방학세미나

2팀

주혜인
심예진
문병철
반경림

INDEX

1

EDA

- 데이터 탐색

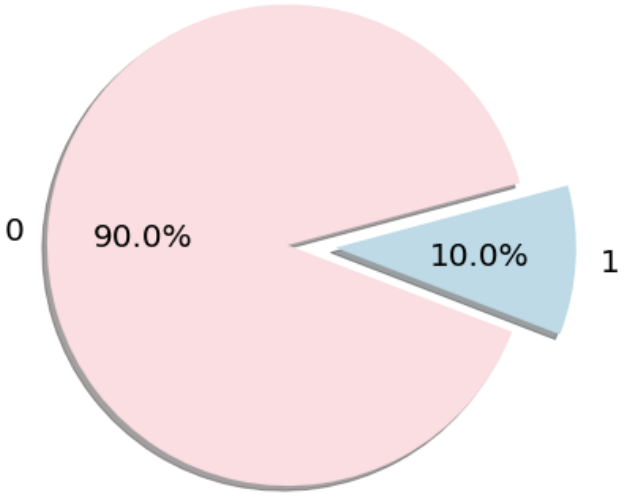
반응 변수

1. 변수 타입

Target = 0과 1로 구성된 범주형 변수

	target
0	0
1	1
2	0
3	0
4	0
...	...
27995	0
27996	0
27998	0
27999	0

2. 클래스 불균형



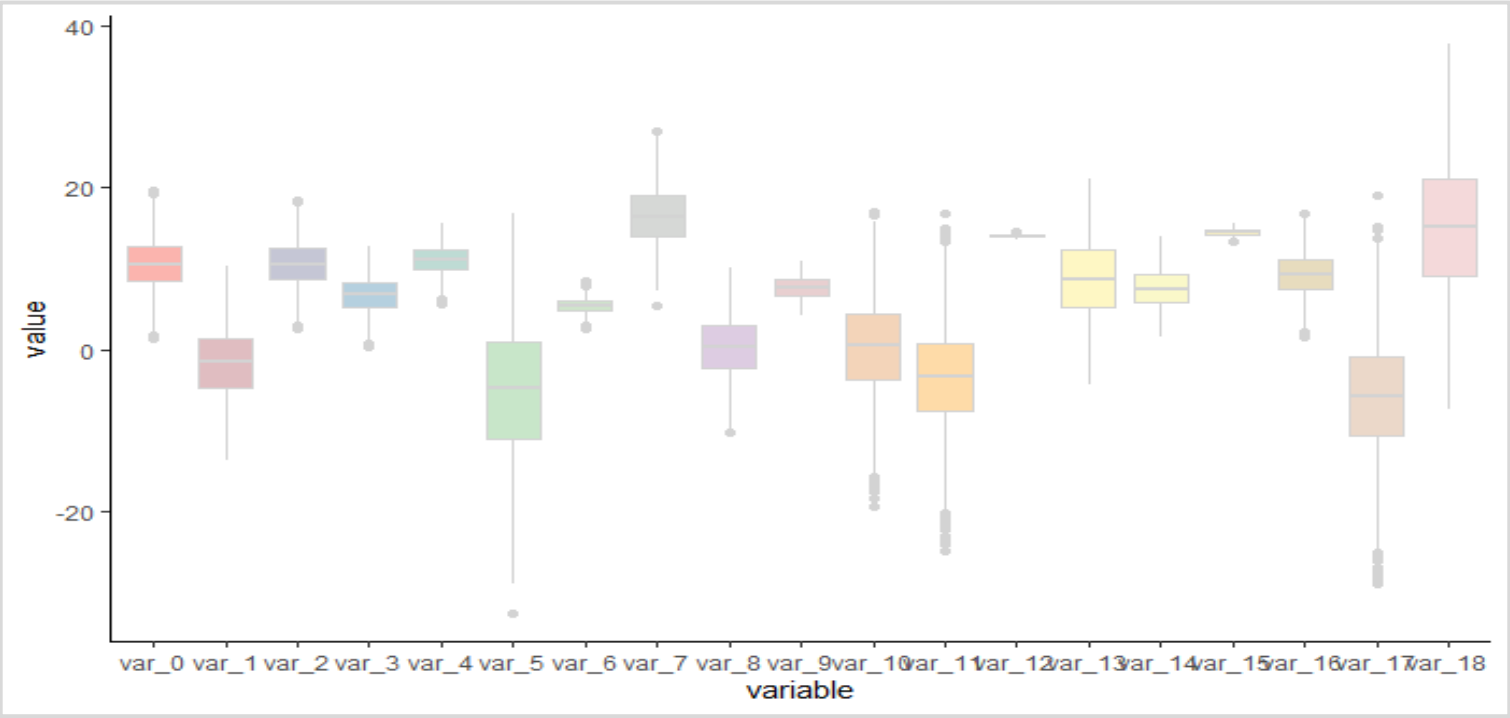
target 변수의 클래스 비율

- 데이터 탐색

설명 변수

1. **변수 타입** 수치형 변수

2. **데이터 분포** boxplot을 통해 데이터 분포 파악



1 EDA

분석 흐름

1. 데이터 불균형 해소를 위한 샘플링

smote, borderline smote, svm smote, randomsampling,…

2. 차원 축소

PCA, SVD, FA, Random Forest , …

3. 모델링

Logistic regression, LGBM, XGBoost, KNN, Decision Tree,…

2

Sampling

