

TalkingData AdTracking Fraud Detection Challenge



CONTENTS

01
Competition
Overview

02
Analysis
environment

03
Trial and Error

04

01

Competition Overview

01. Competition Overview

Description

- 온라인 광고 회사는 클릭만으로도 광고 비용이 상승 할 수 있다.
- 사기성 클릭이 전체 클릭의 90% 이상이 잠재적인 사기성 클릭이다.
- 기존에는 클릭 수가 많은 추적하여 IP와 device를 이용한 블랙리스트를 이용한 후발적 조치만 하였다.
- 좀 더 빠르게 사기꾼을 색출하여 데이터 낭비를 방지하고 싶다.

01. Competition Overview

Description



01. Competition Overview

Evaluation

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

Submission File

For each `click_id` in the test set, you must predict a probability for the target `is_attributed` variable. The file should contain a header and have the following format:

```
click_id,is_attributed
1,0.003
2,0.001
3,0.000
etc,
```

02

Analysis environment

02. Analysis environment

1. 대용량 데이터 작업환경



train.csv
7GB



test.csv
800MB



대용량 데이터!
개인 노트북으로 분석 불가

Azure 가상머신을 사용해보자!

02. Analysis environment

1. 대용량 데이터 작업환경

Q) 어떤 가상머신을 구현해야 할까?

- 요구사항

1. 큰 데이터를 처리할 작업공간
2. 작업공간의 공유
3. R언어 사용의 편리성

- Rstudio Server

웹에서 가상머신 IP주소로
간편하게 Rstudio에 접근 가능
Linux환경필요 (Windows 미지원)



Linux 필요



Linux 기반 OS "Ubuntu" 선택

02. Analysis environment

2. Azure로 가상머신 구현하기

@) Azure에서 Ubuntu 구현하기


1. Linux에 접근하기 위해 PuTTY 라는 프로그램을 설치
2. PuTTYgen 실행 후 키 생성
3. Azure에서 Ubuntu 선택 후 SSH 공개 키에 키 입력

...완성!



PuTTYgen SSH Key



 Ubuntu Server 17.10 VM
[자세한 정보](#)

* 이름

VM 디스크 유형 ⓘ

SSD ▼

* 사용자 이름

* 인증 형식

SSH 공개 키

암호

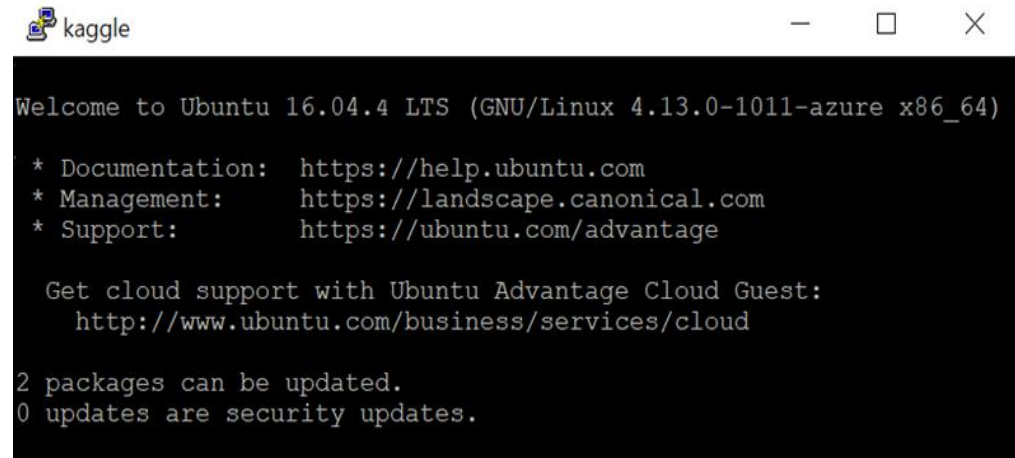
* SSH 공개 키 ⓘ

02. Analysis environment

2. Azure로 가상머신 구현하기

@) Ubuntu에 R 환경 구현하기

1. PuTTY를 실행하여 가상머신 IP로 Ubuntu 실행
2. Linux 명령어로 R과 Rstudio Server설치
3. Azure에서 Rstudio Server의 포트 8787 열어주기
4. 인터넷 주소창에 가상머신 IP:8787로 R server 접근 후 Linux ID로 서버 로그인

