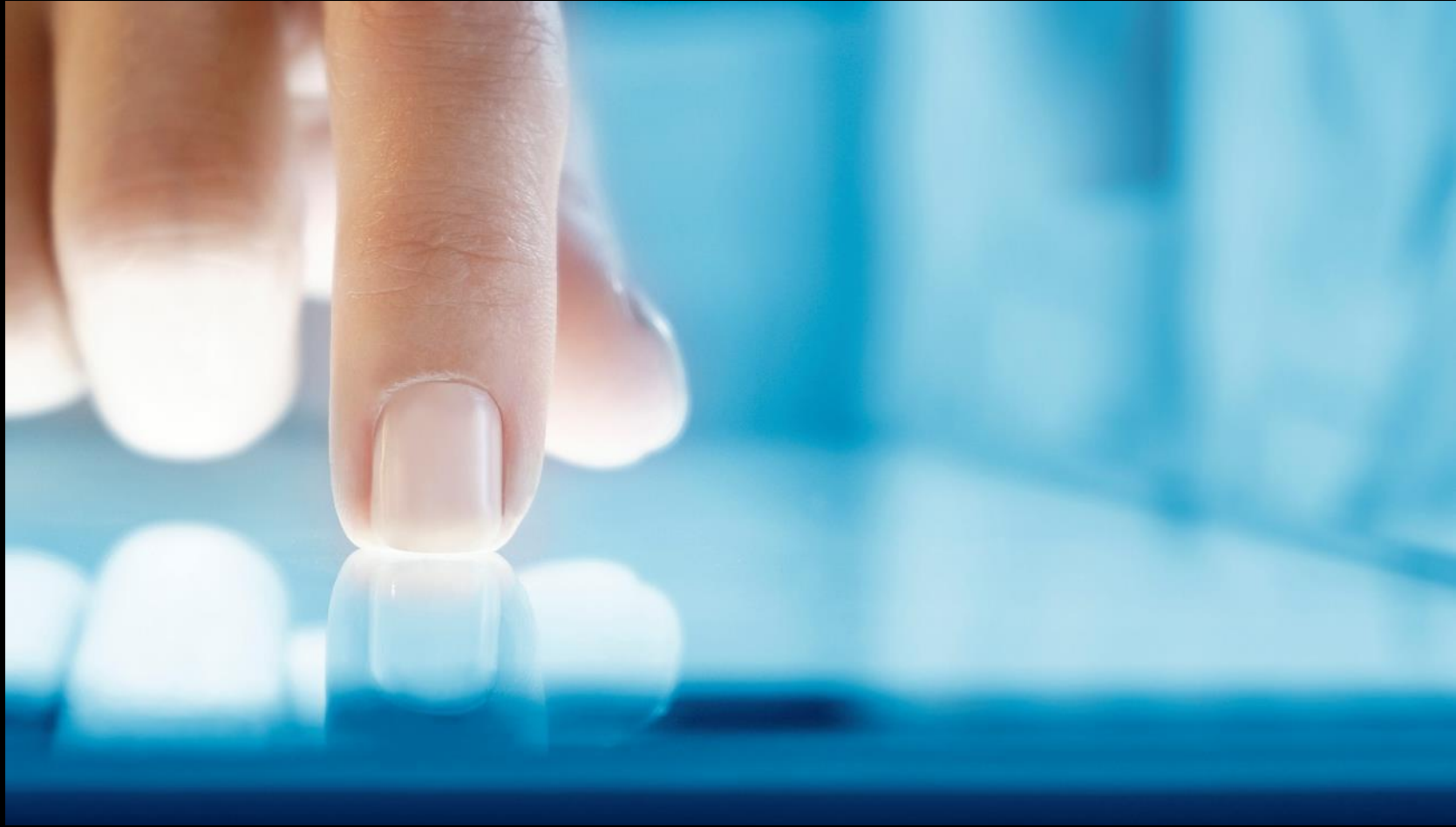
Vicky W. Li, Halley Costantino, John C. Rowland, Leiyu Yue, and Shaloo Gupta are employees of Kantar, which administers the National Health and Wellness Survey and conducted this study and analysis.