2 Sampling

- Sampling 방법

Undersampling

많은 레이블을 가진 데이터 세트를 적은 레이블을

가진 데이터 세트 수준으로 **감소**시키는 기법

Oversampling

적은 레이블을 가진 데이터 세트를 많은 레이블을

가진 데이터 세트 수준으로 **증식**하여 학습에 충분한 데이터를 확보하는 기법

Smote	Borderline Smote	Random Sampling
SVM Smote	ADASYN	K-Means Smote

2 Sampling

Sampling 방법

Undersampling

많은 레이블을 가진 데이터 세트를 적은 레이블을

가진 데이터 세트 수준으로 **감소**시키는 기법

Oversampling

적은 레이블을 가진 데이터 세트를 많은 레이블을

가진 데이터 세트 수준으로 **증식**하여 학습에 충분한 데이터를 확보하는 기법

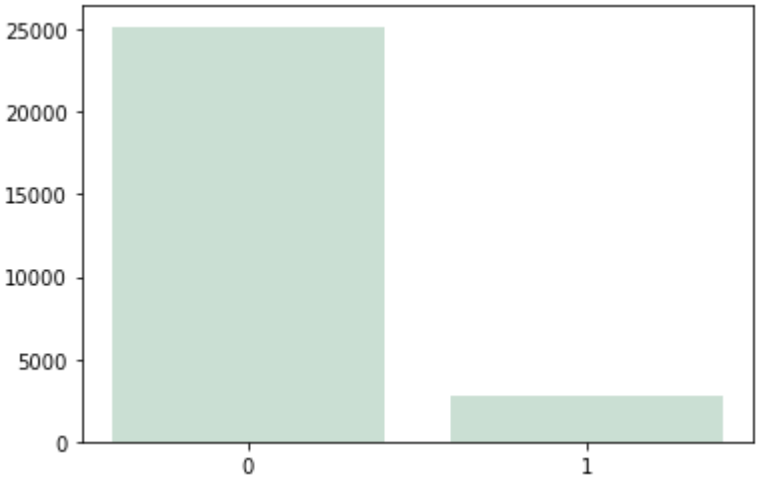
- Smote
- Borderline Smote
- Random Sampling
- SVM Smote
- ADASYN
- K-Means Smote

