4조

ML 프로젝트

Linking Writing Processes to Writing Quality

주제 : 에세이 품질 예측

팀원) 허수영, 권영수, 이지원, 고은채

목 차

Ⅰ. 문제 정의

1 대회 Overview

kaggle

대회 목적

글쓰기 프로세스의

특징을 이용하여

에세이 품질을 예측

평가지표

RMSE

(Root Mean Squared Error)

$$\text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)^{1/2}$$

상금 수여

1) 리더보드 점수 – 3등 이내

2) 효율 점수(시간 단축) – 3등 이내

$$\text{Efficiency} = \frac{\text{RMSE}}{\text{Base} - \min \text{RMSE}} + \frac{\text{RuntimeSeconds}}{32400}$$

요구사항

Code competition

✓ 반드시 캐글 notebook으로 제출해야 함.

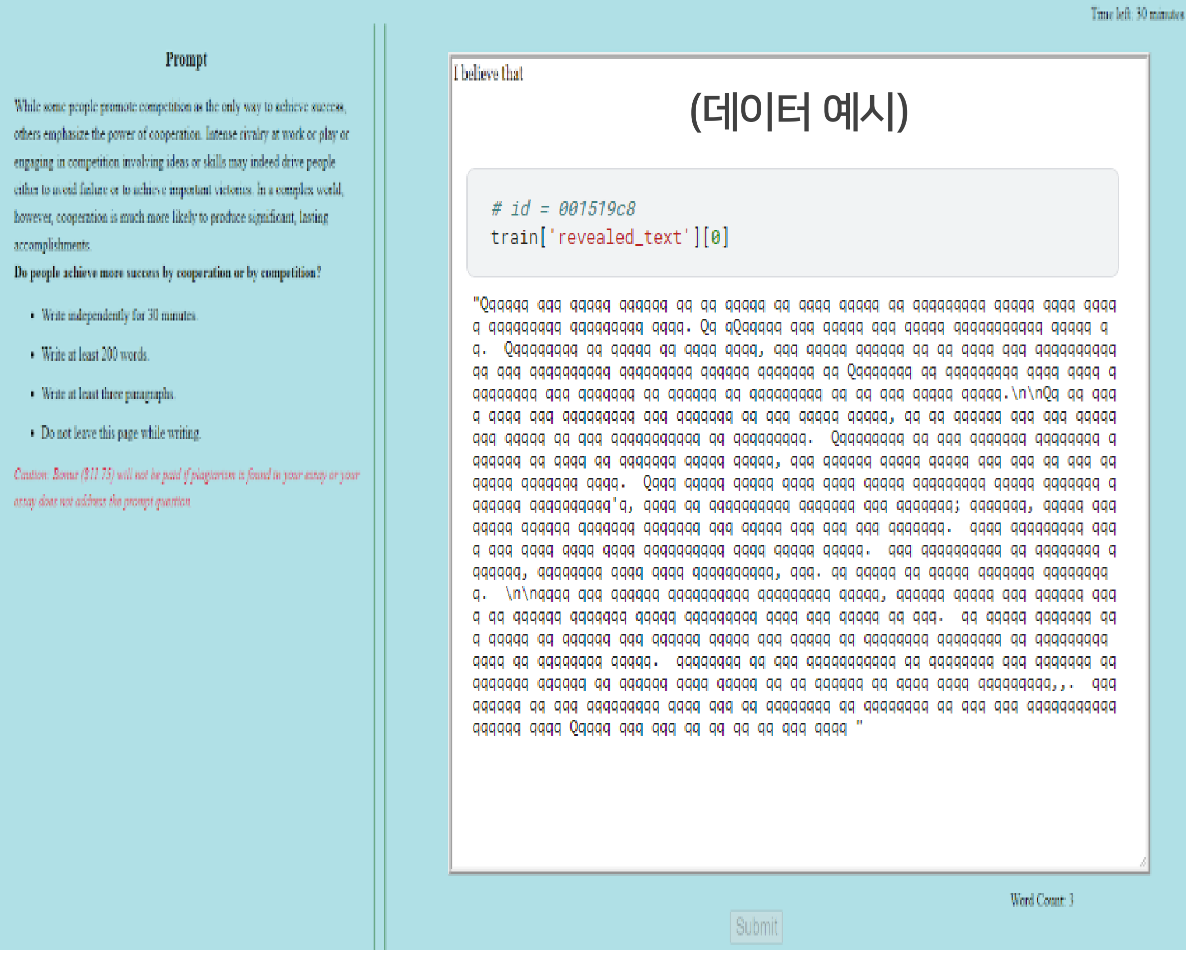
✓ Submission 파일만 제출할 수 없음.

✓ 인터넷 엑세스 사용 안됨

kaggle

데이터 수집 절차

키스트로크 로그 프로그램



User ID	A	B	C	D

↓

에세이 작성: 30분 동안 200 단어 이상의 에세이 작성

키스트로크 로그 프로그램 ⇒ ↓

평가 수행: 글쓰기 과정을 추적하여 에세이를 평가

↓

점수	5.0	4.5	3.0	1.5

1 대회 Overview

데이터 수집 절차

kaggle

데이터 컬럼

Event ID	어떤 이벤트가 발생된 인덱스 값
Down Time /Up Time	키나 마우스를 누르거나 떼었을 때 시간(단위 : milliseconds)
Action Time	키나 마우스가 눌러진 채, 지속된 시간(down time과 up time 차이)
Activity	키나 마우스 활동 범주(고윳값 6개)
Down Event /Up Event	키 또는 마우스 중 어떤 것을 클릭 했는지
Text Change	키나 마우스의 누른 결과로 변경된 텍스트가 있는 경우
Cursor Position	키 또는 마우스를 누른 후 텍스트 커서 위치의 문자 인덱스
Word Count	키 또는 마우스를 누른 후 에세이 단어 개수

