Ensemble of Diverse Gradient Boosting Decision Trees for MOOCs Dropout Prediction

Team: kyazuki&DT@Keio univ. Ohmori Lab

Ikki Tanaka: ikki0407@gmail.com

Shunnosuke Ikeda: know-knew-known@softbank.ne.jp

Our Team

Ikki Tanaka

- 1st year Masters program @Keio univ.
- Research: Wind Speed Prediction for EMS
- Like: Data Analysis, Cycling

Shunnosuke Ikeda

- 4th year undergraduate
- Research: Control

Our Team

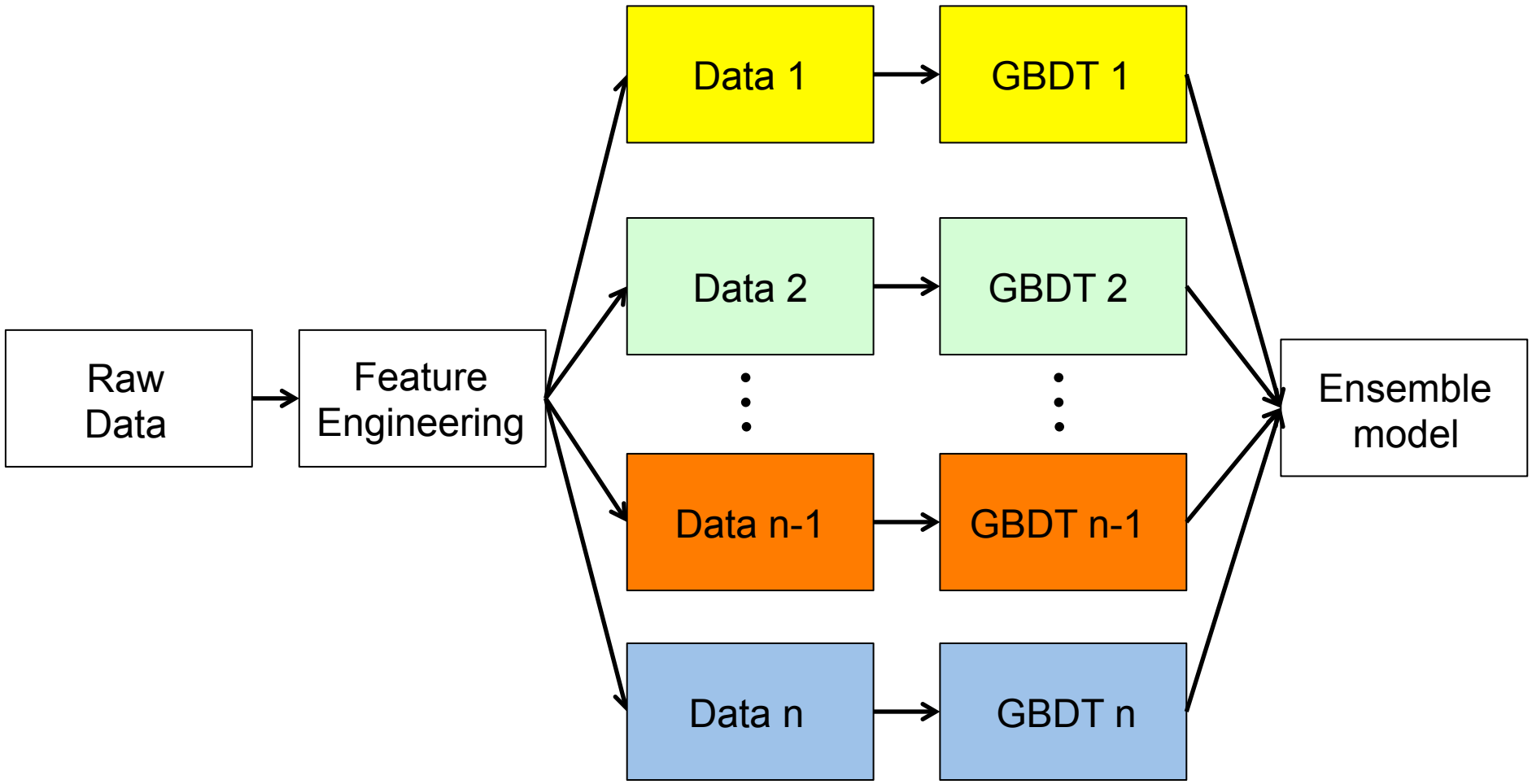
- PC and environment
  - 2 PCs (MacBook Pro Late 2013)
  - Python and R
- Main role of our team
  - Programming and Idea(Ikki)
  - Idea(Shunnosuke)
- Pace producing new models
  - One feature per a day
  - One model takes 8 hours
- How to come up with ideas
  - Visualization, Looking at raw data, Feedback

Python encounters in AUS

Index

1. Flow of Prediction
2. Feature Engineering
   - Basic features
   - Effective features
3. Visualization
4. Prediction Methods
   - Deep Learning
   - Logistic Regression
   - Factorization Machine
   - Gradient Boosting Decision Tree
5. Ensemble
6. Knowledge
7. Conclusions

Overview of Prediction



Feature Engineering

Basic Features

count

- course_id
- event
- username
- source
- object
- category
- module depth
- dropout

etc

Dummy Variables

- course_id
- hour0～23
- day of the week
- 10/2013～08/2014

etc

time

- login time(max, min, mean, std, max – mean, mean – min)
- Interval of login time (max, min, mean, std)
- three divided periods of courses

etc

Effective Features

Groupby username

Sum
mean
std

Within 10 days after course end

- count of courses enrolled by username
- event of courses enrolled by username
- length of overlapping courses enrolled by username

