AI RUSH 2nd place solution summary

@TeamMDL

2nd place

Final Score

Rank	Name	Score
1	NYT	900
2	TeamMDL	770
3	Cheat_Key	640
4	BTSiit	560
5	team_tsubame	500
6	Believe My Boss	480
7	kagglers	460

Contents

- Preprocessing
- Classifiers and method
- Ensemble

Preprocessing

・Raw features

hh, age_range, category of the article

・Created features

read_num,read_num_catX, historyX_category
,id_X
ーthese are created by read_article_ids and article_id

・The number of total features is 48, and 33 features are created from read_article_ids.

・Note that we did not used image features and title. Image features did not played good role for our model, and we did not know how to transform the title into features.

read_num

- How many articles did the user read. (length of read_article_ids)
- It is expected that the user tend to click the article if the user read many articles.

read_num_catX

- The numbers of articles which belong to each category.
- This feature indicates the users preference. If the article’s category is close to the user’s preference, he tends to click it.

historyX_category

- Xth latest category which the user saw. X is from 1 to 10.
- If the tendency of history is similar to the article, the user tends to click it.

id_X

- Xth number of article_id. X is from 1 to 12.
- Assume that 981029012232 is the article_id. In this case, id_1 is 2, id_2 is 3, etc.
- Actually we are not sure why these features work.

Preprocessing detail

- Applied PCA to image features, however it doesn’t worked well for our model.
- We also created and tried many other features and validated. The features explained before performed well.

