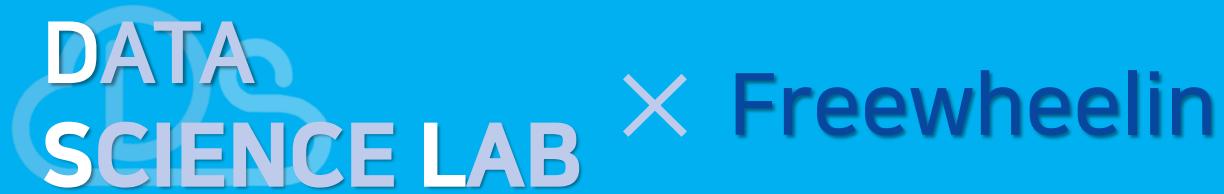


기업연계프로젝트 : 프리월린



CONTENTS

- 01 회사 및 데이터 소개
- 02 **프리월린 C**
 - 학생별 실력 측정 지표 정의
 - 서비스 효과성 검증
- 03 **프리월린 B**
 - Feature extraction for clustering
 - K-means clustering
- 04 **프리월린 A**
 - K-means(DTW) Time series Clustering
 - Embedding 기반 정답률 예측 모델

Freewheelin

수학 문제은행 솔루션 개발 (5억 건의 채점 데이터)



Freewheelin

수학 문제은행 솔루션 개발 (5억 건의 채점 데이터)



매쓰플랫

수업 준비 수업

학습지 교재 테마별 추천학습지

내 학습지 즐겨찾는 문제

전체 초 중 고 과목 전체 학습지 태그 전체

선택 학년 태그 학습지명

기하 기하 공간도형 기출(Killer)
공간도형 기출(Killer) 예 ~ 구의 방정식

수학 I 수학 I 지수와 로그 자동 계산
38문제 | 중 | 거듭제곱근과 지수법칙 ~ 지수의 확장

수학(상) 고등수학 (상) 다항식 자동 계산
50문제 | 중 | 다항식의 연산 ~ 인수정리

매쓰플랫

수업 준비

학습지 교재

내 교재 시중교재

전체 초 중 고 과목 전체 시중교재 교과서

학년 교재명

수학(상) RPM - 고등수학(상) (2023) 공동이 지원

수학(상) 개념원리 - 고등수학(상) (2023) 공동이 지원

수학(상) 라이트쎈 - 고등수학(상) 공동이 지원

Freewheelin

수학 문제은행 솔루션 개발 (5억 건의 채점 데이터)



“매쓰플랫 서비스를 이용하면 학생들의 실력이 오를까?”

대한민국 1등 수학문제은행.
매쓰플랫

22' 중앙일보 주관 고객우수브랜드 대상
문제은행 서비스 부문 4년 연속 수상

지금 10일 무료 체험하기 →

매쓰플랫 블랙프라이데이 최대 70% OFF(-11.24)

“무엇을 보고 학생들의 실력을 알 수 있을까?”

“많은 양의 채점 데이터를 어떤 방식으로 활용할 수 있을까?”

Freewheelin

수학 문제은행 솔루션 개발 (5억 건의 채점 데이터)



“매쓰플랫 서비스를 이용하면 학생들의 실력이 오를까?”

“실력을 표현하는 지표를 시계열 데이터로 수집하자”

“무엇을 보고 학생들의 실력을 알 수 있을까?”

“실력을 표현하는 지표를 새롭게 만들자”

“많은 양의 채점 데이터를 어떤 방식으로 활용할 수 있을까?”

“학생을 clustering 해서 학습 방식 솔루션을 제공하자”



Freewheelin

수학 문제은행 솔루션 개발 (5억 건의 채점 데이터)



"매쓰플랫 서비스를 이용하면 학생들의 실력이 오를까?"

"실력을 표현하는 지표를 시계열 데이터로 수집하자"

매쓰플랫 서비스 고객층이 파악되지 않아 서비스 개선 방향성을 수립하기 어렵다.