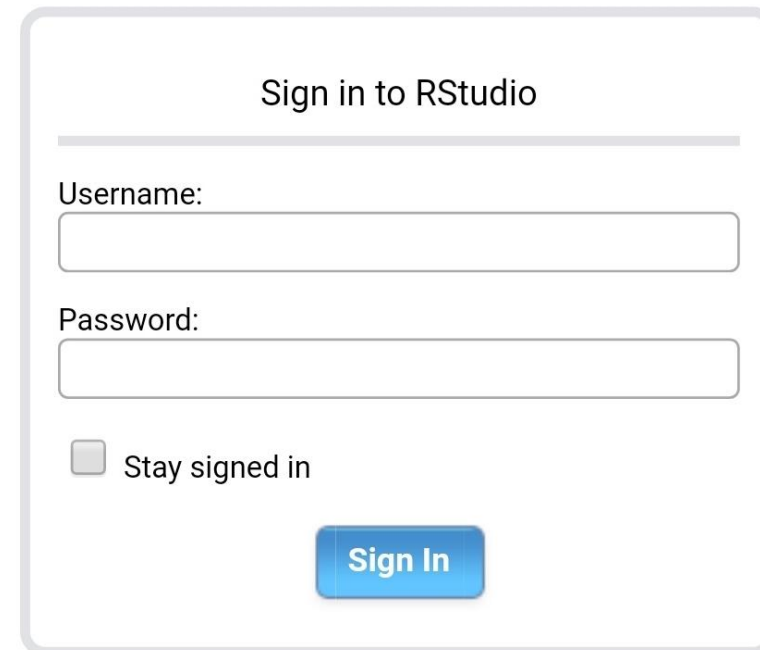


```
kaggle
Welcome to Ubuntu 16.04.4 LTS (GNU/Linux 4.13.0-1011-azure x86_64)

* Documentation:  https://help.ubuntu.com
* Management:    https://landscape.canonical.com
* Support:        https://ubuntu.com/advantage

Get cloud support with Ubuntu Advantage Cloud Guest:
http://www.ubuntu.com/business/services/cloud

2 packages can be updated.
0 updates are security updates.
```



Sign in to RStudio

Username:

Password:

☐ Stay signed in

02. Analysis environment

3. Ubuntu에 R작업환경 구축하기

커널의 여러 기법들을 사용하려 하였으나, 우분투의 기본 지원 R버전이 낮아서 문제가 발생 -> 우분투의 Sources.list 에 R repository 추가해서 R버전 업데이트

You need to **add R's repository** to your system:

1. Use your favorite text editor (I'm using `gedit` as an example) to open `/etc/apt/sources.list` :

```
sudo -H gedit /etc/apt/sources.list
```

2. Add this line to the file (if this is slow, use **another mirror**. You may also want to change `precise` into the codename for your Ubuntu version --- e.g., `trusty` for 14.04):

```
deb http://cran.rstudio.com/bin/linux/ubuntu precise/
```



추가적으로, 패키지 설치를 위하여 Git과 Cmake를 설치
-> LightGBM 등의 모델링 패키지 설치가능

```
library(devtools)  
options(devtools.install.args = "--no-multiarch") # if you  
install_github("Microsoft/LightGBM", subdir = "R-package")
```

02. Analysis environment

3. Ubuntu에 R작업환경 구축하기

...였지만 끝난 것이 아니었다. 문제발생!

- 문제1) 가상머신이 체험판이라 7G데이터를 분석하기에는 메모리가 부족하다..
- 방안)
 1. data.frame이 아닌 data.table의 fread()로 데이터 불러오기.
 2. 필요 없는 변수 없애기. (attributed_time)
 3. 전체 데이터를 작은 데이터로 줄여서 사용하기.
 4. 코드 중간중간에 가비지 컬렉션 gc() 자주 사용해서 메모리 관리하기.
- 문제2) EDA 시각화에 시간이 오래 걸린다.