LASSO (Least Absolute Shrinkage and Selection Operator) and XGBoost (eXtreme Gradient Boosting) Models for Predicting Depression-Related Work Impairment in US Working Adults

Vicky W. Li, Halley Costantino, John C. Rowland, Leiyu Yue, and Shaloo Gupta
Kantar Health, Real World Evidence, New York,
United States of America

April 16, 2021



Contents



Introduction



Methods



Results



Conclusion



1 Introduction

Introduction

- Depression is a very debilitating condition affecting 7.1% of US adults¹
- Work productivity loss among adults with depression has been shown to be prominent^{2, 3, 4}
 - Annual productivity loss per person estimated to be \$5,674 to \$15,150 in 2013, with increasing severity of depression²

The current study

- Predicts total work impairments as a result of absenteeism and presenteeism
- Using recent patient-reported data
- Identifies top risk factors for impairment
- Through both linear and non-linear prediction methods: LASSO (regularized linear regression) and XGBoost (decision-tree-based ensemble algorithm)

Disclaimer: Results of this study should be used only for research purposes, and should not be used to discriminate against certain subpopulation(s) of patients (demographic- or health-wise).



2 Methods



Data

Sample

2019 National Health and Wellness Survey

- Internet-based survey administered to a general population panel
- Representative of US Census on age, gender and ethnicity
- Participants age 18+

Inclusion criteria

- Age 18-64
- Employed (Full-time, Part-time, or Self-employed)
- 0 to 60 hours of work lost or impaired in the past 7 days
- A self-reported physician diagnosis of depression or having experienced depression in the past 12 months
- Analysis N = 12,124

Variables

Outcome: Overall work impairment

- Work Productivity and Activity Impairment questionnaire (WPAI-GH) ⁵
- Measure's absenteeism and presenteeism in past 7 days
- Unit: Hours (missed or impaired)
- hours missed work + hours worked x productivity loss/10
- Productivity loss:

Health problems had no effect on my work	0	1	2	3	4	5	6	7	8	9	10	Health problems completely prevented me from working
--	---	---	---	---	---	---	---	---	---	---	----	--

Predictors (54 in total)

- Socio-demographic characteristics (10)
- Health-related characteristics (10)
- Comorbidities (22)
- Sleep troubles experienced (12)

Models

LASSO

Group LASSO selection with Nesterov's method

- Have categorical variables and interaction terms
- SAS/STAT® software PROC HPGENSELECT
- Regularization parameter $\lambda = \rho^i$, $\rho = 0.8$ and $i = i$ th step
 - Chosen $\lambda = \mathbf{0.012576}$
- Selection criterion: AIC
- Training data: 80%, validation (=test) data: 20%
- Poisson distribution for count data

Model performance

- Root MSE
- Scattered plot: Predicted vs. Actual

Feature importance

- Standardized beta estimates

XGBoost

XGBoost regression

- Python XGBoost
- Booster = “gbtree”, Objective Function = “count:poisson”
- Hyperparameters:
 - N_estimators = 275, learning_rate = 0.01, min_child_weight = 5, max_tree_depth = 5, alpha = 0.05, lambda = 0.9
- Training data: 60%, validation data: 20%, test data: 20%
- Poisson distribution for count data

