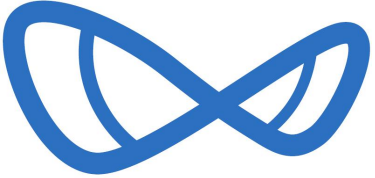


An Effective Ensemble Framework with Multichannel Time Series for User Retention Prediction

Zhiruo Li, Zhihui Cui, Shunyao Wu

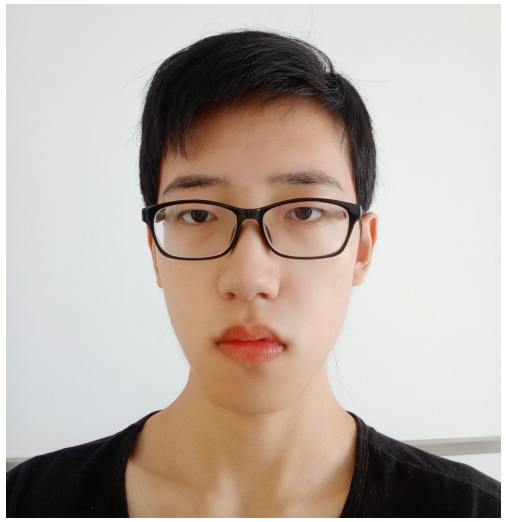
Qingdao University

Feb 25th, 2022

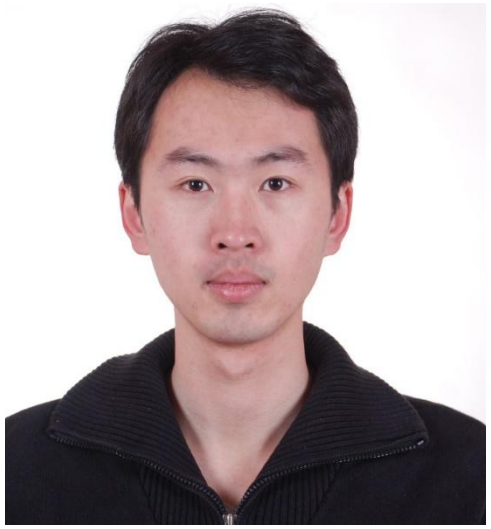
QDU Team Profile

Bioinformatics Group
Qingdao University

Zhiruo Li
Graduate Student
at Qingdao University,
major in applied statistics



Zhihui Cui
Pre-final Year Student
at Qingdao University



Shunyao Wu
Assistant Professor
at Qingdao University

QDU achieved top rankings in 5 data mining competitions, e.g.

- ✓ Ranked 1st in *CIKM AnalytiCup 2019*
- ✓ Ranked 6th in *KDD CUP 2020*

Problem Understanding

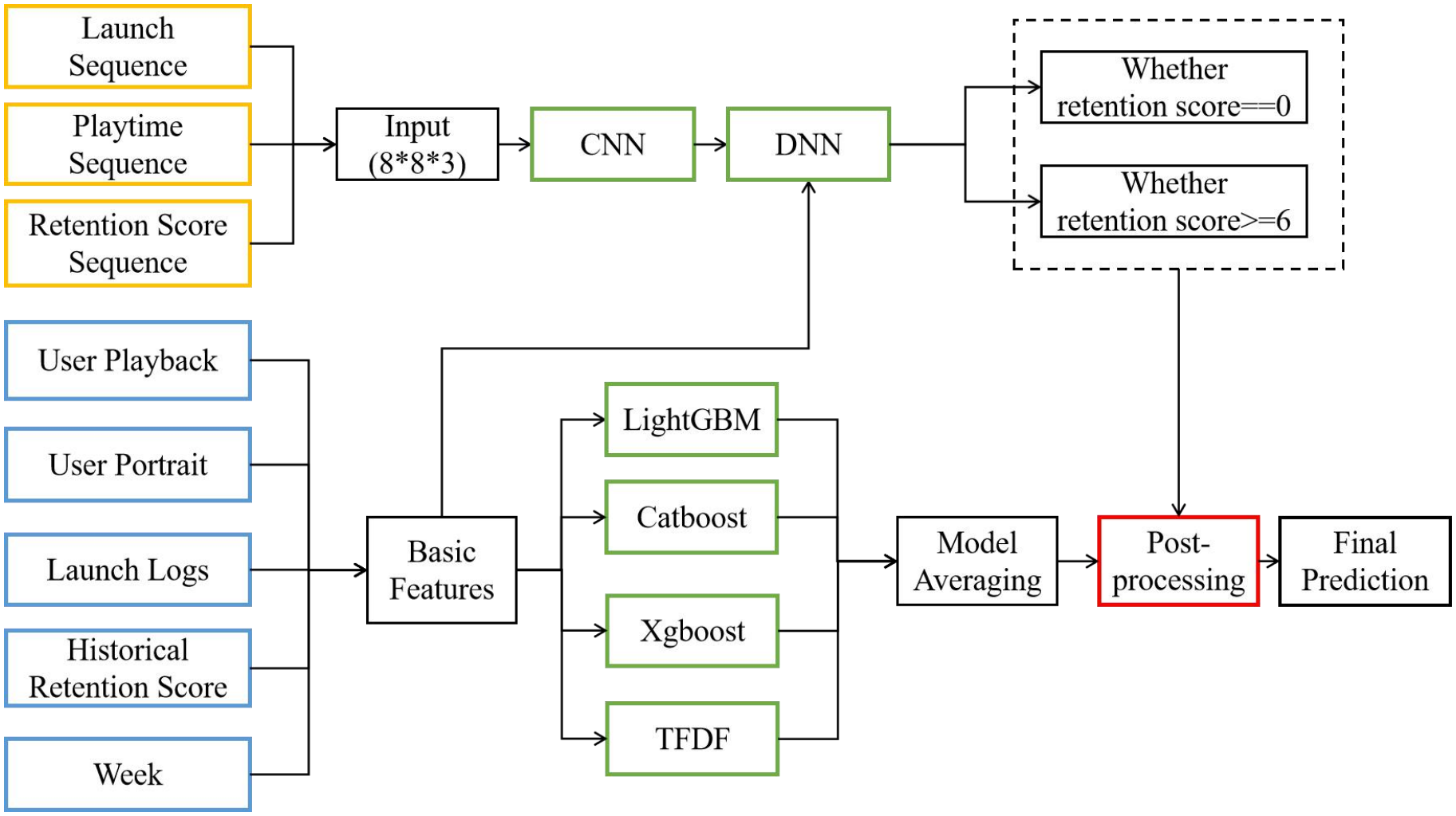
- **Task:** Predict the 7-day retention score of any user
- **Evaluation Metric**