입력 데이터

```python
# id = 0022f953
train['revealed_text'][1]
```

```
'Qqqq qq qqqqqqqqqqq ? Qq qq qqq qqq qqq, qqqqq qqq qqqqq, qq qq qq qqqqqqqq qq
qqqqqq, qqq qqq qqq qq qqqq qqqq qqq qqqqq qq qqq qqqqqqqqqq qq qqq qq. Qqqq qq
q qqqqqq qqqq qq q Qqqq qqq qqqq qqq qqq qq q qqqqqqq qq qqqq q qqqqq qq qqq.
\n\t\n      Qqqqqq qq qqqq qqq qqq qqqq qqq qqqqq qqq qqq qqq qqqqq qqq qqq qqqq
qqq qqqqqqqqqqq qqq qqqqqq q qq qqqqqq Qqq qqqq qqqqq qqqq qqqq q qqqqqq qqqqqq
q qqqqqq qq qqqq qqqq qqqqq qqqq qqqqq qq qqq qqqqqqq, qqq qqq qqqqq qqq. Qqq q
qqqqqq qqqq qq qqqqq qqq qqq qqq qqqq qqq qq qqqqq qqqqqqqqqqqqq.qQqqq qqqqq qq
q qqq qqq qqqqq qqqqqqq qqq qqq qq qqqq q qqqq qqqq qqqqq qqqq qqqqqq qqq qqqqq
q qqqq qqqq qq "qqqqq" qq qqqqq qqq qqqq, qq qqq qq qqqq qqqq qqqq qq qqq qqqq
qq qqqqq qqqq qqq qqqq. \n      Qqqq qqq qqqq qq qqq qqq qqqqqqqq qq qqqqq qqqqq
qqqqq qqqqqq - qqqqqqq, qqqq, qqqqq, qqqq, qqqqqqq qqq qq qqqq. Qqq qqqqqqq qqq
qqqq qq qqqq  qqq qqqqqq qqq qqqq Q qqqq\'q. Q qqqq Q qqq\'q qqqqqqqqqqqqqq "qq
qqqqqqqqqqq" qqq qqqq qqqqq qqqqqqq qqqq q qqqqq qqqqq qqqqqq qqq qqqqq qqq qqq
qqqqq Q qqq, qqqqq qqq qqqq qqqqqq qq qqqq qqq qqqqqq qqqqqq qq qqqq qq qqqq qq
qqq.,Q qq qqqqq qqqQqq qqqqq qqqq qqq qqq qqqqq qqq qqq qqq qqqqqq qq qqq qqqqq
q qqqqqqq qqq qq qqqqqq qqqqq qq qqqq qqq qq qqqqq qqqqq qq qqqqqq qqqqq qq qqq
qqqq-\n      \n     Qqqqqq qq qq qqqqqqqqqq - qqq qq qqqqqq. Qqq qq qqqq qqq "q
qqqQ" qq qqqqq, qqqq qqq qqq qqq qqq qqqqqqqqqq, qqq qqq qqq qqqqqq qqqqqq. Qqq
q qqqq qq qq qqqq qqqqq qq qqqq qq qqqqq qqqqqqq qqqq qq qqqq qqqqqqq qqq qqq q
qqq qqqq, qq qq qqqq qqqqq q, qqq qqqqqqqqq qq qqqq qqqq - qq qqqqq qqqq qq qqqq
qqq qq qq q qqqqqqqqqqqq qqqqqqqqqqqqqqq q Qqqq qqqqqq    qqqqqqqqq qqq qqqqqqq
qqqqqqqqqqqqqqqqqq, '
```

- ✓ 실제로 작성된 단어는 모두 문자열 q로 변환됨
- ✓ 글의 문맥이나 문장력의 우수성을 파악하기 어려움

순수하게 **log에 저장된 작성 패턴**으로만, 평가해야 하는 대회임을 파악

Ⅱ. 데이터 전처리 및 분석

1 접근 순서

kaggle

일반적인 분석 과정

문제 정의

데이터 전처리
및 분석

베이스
모델 적용

성능 개선
(피처 추가)

적용한 분석 과정

문제 정의

피처 생성(groupby)
및 분석

베이스
모델 적용

성능 개선

ID별 log 데이터 파일(csv)

	id	event_id	down_time	up_time	act
0	001519c8	1	4526	4557	
1	001519c8	2	4558	4962	
2	001519c8	3	106571	106571	
3	001519c8	4	106686	106777	
4	001519c8	5	107196	107323	
5	001519c8	6	107296	107400	
6	001519c8	7	107469	107596	
7	001519c8	8	107659	107766	
8	001519c8	9	107743	107852	
9	001519c8	10	107840	107978	
10	001519c8	11	108008	108195	
11	001519c8	12	108104	108259	
12	001519c8	13	108229	108370	
13	001519c8	14	108341	108486	
14	001519c8	15	109296	109438	
15	001519c8	16	109423	109559	
16	001519c8	17	109560	109729	

	id	event_id_max	up_time_max	action_time_sum
0	001519c8	2557	1801969	297243

2 EDA(log 파일)

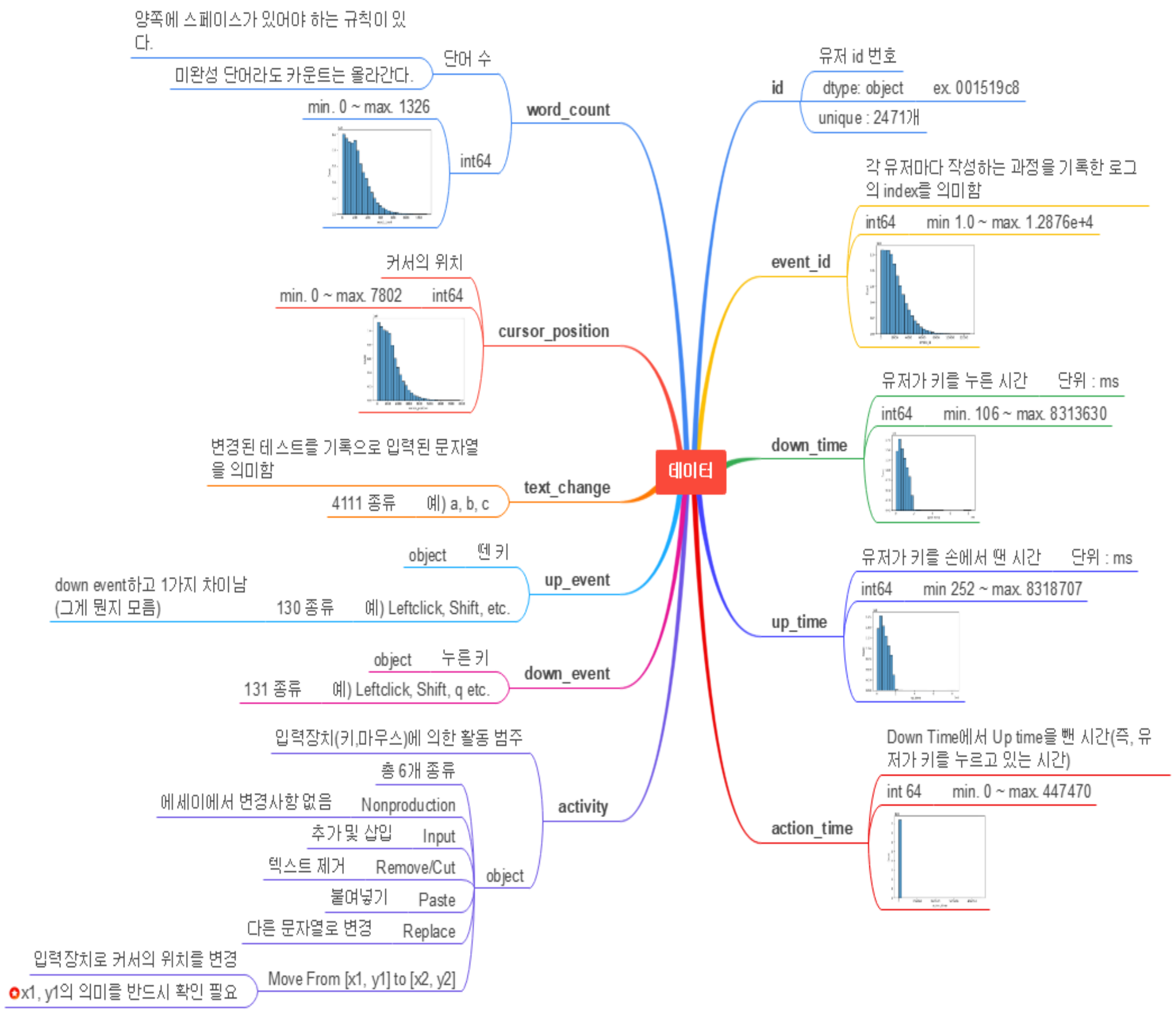
kaggle

데이터 컬럼

데이터 세트 형상: (8405898, 11)

