Predicting Problematic Internet Use

Connor McManus, Sean Shen,
Victoria Uchman, Barret White

Introduction

Project Topic

- **The Problem:** Problematic internet use is a growing concern, especially among youth, with impacts on mental health and learning abilities.
- **The Dataset:** We used the Healthy Brain Network (HBN) dataset, containing clinical, physical activity, internet usage, and demographic data for approximately 5,000 participants aged 5–22.

Our Approach

- **Research Questions:**
  - What drives higher levels of problematic internet use?
  - Can demographic, physical activity, and behavioral markers predict internet usage severity?
- **Methodology:**
  - Machine learning models for classification and feature selection.
  - Evaluation using Quadratic Weighted Kappa (QWK) to balance predictions for ordinal data.

Dataset Overview

- **Scope**: The HBN dataset contains clinical and research data from approximately 5,000 participants aged 5–22. It integrates information on physical activity, fitness, internet usage behavior, and demographic data to study mental health and learning disorders.
- **Objective**: The dataset aims to identify biological markers for improving the diagnosis and treatment of mental health issues and learning disorders.
- **Goal**: Predict participants' Severity Impairment Index (SII), which measures the degree of problematic internet use (scored from 0 = None to 3 = Severe).
- **Data Composition**:
  - **Tabular Data**: Contains participant demographics, physical and fitness measures, and internet usage behavior.
  - **Time-Series Data**: Accelerometer readings tracking motion and daily activity.

Imputation Methods

- KNN Imputation
  - Imputes the mean (or mode, for classification) of the k-nearest neighbors to a missing point
  - Uses Euclidean distance to measure how close a row with a missing value is (the value of interest) to other rows in the data
  - We used a value of 5 for K - have to pick arbitrarily
- Simple median/mode imputation
  - We also tried imputing the median (for numerical values) or mode (for categorical) to replace missing values
  - Only used the column that the missing value occurred in - unlike KNN, this method does not utilize the rest of the data

MICE (Multivariate Imputation by Chained Equations)

- 1st iteration: predict missing values of a column using other columns
  - Other columns temporarily imputed by median and mode
  - Do this for every column
- 2nd iteration: predict missing values again with newly imputed data
  - Better data this time from predictive modeling
  - Many more iterations
- N-th iteration: newly imputed data this iteration is identical to the last iteration
  - Stabilized results, stop iterating
- Strengths:
  - Robust
  - Allows for different modeling techniques (Random Forest)
- Weaknesses:
  - Computationally intensive
  - No guarantee of convergence

MICE in Pictures

a	b	c
NA	3	7
5	0	NA
NA	6	9
3	NA	6

Original data with missing data

a	b	c
median(a)	3	7
5	0	median(c)
median(a)	6	9
3	median(b)	6

Temporary median impute

a	b	c
a1*	3	7
5	0	c2*
a3*	6	9
3	b4*	6

Impute using imputed (median)

a	b	c
a1**	3	7
5	0	c2**
a3**	6	9
3	b4**	6

Impute using imputed (prediction)

Quadratic Weighted Kappa

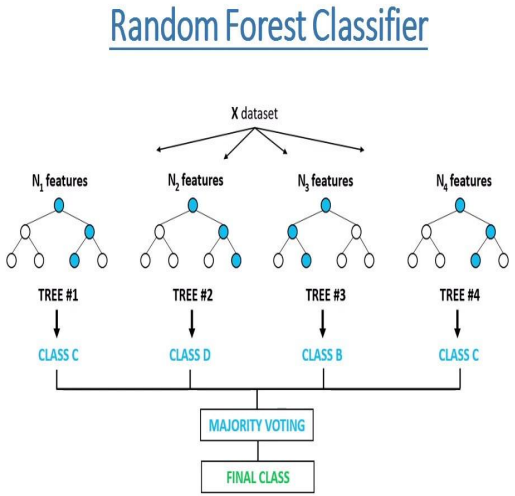
- QWK is a measure of the agreement between two ordinally scaled samples
- This metric varies from 0 (random agreement) to 1 (complete agreement). In the event that there is less agreement than expected by chance, the metric may go below 0.
- Assigns weights for different classes, addressing some class imbalance