매쓰플랫 서비스가 실제로 학생의 성적 향상에 기여하는지 알 수 없다.

학생들의 학업 능력을 파악할 수 있는 단위가 없다.

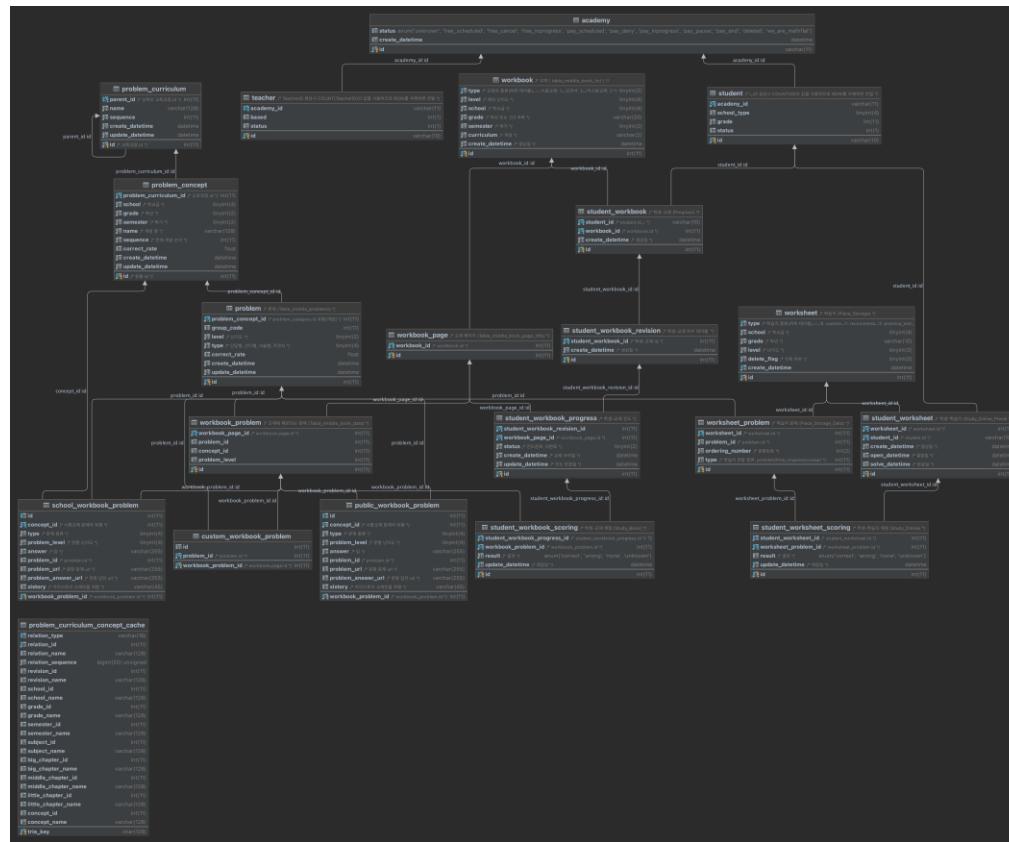


"많은 양의 채점 데이터를 어떤 방식으로 활용할 수 있을까?"

"학생을 clustering 해서 학습 방식 솔루션을 제공하자"

데이터

학생, 학원, 문제, 단원 등 21개의 DataFrame(Table)



problem		
설명	문제	
컬럼명	설명	비고
id	아이디	
problem_concept_id	문제 개념 아이디	problem_concept.id
group_code	그룹 코드	
level	난이도	
type	문제유형	0 - 단답형, 1 - 선다형, 2 - 서술형, 3 - 주관식
correct_rate	출제 수	
create_datetime	최초생성일시	
update_datetime	최종변경일시	

problem_concept		
설명	문제 개념	
컬럼명	설명	비고
id	아이디	
problem_curriculum_id	문제 교육과정 아이디	problem_curriculum.id
school	학교급	1 - 초등학교, 2- 중학교, 3 - 고등학교
grade	학년 또는 고3 과목	
semester	학기	
name	개념명	