	피처	데이터 타입	결측값 개수	고윳값 개수	고윳값
0	id	object	0	2471	[001519c8, 0022f953, 0042269b, 0059420b, 00758...
1	event_id	int64	0	12876	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
2	down_time	int64	0	1836078	[4526, 4558, 106571, 106686, 107196, 107296, 1...
3	up_time	int64	0	1835993	[4557, 4962, 106571, 106777, 107323, 107400, 1...
4	action_time	int64	0	3509	[31, 404, 0, 91, 127, 104, 107, 109, 138, 187,...
5	activity	object	0	50	[Nonproduction, Input, Remove/Cut, Replace, Mo...
6	down_event	object	0	131	[Leftclick, Shift, q, Space, Backspace, ., ,, ...
7	up_event	object	0	130	[Leftclick, Shift, q, Space, Backspace, ., ,, ...
8	text_change	object	0	4111	[NoChange, q, , ., ,, qqq qqqqq => , qqqqq...
9	cursor_position	int64	0	7803	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...
10	word_count	int64	0	1327	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...

- ✓ 데이터 타입은 object와 int 타입으로 구분
- ✓ 결측값은 없음



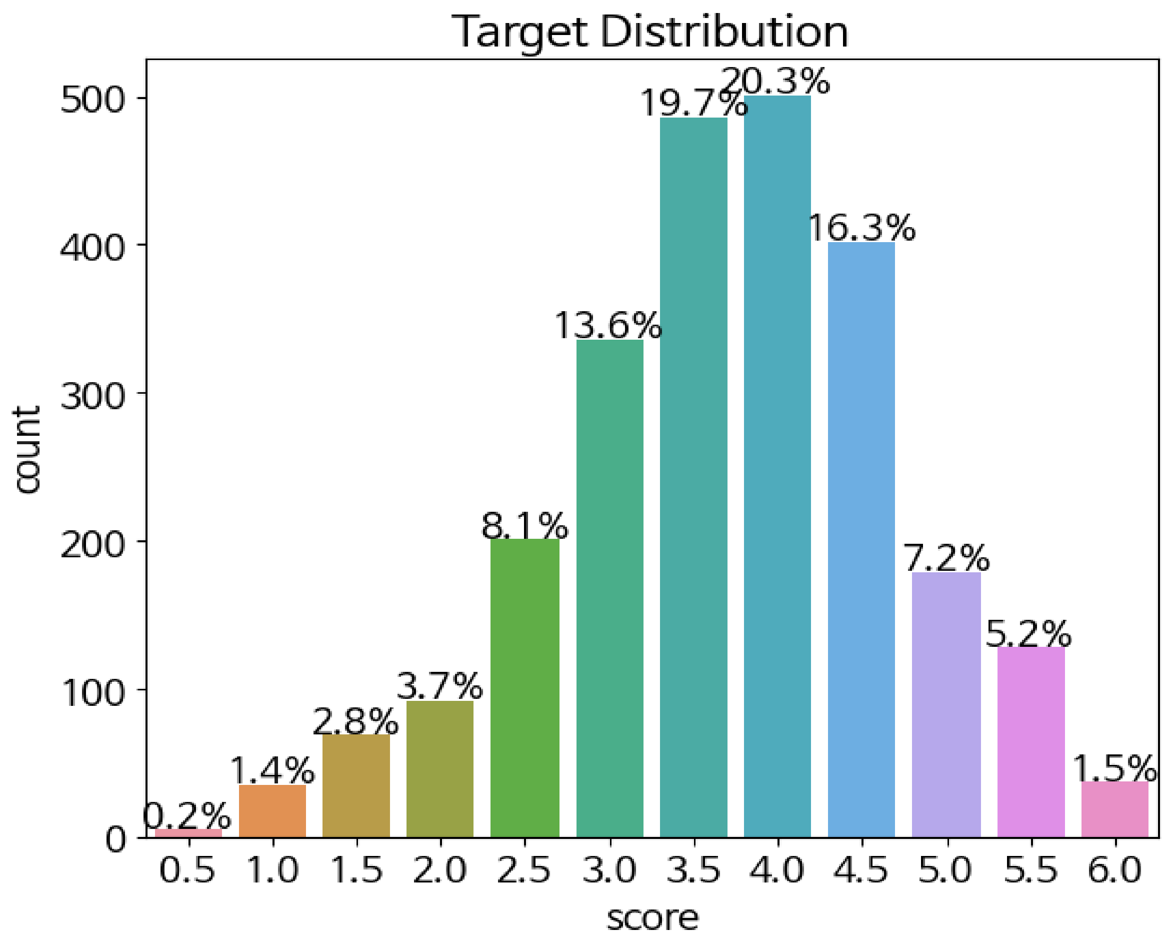
2 EDA(타깃값)

kaggle

타깃값 분포

```
train_original['score'].value_counts()

4.0    501
3.5    486
4.5    402
3.0    336
2.5    201
5.0    179
5.5    128
2.0     92
1.5     69
6.0     37
1.0     35
0.5      5
Name: score, dtype: int64
```



