Project report

Project Title: **Option B: Kaggle Competition: Child Mind Institute — Problematic Internet Use**

Prepared by: Thi Thu Huyen Nghiem

Date: 12/01/2024

Project background:

Child Mind Institute hosts this Kaggle competition to develop a model to predict problematic internet usage among children and adolescents based on their physical activity and fitness data based on some provided data. The Severity Impairment Index (sii) measures participants' problematic internet use. The quadratic weighted kappa, which measures the agreement between two outcomes, evaluates the model. The submission deadline for this competition is 12/19/2024.

My passion for applying machine learning in healthcare drives me to this competition. The potential for real-life impact is both thrilling and inspiring.

https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use

Data preprocessing & Feature Engineering:

Two elements of The Healthy Brain Network dataset (HBN), a clinical sample of about 5000 of 5 – 22-year-olds who have undergone both clinical and research screenings is used for this competition: physical activity data and internet usage behavior data.

There are slightly more girls and boys among participants. Age and sii distributions are pretty balanced between genders. Both are right skew. In both genders, most participants were 8 years old, and most of the sii was 0, equivalent to no impact. Besides that, the grading of parents on how their child is affected when spending online features are only available in the train data set. As shared by the competition, the target sii is derived from these fields. Therefore, these features are removed from the train data to avoid data leakage.

About 91% of features (74 out of 81) have missing values, accounting for 20 -80% of their data value. Therefore, it may need to fill in these missing data beforehand. KNN model is used to impute these missing values of numerical features.

Besides the available numerical features, two categorical features might be useful: physical measures by seasons of participation and a sleep disturbance scale by seasons of participation. Therefore, these two features are converted into indicator variables by get_dummies () methods.

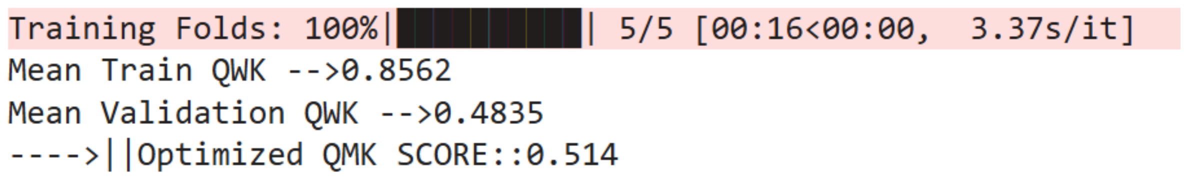
Other features might also important are:

- BMI_Age (BMI/Age) = Physical_BMI (Body Mass Index) / Basic_Demos-Age,
- Internet_Hours_Age = PreInt_EduHx-computerinternet_hoursday (Hours of using computer) x Age,

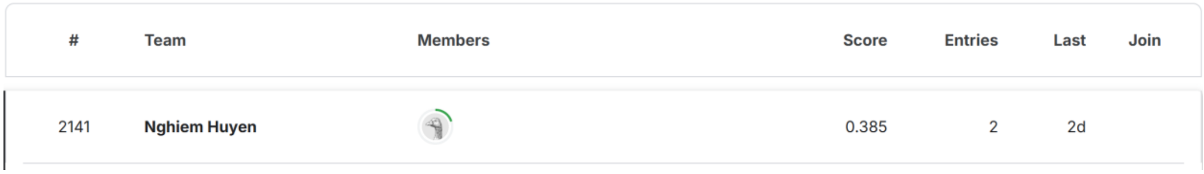
- PAQ_A_BMI (Activity score per body mass index in adolescents) = PAQ_A-PAQ_A_Total/BIA-BIA_BMI,
- PAQ_C_BMI (Activity score per body mass index in children) = PAQ_C-PAQ_A_Total/BIA-BIA_BMI,
- SDS_Age (Sleep disturbance by age) = SDS-SDS_Total_T * Basic_Demos-Age.

Model choice, validation, and performance:

- Model choice:
  - Light gradient boosting
  - XGBoost
- Validation: Stratified K-Fold Cross Validation
  - As the data has a right-skew distribution, so used Stratified K-Fold Cross Validation to ensure that each fold of the dataset contains approximately the same percentage of samples of each class as the complete set.
- Hyperparameter tuning: using Randomized Search CV to tune many hyperparameters
  - Light gradient boosting: n_estimators, learning_rate, max_depth, feature_fraction, bagging_fraction, bagging_freq, lamda_l1, lamda_l2.
  - XGBoost: n_estimators, learning_rate, max_depth, subsample, colsample_bytree, reg_alpha, reg_lambda.
- Performance:



```
Training Folds: 100%|██████████| 5/5 [00:16<00:00,  3.37s/it]
Mean Train QWK -->0.8562
Mean Validation QWK -->0.4835
---->||Optimized QMK SCORE::0.514
```



#	Team	Members	Score	Entries	Last	Join
2141	Nghiem Huyen		0.385	2	2d	

Conclusion:

Participating in this competition is an exciting opportunity to build machine-learning models and gain real experience in the field. Even if the performance is not satisfying enough, many things have been learned: how to deal with real-life data, build a model from scratch, make the model easier to follow, etc. However, there are still many challenges that need to be improved to create a good model, especially in this competition:

- Dealing with an extensive data set with many features, imbalanced distribution, and missing values. Using KNN to impute the missing values but not validate them might cause a later bias in model performance.
- Lacking knowledge in the medical field also limits the ability to add insightful features when doing feature engineering.
- Lacking experiments using actigraphy data, a method of monitoring human rest/activity cycles, to improve the model's performance.
- Both LightGBM and XGBoost are ensemble learning using a boosting algorithm. Therefore, they share the same pros and cons with the data set. Other machine learning models should be applied to utilize other advanced algorithms.

Besides that, some questions that need to address to improve the performance:

- For this competition:
    - How to add and utilize actigraphy data?
    - What else can be done to improve the performance of the LightGBM and XGBoost models?
    - Will the neuro network work better in this case?
- In general, when building a model from scratch:
    - Shall I start from the less complicated model, such as linear regression, and then increase the level of complication if the performance is not good enough?
    - Or shall I start with some model in the middle of the complication ladder, and then whether to move up or down on the model complication ladder depends on how the model performs?