Model performance

- Root MSE
- Scattered plot: Predicted vs. Actual

Feature importance

- Gain across splits where the feature was used

3 Results



Descriptive Results

Socio-demographic and health behavior variables

Age (years), Mean ± SD	39.7 ± 13.7
Sex	
Female	7,737 (63.8%)
Male	4,387 (36.2%)
Census 9 region	
South Atlantic	2,529 (20.9%)
East North Central	1,873 (15.4%)
Pacific	1,764 (14.5%)
Middle Atlantic	1,690 (13.9%)
West South Central	1,266 (10.4%)
Mountain	935 (7.7%)
West North Central	809 (6.7%)
East South Central	672 (5.5%)
New England	586 (4.8%)
Race/ethnicity	
White	7,768 (64.1%)
Hispanic	1,723 (14.2%)
Black	1,344 (11.1%)
Other/Mixed	650 (5.4%)
Asian	639 (5.3%)
Marital status	
Married/living with partner	5,800 (47.8%)
Single	4,560 (37.6%)
Divorced/Separated/Widowed	1,764 (14.5%)
Education	
Highschool or less	1,696 (14.0%)
Some college/associates degree	4,164 (34.3%)
College degree	4,033 (33.3%)
Graduate degree	2,231 (18.4%)

Annual household income	
<\$25K or decline to answer	2,009 (16.6%)
\$25K to <\$50K	3,177 (26.2%)
\$50K to <\$75K	2,478 (20.4%)
\$75K to <\$100K	1,796 (14.8%)
\$100K to <\$150K	1,643 (13.6%)
\$150K to <\$200K	571 (4.7%)
\$200K or more	450 (3.7%)
No. of children living in the same household	
0	7,852 (64.8%)
1 or 2	3,695 (30.5%)
3+	577 (4.8%)
Caregiver of an adult relative	2,760 (22.8%)
Employment status	
Employed full time	8,160 (67.3%)
Employed part time	2,415 (19.9%)
Self-employed	1,549 (12.8%)
Health insurance	10,707 (88.3%)
Healthcare provider visits in past 6 months	
0 times	2,027 (16.7%)
once	1,866 (15.4%)
twice or more	8,231 (67.9%)
Currently use an Rx for depression	4,411 (36.4%)

Body mass index (BMI) category	
Underweight	430 (3.5%)
Normal	3,938 (32.5%)
Overweight	3,193 (26.3%)
Obese	4,116 (33.9%)
Declined to answer	447 (3.7%)
Alcohol use	
Do not drink	2,979 (24.6%)
Once a week or less often	5,749 (47.4%)
Multiple times week or daily	3,396 (28.0%)
Smoking behavior	
Never smoked	6,614 (54.6%)
Former smoker	3,032 (25.0%)
Current smoker	2,478 (20.4%)
Days exercised in past month, Mean ± SD	7.6 ± 8.4
Have taken steps to prevent diabetes or CVD	5,173 (42.7%)
PHQ-9 score	
None/minimal (0-4)	3,638 (30.0%)
Mild depression (5-9)	3,734 (30.8%)
Moderate depression (10-14)	2,257 (18.6%)
Moderately severe depression (15-19)	1,494 (12.3%)
Severe depression (20-27)	1,001 (8.3%)
GAD-7 score	
None/Minimal (0-4)	4,537 (37.4%)
Mild anxiety (5-9)	4,007 (33.1%)
Moderate anxiety (10-14)	2,111 (17.4%)
Severe anxiety (15-21)	1,469 (12.1%)

Descriptive Results

Comorbidities diagnosed with

Anxiety, panic disorder, or PTSD	5,861 (48.3%)
Allergies, nasal allergies, or hay Fever	5,039 (41.6%)
Sleep Disorders	3,648 (30.1%)
Headache or Migraine	3,471 (28.6%)
Pain	3,165 (26.1%)
Hypertension	2,520 (20.8%)
High cholesterol	2,330 (19.2%)
GERD/heartburn	2,185 (18.0%)
Arthritis	1,617 (13.3%)
Asthma	1,435 (11.8%)
Urinary tract infection	1,384 (11.4%)
Thyroid problems	1,132 (9.3%)
Yeast infection	1,059 (8.7%)
Eczema	1,036 (8.5%)
Diabetes	1,029 (8.5%)
Irritable bowel syndrome	930 (7.7%)
Ulcers, ulcerative colitis, or Crohn's	905 (7.5%)
COPD, emphysema, or chronic bronchitis	702 (5.8%)
Cancer	547 (4.5%)
Liver or Renal diseases	521 (4.3%)
CHF, CVD, or peripheral vascular diseases	431 (3.6%)
Diagnosed AIDS or HIV	84 (0.7%)