✓ 타깃값은 score이며, 에세이별 점수

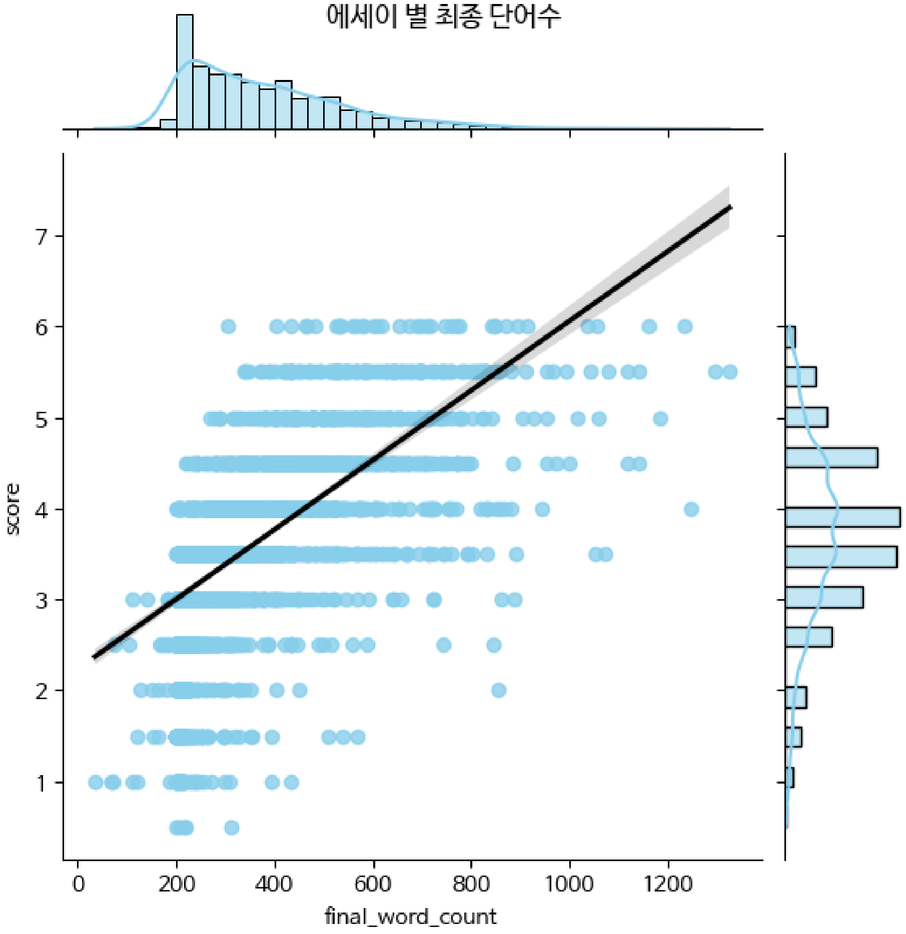
✓ 타깃값의 불균형(Target Imbalance)을 확인

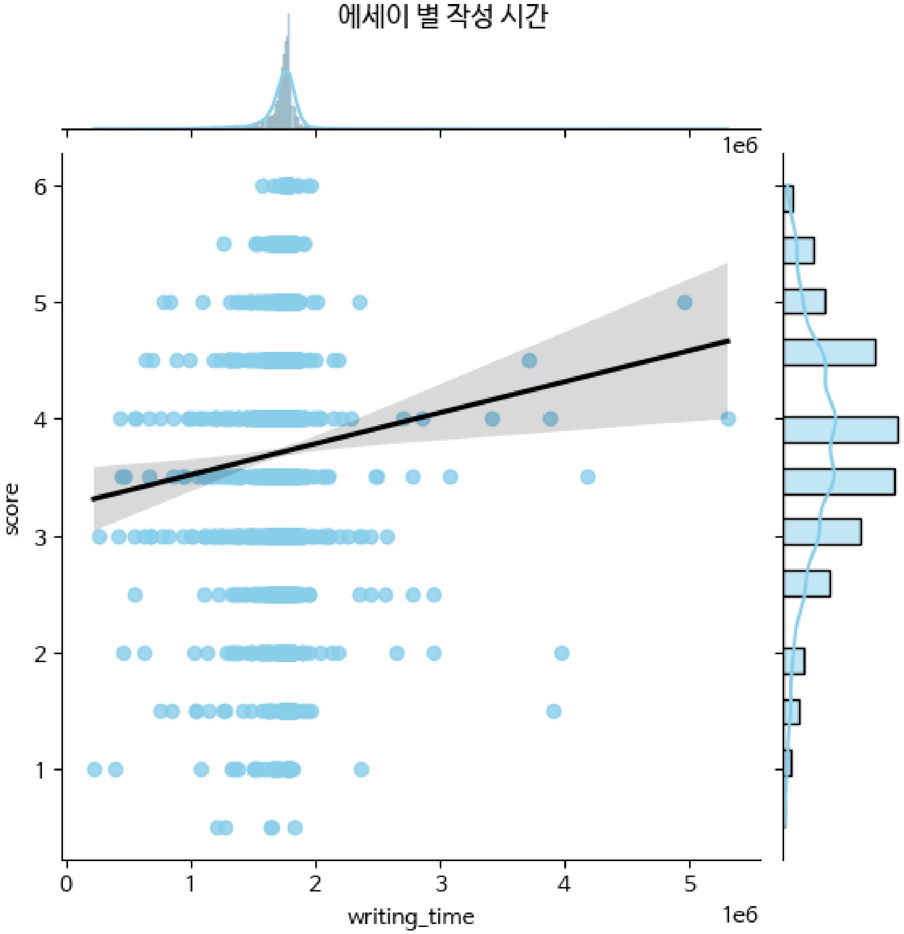
※ Test data set의 ID는 총 3개

2 EDA(train_data)

kaggle

ID별 EDA 시각화





- ✓ 단어 수가 증가할수록 score 점수도 증가하는 경향을 보임

- ✓ 작성시간이 증가함에 따라 score가 증가한다는 경향이 명확하지 않음
- ✓ 이상치 확인

Ⅲ. 베이스모델 적용

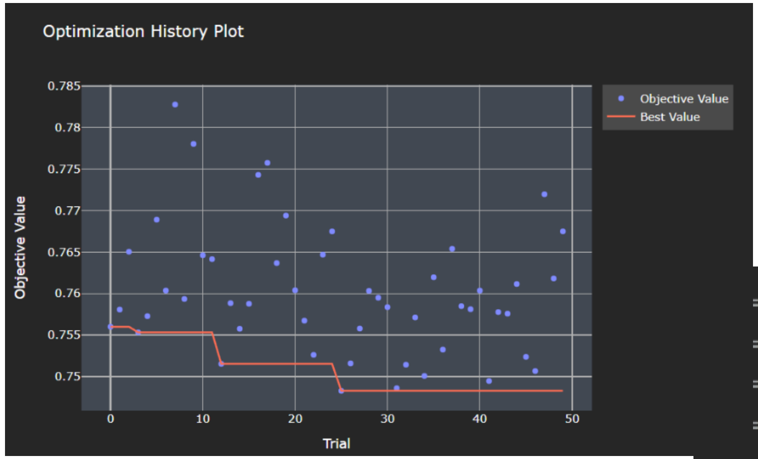
1 첫 번째 베이스라인 모델

kaggle

XGBoost

- ✓ 팀원 개개인이 만든 피처들을 사용해서 베이스라인 구축 완료
- ✓ RMSE : 0.734

	num_event_id	total_action_time	average_action_time	median_action_time	writing_time	num_of_P_Burst	Proportion_of_P-Bursts	Median_RPT	Median_PRT	Median_PPT	Median_RRT	word_count	score
num_event_id	1	0.85	-0.17	-0.14	0.19	0.16	-0.66	-0.19	-0.16	-0.18	-0.18	0.79	0.59
total_action_time	0.85	1	0.32	0.32	0.16	0.12	-0.61	-0.16	-0.14	-0.15	-0.15	0.73	0.55
average_action_time	-0.17	0.32	1	0.93	-0.0075	-0.083	0.071	0.063	0.041	0.047	0.054	-0.055	-0.049
median_action_time	-0.14	0.32	0.93	1	-0.026	-0.13	0.06	0.072	0.049	0.058	0.062	-0.019	-0.023
writing_time	0.19	0.16	-0.0075	-0.026	1	0.33	0.15	0.057	0.053	0.052	0.057	0.1	0.064
num_of_P_Burst	0.16	0.12	-0.083	-0.13	0.33	1	-0.26	-0.27	-0.23	-0.25	-0.25	0.019	0.097
Proportion_of_P-Bursts	-0.66	-0.61	0.071	0.06	0.15	-0.26	1	0.38	0.34	0.36	0.36	-0.55	-0.3
Median_RPT	-0.19	-0.16	0.063	0.072	0.057	-0.27	0.38	1	0.98	0.99	0.99	-0.15	-0.096
Median_PRT	-0.16	-0.14	0.041	0.049	0.053	-0.23	0.34	0.98	1	0.99	0.99	-0.12	-0.071
Median_PPT	-0.18	-0.15	0.047	0.058	0.052	-0.25	0.36	0.99	0.99	1	1	-0.13	-0.084
Median_RRT	-0.18	-0.15	0.054	0.062	0.057	-0.25	0.36	0.99	0.99	1	1	-0.13	-0.08
word_count	0.79	0.73	-0.055	-0.019	0.1	0.019	-0.55	-0.15	-0.12	-0.13	-0.13	1	0.64
score	0.59	0.55	-0.049	-0.023	0.064	0.097	-0.3	-0.096	-0.071	-0.084	-0.08	0.64	1




```
========= Fold 1 =========
========= Fold 2 =========
========= Fold 3 =========
========= Fold 4 =========
========== Fold 5 ==========
========== Fold 6 ==========
========== Fold 7 ==========
========== Fold 8 ==========
========== Fold 9 ==========
========== Fold 10 ==========