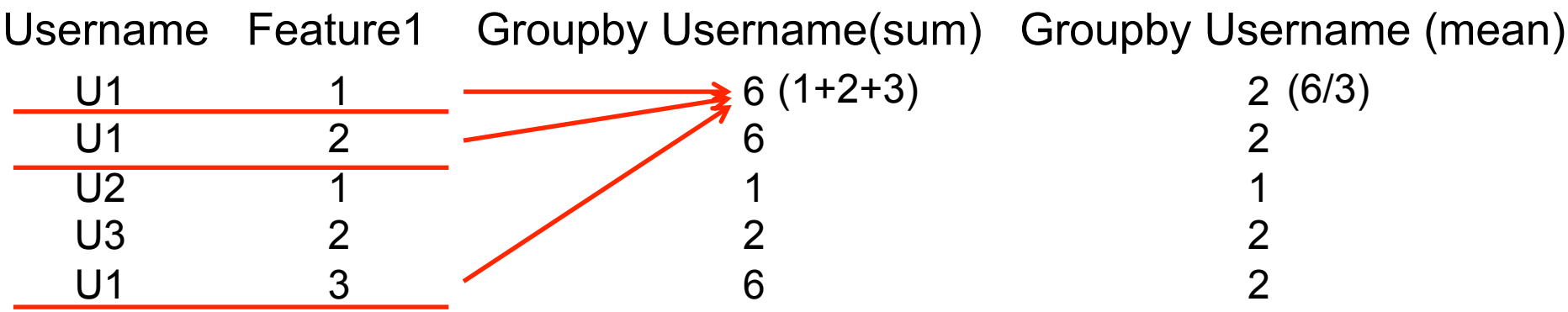
etc

Basic Features

Name	Feat. No	Feature Description
course_id	Feat. 1～39	Dummy variables of 39 courses
weekday	Feat. 123～129	Dummy variables of weekday
count of module depth	Feat. 136～140	Mean of the count of the module depth in object.csv
chapter count	Feat. 501～696	Count of the object whose category is chapter in log data
past non drop rate	Feat.740	Mean of the past non-dropout rate of username
interval of login	Feat.1105～1108	Mean of the login interval(Max, Min, Std)

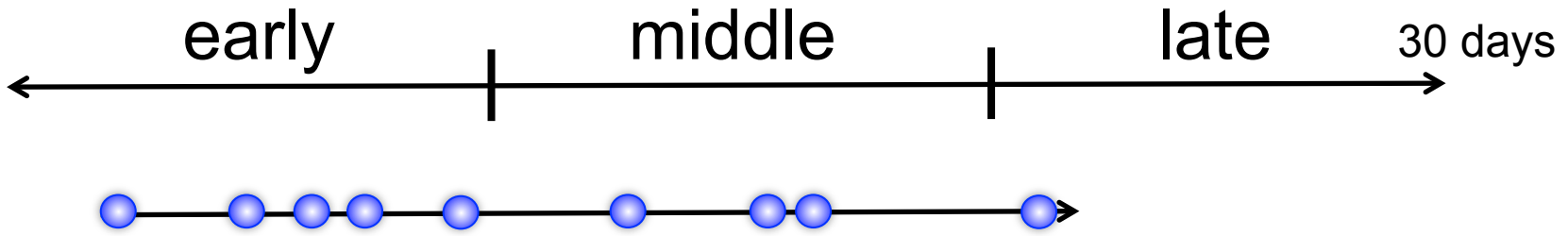
Features based on username

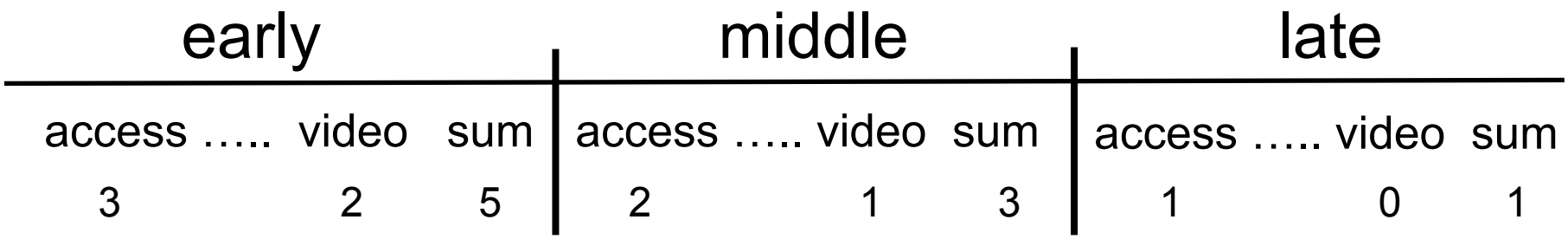
- Sum of basic features by username
- Mean of basic features by username
- Std of basic features by username



Divide course periods into three

- The count of the event within divided periods
- The sum of the event within divided periods



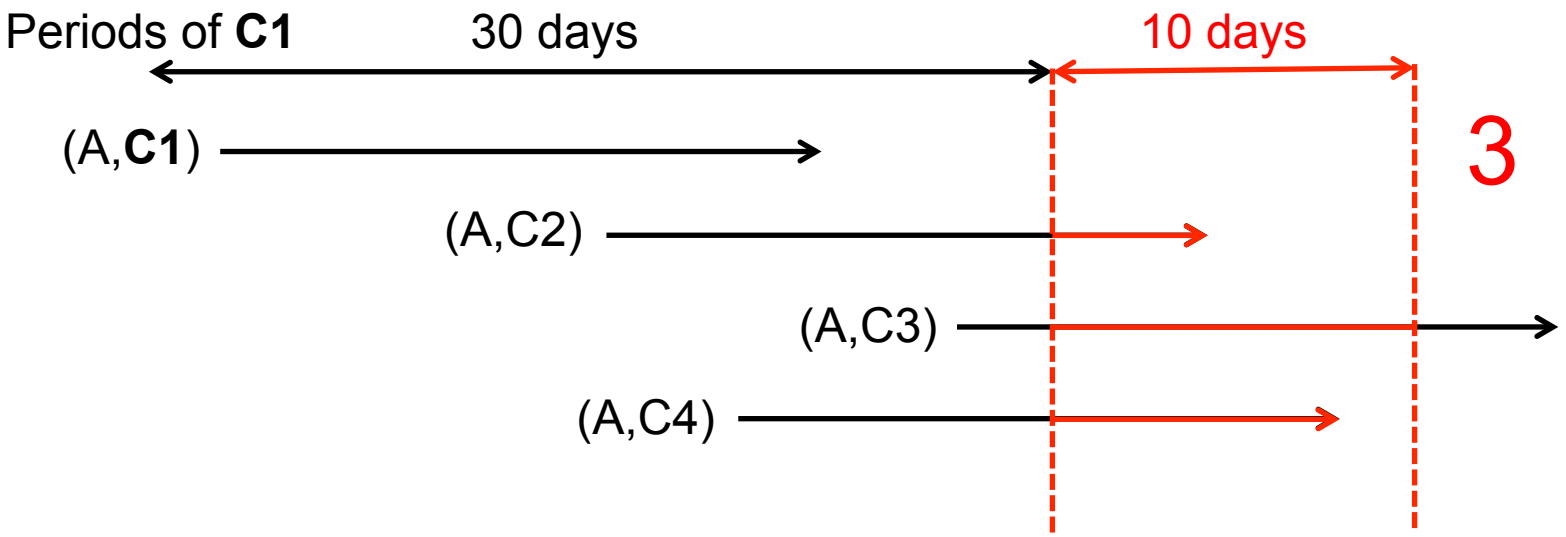


Features for 10 days

"10 days" means the periods of the dropout definition

- Count of the enrolled courses by same user for 10 days

Ex) Let username "A", course1 be (A, C1)