커널 EDA를 참고하고 train_sample데이터로 확인해보기

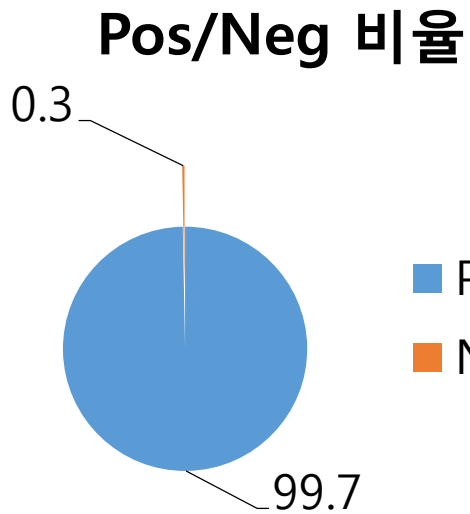
=> 비록 모델링 자체에 1억 8천건의 데이터를 전부 사용하지는 못하고 있지만 데이터를 불러오고 모델링 할 수 있는 환경 구축 성공!

03

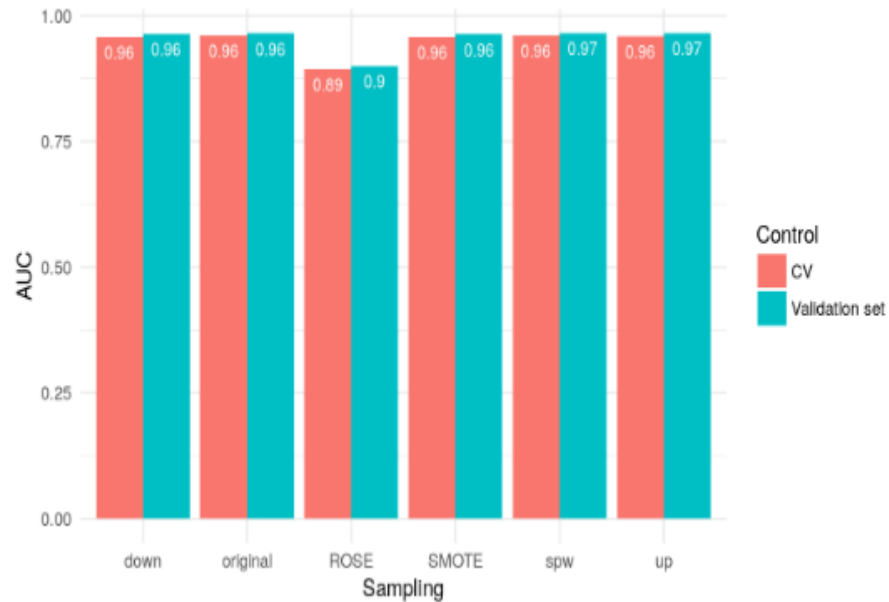
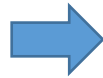
Trial and Error