Loss : 0.7559
```

BaseModel_Ver01_Writing Quality - Version 5

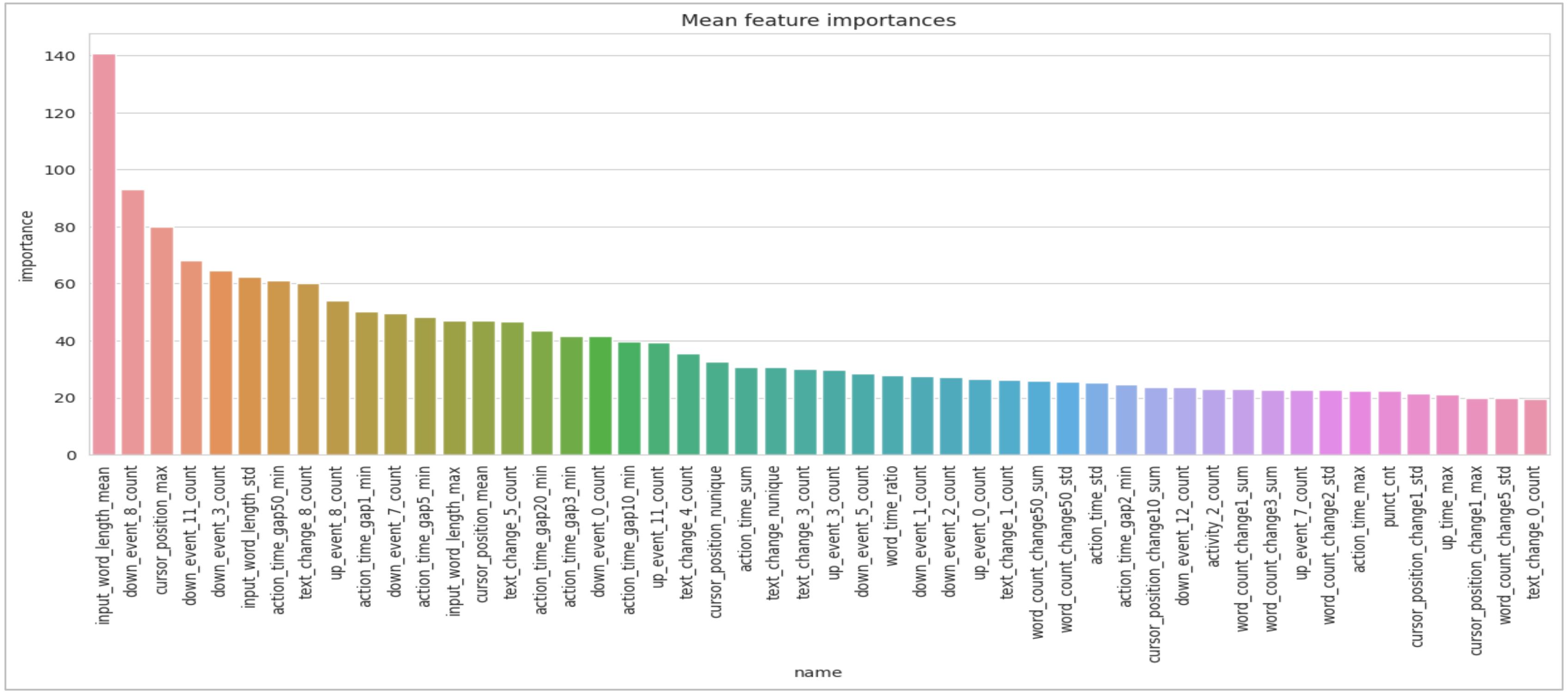
Succeeded · 6d ago

0.734

2 두 번째 베이스라인 모델

LightGBM(캐글러)

- ✓ 상위 캐글러의 코드에서 사용된 피처들을 도입하여 시도
- ✓ 다양한 피처를 사용해서 성능이 많이 향상됨
- ✓ RMSE : 0.611



BaseModel_Ver02_Writing Quality - Version 1

Succeeded · 4d ago

0.611

Ⅳ. 성능 개선

1 추가 피처 생성(논문참조)

KeyStroke Measure

kaggle

The Wolf Of SUTD (TWOS): A Dataset of Malicious Insider Threat Behavior Based on a Gamified Competition

Athul Harilal*, Flavio Toffalini, Ivan Homoliak, John Castellanos, Juan Guarnizo, Soumik Mondal
ST Electronics-SUTD Cyber Security Laboratory, Singapore University of Technology and Design, Singapore
{athul_harilal, ivan_homoliak, mondal_soumik}@sutd.edu.sg
{flavio_toffalini, john_castellanos, juan_guarnizo}@mymail.sutd.edu.sg

Martín Ochoa
Department of Applied Mathematics and Computer Science, Universidad del Rosario, Bogotá, Colombia
martin.ochoa@urosario.edu.co

Abstract

In this paper we present the TWOS dataset that contains realistic instances of insider threats based on a gamified competition. The competition simulated user interactions in/among competing companies, where two types of behaviors (normal and malicious) were incentivized. For the case of malicious behavior, we designed sessions for two types of insider threats (masqueraders and traitors). The game involved the participation of 6 teams consisting of 4 students who competed with each other for a period of 5 days, while their activities were monitored considering several heterogeneous sources (mouse, keyboard, process and file-system monitor, network traffic, emails and login/logout). In total, we obtained 320 hours of active participation that included 18 hours of masquerader data and at least two instances of traitor data. In addition to expected malicious behaviors, students explored various defensive and offensive strategies such as denial of service attacks and obfuscation techniques, in an effort to get ahead in the competition.