Sleep problems experienced

Experienced sleep problems	9,683 (79.9%)
Difficulty falling asleep	5,431 (44.8%)
Daytime sleepiness	4,417 (36.4%)
Poor quality of sleep	4,409 (36.4%)
Waking up several times during the night	4,386 (36.2%)
Waking up too early (such as before the alarm clock)	4,000 (33.0%)
Waking up during the night and not being able to get back to sleep	3,999 (33.0%)
Waking up to go to bathroom	3,901 (32.2%)
Night sweats/hot flashes	2,310 (19.1%)
Leg cramps/leg problems	1,902 (15.7%)
Pain	1,819 (15.0%)
Difficulty staying awake	1,705 (14.1%)
Sleep apnea (temporary absence of breathing)	1,249 (10.3%)
Other	148 (1.2%)

Model Results

LASSO (17 predictors)

RMSE in test dataset = 10.82

Predictors (in decreasing importance)	Standardized Estimates	Association (+/-)
PHQ-9 score	0.1650 +	
Employment status	-0.1413 -	(Part time)
GAD-7 score	0.1275 +	
Pain	0.1000 +	
Gender	-0.0964 +	(Male)
Waking up from pain	0.0836 +	
No. of doctor visits in past 6 months	0.0772 +	
No. of days exercised in past month	-0.0760 -	
Sleep disorders	0.0756 +	
Headache or migraine	0.0653 +	
Smoking status	0.0606 +	(Current)
Race/ethnicity	0.0584 -	(Whites)
Being a caregiver for an adult relative	0.0530 +	
Arthritis	0.0480 +	
No. of children	0.0449 +	
Age	-0.0400 -	
Taken steps to prevent diabetes or CVD	0.0260 +	

XGBoost (showing top 17 features)