Variations of features for 10 days

- Count of the enrolled courses for 20 days

- Count of the access day of the enrolled courses for 10 days

1	2	3	4	5	6	7	8	9	10	Sum
0	0	1	1	1	1	0	0	1	1	6

- Mean of the access day of the enrolled courses for 10 days(Max, Min, Std)

Max	min	mean	std
10	3	6.16	2.54

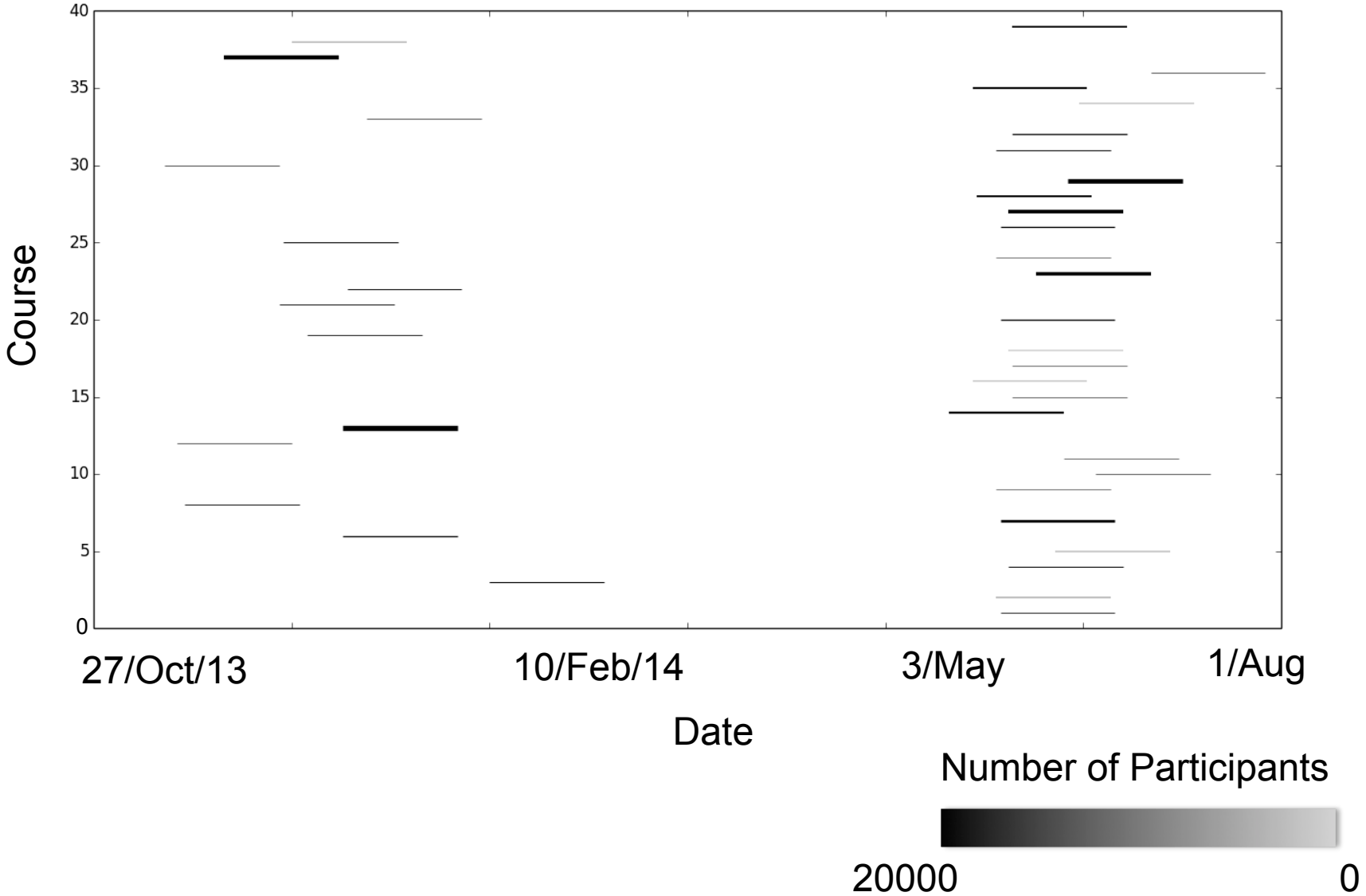
- Count of the event of the enrolled courses for 10 days