03. Trial and Error

Imbalance data



■ POS
■ NEG

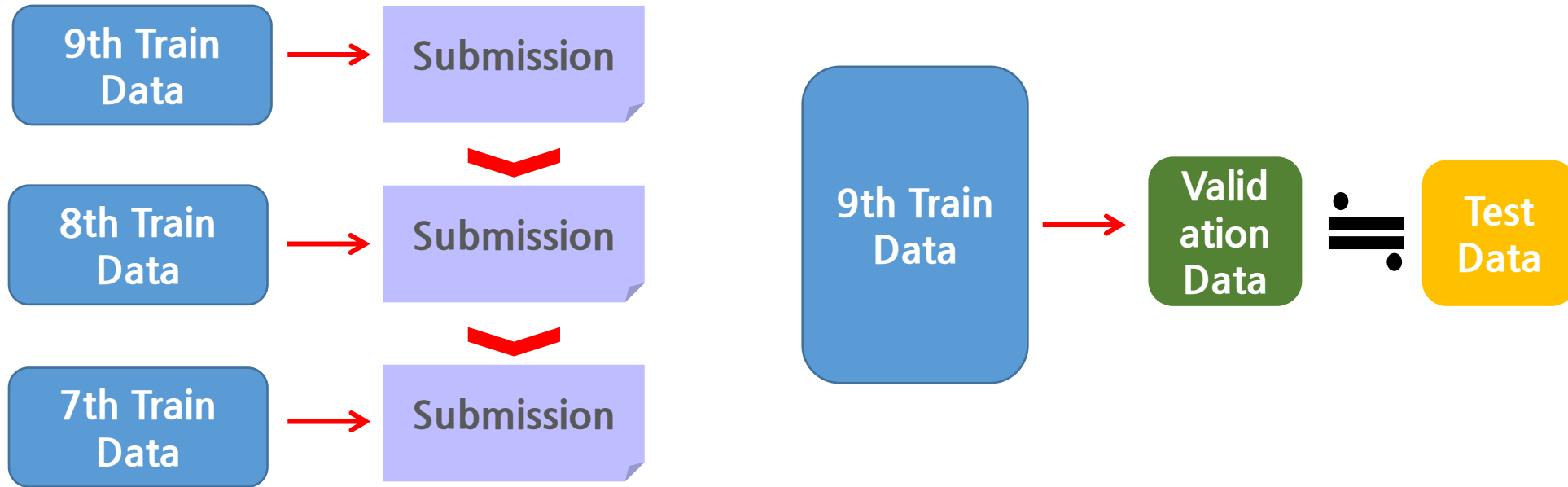


Parameter :
scale_pos_weight= 99.7

- 불균형 label : 0.3% : 99.7%
 - > 불균형 label을 가진 데이터를 학습하면 특정 label에 과적합 될 가능성이 있다.
 - > but sampling 후 학습 결과를 비교해보면 차이가 없다는 것을 알 수 있다.
 - > sampling을 하지 않아도 모델의 파라미터를 제어하면 비슷한 효과를 낼 수 있다.

