Problematic Internet Use Classification

Michael Lu
December 13, 2024
Brown University, Department of Computer Science
https://github.com/mdlu02/InternetUseClassification

Background and Review

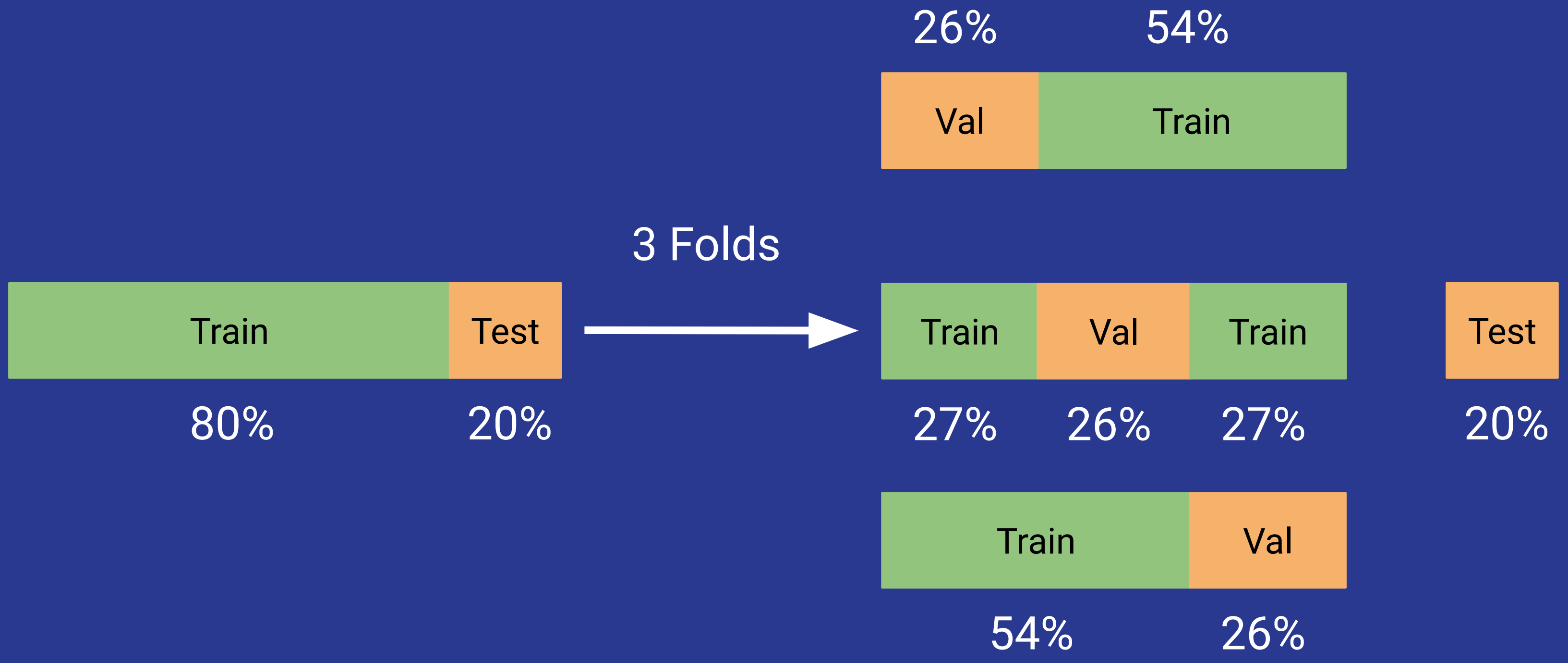
Problem & Goal

- Increased internet use associated with mental health problems.
- Predict severity of negative internet use.
- Tabular and time series data for ~5000 5-22 year-olds.
- Classification vs. Regression

Challenges

- Significant amount of missing data.
- Highly imbalanced dataset.
- Time series and tabular data.