$$Score = 100 * (1 - \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{7})$$

F_t : forecast value, A_t: actual value

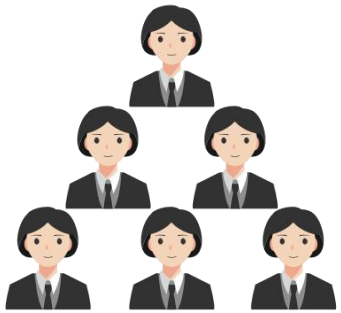
- **Data**
  - user playback data
  - user protrait date
  - video related data
  - app launch logs
  - user interaction data

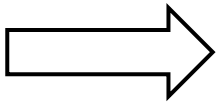
Our Framework

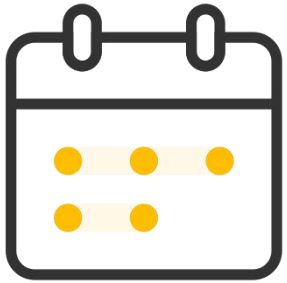


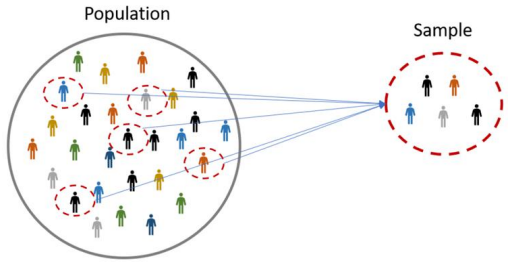
Bioinformatics Group
Qingdao University

Preprocessing









Training Set

Select Users

only users from the online test set

Extend Dates

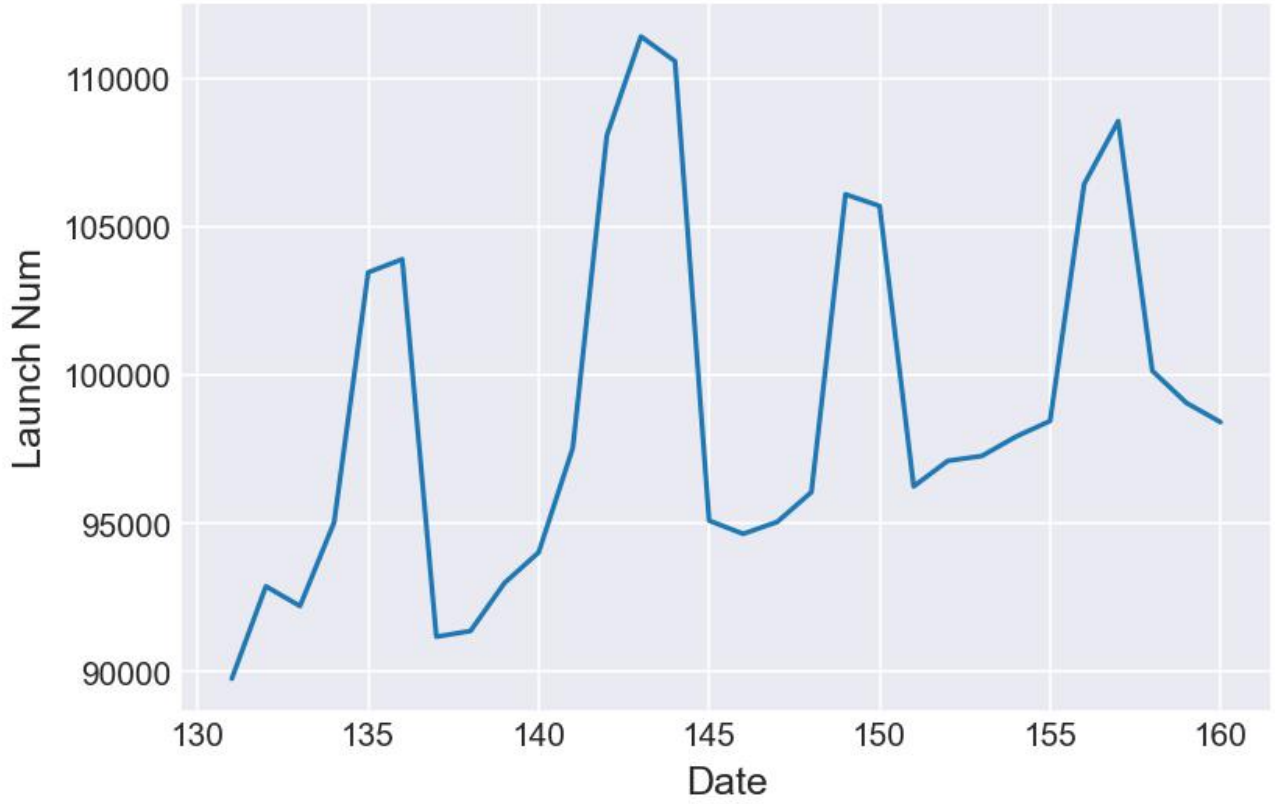
all available dates, i.e. $[date\_min, date\_max - 7] \cup [131, 153]$