access	.....	Video	Sum
5		2	25

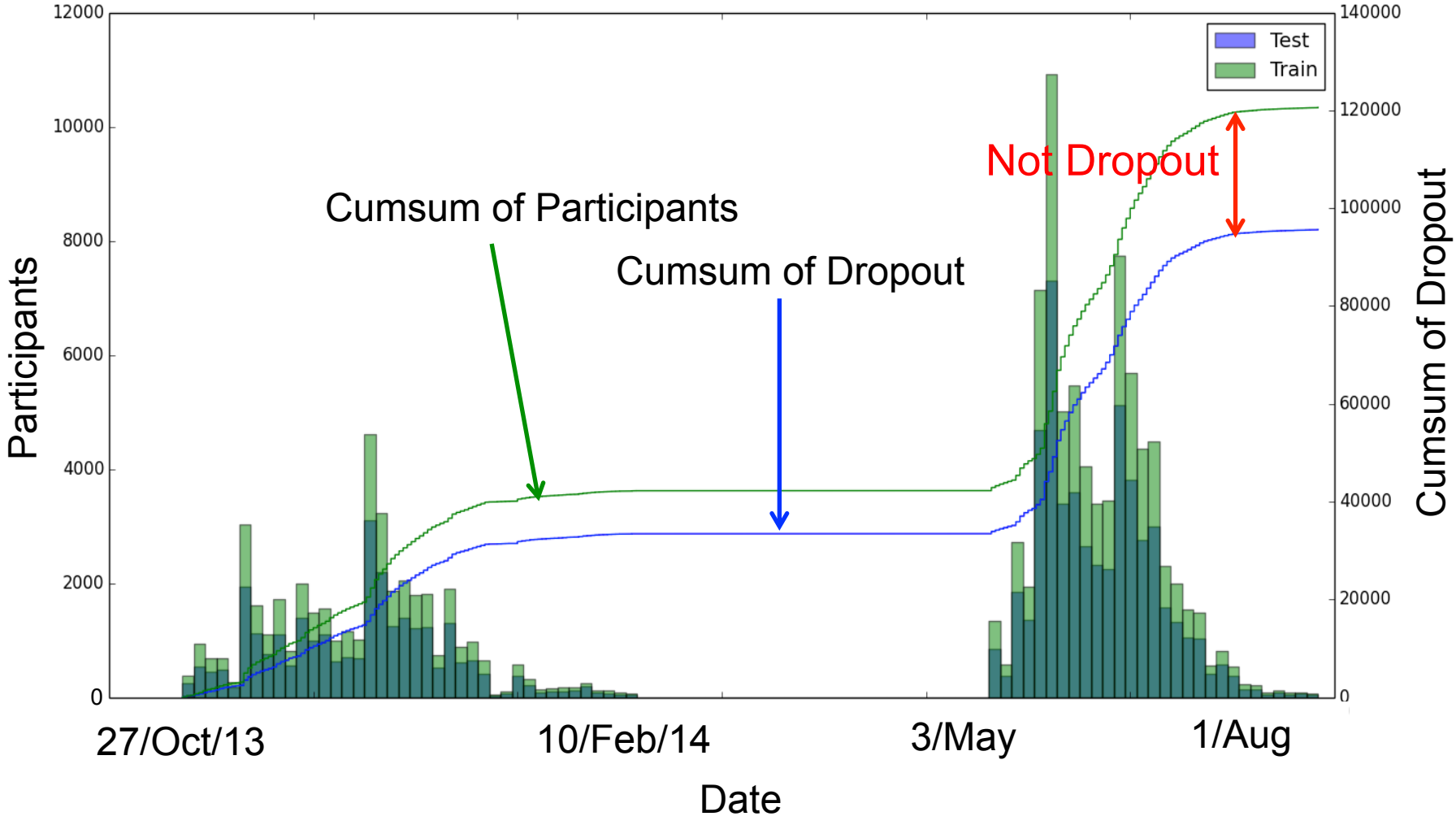
- Percentage of the event of the enrolled course for 10 days

access	.....	Video
0.2	.....	0.08

Periods and Participants of Courses



Similar distributions between train and test



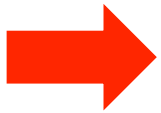
Prediction Methods

- Deep Learning

  Best NN was single layer NN

  - 3,4 layers
  - Dropout, minibatch, ReLU, …
  - Lasagne, caffe

- Logistic Regression
  - Scikit-learn
  - Vowpal wabbit (feature interaction)



Not Good

0.89XX-0.90XX

- Factorization Machine
  - LibFM, LibFFM

Great Model

- Gradient Boosting Decent Tree
  - XGboost
  - Random search for parameters
  - 5-fold cross validation