2 Sampling

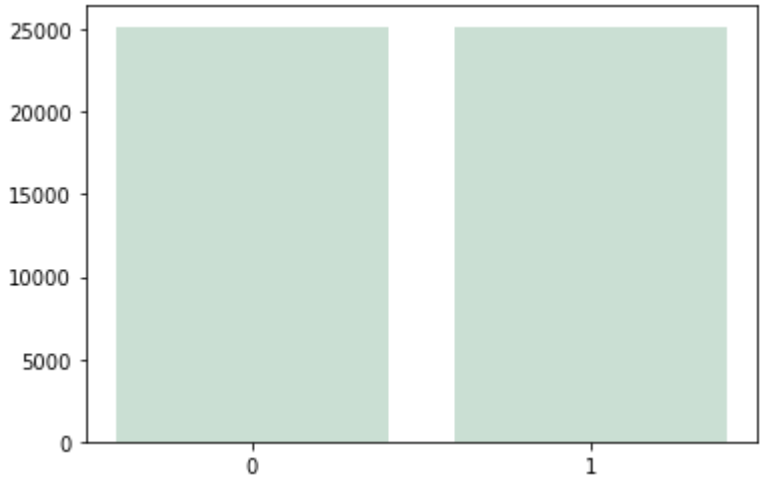
- Smote

SMOTE

낮은 비율의 클래스에서 임의로 선택한 샘플과 K개의 최근접 이웃 간의 차에
0~1사이의 임의의 값을 곱하는 방식으로 새로운 데이터 합성



28000 rows X 200 columns



50382 rows X 200 columns

2 Sampling

- Smote

SMOTE

낮은 비율의 클래스에서 임의로 선택한 샘플과 K개의 최근접 이웃 간의 차에
0~1사이의 임의의 값을 곱하는 방식으로 새로운 데이터 합성

28000 rows X 200 columns

단순히 minority class에서 랜덤하게 샘플링

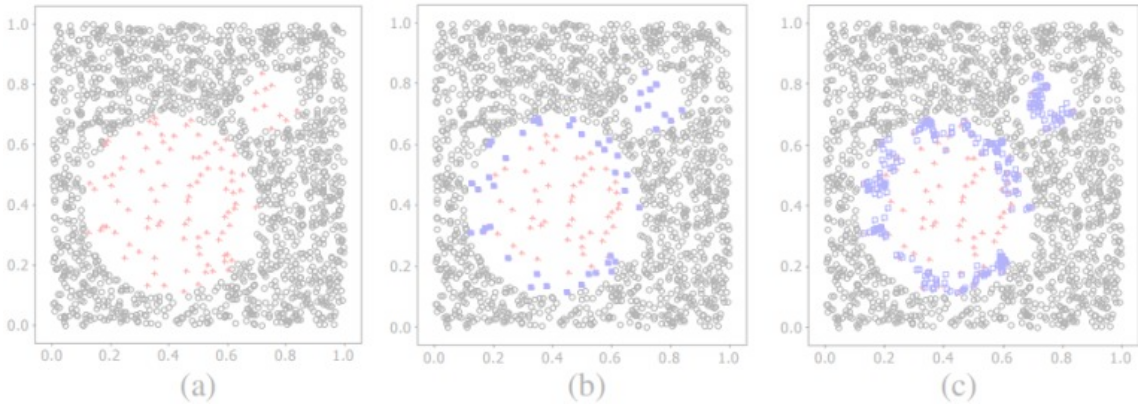
50382 rows X 200 columns

- Borderline Smote

Borderline SMOTE

클래스 범위 내에서만 데이터를 생성하여

특정 분포에만 **중복적으로 데이터가 발생되는 SMOTE를 개선**



다른 class와의 borderline에 있는 샘플들을 늘림으로써

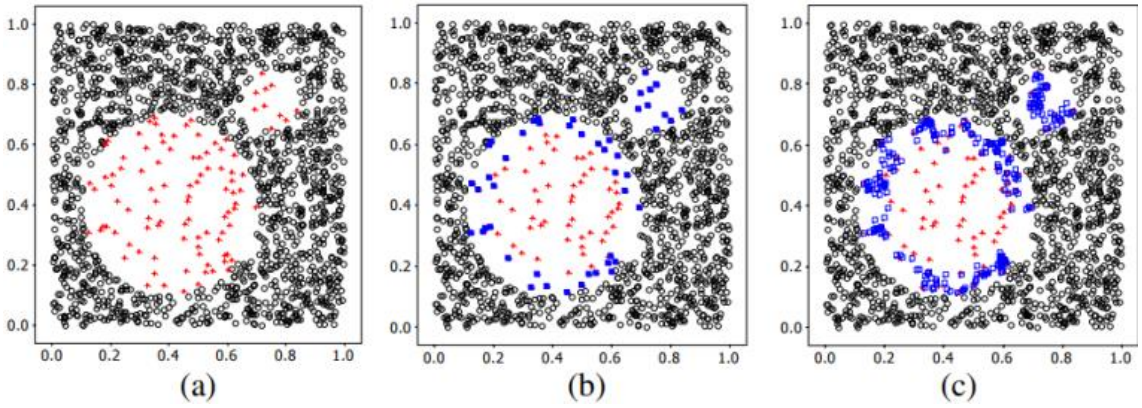
분류하기 더 어려운 부분에 집중

- Borderline Smote

Borderline SMOTE

클래스 범위 내에서만 데이터를 생성하여

특정 분포에만 **중복적으로 데이터가 발생되는 SMOTE를 개선**



다른 class와의 borderline에 있는 샘플들을 늘림으로써

분류하기 더 어려운 부분에 집중

