Daniel Liu and Pele Yu

MIS Data Challenge Submission

Using gradient boosted forest models to examine the impact of ACEs in predicting mental well-being

Introduction: Adverse Childhood Events (ACEs) were first found to be correlated with poor health outcomes in adulthood 1998.[1] Since then, further work has reinforced their effect on both long-term mental and physical well-being.[2,3] The Behavioral Risk Factor Surveillance System (BRFSS) is a yearly telephone survey that interviews over 400,000 adults a year to collect information on personal health behaviors, including ACES, as well as measures of physical and mental well-being.[4] We utilize this dataset to predict mental well-being for patients, examine the effect of ACEs on mental health in recent data, and identify potentially modifiable risk factors to improve mental health.

Methods: The data used was from the 2020 BRFSS survey. Of the 401,958 patients, approximately 67 % were missing ACEs information, which was decided by whether the state implemented the module or not. Given that fact and the high amount of missing data, imputation was not attempted and only patients with ACE data were retained. This brought the patient count down to 129,386.

The target variable was a computed variable for the number of days mental health was not good over the last month (0, 1-13, or 14+). There was also a category in the target variable for missing/not asked/refused. As it would not be useful to predict this outcome, these patients were also removed from the dataset and ~126,998 patients remained.

As there were three target categories, this was treated as a multi-classification task (see supplementary table 1 for full breakdown). Gradient boosted forest models have historically performed well in this type of task, so XGBoost[5] and CatBoost[6], which is designed specifically for categorical data, were selected. Other than converting initial data to an importable format, Orange 3.31 was used for the remainder of the project.[7] Files available at: https://github.com/danielsdliu/mis-data-challenge

Initial features included all ACEs, measures of physical well-being, risk factors, and demographics. The data was split using into a training and testing set using a 7:3 ratio. Feature selection was performed using information gain and the top 60% of features were retained. The final dataset was only missing 0.2% of the data, likely related to different versions of the ACEs module. Final features included:

- ACEs:
  - Living with someone: depressed/mentally ill; is a drinker/alcoholic; illegal drug user/prescription drug abuser; anyone who served time or sentenced to serve time
  - Experienced: verbal abuse, being touched sexually, being forced to touch someone sexually, sexual abuse, physical abuse; parents divorcing; domestic violence
- Physical well-being: days out of 30 in good physical health, physical activity in leisure time
- Demographics: Age, Income bracket, sex, level of education completed
- Risk Factors: availability of healthcare coverage, smoker, asthma status, number of children

XGBoost and CatBoost were compared using F1, precision, recall. AUC and accuracy are reported as well. Models were also evaluated with Shapley values as well as feature importance.

Results: CatBoost outperformed XGBoost in all aspects and the performance metrics are reported below. See supplementary table 2 for full classification performance results.

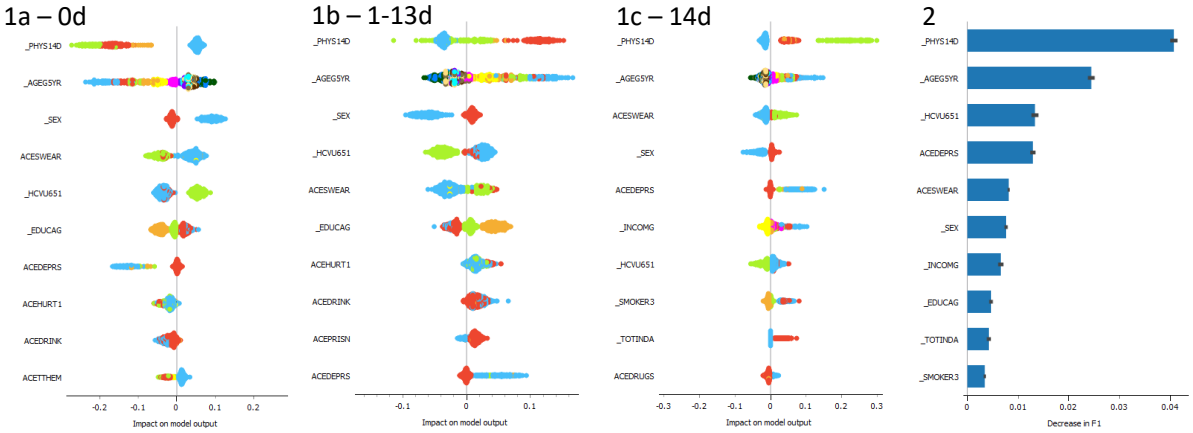
Catboost (target)	F1	Precision	Recall	AUC	Accuracy
0 poor mental health days	0.813	0.721	0.933	0.757	0.720
1-13 poor mental health days	0.271	0.463	0.192	0.692	0.769
14+ poor mental health days	0.339	0.509	0.255	0.788	0.879