Sampling

70% of Retention score 0

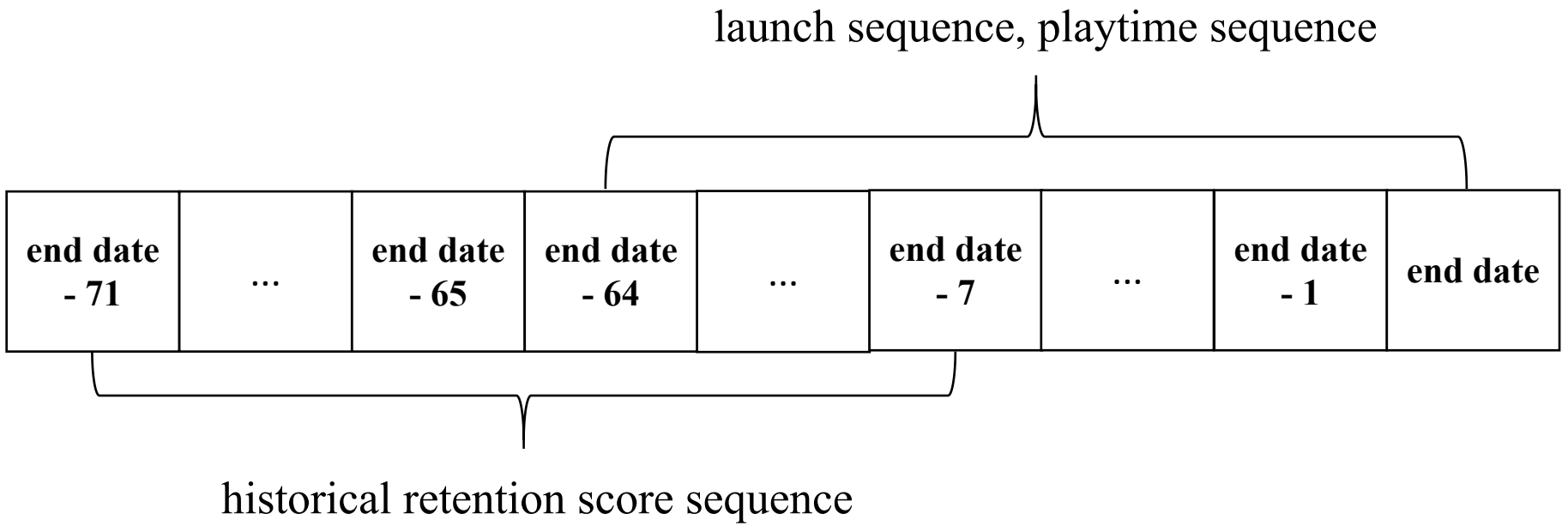
Exploratory Data Analysis

- Cyclical pattern, every week
- Example
  - ✓ date 143: Saturday
  - ✓ date 144: Sunday
  - ✓ date 145: Monday