Student

Student_id (어떤 학생)	Academy_id (어떤 학원)	Schooltype (초/중/고)	Grade (학년)	...
I7951	D0409	3	3	

Student

Student_id (어떤 학생)	Academy_id (어떤 학원)	Schooltype (초/중/고)	Grade (학년)	...
I7951	D0409	3	3	

Student_worksheet_scoring

Student_id (어떤 학생)	Problem_id (문제 정보)	Date (푼 날짜)	Result (채점결과)
I110			
I7952			
I7951	632513	2021-04-21	wrong
I5517			

Student

Student_id (어떤 학생)	Academy_id (어떤 학원)	Schooltype (초/중/고)	Grade (학년)	...
I7951	D0409	3	3	

Student_worksheet_scoring

Student_id (어떤 학생)	Problem_id (문제 정보)	Date (푼 날짜)	Result (채점결과)
I110			
I7952			
I7951	632513	2021-04-21	wrong
I5517			

Problem

Problem_id (문제 정보)	Curriculum (단원)	Correct_rate (정답률)	...	Level (난이도)
241561				
235671				
632513	미분의 조건	55.22%	...	4
732512				
123152				

Student

Student_id (어떤 학생)	Academy_id (어떤 학원)	Schooltype (초/중/고)	Grade (학년)	...
I7951	D0409	3	3	

Student_worksheet_scoring

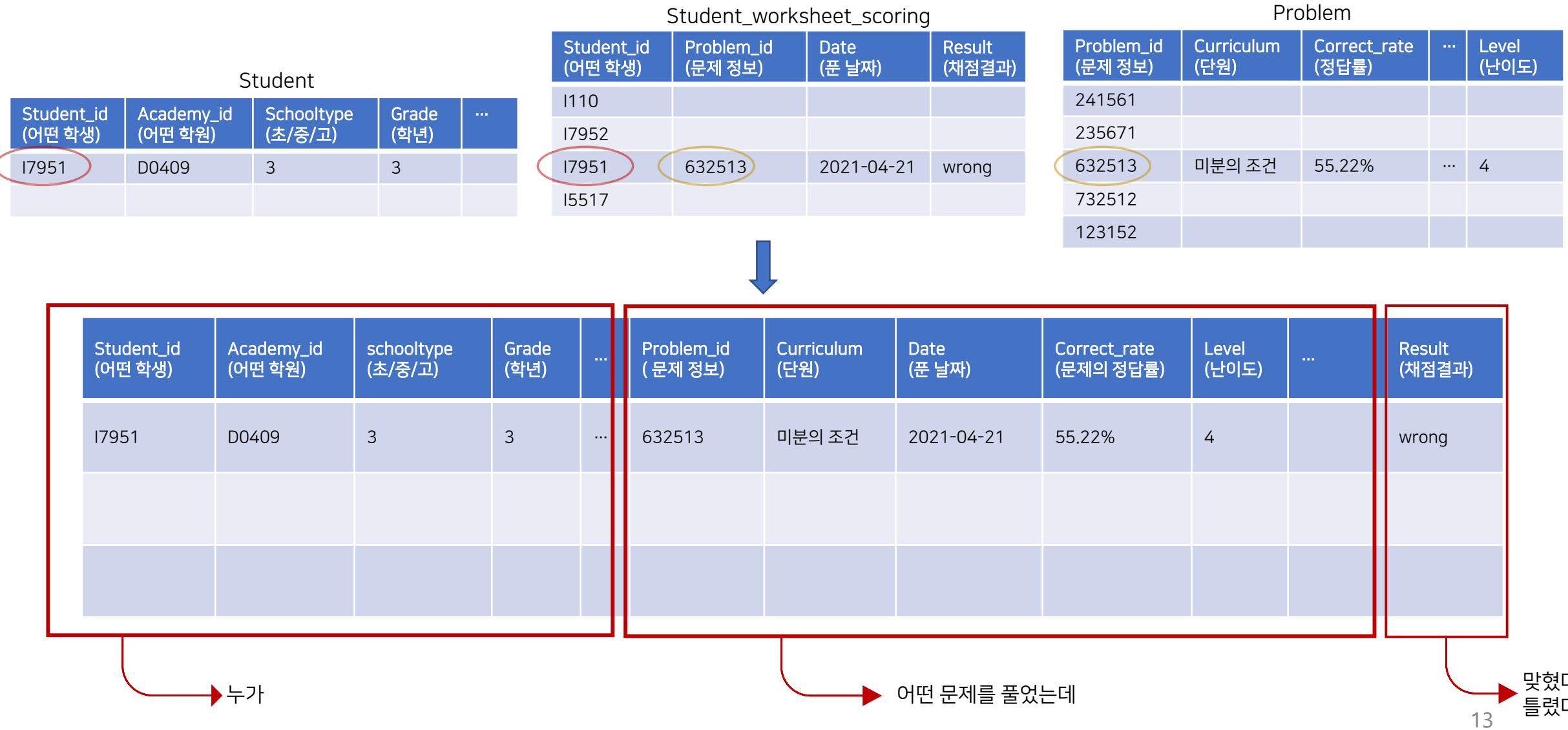
Student_id (어떤 학생)	Problem_id (문제 정보)	Date (푼 날짜)	Result (채점결과)
I110			
I7952			
I7951	632513	2021-04-21	wrong
I5517			