- Parameters

  max_depth: 7

  num_round: 1550

  eta: 0.01

  subsample: 0.59

  colsample_bytree: 0.84

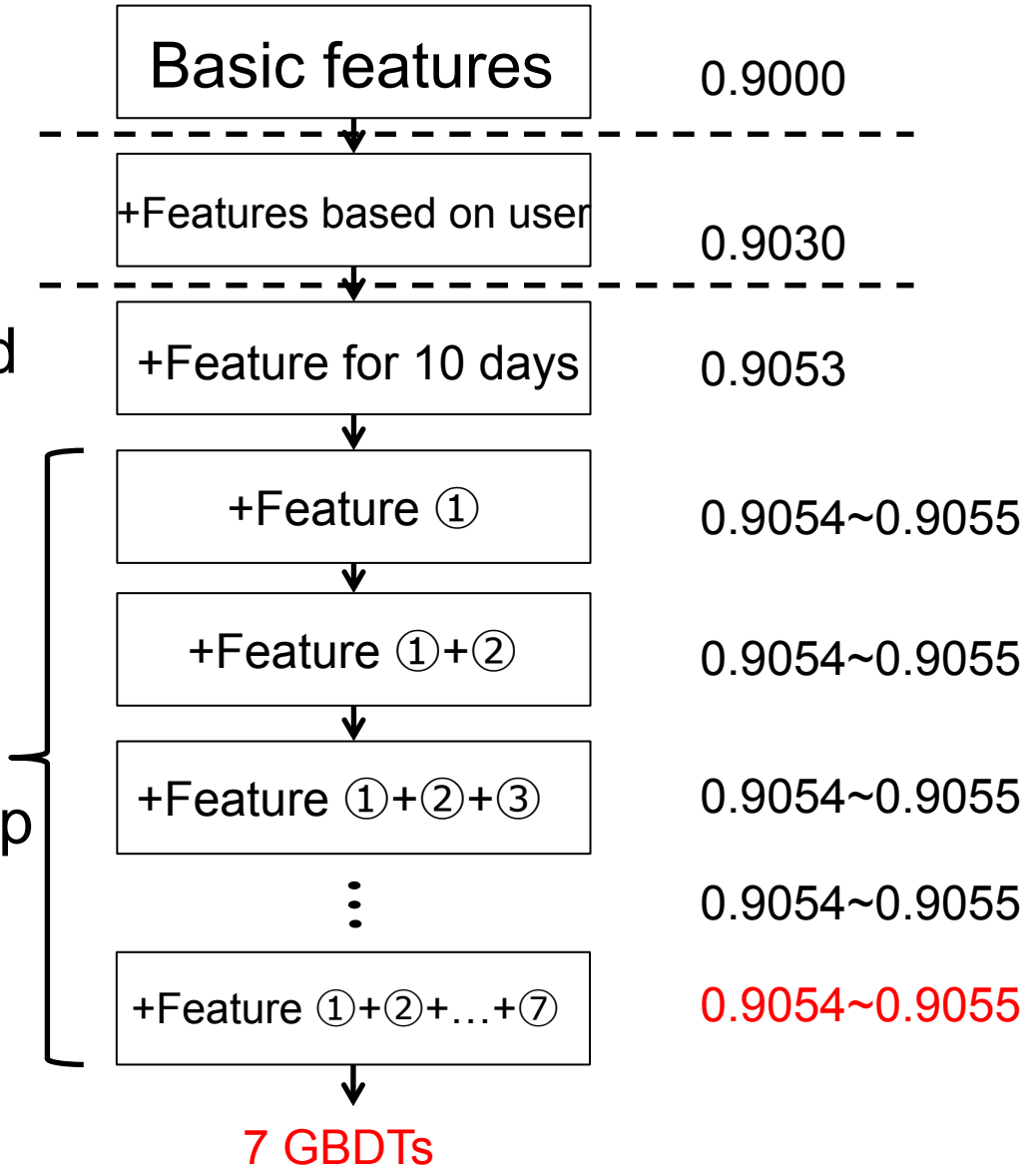
  gamma: 2.79

  min_child_weight: 3

First Flow chart

"Feature for 10 days" means count of enrolled courses for 10 days

features①-⑦ are noticed just before the end date of KDDCup



Features just noticed before end date

①Count of the enrolled courses for 20 days

②Count of each day of the enrolled courses for 10 days

③Mean of the day of the enrolled courses for 10 days (Max, Min, Std)

④Count of the Sessions of the enrolled courses for 10 days

⑤Percentage of the sessions of the enrolled courses for 10 days

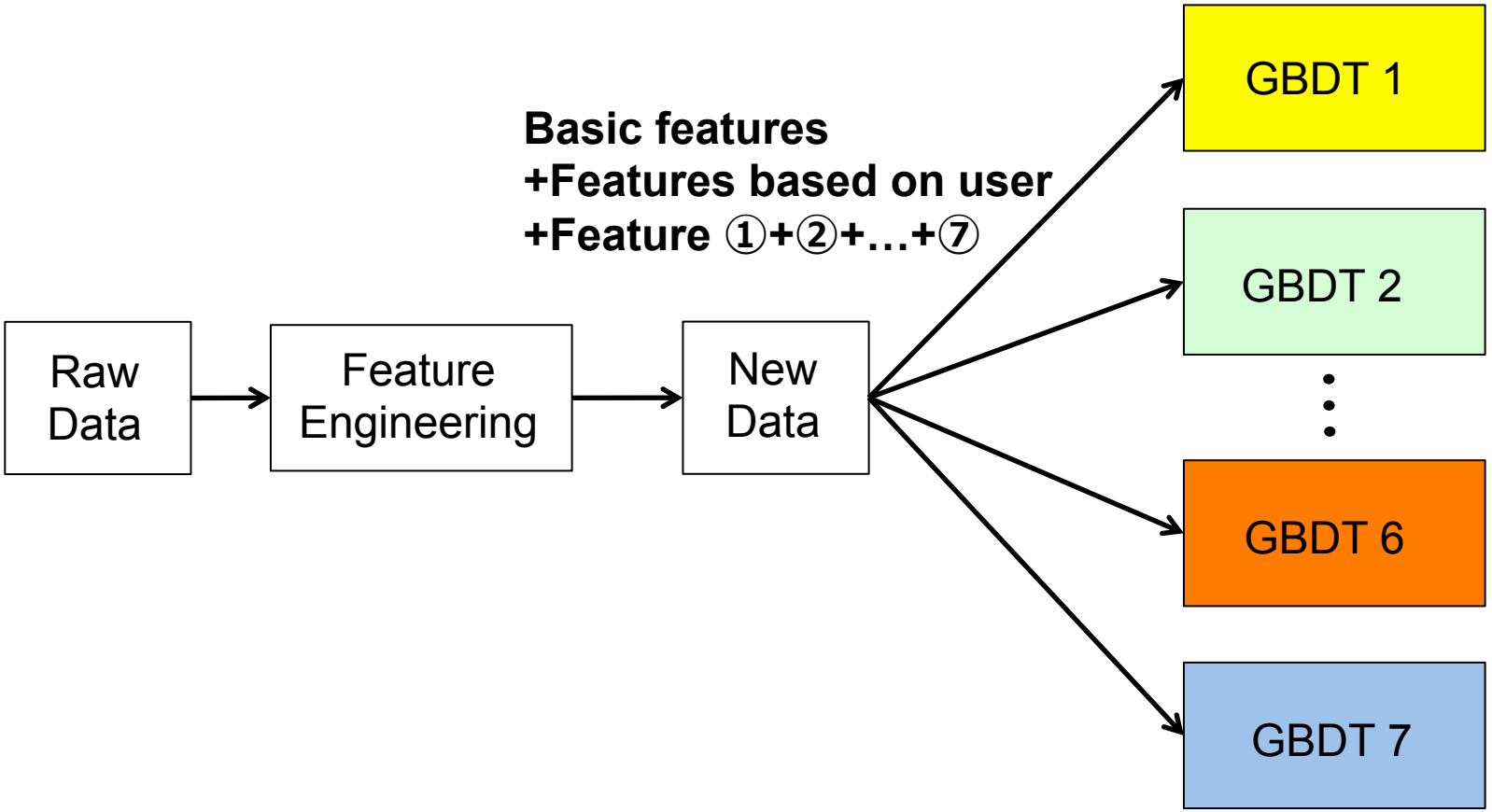
⑥Mean of the session interval(Max, Min, Std)