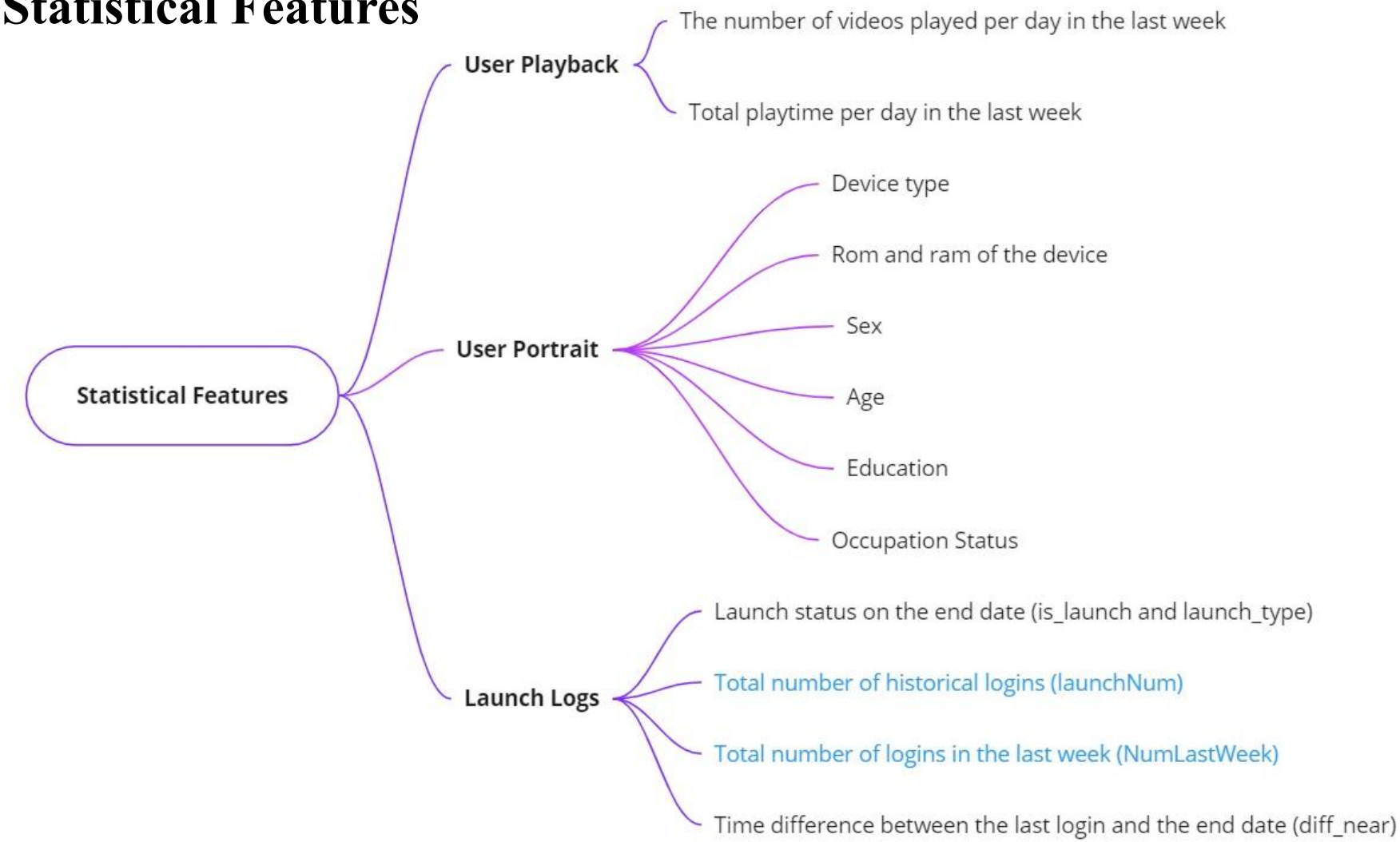
Feature Engineering

➢ Multichannel Sequences



Feature Engineering

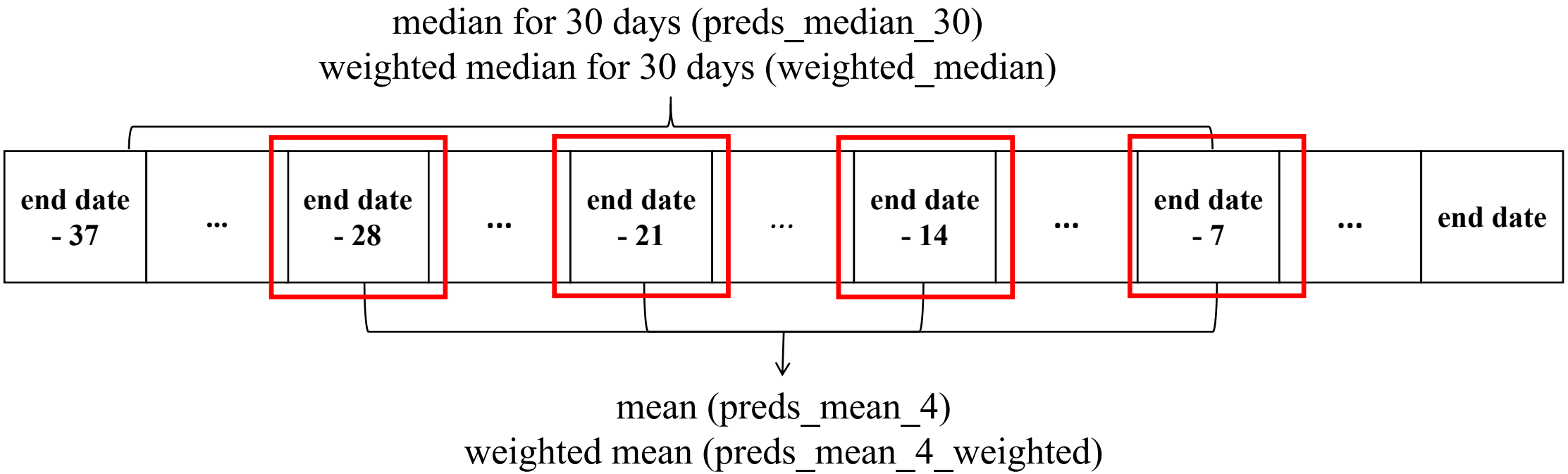
➢ Statistical Features



Feature Engineering

➢ Statistical Features

✓ Historical Retention Score



Feature Engineering

➢ Statistical Features

✓ Historical Retention Score

Feature Name	Stage A	Stage B
preds_mean_4	84.522	84.581
preds_mean_4_weighted	**84.994**	**85.013**
preds_median_30	83.778	83.906
weighted_median	83.123	83.268