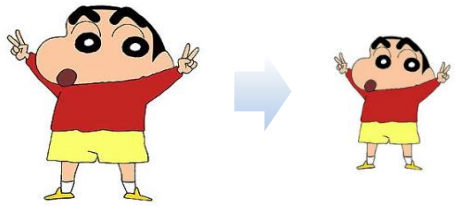
3

차원축소

3 차원 축소

- 차원축소 방법 선정

차원축소



PCA

FA

SVD

Stepwise Selection

LDA

Forward Selection

RandomForest

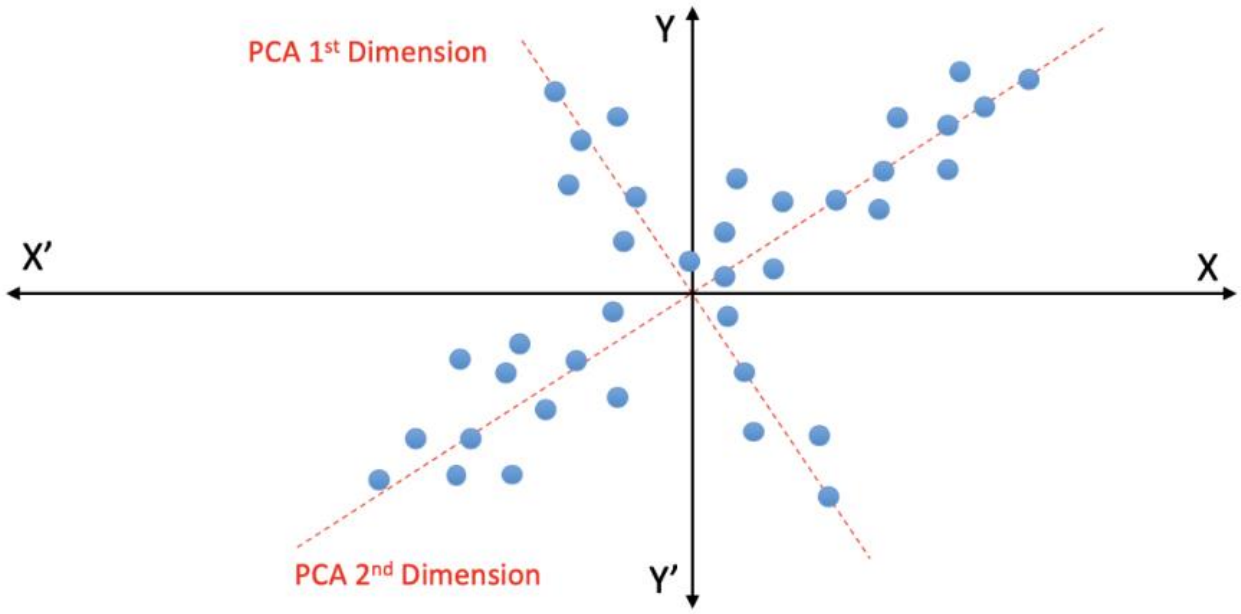
시간 복잡도(계산 시간) 공간 복잡도(저장하는 변수의 양) ↓

고차원 데이터 → 저차원 데이터 (Overfitting 방지)

3 차원 축소

- PCA

PCA Principal Component Analysis = 주성분 분석



데이터의 분산을 최대한 보존하면서 서로 직교하는 새 축을 찾아

고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환

3 차원 축소

- PCA

PCA Principal Component Analysis = 주성분 분석

var_0	var_1	var_2	var_3	⋯	var_196	var_197	var_198	var_199
5.0702	-0.5447	9.5900	4.2987	⋯	6.6576	9.2553	14.2914	-7.6652
16.3699	1.5934	16.7395	7.3330	⋯	9.6846	9.0419	15.6064	-10.8529
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
9.7148	-8.6098	13.6104	5.7930	⋯	6.7980	10.0342	15.5289	-13.9001