Problem

Problem_id (문제 정보)	Curriculum (단원)	Correct_rate (정답률)	...	Level (난이도)
241561				
235671				
632513	미분의 조건	55.22%	...	4
732512				
123152				

↓

Student_id (어떤 학생)	Academy_id (어떤 학원)	schooltype (초/중/고)	Grade (학년)	...	Problem_id (문제 정보)	Curriculum (단원)	Date (푼 날짜)	Correct_rate (문제의 정답률)	Level (난이도)	...	Result (채점결과)
I7951	D0409	3	3	...	632513	미분의 조건	2021-04-21	55.22%	4		wrong



CONTENTS

- 01 회사 및 데이터 소개
- 02 **프리월린 C**
 - 학생별 실력 측정 지표 정의
 - 서비스 효과성 검증
- 03 **프리월린 B**
 - Feature extraction for clustering
 - K-means clustering
- 04 **프리월린 A**
 - K-means(DTW) Time series Clustering
 - Embedding 기반 정답률 예측 모델

프리월린 C

- 1) 학생별 실력 측정 지표 정의
- 2) 서비스 효과성 검증

#1 문제 정의

Problem Defining

Problem Defining

What is the problem?

학생들의 수학 실력을 정량화할 수 있을까? 매쓰플랫의 효과일까?

매쓰플랫 서비스가 학생들의 실력 향상에 효과적인가?

Problem1

성적 = 실력?

데이터의 부재

1. 내신 성적 / 모의고사 성적 데이터의 부재
2. But 1, 내신 수준은 학교마다 상이
3. But 2, 내신, 모의고사 성적이 다른 경우 어떤 성적을 신뢰할까?

→ 우선 로깅을 제안할 예정

Problem2

정답률 = 실력?

난이도 & 문제 풀이량 등 고려 x

summary 평가결과요약



1. 새로운 유형만 푼 경우?
2. 새로운 챕터에 진입한 경우?
3. 난이도가 높은 문제만 푼 경우?

Problem3

매쓰플랫 때문?

너무 많은 다양성

1. 각 학원마다 학생들의 실력이 다르다.
2. 선생님마다 강의력 / 관리 수준이 다르다.
3. 각 학원마다 관리하는 방식이 다르다.

Problem Defining

What is the problem?

학생들의 수학 실력을 정량화할 수 있을까? 매쓰플랫의 효과일까?

매쓰플랫 서비스가 학생들의 실력 향상에 효과적인가?

Quest1

실력 지표 정의

Problem1

성적 = 실력?

데이터의 부재

1. 내신 성적 / 모의고사 성적 데이터의 부재
2. But 1, 내신 수준은 학교마다 상이
3. But 2, 내신, 모의고사 성적이 다른 경우 어떤 성적을 신뢰할까?