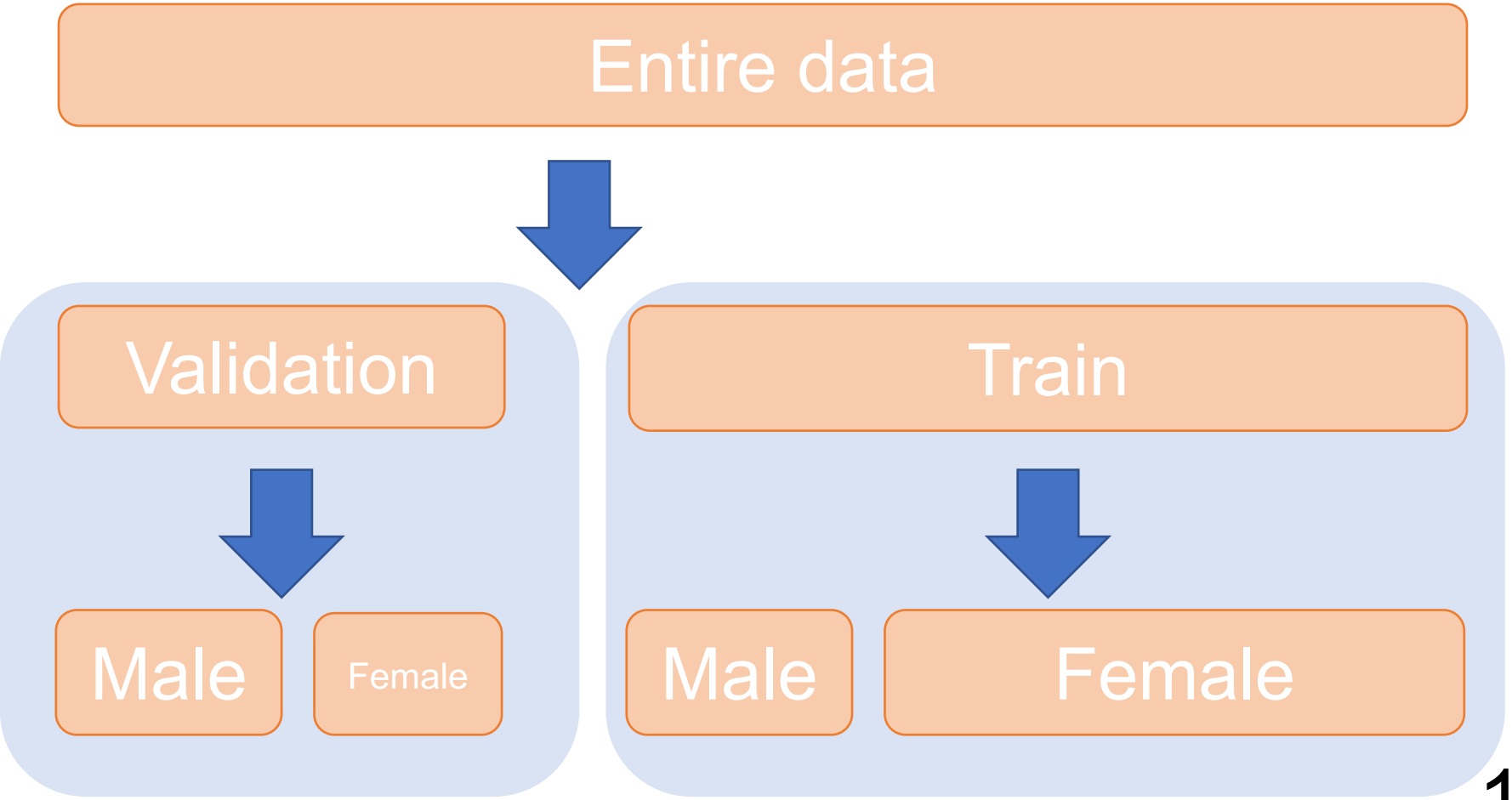
Classifier

Used: XGB LGBM

- We also tried CatBoost but didn’t work well.
- Used optuna to tune the hyperparameters.

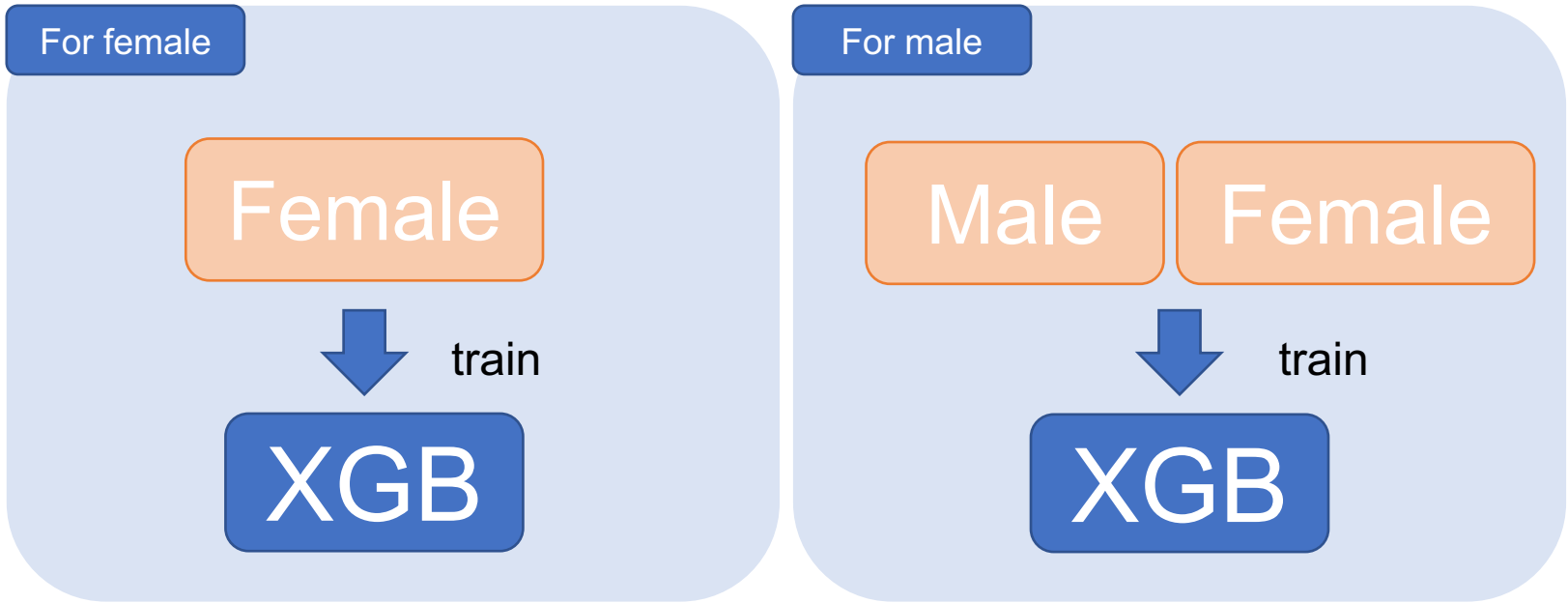
Method

- First, split the data.



Make two classifiers

- Second, make two kinds of classifiers. One for female, the other for male.



- This is based on the intuition that interesting articles for men and women are different. Ex: women tend to be more interested in parenting than men.
- Note that male classifiers are trained the whole training samples because the male samples are a little few.