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$$

Random Forest

Pros:

- Provides feature selection and importance
- Can be optimized for QWK
- Flexible and performs well in high dimensions



Cons:

- Treats ordinal variables as categorical
- Can suffer if ordinal classes are imbalanced

Our Random Forest Model obtained a CV 5 fold mean QWK of 0.4204

XGBoost

- Extension of gradient boosting
  - Builds upon the pseudo-residuals, rather than the standard residuals, to slowly built a series of sequential trees
  - Pseudo-residuals: gradients of the loss function with respect to the previous tree's predictions
    - Includes information on the direction in which the tree's predictions need to be adjusted
- XGBoost implements regularization (ridge regression, in this case) to control the complexity of the added trees
- Primary tuning parameters: number of trees, shrinkage, and depth
  - Can also tune the minimum loss reduction required to split a tree, how many features are used to build each tree, and the weights of each observation needed to be in a node

XGBoost Results

- Used cross validation to iterate over all possible combinations of parameters
  - Maximum QWK found at a maximum depth of 3, a shrinkage parameter (eta) of 0.1, and 100 rounds to build the tree

	eta <dbl>	max_depth <dbl>	gam... <dbl>	colsample_bytree <dbl>	min_child_weight <dbl>	subsample <dbl>	nrounds <dbl>	QWK <dbl>	QWKSD <dbl>
146	0.10	3	0	0.7	1	0.7	100	0.4159484	0.03822024
149	0.10	3	0	0.7	1	0.8	100	0.4130952	0.03673798
152	0.10	3	0	0.7	5	0.7	100	0.4082625	0.03376811
155	0.10	3	0	0.7	5	0.8	100	0.4211356	0.03263789
158	0.10	3	0	0.8	1	0.7	100	0.4194452	0.04037843

Bidirectional Stepwise QWK

- Stepwise framework but with QWK (No AIC, BIC)
- polr() for multinomial ordinal regression
  - Formula, data, method/link function (logistic, probit)
- Uses k-fold cv to prevent overfit (QWK does not balance fit and complexity)
- Algorithm
  1. Start with no predictors
  2. Add/Remove a predictor, k-fold CV
     a. Improve QWK, then keep the addition/removal
     b. Otherwise, go to the next iteration
  3. Stop when cross-validated QWK does not improve

Stepwise Strengths and Weaknesses

- Strengths:
  - Interpretability of output
  - Intuitive algorithm
  - Flexible that added predictors can be removed if not useful later
- Weaknesses
  - Computationally heavy (CV for every change in predictors)
  - No direct parameter to control model complexity
- Potential improvement
  - Add a penalization term in QWK objective function (like LASSO)
  - Optimize QWK of a single link function instead of averaging
  - Using PCA to produce features

Output - Bidirectional Stepwise QWK

```
$selected_predictors
 [1] "Basic_Demos.Sex"                          "PreInt_EduHx.computerinternet_hoursday"
 [3] "Physical.Height"                          "Physical.Weight"
 [5] "Physical.HeartRate"                       "FGC.FGC_CU"
 [7] "FGC.FGC_SRL"                              "FGC.FGC_TL"
 [9] "SDS.SDS_Total_Raw"                        "SDS.SDS_Total_T"
[11] "FGC.FGC_SRL_Zone"                         "BIA.BIA_Frame_num"
[13] "CGAS.CGAS_Score"                          "Physical.BMI"
[15] "Physical.Diastolic_BP"                    "FGC.FGC_SRR"
```