⑦The number of the sessions and the count of the events within divided periods

Model Training

We trained 7 GBDTs using all features

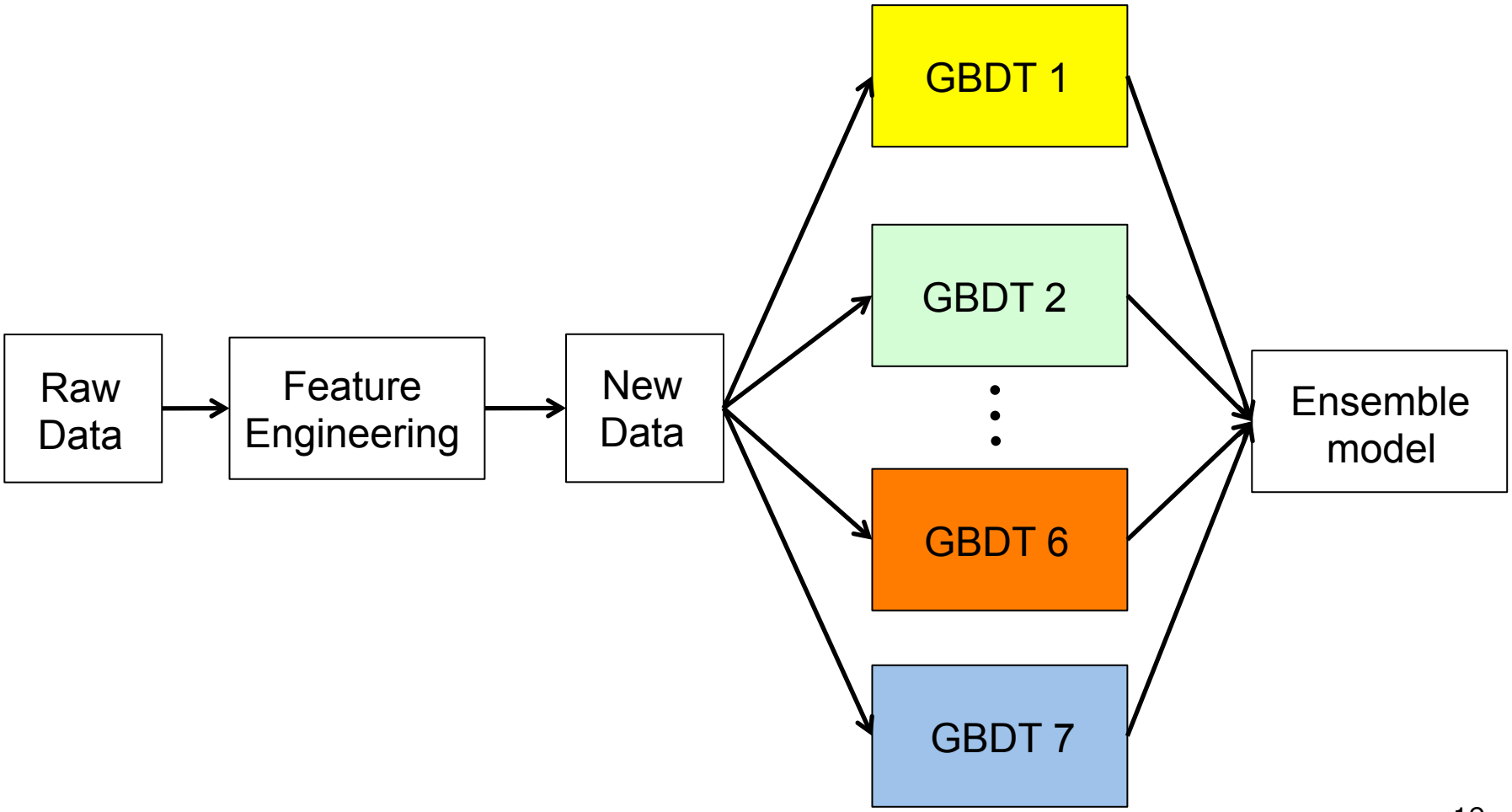
Score of these models were about 0.9054~0.9055



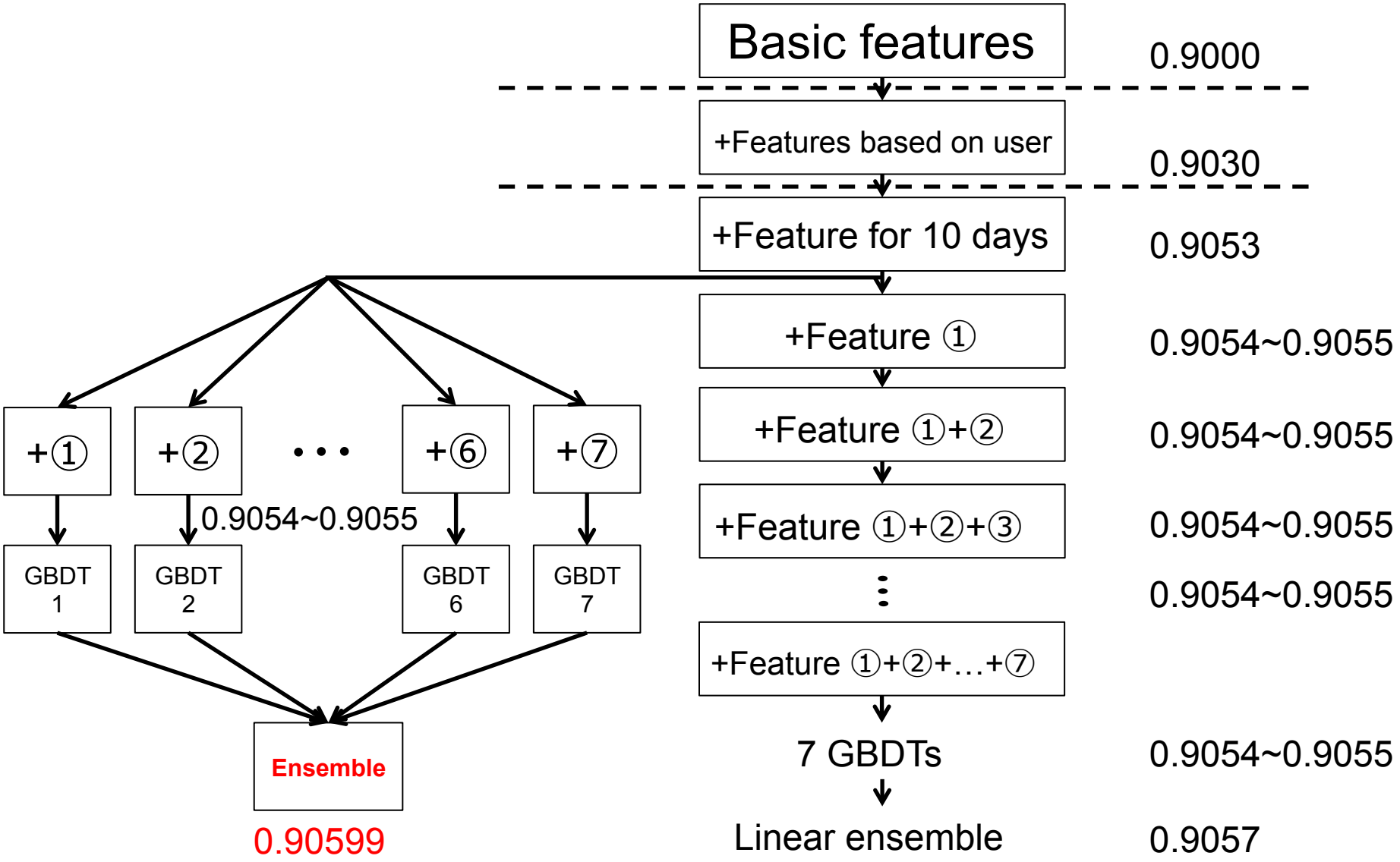
Linear Ensemble of 7 GBDTs

Linear Ensemble of 7 GBDTs (All features)