Split + Crossfold Validation



Train | Test

80% | 20%

3 Folds

26% | 54%

Val | Train

Train | Val | Train

27% | 26% | 27%

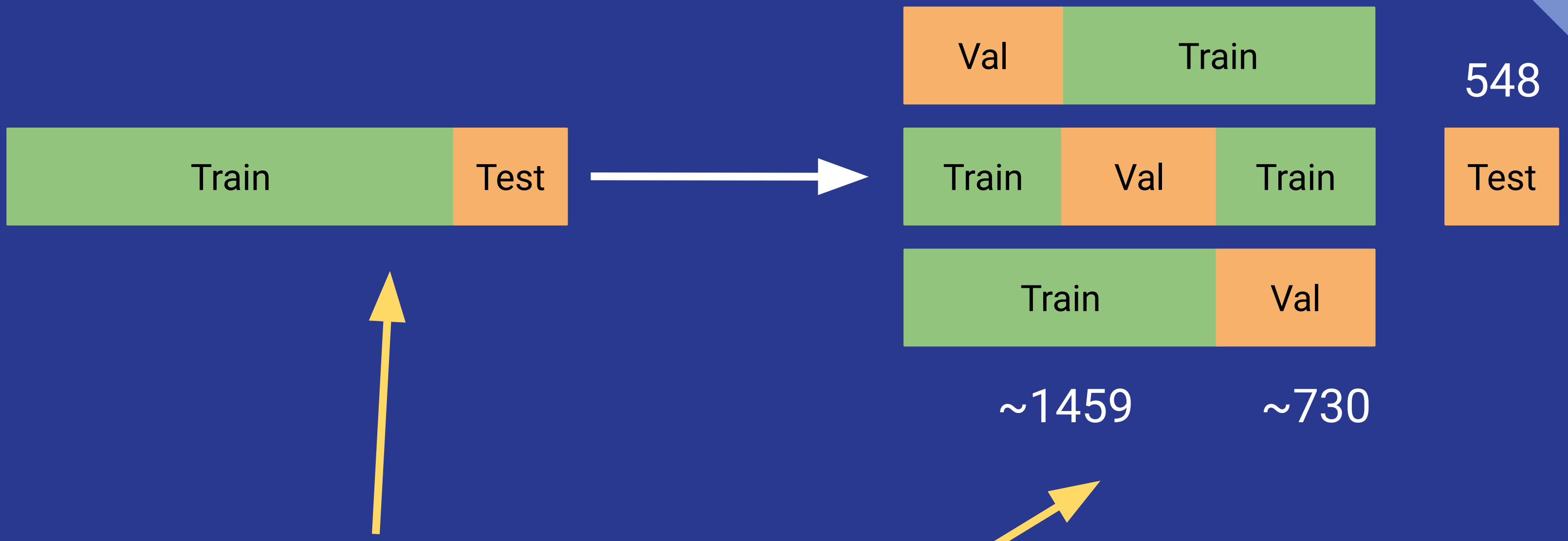
Test

20%

Train | Val

54% | 26%

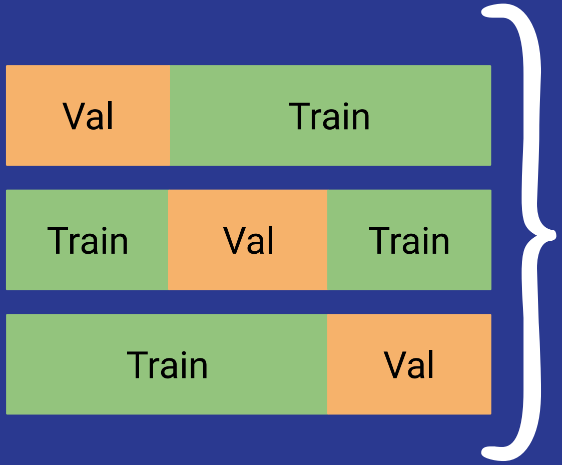
Split + Crossfold Validation



Stratified by SII, age group, sex, and the existence of time series data.