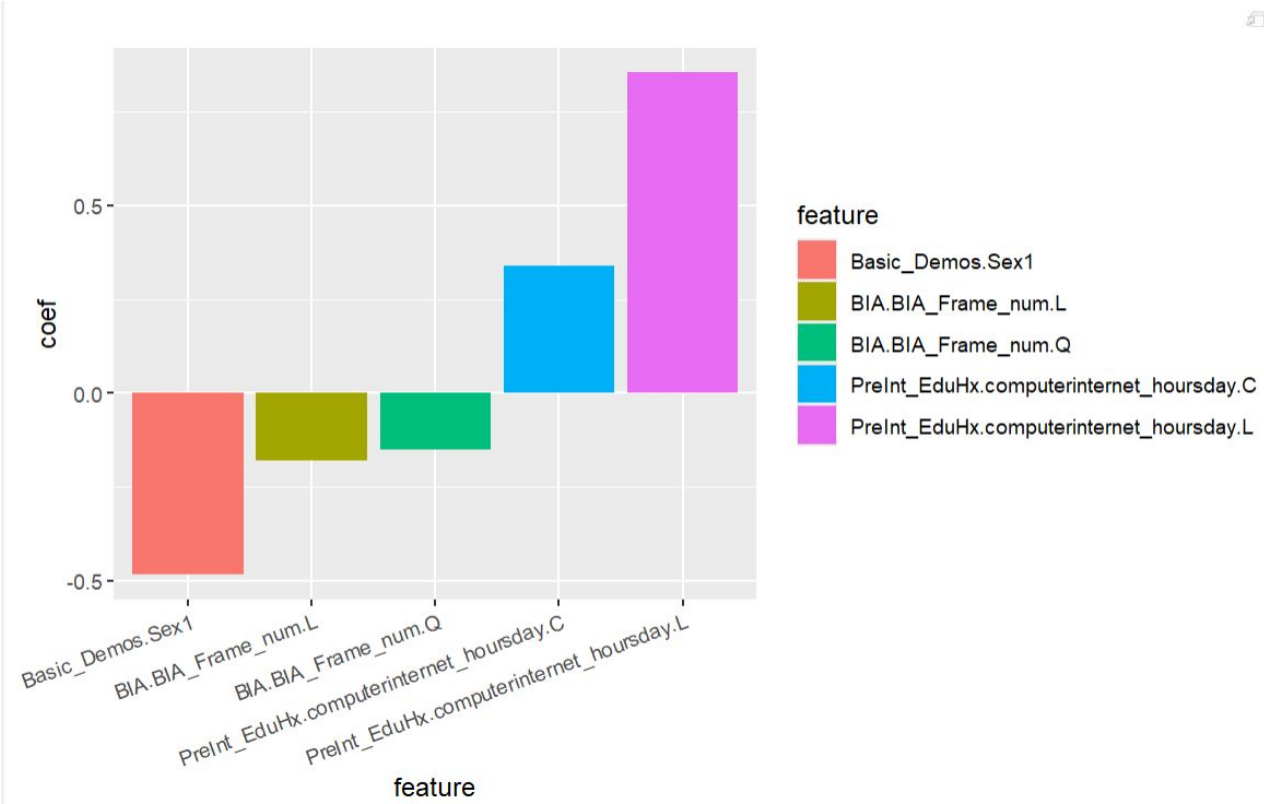
Training cv (5-fold)

method	QWK
cauchit	0.3938808
cloglog	0.3114185
logistic	0.3793043
loglog	0.3188933
probit	0.3596778

Full dataset cv (5-fold)

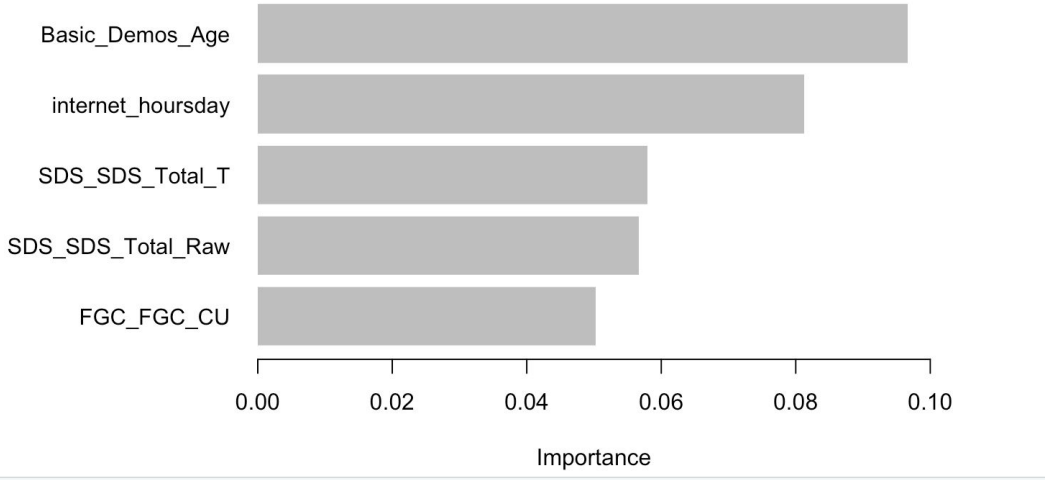
method	QWK
cauchit	0.3926447
cloglog	0.3058249
logistic	0.3672446
loglog	0.3224766
probit	0.3582314