200 columns

var_0	var_1	var_2	var_3	⋯	var_183	var_184	var_185	var_186
1.770103	0.334118	0.008692	1.454354	⋯	-1.421193	-1.274556	1.149587	-0.055090
-1.659048	-0.029323	-0.732577	-0.023496	⋯	0.647421	-0.076577	0.629799	0.490878
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.187560	0.310526	1.453793	-0.320839	⋯	-0.542791	-0.796318	1.246547	0.385255

187 columns

95% 설명력을 지니는 변수로 차원 축소

4

Data set 선정

4 Data set 선정

- Data set 조합

Baseline Model

LogisticRegression(random_state=0)

PCA X	PCA O
SMOTE F1 score : 0	SMOTE F1 score : 0.2696
Borderline SMOTE F1 score : 0	Borderline SMOTE F1 score : 0.0733

4 Data set 선정

- Data set 조합

Baseline Model

LogisticRegression(random_state=0)

PCA X	PCA O
SMOTE F1 score: 0	SMOTE F1 score : 0.2696
Borderline SMOTE F1 score: 0	Borderline SMOTE F1 score : 0.0733

5

모델링

5 모델링

- 지난 5일간의 여정…

modelling for imbalanced data

train data + PCA

Balanced Bagging	Easy Ensemble
Random Forest	Logistic Regression → cutoff value 조정

5 모델링

- 지난 5일간의 여정…

SMOTE/Borderline SMOTE + PCA

MLP (Multi-Layer Perceptron)	SVM	XGBoost
CatBoost	LGBM	KNN
Random Forest	Logistic Regression	Decision Tree

5 모델링

- 지난 5일간의 여정…

SMOTE/Borderline SMOTE + PCA

MLP (Multi-Layer Perceptron)	SVM	XGBoost
CatBoost	LGBM	KNN
Random Forest	Logistic Regression	Decision Tree

5 모델링

- 지난 5일간의 여정…

catboost

· 범주형 변수를 처리하는데 중점을 둔 알고리즘

· 오버피팅을 막기 위해 ordering principle, feature combination 등의 방법 사용

과연 f1 score는??

LGBM

· Gradient Boosting 프레임워크로 Tree기반 학습 알고리즘

· 트리가 수직적으로 확장한다는 점에서 기존의 다른 알고리즘과 차별화

과연 f1 score는??

5 모델링

- 지난 5일간의 여정…

catboost

iterations=11

max_depth=10

eval_metric='Accuracy'

learning_rate=0.2

loss_function='MultiClass'

random_state=0

f1 score : 0.4242

Kaggle: 0.3986

LGBM

learning_rate=0.01

min_data_in_leaf=91

num_iterations=100

max_depth=-1, boosting=gbdt

objective='binary', random_state=0

metric='auc', num_leaves=51

is_training_metric='True'

f1 score : 0.4305

Kaggle: 0.4016

5 모델링

- 지난 5일간의 여정…

Logistic Regression

종속변수가 범주형인
데이터를 대상으로 하는 분류
(classification) 기법

과연 f1 score는??

SVM

주어진 데이터가 어느 범주에 속할지
판단하는 이진 선형 분류 모델

초평면을 이용

과연 f1 score는??

5 모델링

- 지난 5일간의 여정…

Logistic Regression

C=0.00069519279617

penalty='l2'

solver='saga'

max_iter=100

random_state=0

f1 score: 0.4105

Kaggle: 0.375

SVM

C=0.2

random_state=0

kernel='poly'

f1 score: 0.3685

Kaggle: 0.3857

6

최종결과

6 최종결과

- 최종모델선정

C=0.2 , random_state=0 , kernel='poly', degree = 3

Kaggle score : **0.41891**

6 한계 및 의의

한계

1. Grid Search 오류 발생으로 손 튜닝
2. 0.5를 넘지 못했..
3. 차원축소를 활용하였지만 변수를 많이 줄이지 못하였다
4. 코로나로 인하여 온라인으로 진행
5. 수적 열세

6 한계 및 의의

- 의의

1. 샘플링 / 차원축소 / 모델링 기법 아는 거 총출동!

그 누구보다 많이 했다고 생각하는데...