Cross Validation

For each ML model and associated parameter set:



Val | Train

Train | Val | Train

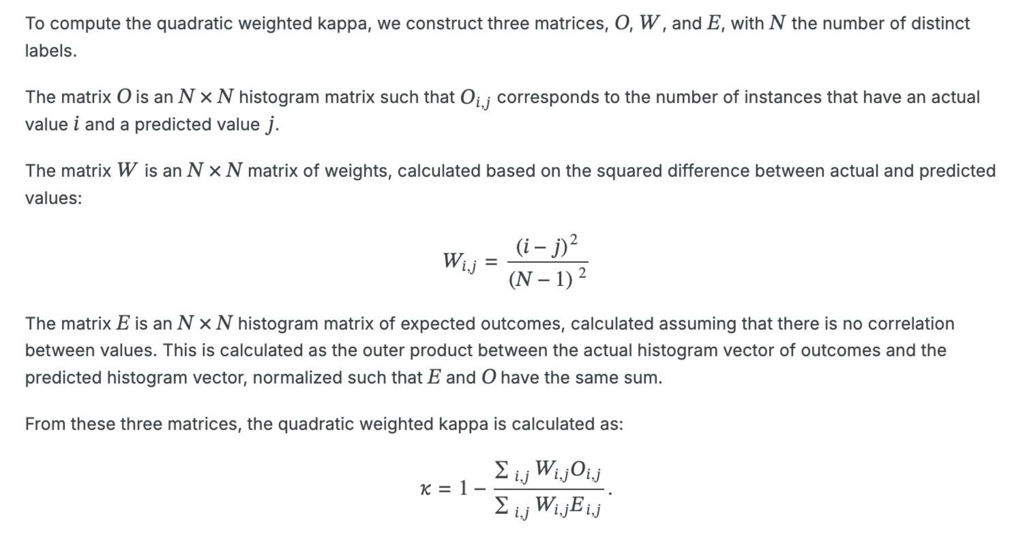
Train | Val

Mean k-fold quadratic weighted kappa score for each model + param set

Cross Validation - QWK

Quadratic Weighted Kappa:

- Ranges from -1 to 1
- Measures agreement between two outcomes
  - <0 is worse than random
  - 0 is random/baseline agreement
  - 1 is perfect agreement
  - Current 1st place achieves 0.5 QWK
- Baseline score is always 0 even if the data is unbalanced!



To compute the quadratic weighted kappa, we construct three matrices, O, W, and E, with N the number of distinct labels.

The matrix O is an $N \times N$ histogram matrix such that $O_{i,j}$ corresponds to the number of instances that have an actual value i and a predicted value j.

The matrix W is an $N \times N$ matrix of weights, calculated based on the squared difference between actual and predicted values:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^{2}}$$

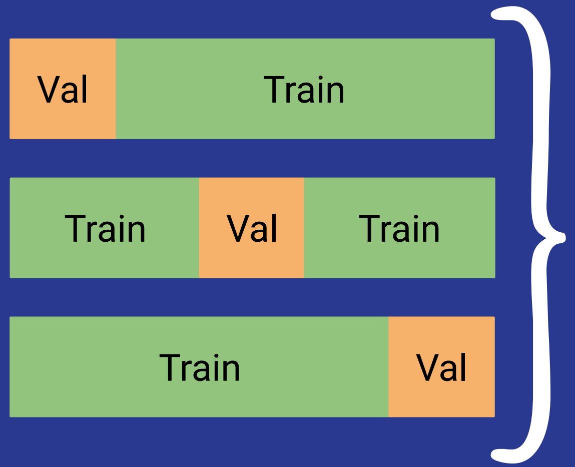
The matrix E is an $N \times N$ histogram matrix of expected outcomes, calculated assuming that there is no correlation between values. This is calculated as the outer product between the actual histogram vector of outcomes and the predicted histogram vector, normalized such that E and O have the same sum.

From these three matrices, the quadratic weighted kappa is calculated as:

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}.$$

Cross Validation

For each ML model and associated parameter set:



Val | Train

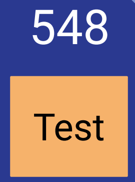
Train | Val | Train

Train | Val

Select best params for each model by val QWK.