Furthermore, we illustrate the potential use of the TWOS dataset in multiple areas of cyber security, which does not limit to malicious insider threat detection, but also areas such as authorship verification and identification, continuous authentication, and sentiment analysis. We also present several state-of-the-art features that can be extracted from different data sources in order to guide researchers in the analysis of the dataset. The TWOS dataset is publicly accessible for further research purposes.

Keywords: malicious insider threat, masquerader, traitor, multiplayer game, user behavior monitoring, feature extraction, authorship verification, continuous authentication, sentiment analysis.

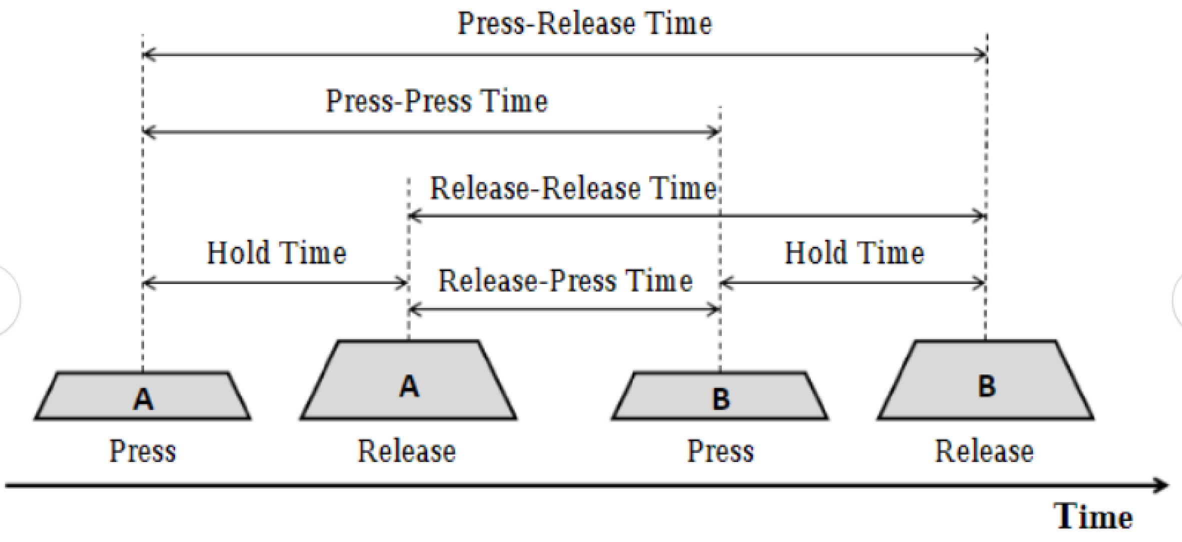


Figure 10: Low-level keystroke dynamics features

1 추가 피처 생성(논문참조)

KeyStroke Measure

RESEARCH REPORT

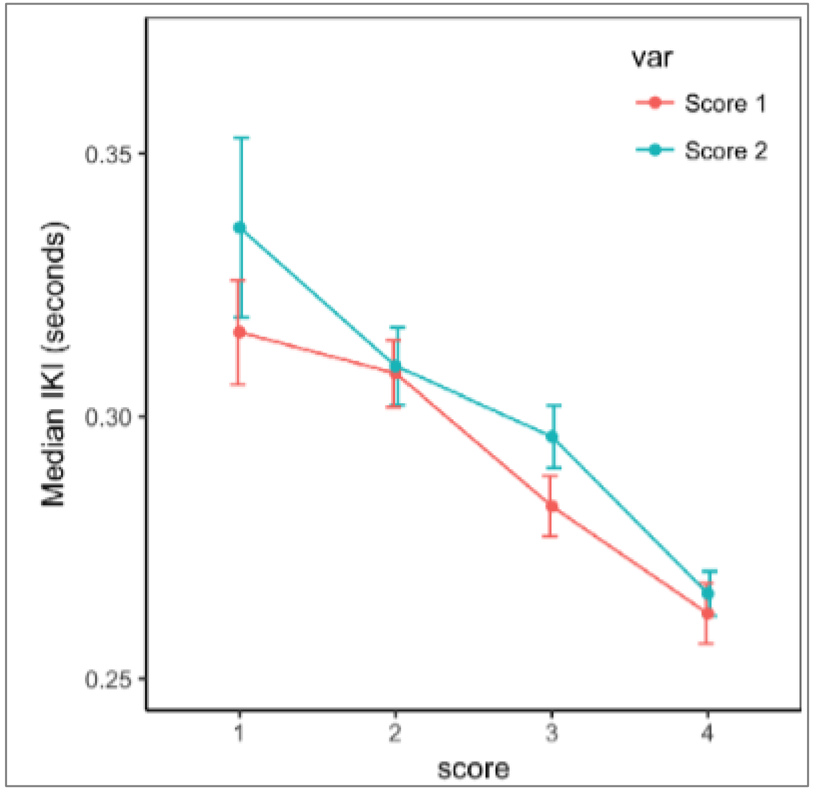
Analysis of Keystroke Sequences in Writing Logs

Mengxiao Zhu, Mo Zhang, & Paul Deane

Educational Testing Service, Princeton, NJ

The research on using event logs and item response time to study test-taking processes is rapidly growing in the field of educational measurement. In this study, we analyzed the keystroke logs collected from 761 middle school students in the United States as they completed a persuasive writing task. Seven variables were extracted from the keystroke logs and compared with different score and gender groups. Group comparisons were also made using methodologies borrowed from sequence mining. Students' composition strategies over the course of the writing process were also investigated. The findings of this study have implications for gaining deeper understanding of observed group differences and for designing interventions to close the achievement gaps among population groups.