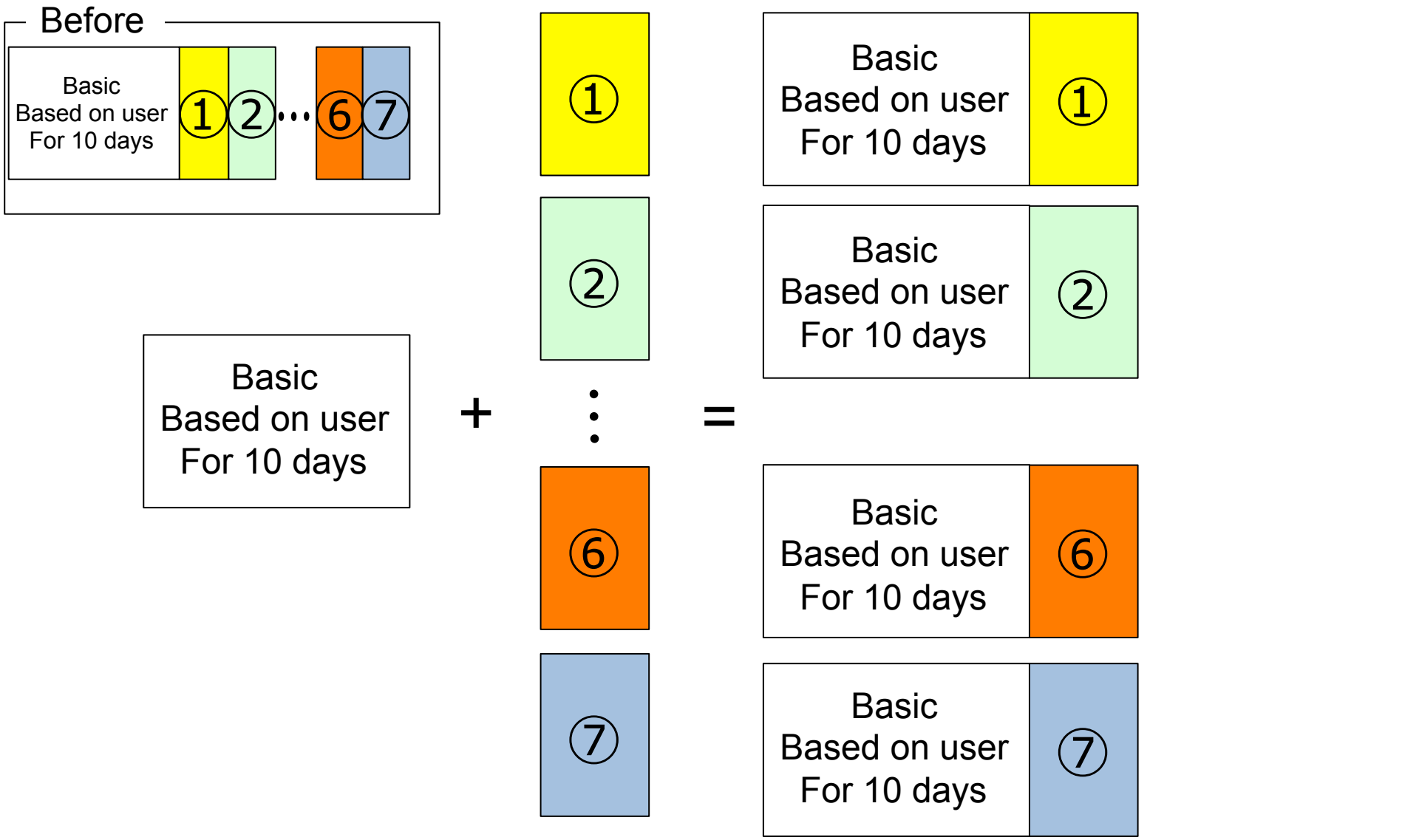
About 0.9057(0.0002up)