Train on full CV data and test on hold out test set. Track:

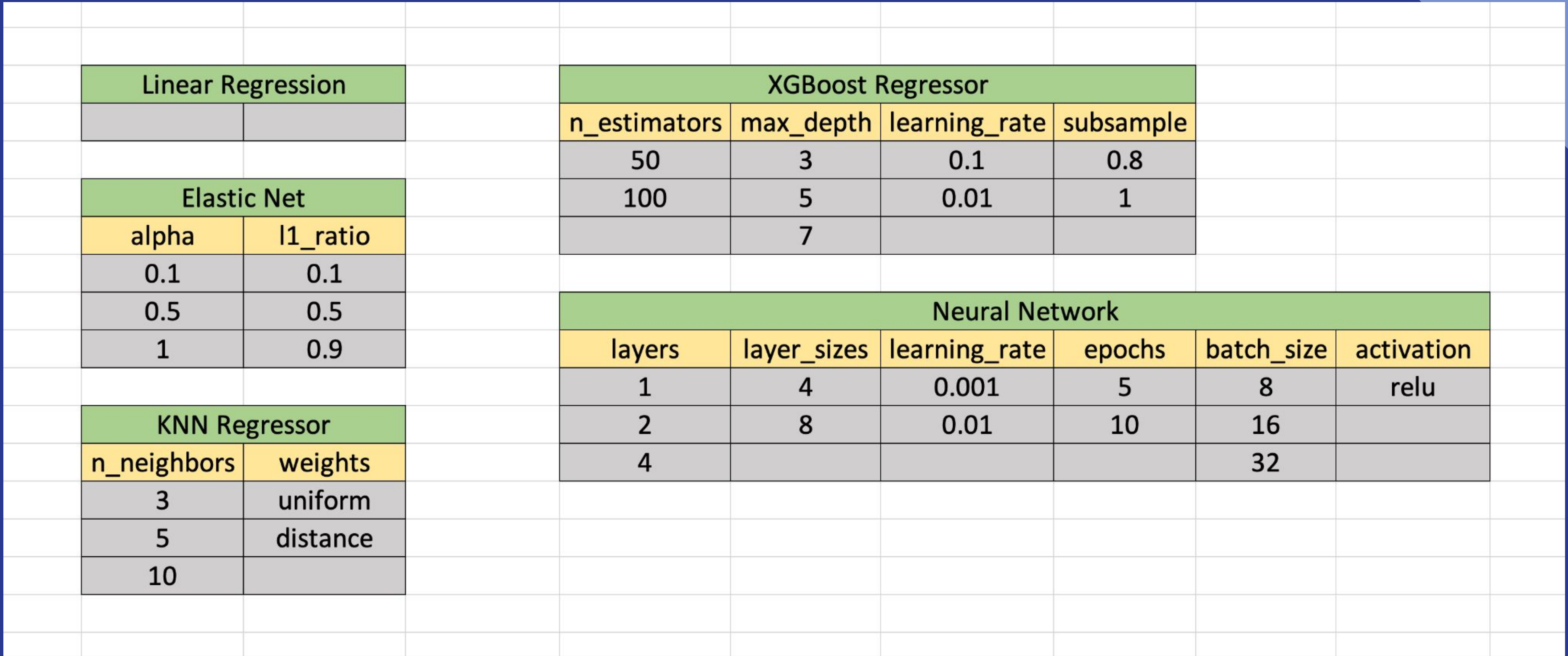
- RMSE, QWK, Weighted F1
- Confusion matrix, AUC-ROC, and Precision-Recall Curves