03. Trial and Error

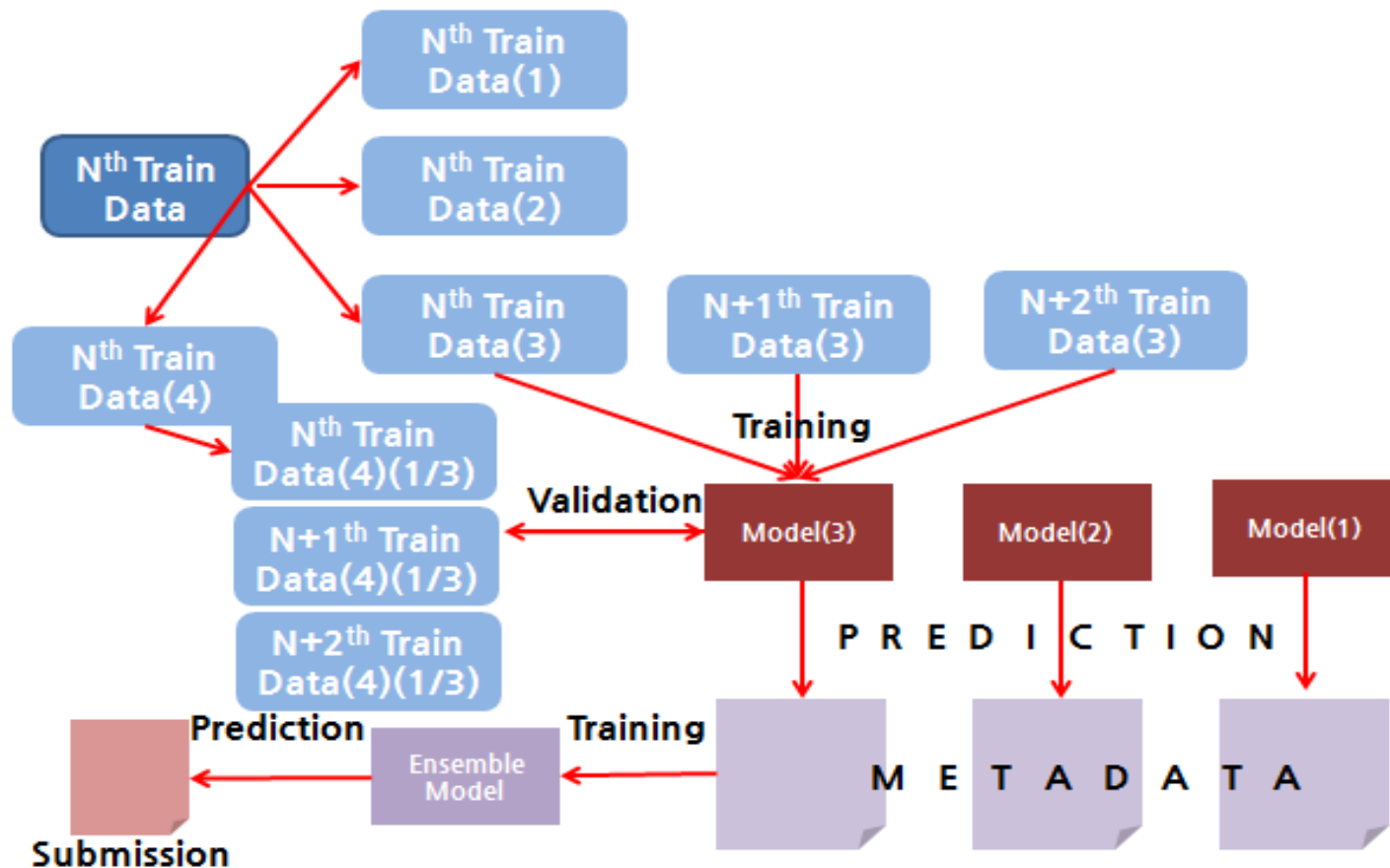
Train, Validation dataset



- **Train data** : 보통 데이터가 많으면 예측률이 높은 모델을 만들 수 있다.
-> but 전체 데이터를 활용할 수 있는 memory가 부족
-> solution 1 : test set과 가장 가까운 데이터를 활용
- **Validation data** : test data와 가장 유사한 조건의 9th data 일부 활용