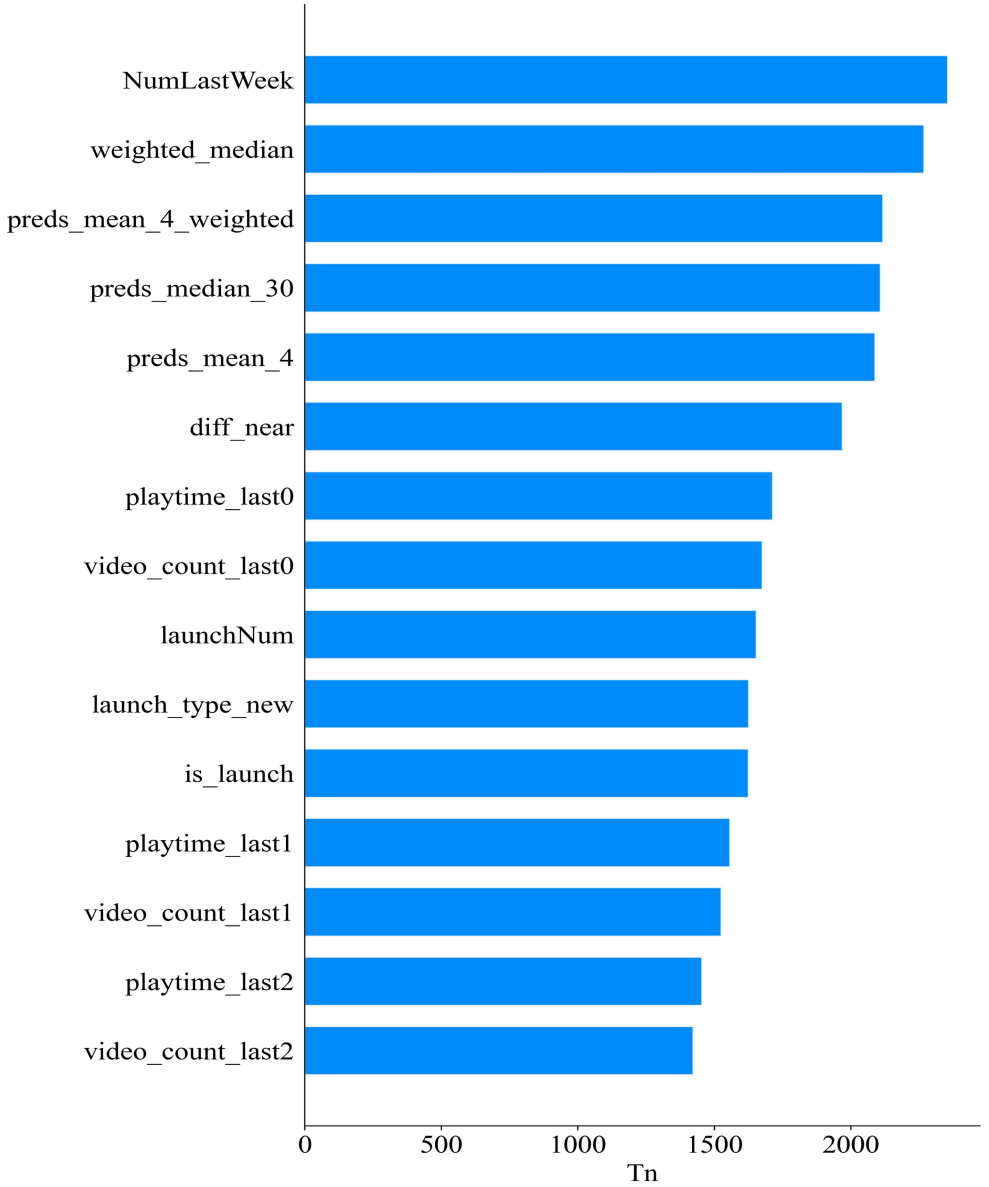
Feature Engineering

Bioinformatics Group
Qingdao University

➢ Feature Selection

- ✓ Mean variance test (MvTest)
- ✓ Distribution free test of Independence
- ✓ Test statistic:

$$T_n = n\widehat{MV}(X|Y)$$
$$= \sum_{r=1}^{R} \sum_{i=1}^{n} \widehat{p_r} * [\widehat{F_r}(X_i) - \widehat{F}(X_i)]^2$$