548

Test

ML Models + Hyperparameters



Linear Regression	

Elastic Net	
alpha	l1_ratio
0.1	0.1
0.5	0.5
1	0.9

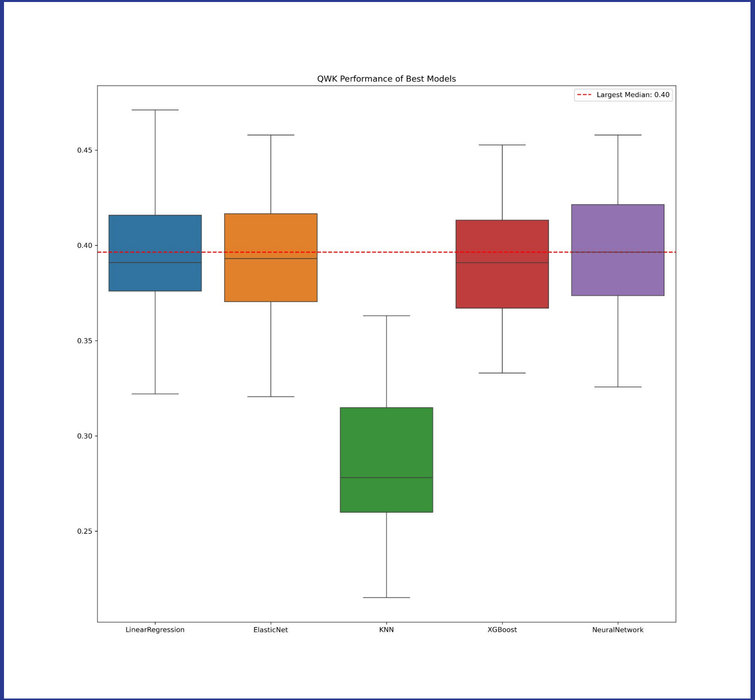
KNN Regressor	
n_neighbors	weights
3	uniform
5	distance
10	

XGBoost Regressor			
n_estimators	max_depth	learning_rate	subsample
50	3	0.1	0.8
100	5	0.01	1
	7		

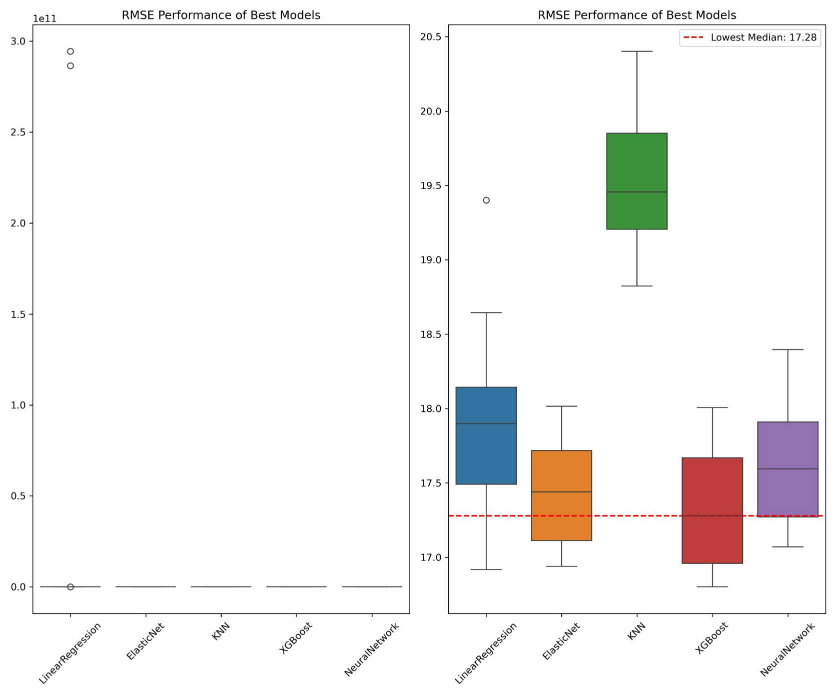
Neural Network					
layers	layer_sizes	learning_rate	epochs	batch_size	activation
1	4	0.001	5	8	relu
2	8	0.01	10	16	
4				32	

Results - Best Models

QWK Performance (larger is better)

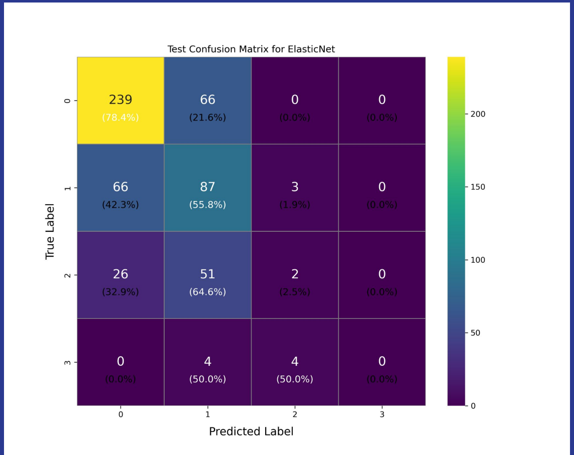


RMSE Performance (smaller is better)