RMSE in test dataset = 10.67

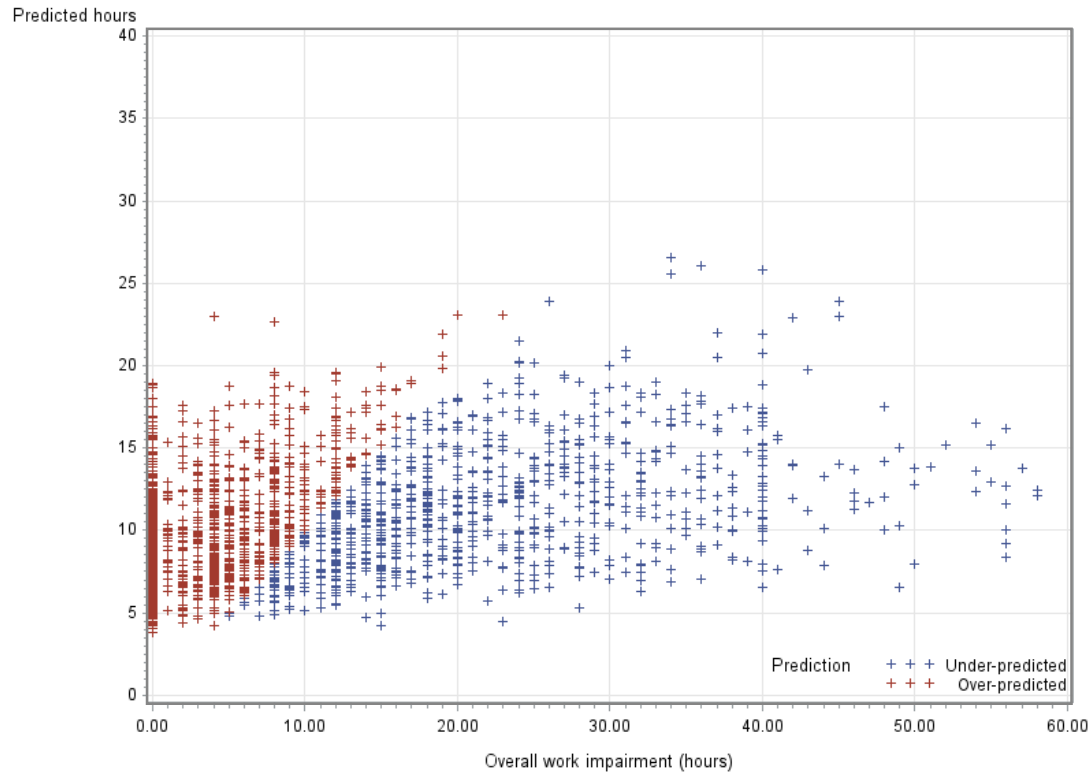
Feature	Importance
PHQ-9 score	0.1176
Pain	0.1139
Employment status	0.0576
GAD-7 score	0.0464
Headache or migraine	0.0454
Liver or Renal diseases	0.0394
Waking up from pain	0.0324
Sleep disorders	0.0298
Being a caregiver for an adult relative	0.0286
Taken steps to prevent diabetes or CVD	0.0266
CHF, CVD, or peripheral vascular diseases	0.0254
Arthritis	0.0236
GERD or heartburn	0.0205
Smoking status	0.0176
No. of children	0.0172
High Cholesterol	0.0161
No. of doctor visits in the past 6 months	0.0133

- Top variables selected by both
- Socio-demographic or health behavior variables