Keywords Keystroke log; sequence analysis; writing assessment



kaggle

XGBoost

➢ 하이퍼파라미터 범위 (Optuna를 통해 최적의 하이퍼파라미터를 추출)

```
param = {
    'lambda': trial.suggest_float('lambda', 1e-3, 0.1),
    'alpha': trial.suggest_float('alpha', 1e-3, 1.0),
    'colsample_bytree': trial.suggest_float('colsample_bytree', 0.4, 1),
    'subsample': trial.suggest_float('subsample', 0.4, 1),
    'learning_rate': trial.suggest_float('learning_rate',0.0001, 0.1),
    'n_estimators': trial.suggest_int('n_estimators', 100, 1000),
    'max_depth': trial.suggest_int('max_depth', 4,8),
    'min_child_weight': trial.suggest_int('min_child_weight', 2, 50),
}
```

➢ 최적의 하이퍼파라미터를 교차검증 진행

```
model = xgb.XGBRegressor(reg_lambda=0.062039020636607344,
                         alpha=0.892907254615829,
                         colsample_bytree=0.5927968006434249,
                         subsample=0.5758791677351336,
                         learning_rate=0.09032689672187355,
                         n_estimators=547,
                         max_depth=5,
                         min_child_weight=33)
```

2 모델 훈련 및 성능검증

kaggle

모델별 예측 결과

➢ 하이퍼파라미터 범위 (Optuna를 통해 최적의 하이퍼파라미터를 추출)

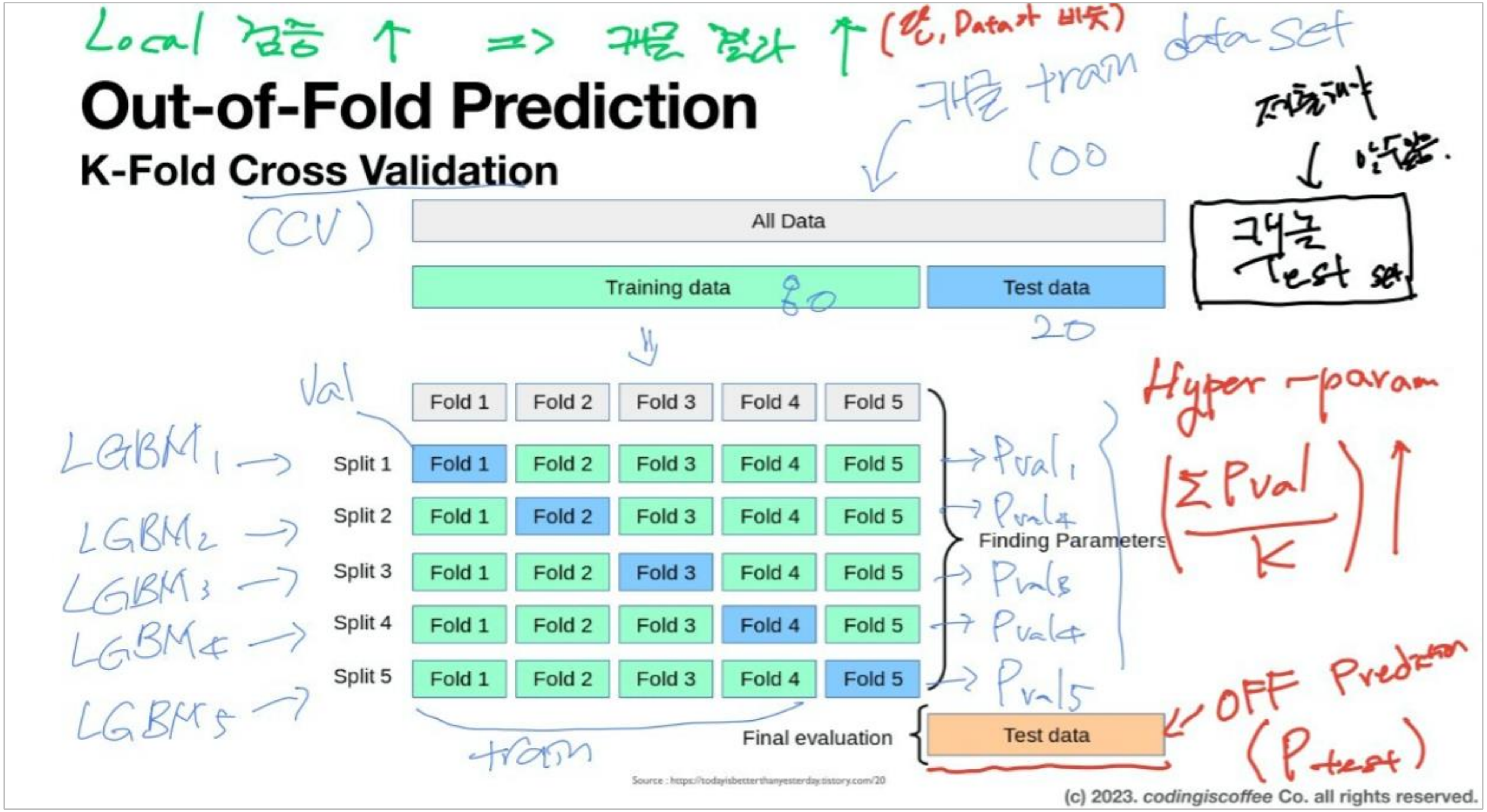
```
param = {
    'metric': 'rmse',
    'random_state': 42,
    'n_estimators': 10000,
    'reg_alpha': trial.suggest_float('reg_alpha', 1e-3, 10.0, log=True),
    'reg_lambda': trial.suggest_float('reg_lambda', 1e-3, 10.0, log=True),
    'colsample_bytree': trial.suggest_float('colsample_bytree', 0.5, 1),
    'subsample': trial.suggest_float('subsample', 0.5, 1),
    'learning_rate': trial.suggest_float('learning_rate', 1e-4, 0.1, log=True),
    'num_leaves' : trial.suggest_int('num_leaves', 2, 32),
    'min_child_samples': trial.suggest_int('min_child_samples', 1, 100)
}
```

➢ 최적의 하이퍼파라미터를 교차검증 진행

```
model = lgb.LGBMRegressor(num_leaves=18,
#                          max_depth=15,
                           learning_rate=0.023691696274555238,
                           n_estimators=10000,
                           subsample=0.6377463608066083,
                           min_child_samples=43,
#                          feature_fraction=0.75,
                           reg_alpha=0.3381890369449931,
                           reg_lambda=0.0022112993176679648,
                           colsample_bytree=0.5716208570394763,
                           random_state=42,
                           verbose=20,
                           metric=None)
```