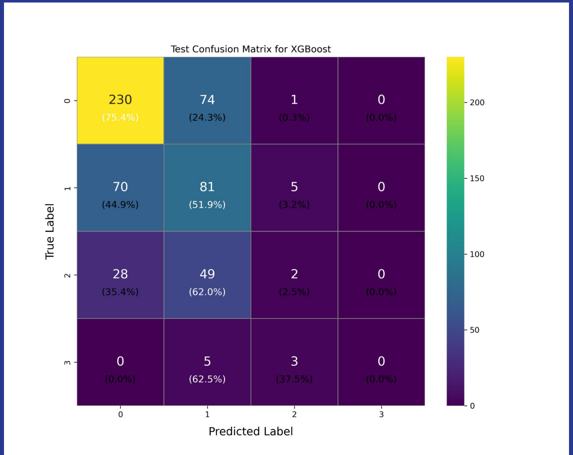


Results - Confusion Matrices

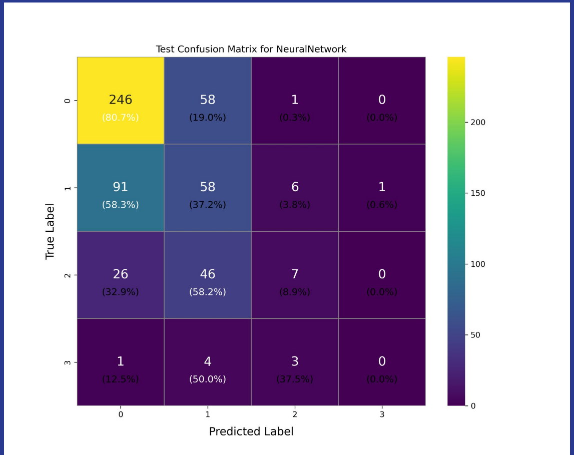
Elastic Net



XGBoost

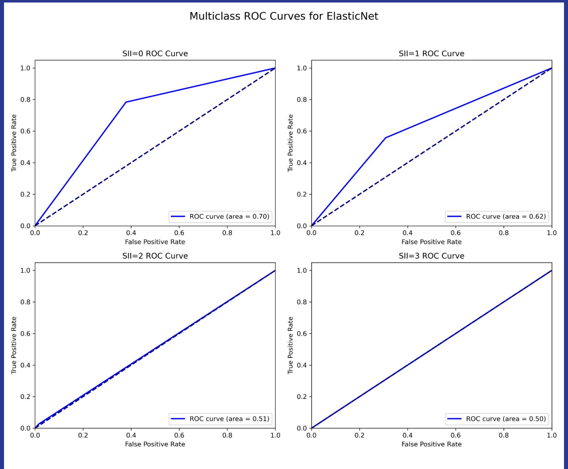


Neural Network

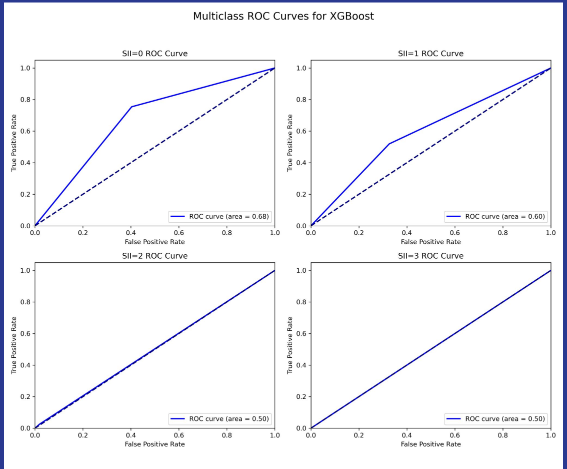


Results - AUC-ROC Curves

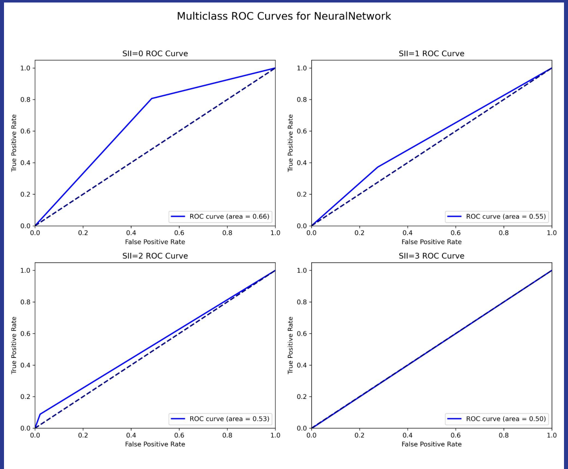
Elastic Net



