**Appendix**

Supervised Learning Experiments Methodological Details

Elastic Net (ENET) is an extension to the standard logistic regression, adding penalties to the loss function during training with the aim of encouraging simpler models with smaller coefficient values. ENET was chosen as it is a well-known standard in the biomedical literature performing robustly in a wide range of applications; however, without basis functions and feature engineering, it can only fit a linear decision boundary in feature space and cannot account for interactions. Extreme Gradient Boosting (XGBoost) is based on an ensemble of classification (or regression) trees where, with each iteration, a new tree is fit to the pseudo-residuals from the previous tree(s), i.e. the negative gradient of a differentiable loss function (logistic loss in our case) evaluated at the previous model’s predictions (1). XGBoost was chosen due to its state-of-the-art performance on structured/tabular data sets (2), being multiple-time winner of Kaggle competitions.

Each training set from the inner loop of nested cross-validation underwent the following pre-processing steps. 1) As ENET requires a fully observed data set, MissForest algorithm was used for imputation (3). While there is not a clear-cut value for when a variable’s missing rate is too high for reliable imputation, the authors of MissForest demonstrated its robustness in regimes where a maximum of 30 % of the data matrix was randomly replaced with missing values. At the same time, even classification algorithms capable of directly handling missing values are not immune to disproportionately high missing rates. As a result, since EDA showed that a group of variables had missing rates > 85% and, moreover, had multiple entries not properly coded (i.e. a qualitative judgement was expressed in place of the actual score for naturally numerical variable), we decided to exclude these variables from the analysis. These included: Apgar score at minute 1 and at minute 5 after birth, gestational age, newborn’s skull circumference, length, and weight. 2) Regularization in ENET is sensitive to the scale features are measured at, while this is not an issue with tree-based models. As a result, continuous features were rescaled to have zero mean and one standard deviation. 3) During EDA we verified that some early AAO definitions produced some mild to moderate imbalance in the target variable distribution. This can pose a challenge to predictive modelling. In cases where the ratio minority/majority class was <3/7, Synthetic Minority Oversampling Technique-Encoded Nominal and Continuous (SMOTE-ENC) (4)was used to produce synthetic observations for the minority class. The number of synthetically generated instances for the minority class was chosen so that the proportion minority/majority class was equal to 3/7. Where applicable, i.e. step 1) and 2), the model learned from the training data was subsequently fit to the validation set. The best model under each algorithm from the inner 5-fold cross-validation was then refit to the corresponding outer training set before generalisation estimation on the corresponding test set. As with the inner loop, step 1) and 2) of the preprocessing pipeline were applied on the outer test set using parameters learned from the corresponding outer training set. Furthermore, the data matrix underwent steps 1) and 2) prior to Deep Embedding Clustering in the unsupervised learning experiments.

**References**

1. Schapire RE: The Boosting Approach to Machine Learning: An Overview [Internet], inDenison DD, Hansen MH, Holmes CC, et al., editorsNonlinear Estimation and Classification. New York, NY, Springer, 2003, pp 149–171.[cited 2022 Jan 28] Available from: https://doi.org/10.1007/978-0-387-21579-2\_9

2. Shwartz-Ziv R, Armon A: Tabular Data: Deep Learning is Not All You Need [Internet]2021[cited 2022 Jan 28] Available from: https://openreview.net/forum?id=vdgtepS1pV

3. Stekhoven DJ, Bühlmann P: MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 2012; 28:112–118

4. Mukherjee M, Khushi M: SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. Appl Syst Innov 2021; 4:18

**Supplementary Table 1**. **Variables from the FACE-BD cohort considered for this study.** The rows shaded in grey are the variables used as predictors in the supervised learning experiments and as clustering variables in the unsupervised learning experiments. BD: Bipolar Disorder; entries marked with \* are individual items of the questionnaire Childhood Trauma Questionnaire; WURS: Wender Utah Rating Scale; nan: missing value number

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **nan** | **mean (sd)** | **yes/no** |
| **Maternal Age at Birth (Years)** | **828** | 28.10 (5.92) | **/** |
| **Paternal Age at Birth (Years)** | 893 | 30.89 (6.78) | **/** |
| **Not at Term Delivery** | **721** | **/** | 445/2832 |
| **Caesarean Section** | **0** | **/** | 273/3725 |
| **Neonatal Hospitalization** | **0** | **/** | 230/3768 |
| **Physical Abuse\*** | **197** | 6.58 (3.05) | **/** |
| **Emotional Neglect\*** | **196** | 12.24 (5.06) | **/** |
| **Physical Neglect\*** | 193 | 7.02 (2.75) | **/** |
| **Sexual Abuse\*** | 196 | 7.21 (2.76) | **/** |
| **Emotional Abuse\*** | **194** | 10.30 (5.26) | **/** |
| **Physical Neglect\*** | 193 | 7.02 (2.75) | **/** |
| **Sexual Abuse\*** | 196 | 7.21 (2.76) | **/** |
| **Emotional Abuse\*** | **194** | 10.30 (5.26) | **/** |
| **Dyslexia** | **1031** | **/** | 181/2786 |
| **Dysorthography** | **1002** | **/** | 86/2910 |
| **Dyscalculia** | **991** | **/** | 35/2972 |
| **Dysphasia** | **938** | **/** | 12/3055 |
| **Dyspraxia** | 962 | **/** | 51/2985 |
| **Speech Acquisition Delay** | **962** | **/** | 51/2985 |
| **Stuttering** | 917 | **/** | 43/3038 |
| **Gait Acquisition Delay** | **968** | **/** | 42/2988 |
| **Early Childhood Febrile Seizures** | **952** | **/** | 42/3004 |
| **WURS Total score** | 299 | 69.99 (29.39) | **/** |
| **BD Family History** | **0** | **/** | 715/3283 |
| **Cannabis Misuse Predating BD** | **518** | **/** | 286/3194 |
| **Age at Onset (Years)** | **0** | 23.71 (9.60) | **/** |

**Supplementary Figure 1**. **Sketch of the supervised learning workflow used in this study**. In brief, from the original data matrix cases with missing values (NAs) under the target variables, i.e. age at onset, were removed from further analyses. Exploratory data analysis (EDA) was subsequently carried out on a 80% split of the data matrix derived with stratified random resampling, where stratification was with respect to the early vs non-early age at onset definition at hand. Predictors with missing rate greater than 85% at the EDA were dropped from further analyses. For each classifier, i.e. Elastic Net Regression and Extreme Gradient Boosting, generalisation performance estimates were derived with nested cross-validation. In the outer resampling loop, ten pairs of training/test tests were produced. In each of these outer training sets the optimum configuration of hyperparameters of the classification algorithms were selected through a grid search while optimizing the area under the Receiver Operator Curve (AUROC) in a 5-fold cross-validation. The so-tuned classification algorithms were then fitted on each outer training set and their performance was evaluated on the outer test sets. Each training set from the inner loop underwent the following pre-processing steps: missing value imputation with MissForest algorithm, Standardization, and, in case of minority/majority class imbalance <3/7, up-sampling with SMOTE-ENC. The best model from the 5-fold inner resampling loop was then refitted on the whole outer training set and a generalization performance estimate was derived from the corresponding outer test set.

Diagram

Description automatically generated