03. Trial and Error

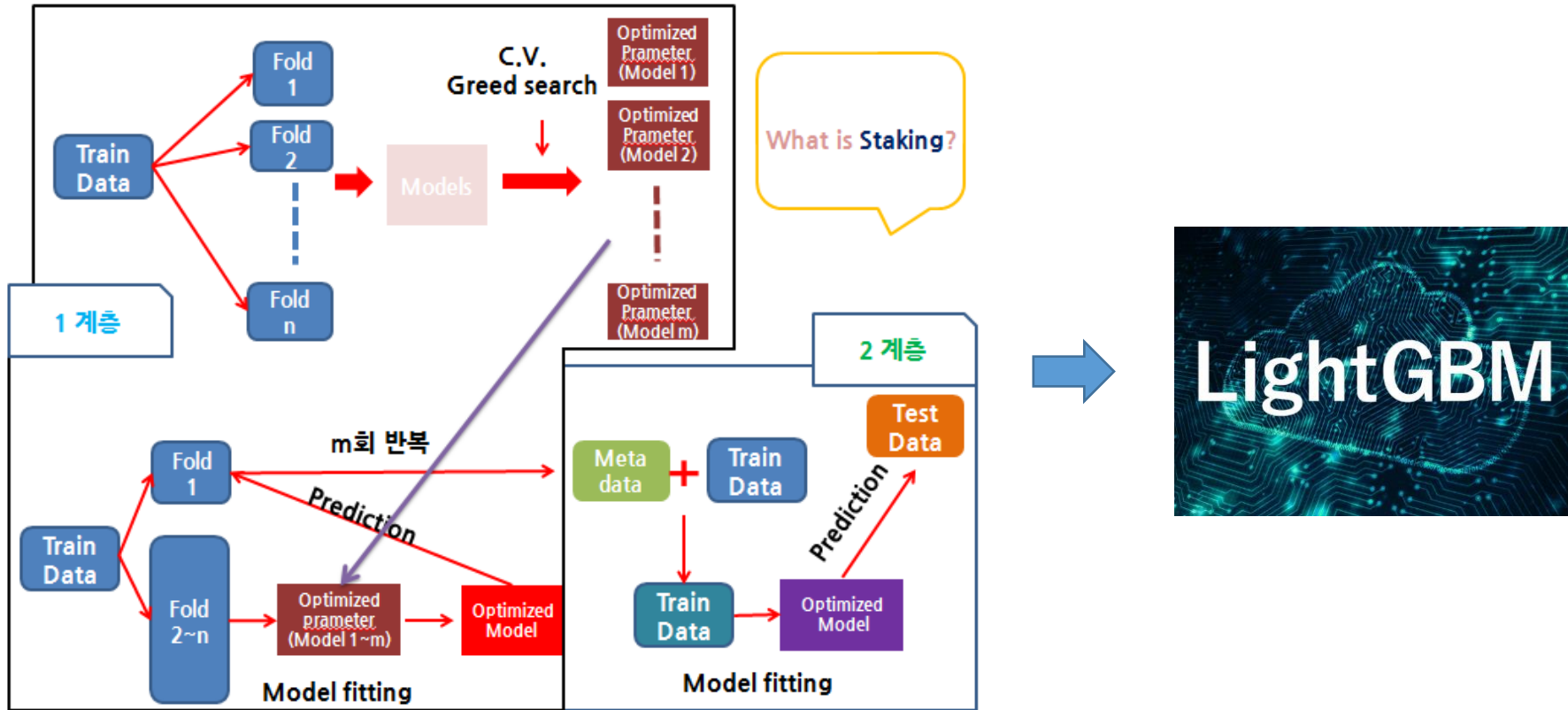
Train, Validation dataset



- **Train data** : solution 2 : staking과 유사한 방법으로 전체 data 학습

03. Trial and Error

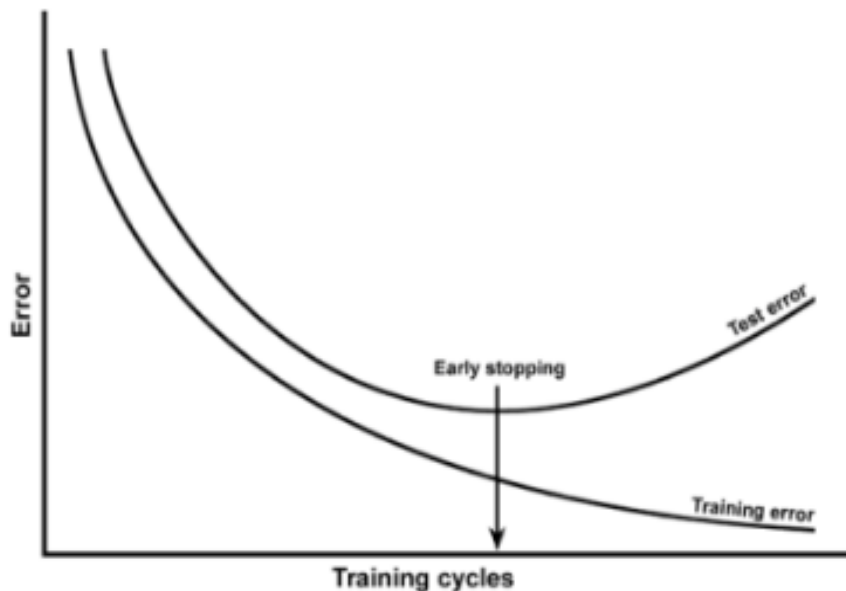
Model selection



- 계획은 stacking을 위해 여러 모델을 염두에 두고 진행.
-> but single model의 성능을 확보하는 것이 생각보다 어려웠고, 빠른 결과를 보기 위해 학습 속도가 제일 빠른 lgbm으로 최종 모델 변경.

03. Trial and Error

Parameter tuning



TalkingData Wordbatch FM_FTRL LB:0.9769

9d ago 0.9769



- Validation score와 test score의 차이를 줄이기 위해(정규화) :
nrounds , early_stopping_rounds , eval_freq 을 조정해 보았지만
유의미한 결과가 나오지 않음. Parameter tuning 문제 보다는 모델
의 한계 혹은 train data의 한계라고 결론 지음 -> blending을 이용

03. Trial and Error

Feature Engineering



VS



빠른 처리 속도,
다양한 응용 가능

직관적인 문법, 응용도 쉬움

시간 관련 데이터 연산이 쉬움

- Feature engineering을 위해 여러 번의 시도가 필요(시간 소요) :
dplyr은 익숙한 tool이지만, 통상 30개의 feature를 만드는데 40분정도 소요 -> data.table을 이용하면 절반의 시간으로 가능하다.
- 시간 관련 feature를 다룰 때 유용한 툴 **lubridate** :
Ymd_hms(시간을 나타낸 문자열) +/- hours() , Ymd_hms()를 통해
시간 데이터 연산이 가능

Q & A



Thank you