XGBoost

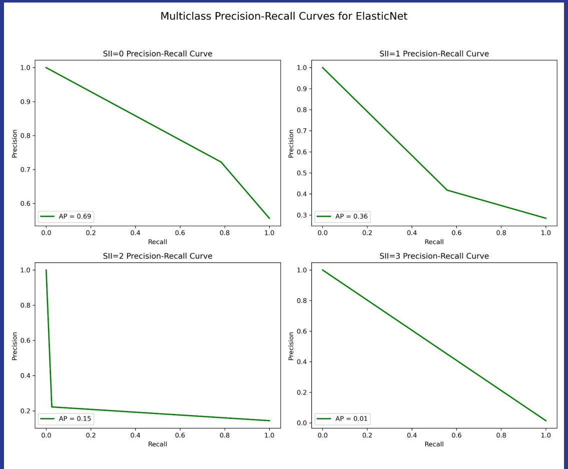


Neural Network

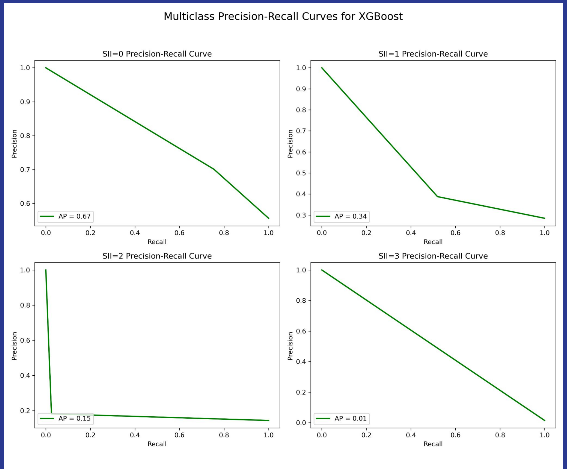


Results - Precision-Recall Curves

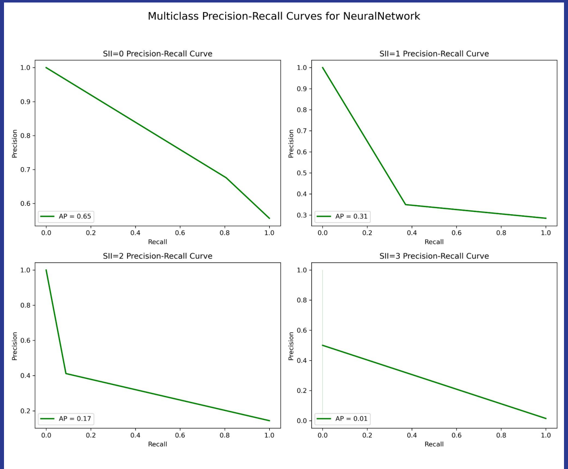
Elastic Net



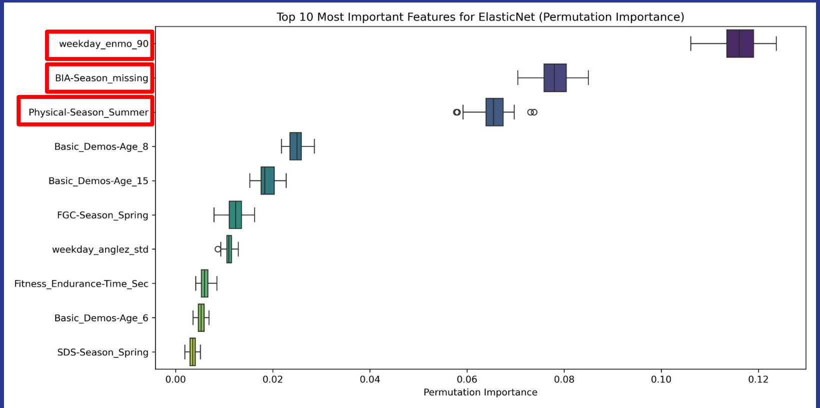
XGBoost

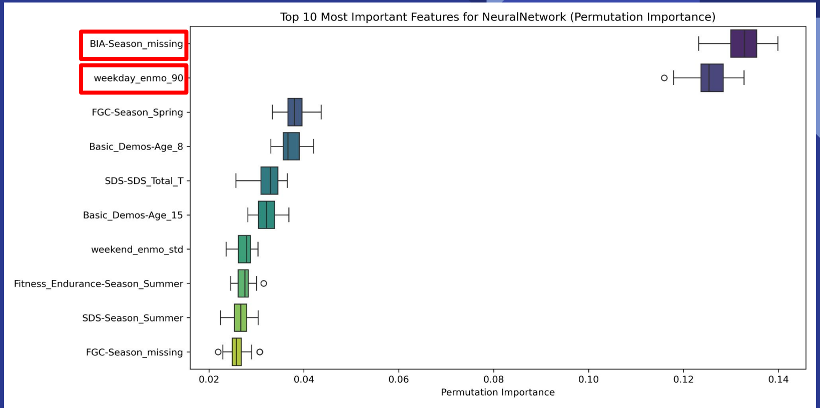


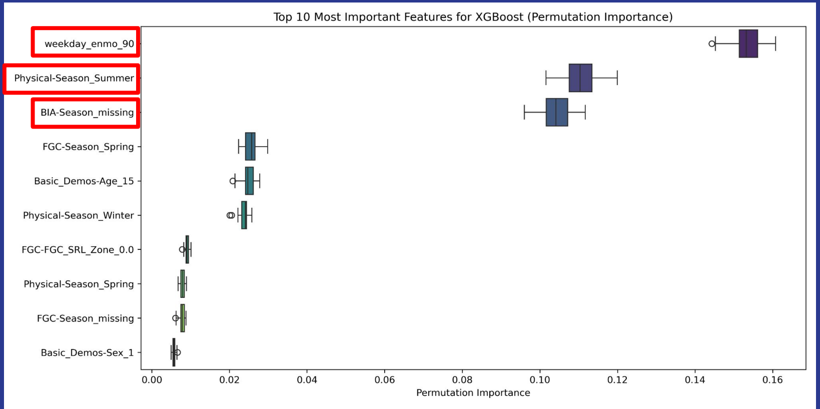