Features with the Highest impact in Stepwise

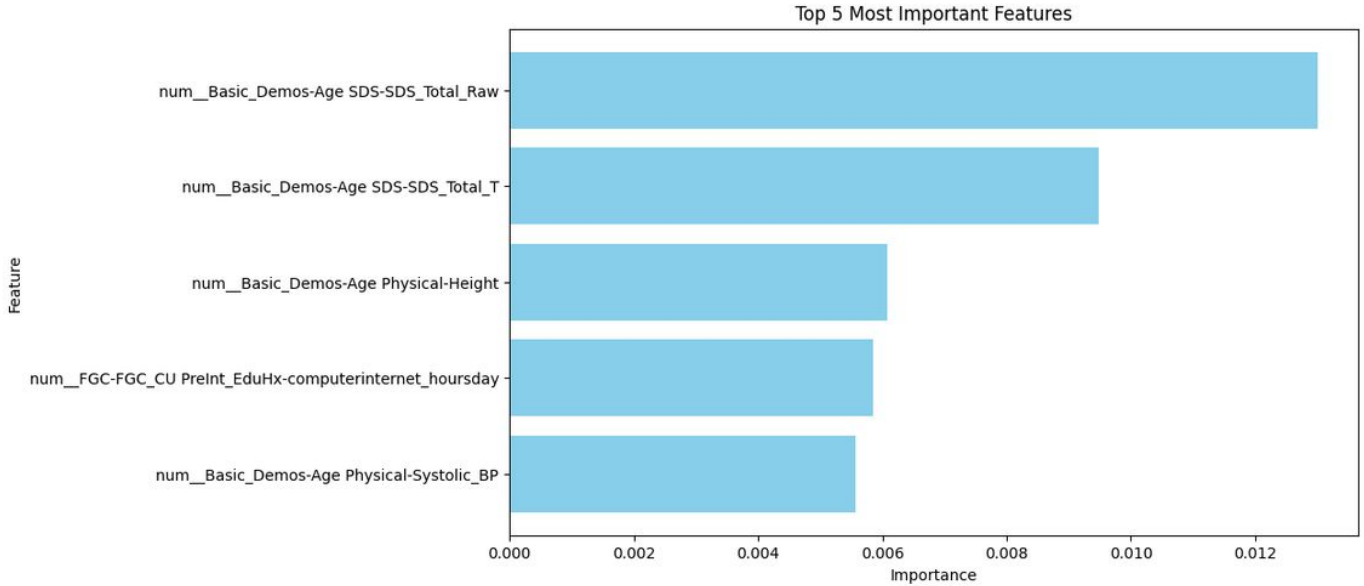


XGBoost Feature Selection

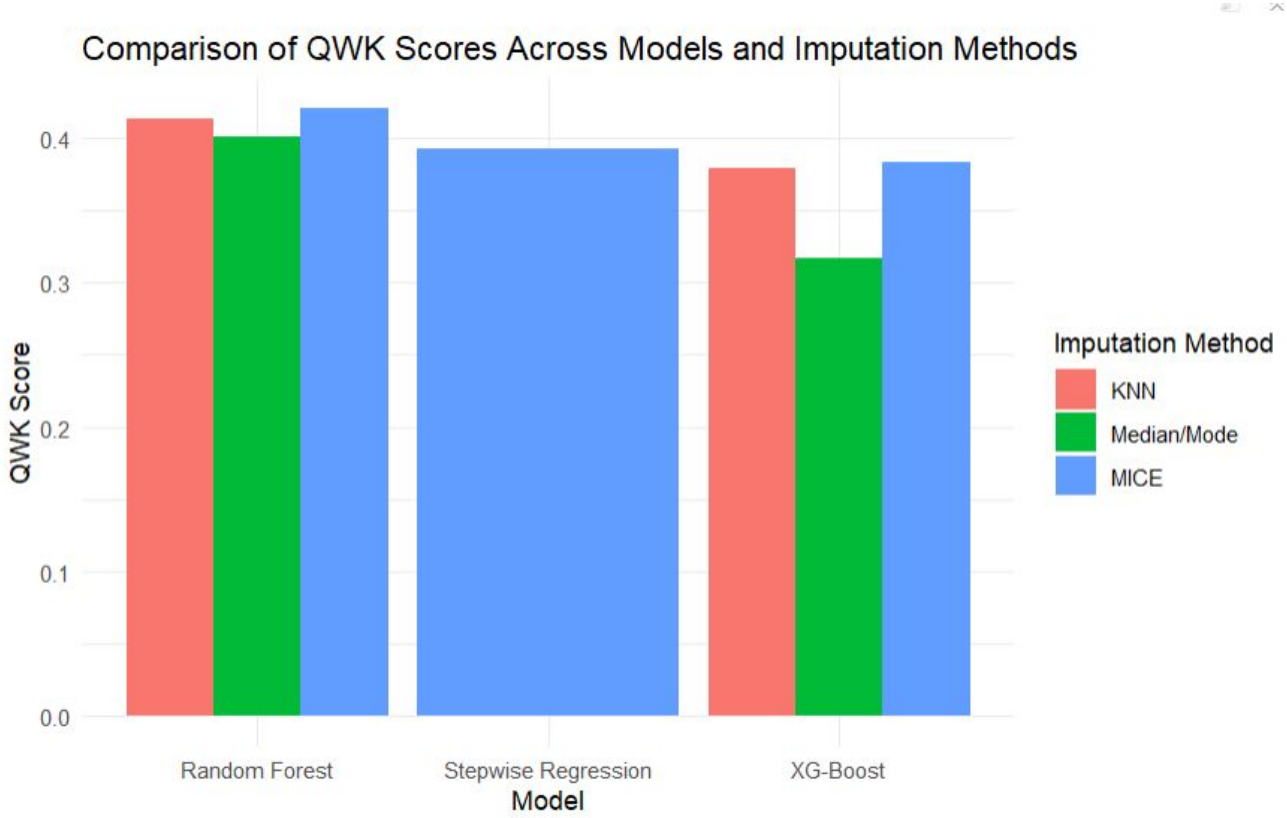
- Can perform feature selection: sum the amount each split improves the model performance, weighted by the number of observations in each node
- Top 5 factors shown in terms of importance shown below



Most Predictive Features



Comparing Models



Conclusion

Summary of Findings

- **Top Predictors:** Identified key features such as symptom severity scores (SDS raw & standardized), height, daily computer/internet usage, and systolic blood pressure.

- **Model Performance:**
  - **Random Forest:** Robust feature selection with a 5-fold CV QWK of 0.420.
  - **XGBoost:** QWK with optimized parameters (depth = 3, shrinkage = 0.1).
  - **Bidirectional Stepwise QWK:** Provided interpretable insights into ordinal relationships.

- **Insights Gained:**
  - **MICE Imputation:** Improved dataset quality, reducing bias from missing values.

- **Possible Future Directions:**
  - **Refining Metric:** Penalize terms in QWK for better balancing of fit and complexity
  - **Expanded Analysis:** Include time-series data for longitudinal insights
  - **Practical Applications:** Inform clinicians and policymakers to tackle problematic internet use

Bibliography

Child Mind Institute. (n.d.). *Child mind institute - problematic internet use*. Kaggle. https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/data

“What Is Xgboost? An Introduction to XGBoost Algorithm in Machine Learning: Simplilearn.” *Simplilearn.Com*, Simplilearn, 7 Nov. 2023, www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article.

MIT School of Distance Education. “Data Imputation Techniques: Handling Missing Data in Machine Learning.” *MIT School of Distance Education Blog*, blog.mitsde.com/data-imputation-techniques-handling-missing-data-in-machine-learning/