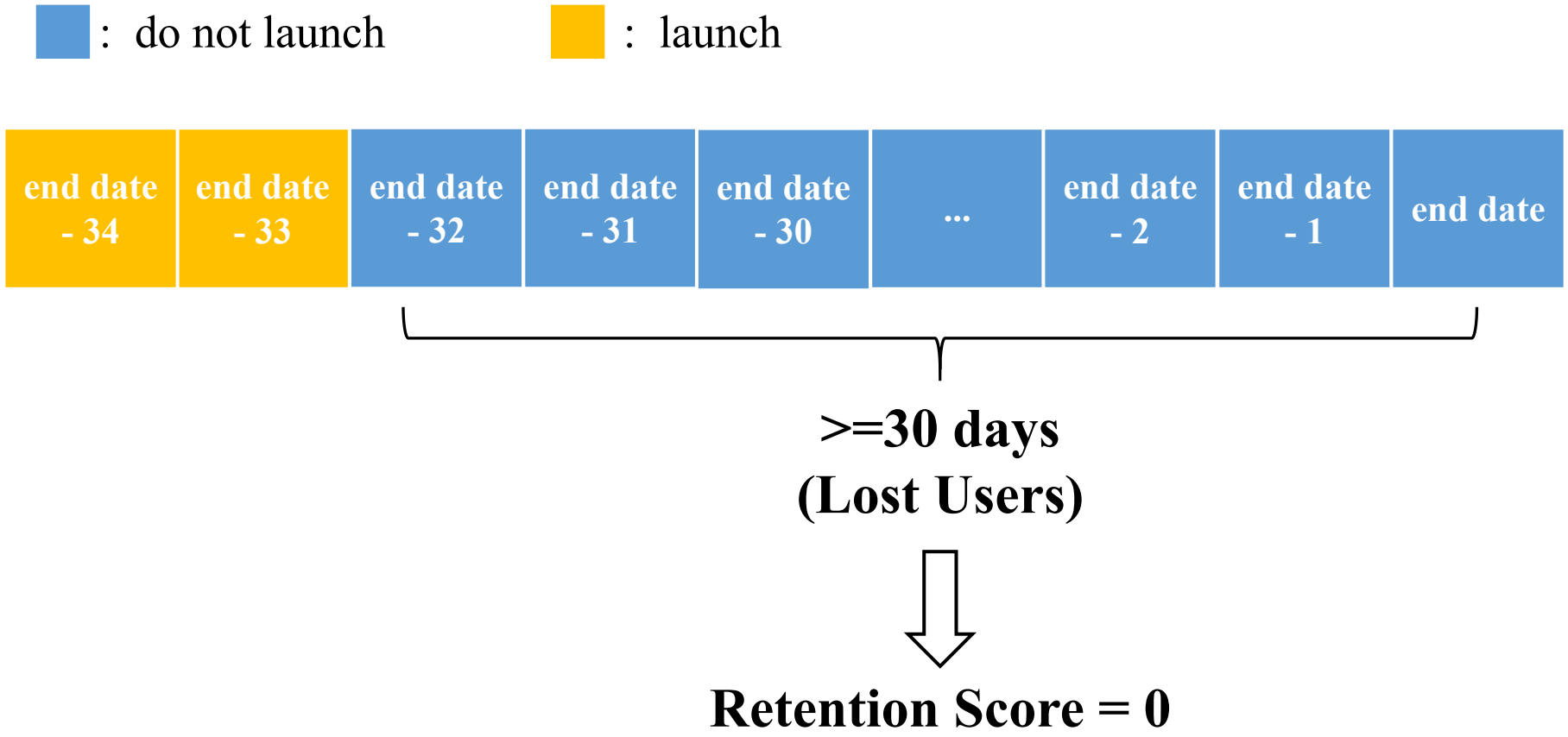


Model Averaging

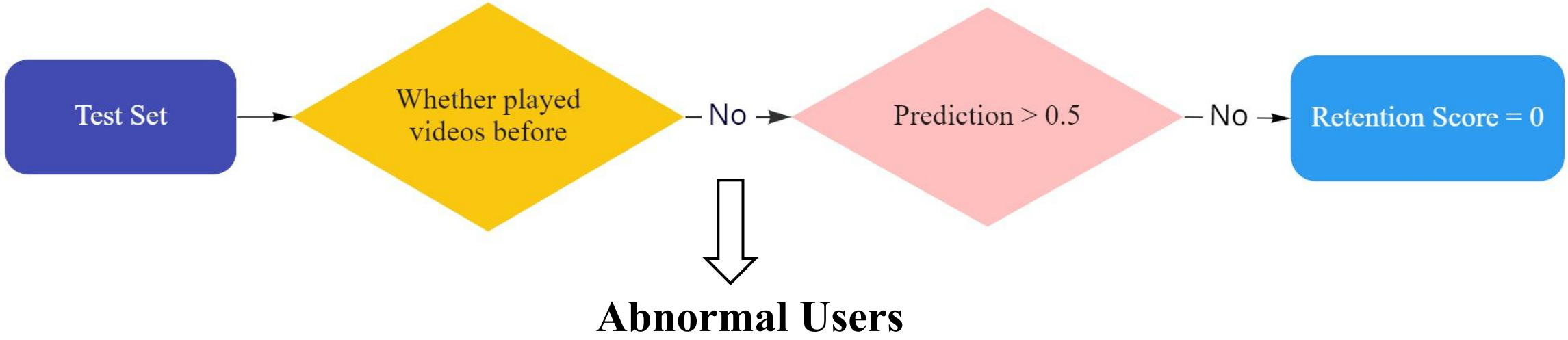
- Assemble 4 tree-based models by the harmonic mean
  - ✓ LightGBM
  - ✓ Catboost
  - ✓ Xgboost
  - ✓ Tensorflow Decision Forest (TFDF)

$$\hat{y} = \frac{4}{\frac{1}{\hat{y}_{lgb} + 0.00001} + \frac{1}{\hat{y}_{cbt} + 0.00001} + \frac{1}{\hat{y}_{xgb} + 0.00001} + \frac{1}{\hat{y}_{tfdf} + 0.00001}}$$