2. 하루도 빠짐없이 매일 회의 (**희로애락** 공유)

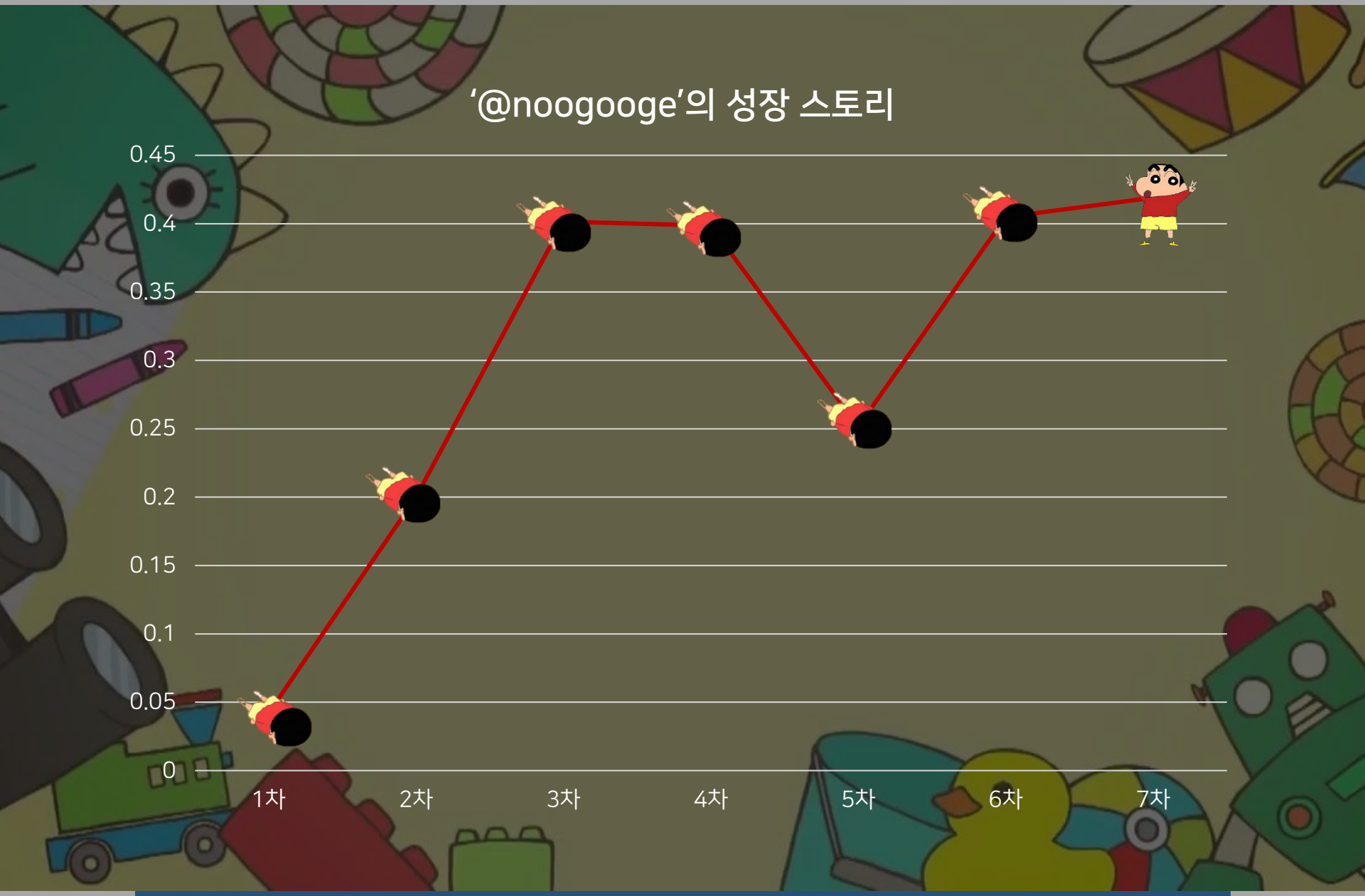
4팀 중 제일 먼저 회의 시작! (아마도..)

