WiDS 2025 Datathon Report

Chang Lu

May 5, 2025

Abstract

This report describes machine learning pipelines developed to predict ADHD diagnosis and biological sex based on a combination of functional connectome features and subject metadata. After preprocessing and dimensionality reduction, multiple models were trained and compared using cross-validation. The best performing models were used to make final predictions on a test set.

1 Introduction

The WIDS Datathon project gives me data provided by the Healthy Brain Network (HBN), a signature scientific initiative of the Child Mind Institute, in collaboration with the Ann S. Bowers Women's Brain Health Initiative (WBHI), Cornell University, and UC Santa Barbara. The aim of this competition is to predict both the ADHD diagnosis and the sex of participants using a combination of functional brain imaging data and socio-demographic, emotional, and parenting information.

The dataset includes:

- **Target labels**: ADHD diagnosis (binary; 0=Other/None, 1=ADHD) and biological sex (binary; 0=Male, 1=Female),
- **Functional connectome matrices**: Time-series correlations from fMRI scans representing brain region connectivity,
- **Metadata**: Socio-demographic (e.g., race, site, education), emotional (e.g., SDQ scores), and parenting survey data.

My goal is to preprocess categorical metadata, merge them with the quantitative and functional features, and train predictive models. The final objective is to submit predicted values of ADHD diagnosis and sex on a test set for evaluation(By kaggle).

2 Data Overview

The training data includes:

- **Functional connectome features**: 19,900 continuous variables.
- **Quantitative metadata**: 18 columns, including psychological test scores.
- **Categorical metadata**: 9 columns, such as demographic indicators.
- **Targets**: ADHD_Outcome and Sex_F.

Missing data was handled by imputing medians for numerical variables and adding a `‘‘MISSING’’` level for categoricals.

3 Preprocessing Pipeline

The full preprocessing pipeline is as follows:

- Categorical: Impute with ‘‘MISSING’’, one-hot encode.
- Quantitative: Median imputation and standard scaling.
- Functional connectome: Median imputation, scaling, and PCA to 300 components.

`ColumnTransformer` and `Pipeline` from scikit-learn were used to modularize this process.

4 Modeling Strategy

The following models were evaluated:

- Logistic Regression (baseline)
- Random Forest with univariate feature selection (k=300)
- XGBoost with PCA
- Tuned Multi-layer Perceptron (MLP)

All models used 5- or 10-fold cross-validation. A subset of 600 rows was used for fast experimentation.

5 Model Comparison

5.1 ADHD Prediction (ROC-AUC)

```
Logistic Regression: mean=0.76855
RF                 : mean=0.82698
XGBoost+PCA        : mean=0.82966
MLP Tuned          : mean=0.56547
```

5.2 Sex Prediction

Model	Accuracy	Precision	Recall	ROC AUC
Logistic	0.7263	0.6148	0.5409	0.7547
RF	0.6768	0.8000	0.0769	0.7093
XGB+PCA	0.6735	0.5685	0.1995	0.6524
MLP Tuned	0.6760	0.5434	0.3462	0.6774

6 Final Model and Test Prediction

The best model for ADHD (XGBoost + PCA) and for Sex (Logistic Regression) were trained on the full training set. Final predictions on the test set were saved to `wids_datathon_final_submission.csv`.

7 Conclusion

Random forest and XGBoost models consistently outperformed simpler baselines for ADHD classification. For predicting biological sex, logistic regression gave the most balanced performance. Deep learning (MLP) underperformed, likely due to the small sample size relative to feature space.