Post-processing

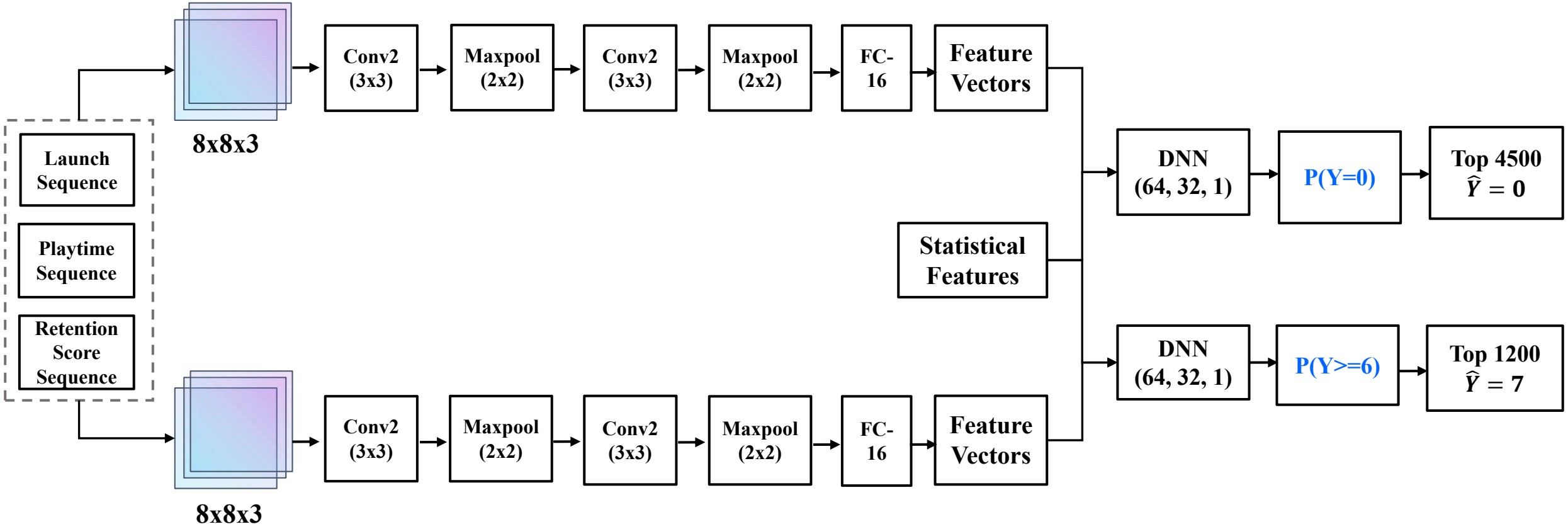


Post-processing



Post-processing

- CNN models, adjust predictions

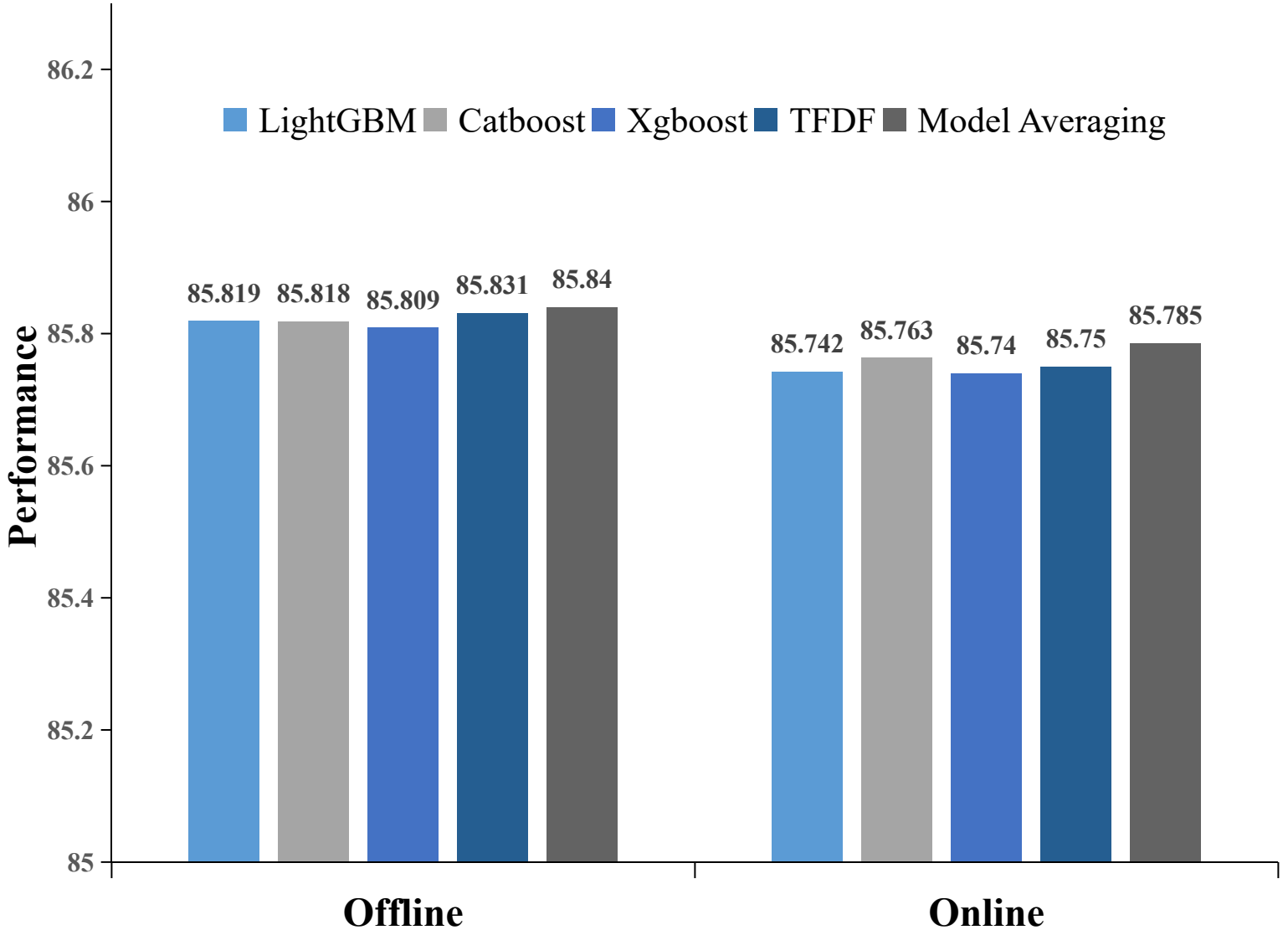


Post-processing

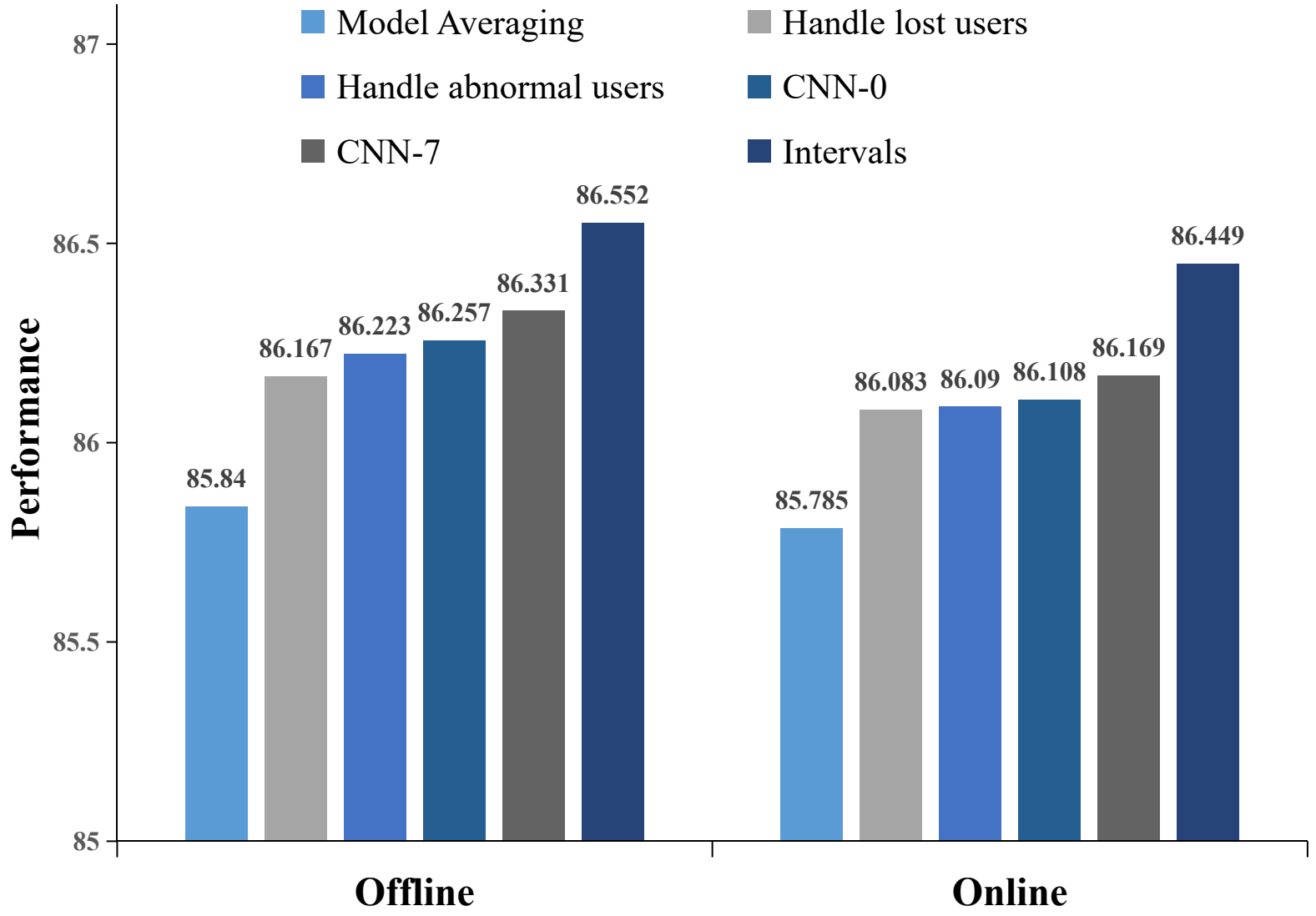
- Adjust the predictions by intervals

Interval	Adjusted Prediction
(0, 0.5)	0
(0.6, 1.4)	1
(1.55, 2.4)	2
(2.55, 3.4)	3
(3.55, 4.4)	4
(4.55, 5.2)	5
(5.55, 6.2)	6

Performance



Performance



Thank you !

If you would like to join us for a master's or doctoral degree, please contact the Founding Director Prof. Xiaoquan Su (suxq@qdu.edu.cn).