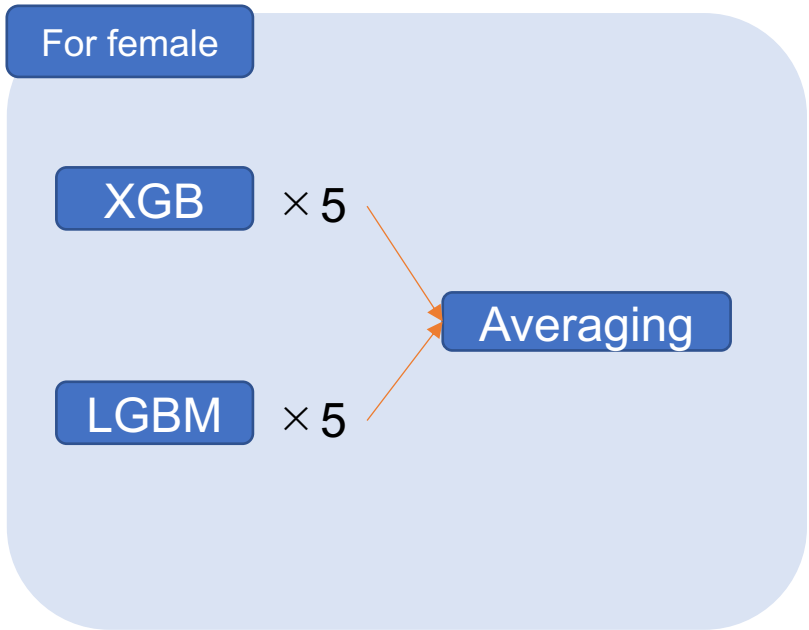
Training detail

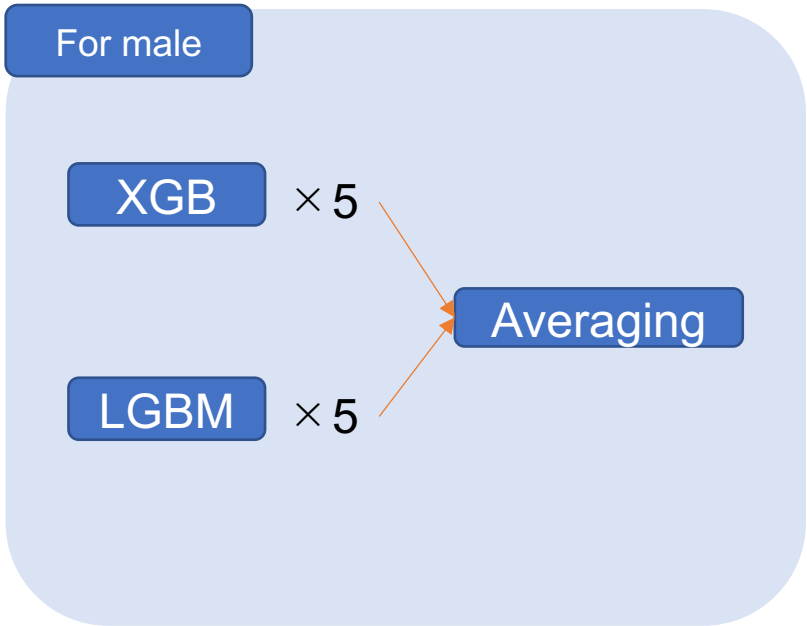
- Used under sampling to address the imbalanced nature of the data.
- Optuna a little improved the classifiers.
- Tried probability calibration but it spoiled the score.
- Tried transforming the article_id by hash function. This was not bad but transformation by categories was better.
- Tried simple neural network

(3 layers) but almost all its outputs are 0s.

Ensemble

- Prepared 5 XGBs and 5 LGBMs for each gender, so the number of total classifiers is 20. And averaged 10 classifiers' output for each gender.





Ensemble detail

- We also tried logistic regression and hard voting. Logistic regression fairly lowered the score. Hard voting performed well however the averaging was a little better than it.

Final score

Rank	Name	Score	Recorded	Count
1 -	NYT	0.2558839627805145	6 days ago	47
2 -	TeamMDL	0.2464982711129344	3 days ago	76
3 -	Cheat_Key	0.24581491412421191	3 days ago	78
4 -	BTSiit	0.24457268480351743	3 days ago	60
5 -	kagglers	0.23925949569103094	3 days ago	51
6 -	Conundrum	0.23919952913478515	5 days ago	7
7 -	Believe My Boss	0.2385804888477161	3 days ago	67

Appendix: 1st round summary

- We took 4th place.
- Trained Vgg, Inception, ResNet50, and ResNet18 respectively.
- Data augmentation by gray scaling, horizontal flipping, and random cropping.
- Ensembled them by simply averaging the outputs. Note that the implementation is fairly difficult because of NSML usability...