→ 우선 로깅을 제안할 예정

Quest2

서비스 효과성 검증

Problem3

매쓰플랫 때문?

너무 많은 다양성

1. 각 학원마다 학생들의 실력이 다르다.
2. 선생님마다 강의력 / 관리 수준이 다르다.
3. 각 학원마다 관리하는 방식이 다르다.

Problem2

정답률 = 실력?

난이도 & 문제 풀이량 등 고려 x

summary 평가결과요약



1. 새로운 유형만 푼 경우?
2. 새로운 챕터에 진입한 경우?
3. 난이도가 높은 문제만 푼 경우?

Problem Defining

Various Considerations

5W1H

실력 측정 지표는 무엇인가?

Who

What

Where

When

Why

How

- 1. 매쓰플랫이
- 2. 선생님이
- 3. 학생이
- 4. 학원이

- 1. 학생 수학 실력을

- 1. 회사에서
- 2. 학원에서
- 3. 집에서

- 1. 문제를 푼 후
- 2. 시험이 끝난 후
(매일/분기/연)
- 3. 시간이 지난 후

- 1. 서비스 효과성을
확인하기 위해서
- 2. 실력 향상을 확인
하기 위해서

해결해야 하는 지점!

Problem Defining

Various Considerations

5W1H

실력 측정 지표는 무엇인가?

Who

What

Where

When

Why

How

- 1. 매쓰플랫이
- 2. 선생님이
- 3. 학생이
- 4. 학원이

- 1. 학생 수학 실력을

실력 지표 for whom?

- 1. 회사에서
- 2. 학원에서
- 3. 집에서

- 1. 문제를 푼 후
- 2. 시험이 끝난 후
(매일/분기/연)
- 3. 시간이 지난 후

- 1. 서비스 효과성을
확인하기 위해서
- 2. 실력 향상을 확인
하기 위해서

해결해야 하는 지점!

#2 서비스 분석

Service Analysis

Service Analysis

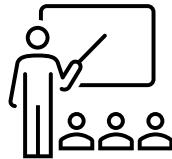
Market Players Analysis

Customer

학생



선생님



학부모



Company

Services

1. 선생님 대상 서비스 (매쓰플랫)
2. 학생 대상 서비스 (풀리 - 런칭 준비중)

Strength

1. 다양한 문제은행 자산
2. 선생님 대상 서비스 업계 선두
3. 실제 학원 선생님들이 만드는 서비스

Weakness

1. 데이터 Team 부재
2. 데이터 기반 의사결정 ↓
3. Data Analytics / Data Science ↓

Competitor

학생 대상 서비스들

1. 선생님 매칭 서비스
2. 문제 추천/풀이 서비스
3. 강의 서비스 (온/오프라인)

선생님 대상 서비스들

1. 문제 추천 서비스
2. 채점 / 성적 관리 서비스
3. 학생 - 학부모 연락 편리화 서비스

학부모 대상 서비스들

1. 자녀 성적 확인 / 관리
2. 자녀 학원 스케줄 관리

Service Analysis

Market Players Analysis

Customer

학생



선생님



학부모



실력 지표 for whom?

Services

1. 선생님 대상 서비스 (매쓰플랫)
2. 학생 대상 서비스 (풀리 - 런칭 준비중)

Strength

1. 다양한 문제은행 자산
2. 선생님 대상 서비스 업계 선두
3. 실제 학원 선생님들이 만드는 서비스

Weakness

1. 데이터 Team 부재
2. 데이터 기반 의사결정 ↓
3. Data Analytics / Data Science ↓

Competitor

학생 대상 서비스들

1. 선생님 매칭 서비스
2. 문제 추천/풀이 서비스
3. 강의 서비스 (온/오프라인)