Flow chart to final model



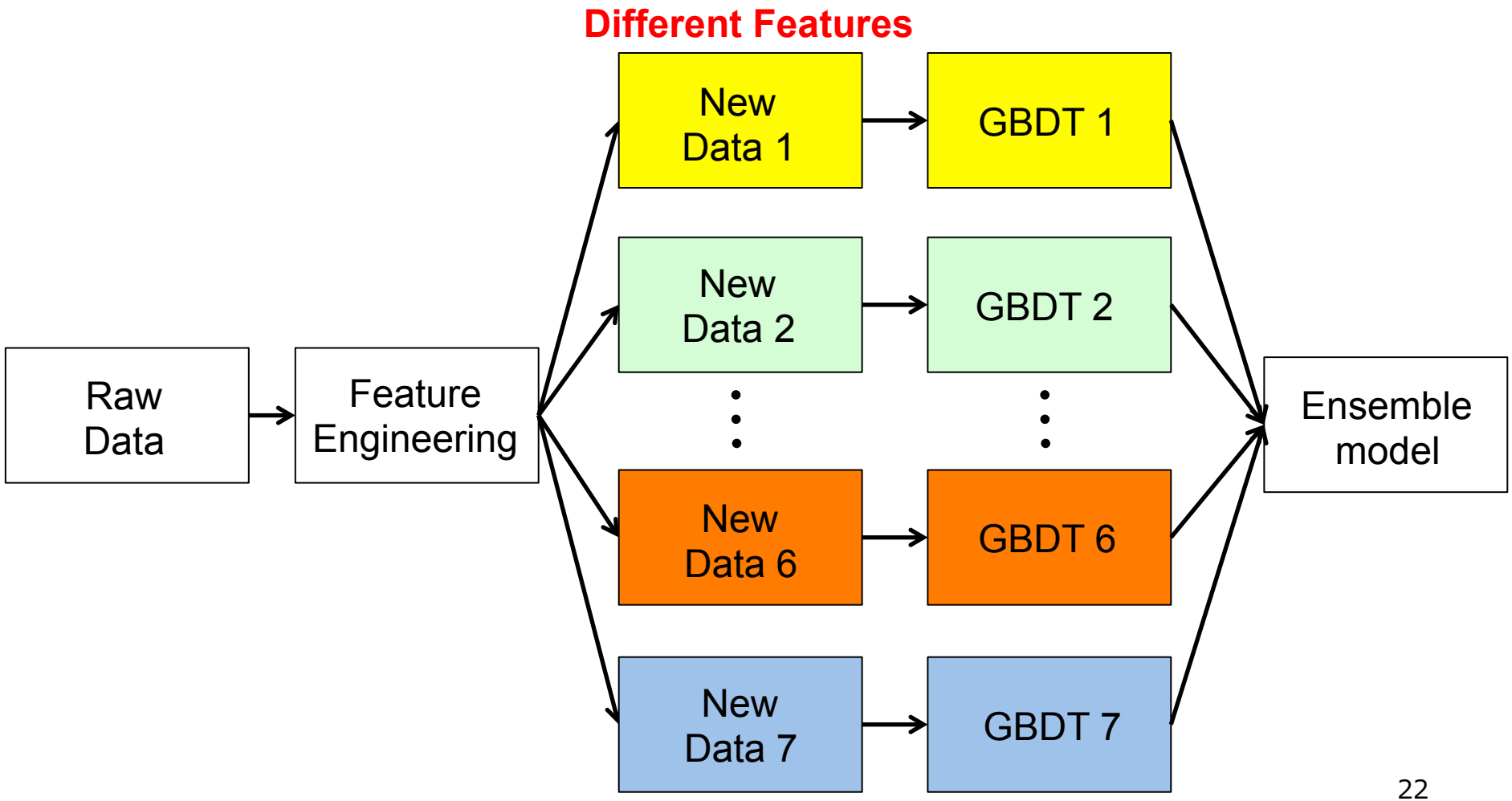
New data with partly different features



Diverse DBDT

Linear Ensemble of 7 GBDTs (partly different features)

About 0.90599(0.0005up)



Summary of ensemble

7 GBDTs (All features, 0.9054-0.9055)

⬇

Ensemble 0.9057 (0.0002UP)

⬇

Our single and ensemble model score had reached the ceiling

⬇

Any good ensemble??

⬇

We created 7 train data with partly different features that did not change its scores(0.9054-0.9055)

⬇

Ensemble 0.90599 (0.0005UP)

What worked out and what didn't

What worked out

- We could come up with new ideas one after another after team merge

What didn't

- We should have used the same language
- Stacking
- Feature selection

Knowledge

MOOCS

- MOOCsでは全体の情報と同様に各個人の情報にも非常に依存している(Groupby user, userの過去参加数etc)

- LRの回帰係数から単なるアクセス数だけでなくその割合が重要だとわかった(videoを見た割合よりもproblemを解いた割合が大きいuserの方が離反しにくい傾向)

コンペ

- 特徴量生成が鍵になった
- XGBoostの相性が良かった
- 世界が相手で毎日が非常にストレスフルだった
- マンパワー・マシンパワーは強いと思った
- 意味がわかる特徴量を使い、シンプルな予測で勝つことができたのは良かった
- すごい人たちに会えて有意義だった(学生のメリット)

Conclusion

- Feature engineering was the main key to improve the score
- “Group-by” username and the feature for 10 days were very effective.
- We came up with a new idea of creating several new train data with partly different features and ensembling GBDTs with these data
- This ensemble was better than that with all features in KDDCup 2015

Appendix. Feature Engineering(detail)

Feat.1～39：course_id（dummy variable)
Feat.40～46：count of event(unique)
Feat.47～52：count of object(mean, sum, std)
Feat.53～56：count of source in problem or access
Feat.57～68：count of username, course_id
Feat.59～60：count of source
Feat.61～67：count of event
Feat.68～98：day1～31（dummy variable)
Feat.99～122：hour0～23（dummy variable)
Feat.123～129：day of the week（dummy variable)
Feat.130～135：count of category
Feat.136～140：count of module depth
Feat.141：source(browser + server)
Feat.142～146：time(min, max, mean, std, length)
Feat.147：max time – mean time
Feat.148：mean time – min time
Feat.149：count of other course for 10 days
Feat.150：overlapped length for 10 days
Feat.151～160：the actual day for 10 days
Feat.161：for 11 -20 days
Feat.162：for 20 days
Feat.163～264：sum of Feat.40～140 by username
Feat.265～377：mean of Feat.40～148 by username
Feat.378～489：std of Feat.40～148 by username
Feat.490～493：start time of chapter(min, max, mean, std)
Feat.494～496：max time-mean time, mean time – min time, max time – min time(chapter)
Feat.497～500：ratio of dropout, dropcount(course, usernamecount)
Feat.501～696：chapter count
Feat.697：sum of chapter count
Feat.698～712：percentage of source, event, category
Feat.713～716：interval of time(max, min, mean, std)
Feat.717～720：actual login day, month, hour, weekday
Feat.721～725：start(max, min, mean,max-mean,min-mean)

Appendix. Feature Engineering(detail)

Feat.726 ～738：count of category par course

Feat.739：past enrollment count

Feat.740：past non droprate

Feat.741～742：count of other courses(max time, min time)

Feat.743～746：length from semester start or end to time

Feat.747：mean stay times

Feat.748～749：count of source(par day)

Feat.750～756：count of event(par day)

Feat.757～765：10/2013～08/2014(dummy variables)

Feat.766～774：sum of month by username

Feat.775～777：count of people login at the same time(min,max,mean)

Feat.778～779：count of people login at the same month(min,max,mean)

Feat.780：count of course/sum count

Feat.781 average length of time(par day)

Feat.782～886：sum of Feat.490～500,697～781 by username

Feat.887～991：mean of Feat.490～500,697～781 by username

Feat.992～1095：std of Feat.477～768 by username

Feat.1096～1097: count of children in course

Feat.1098: count of children

Feat.1099：percent of children in course

Feat.1100～1104：mean of Feat.494～496 by username

Feat.1105～1108：interval of login(max,min,mean,std)

Feat.1109～1111：periods divided into three

Feat.1112～1132：event in divided periods