Daniel Liu and Pele Yu

MIS Data Challenge Submission

Figure 1a-c: Shapley plots for feature importance. The Orange legend and BRFSS data dictionary are required for interpretation. The colors correspond with categories, with the lowest being light blue.

Fig 2: The effect of feature permutation on F1 score.



Discussion: The model is better at identifying patients in good mental health than those that have poor mental, which was expected given the imbalance. It would be difficult to recommend as a screening tool given the poor recall. However, it could trigger a workflow of offering validated screening tools like PHQ-9 or GAD-7 to patients that are potentially at risk for mental health problems as identified by the model. All ACEs were important enough to be retained in the feature selection stage and ranked prominently in the top 10 features for each target class. Experiencing an ACE led to less likelihood of being in good mental health and increased the chances of having poorer mental health days, confirming the effect of ACEs on long-term health. Good physical health was the most important feature overall. Of the final features, smoking, availability of healthcare coverage, and good physical health (to some extent) are modifiable risk factors clinicians can impact.

References:

1. Felitti, V. J. et al. Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults - The Adverse Childhood Experiences (ACE) Study. Am J Prev Med 14, 245–258 (1998).
2. Bellis, M. A. et al. Life course health consequences and associated annual costs of adverse childhood experiences across Europe and North America: a systematic review and meta-analysis. Lancet Public Heal 4, e517–e528 (2019).
3. Petruccelli, K., Davis, J. & Berman, T. Adverse childhood experiences and associated health outcomes: A systematic review and meta-analysis. Child Abuse Neglect 97, 104127 (2019).
4. CDC - BRFSS - Survey Data & Documentation. https://www.cdc.gov/brfss/data_documentation/index.htm (n.d.).
5. Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). 2016. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
6. Dorogush, A. V., Ershov, V. & Gulin., A. CatBoost: gradient boosting with categorical features support. Workshop on ML Systems at NIPS 2017 (n.d.).
7. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349−2353.

Daniel Liu and Pele Yu

MIS Data Challenge Submission

Supplementary Tables

Supplementary Table 1: Distribution of target variable

Targets	Count	% of total
0 days	83005	65%
1-13 days	28459	22%
14+ days	15534	12%
Grand Total	**126998**	**100%**

Supplementary Table 2:

Target 1 = 0 poor mental health days in last 30 days					
Target 2 = 1-13 poor mental health days					
Target 3 = 14+ poor mental health days					
Test Data					
	F1	Precision	Recall	AUC	CA
Target 1					
xgboost	0.810	0.703	0.954	0.743	0.707
Catboost	0.814	0.718	0.939	0.759	0.720
Target 2					
xgboost	0.220	0.466	0.144	0.677	0.771
Catboost	0.265	0.479	0.183	0.697	0.772
Target 3					
xgboost	0.281	0.530	0.191	0.768	0.880
Catboost	0.332	0.504	0.248	0.786	0.878
Train data					
	F1	Precision	Recall	AUC	CA
Target 1					
xgboost	0.810	0.704	0.953	0.746	0.708
Catboost	0.816	0.721	0.941	0.766	0.723
Target 2					
xgboost	0.220	0.460	0.144	0.680	0.770
Catboost	0.272	0.489	0.188	0.704	0.774
Target 3					
xgboost	0.298	0.547	0.205	0.780	0.882
Catboost	0.360	0.539	0.270	0.801	0.882