Neural Network

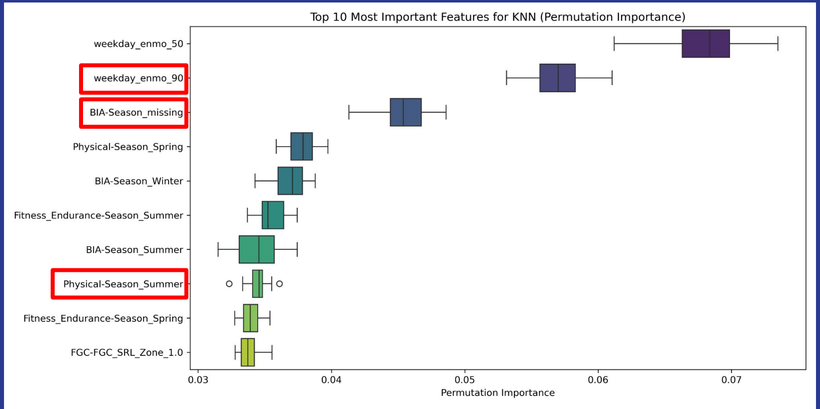


Interpretation - Permutation Importance









Outlook

- Expand parameter grid to maximize performance of each model.
- Try classification approach to allow class weighting to help with imbalance.
- Use SHAP to investigate local performance.