3. 단체 사진 & 인스타 팔로우

4. 우리 팀은 4인이라 뒷풀이도 할 수 있지롱

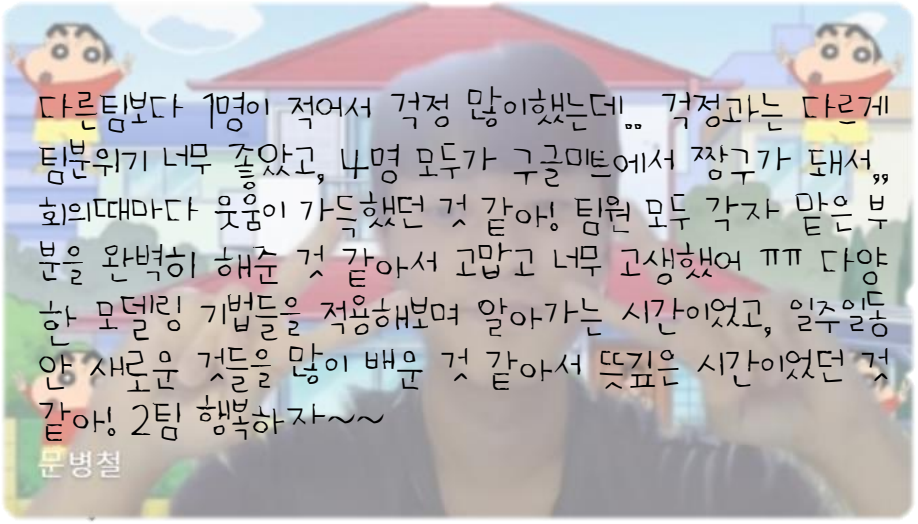
점심 한정

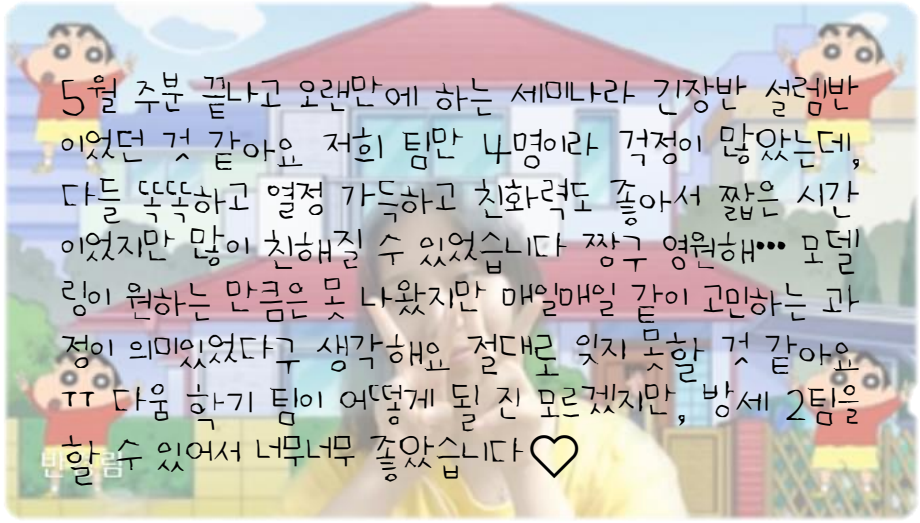
6 제 5의 팀원

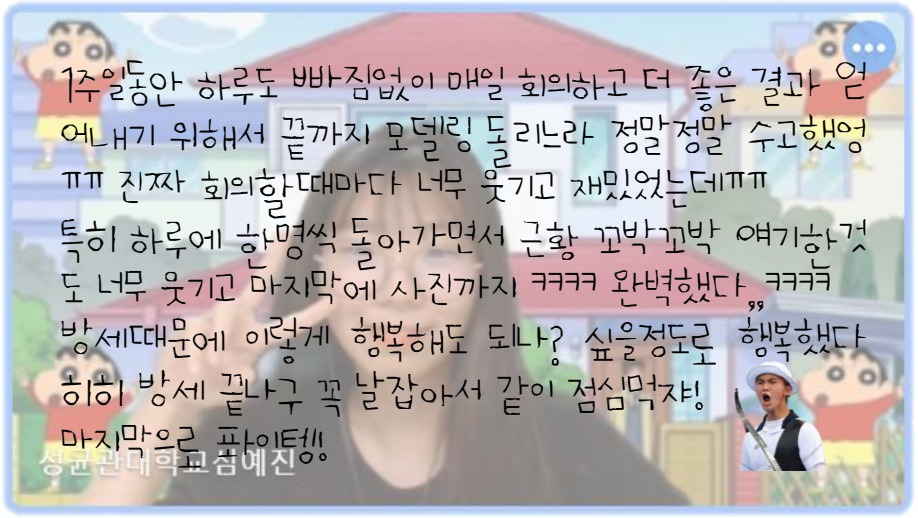

'@noogooge'의 성장 스토리
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1차
2차
3차
4차
5차
6차
7차