kaggle

캐글 대회를 위한 OOF Prediction 전략



※ 출처 : 김용담 강사님 강의자료

3 예측 및 결과 제출

모델별 예측 결과

2023-10-19 기준

kaggle

XGBoost

91	Kwonys	0.604	14

XGB_test_10_19(14:06) - Version 25
Succeeded · 1h ago
0.604

LightGBM

LGBM_test_10_16(12:55) - Version 15
Succeeded · 2d ago · + penalty1
0.605

3 예측 및 결과 제출

캐글 결과 제출

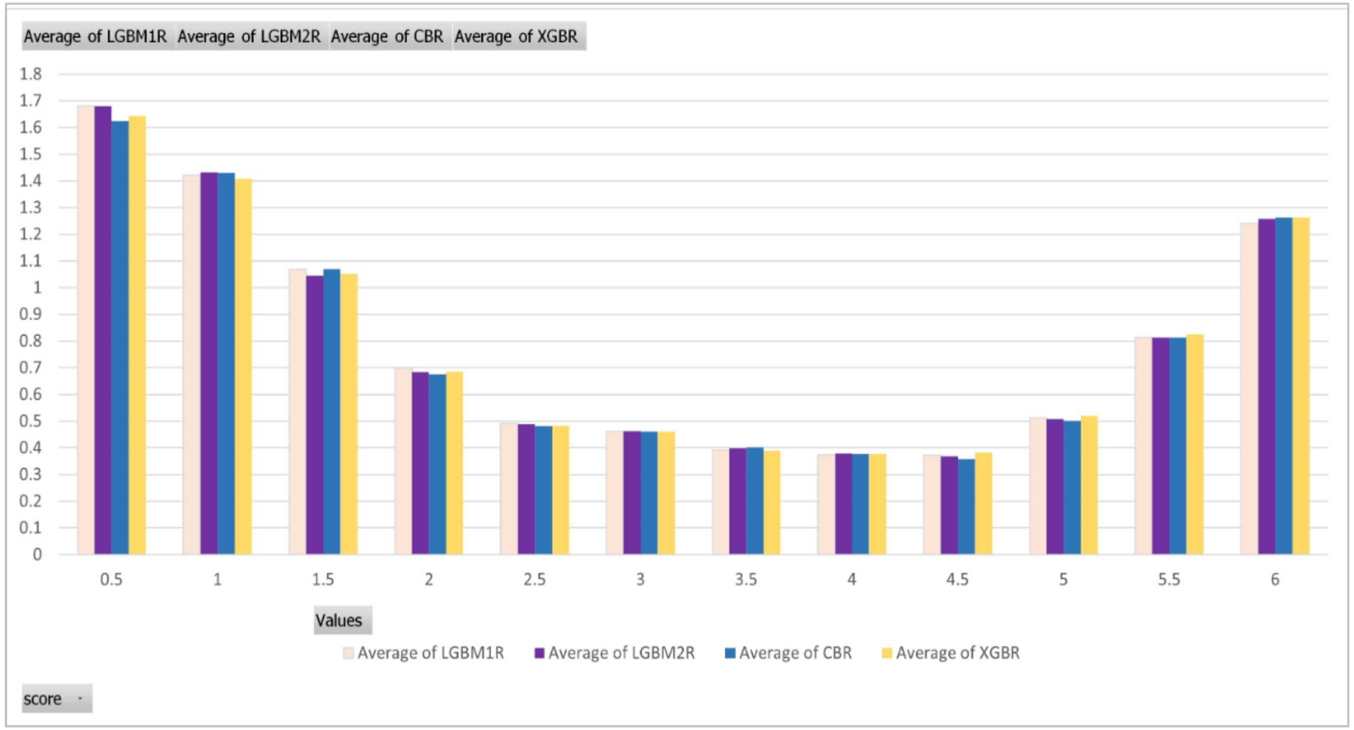
kaggle

Efficiency Score

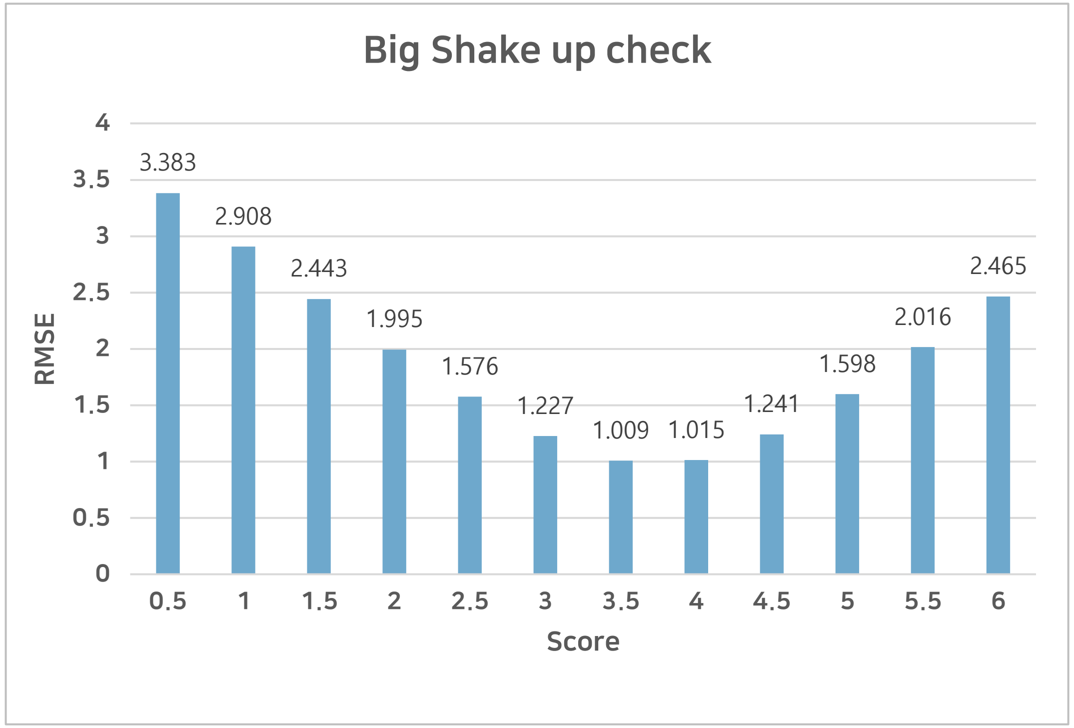
	TeamName	PublicScore	DateSubmitted
EfficiencyRank			
1	Rib~	0.594	Wed Oct 18 03:45:45 2023
2	Marlon Flügge	0.601	Mon Oct 16 17:07:45 2023
3	【Z Lab数据实验室】最菜选手	0.605	Sun Oct 8 12:41:45 2023
4	Joseph Josia	0.604	Fri Oct 6 16:19:24 2023
5	Stochoshi G	0.601	Thu Oct 12 20:08:06 2023
6	sfnga	0.601	Tue Oct 3 19:21:01 2023
7	3sigma	0.610	Tue Oct 10 12:27:21 2023
8	suk1yak1	0.612	Wed Oct 4 14:38:50 2023
9	Ryota	0.598	Wed Oct 18 19:11:16 2023
10	koyarocow	0.606	Thu Oct 12 10:18:21 2023
11	shige_skywalker	0.605	Thu Oct 12 07:42:24 2023
12	The Nam	0.604	Tue Oct 17 20:11:34 2023
13	Soo.Y	0.606	Tue Oct 17 09:54:25 2023
14	chimuichimu	0.601	Sun Oct 15 02:49:56 2023
15	Kwonys	0.605	Mon Oct 16 05:26:09 2023
16	Eunchae Koh	0.605	Mon Oct 16 09:14:01 2023
22	JJJI WON	0.607	Tue Oct 17 07:47:05 2023

4 한계점

kaggle



Average of LGBM1R
Average of LGBM2R
Average of CBR
Average of XGBR
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
Values
Average of LGBM1R
Average of LGBM2R
Average of CBR
Average of XGBR
score



Big Shake up check
RMSE
4
3.5
3
2.5
2
1.5
1
0.5
0
3.383
2.908
2.443
1.995
1.576
1.227
1.009
1.015
1.241
1.598
2.016
2.465
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
Score

경청해주셔서 감사합니다.