Model Results – Predicted vs. Actual Values

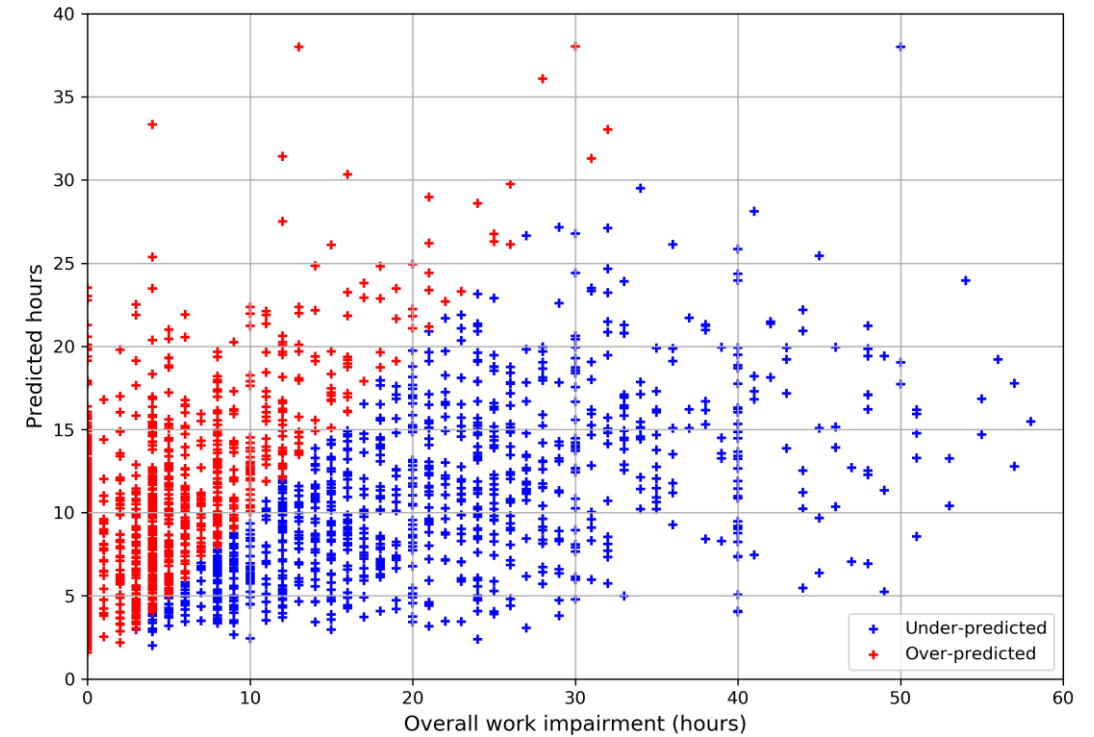
LASSO

Tend to overpredict at 0 hours, and underpredict at 20 hours or above



XGBoost

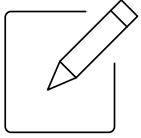
Tend to overpredict at 0 hours, and underpredict at 30 hours or above



4 Conclusion



Conclusion



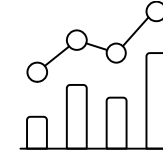
Model performance

- XGBoost performed slightly better than LASSO in terms of RMSE
- Predicted values off by +/- 11 hours on a scale of 0 to 60 hours of work impaired
- Different predictors were selected
 - LASSO: more socio-demographic and health behaviour variables
 - XGBoost: more comorbidity variables



Future improvements

- Including features from a more relevant timeframe
 - Measure **recent** well-being
 - Health experience in the past 30 days, instead of ever being diagnosed with X condition
 - Whether comorbidities were being effectively treated/controlled
- Occupation



Limitations

- Cross-sectional data, no causal inferences can be made
- Self-reported diagnoses are not verified through clinician or medical claims

References

1. <https://www.nimh.nih.gov/health/statistics/major-depression.shtml>
2. Chow W, Doane MJ, Sheehan J, Alphas L, and Le H. Economic Burden Among Patients With Major Depressive Disorder: An Analysis of Healthcare Resource Use, Work Productivity, and Direct and Indirect Costs by Depression Severity. AJMC: Clinical Brief, February 2019.
3. Stewart WF, Ricci JA, Chee E, Hahn SR, Morganstein D. Cost of Lost Productive Work Time Among US Workers With Depression. *JAMA*. 2003;289(23):3135–3144. doi:10.1001/jama.289.23.3135
4. Arne Beck, A. Lauren Crain, Leif I. Solberg, Jürgen Unützer, Russell E. Glasgow, Michael V. Maciosek and Robin Whitebird. Severity of Depression and Magnitude of Productivity Loss. *The Annals of Family Medicine* July 2011, 9 (4) 305-311; doi: <https://doi.org/10.1370/afm.1260>.
5. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmacoeconomics*. 1993;4(5):353-365.
6. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Int Med*. 2001;16(9):606-613.
7. Lowe B, Decker O, Muller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical care*. Mar 2008;46(3):266-274. PMID: 18388841.
8. Statistical Analysis Software (SAS) Version 9.4. SAS Institute, Cary NC.
9. <https://xgboost.readthedocs.io/en/latest/parameter.html>



Questions?

Vicky Li, Senior Biostatistician: Vicky.Li1@kantar.com

Halley Costantino, Biostatistician: Halley.Costantino@kantar.com

Jack Rowland, Biostatistician: Jack.Rowland@kantar.com

Austin Yue, Biostatistician : Austin.Yue@kantar.com

Shaloo Gupta, Director: Shaloo.Gupta@kantar.com



All research is conducted in accordance with the requirements of our Quality System, which confirms to ISO 20252:2019 the International Standard for Market Research, Certification Number : 1019

© Copyright 2021 Kantar. All rights reserved.

No copies or part copies of this document shall be made without the express permission of Kantar.

<https://www.kantar.com/-/media/project/kantar/global/policies/2020/health-division-report-terms-and-conditions-en>