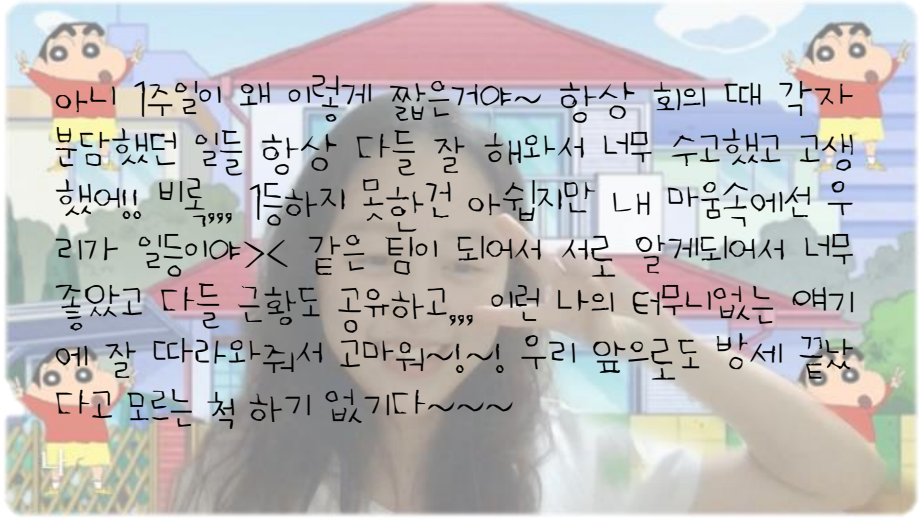
6 최종결과

소감ㅠㅠ


다른팀보다 1명이 적어서 걱정 많이했는데.. 걱정과는 다른게 팀분위기 너무 좋았고, 4명 모두가 구글미트에서 짱구가 돼서,, 회의때마다 웃음이 가득했던 것 같아! 팀원 모두 각자 맡은 부분을 완벽히 해준 것 같아서 고맙고 너무 고생했어 ㅠㅠ 다양한 모델링 기법들을 적용해보며 알아가는 시간이었고, 일주일동안 새로운 것들을 많이 배운 것 같아서 뜻깊은 시간이었던 것 같아! 2팀 행복하자~~
문병철


5월 주분 끝나고 오랜만에 하는 세미나라 긴장반 설렘반 이었던 것 같아요 저희 팀만 4명이라 걱정이 많았는데, 다들 똑똑하고 열정 가득하고 친화력도 좋아서 짧은 시간이었지만 많이 친해질 수 있었습니다 짱구 영원해••• 모델링이 원하는 만큼은 못 나왔지만 매일매일 같이 고민하는 과정이 의미있었다구 생각해요 절대로 잊지 못할 것 같아요 ㅠㅠ 다음 학기 팀이 어떻게 될 진 모르겠지만, 방세 2팀을 할 수 있어서 너무너무 좋았습니다♡


1주일동안 하루도 빠짐없이 매일 회의하고 더 좋은 결과 얻에내기 위해서 끝까지 모델링 돌리느라 정말정말 수고했엉 ㅠㅠ 진짜 회의할때마다 너무 웃기고 재밌었는데ㅠㅠ
특히 하루에 한명씩 돌아가면서 근황 꼬박꼬박 얘기하는것도 너무 웃기고 마지막에 사진까지 ㅋㅋㅋㅋ 완벽했다,, ㅋㅋㅋㅋ
방세때문에 이렇게 행복해도 되나? 싶을정도로 행복했다 히히 방세 끝나구 꼭 날잡아서 같이 점심먹자!
마지막으로 파이팅!!
성균관대학교 심예진


아니 1주일이 왜 이렇게 짧은거야~ 항상 회의 때 각자 분담했던 일들 항상 다들 잘 해와서 너무 수고했고 고생했엉!! 비록,,, 1등하지 못한건 아쉽지만 내 마음속에선 우리가 일등이야>< 같은 팀이 되어서 서로 알게되어서 너무 좋았고 다들 근황도 공유하고,,, 이런 나의 터무니없는 얘기에 잘 따라와줘서 고마워~!~! 우리 앞으로도 방세 끝났다고 모르는 척 하기 없기다~~~
나

THANK YOU