선생님 대상 서비스들

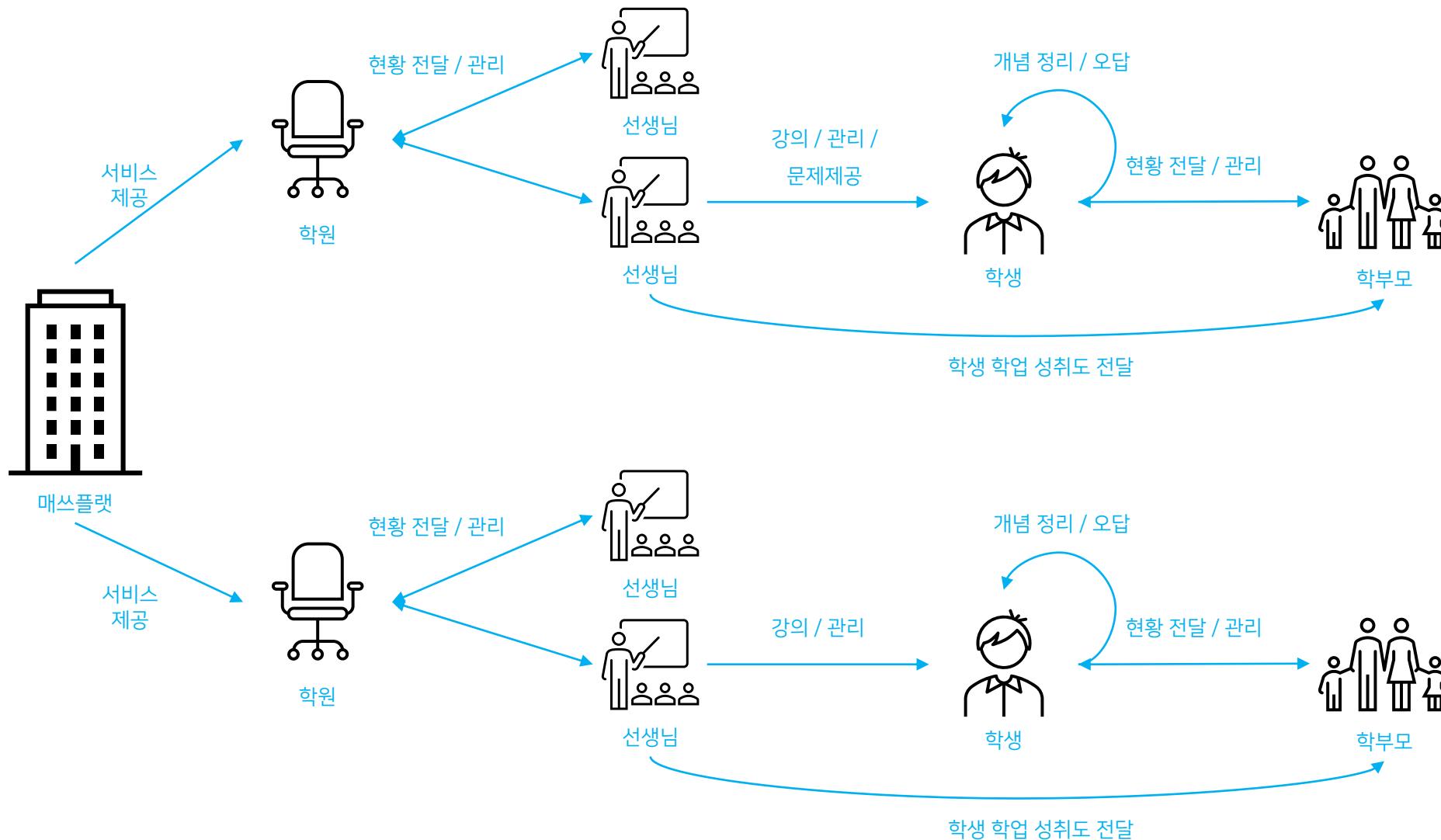
1. 문제 추천 서비스
2. 채점 / 성적 관리 서비스
3. 학생 - 학부모 연락 편리화 서비스

학부모 대상 서비스들

1. 자녀 성적 확인 / 관리
2. 자녀 학원 스케줄 관리

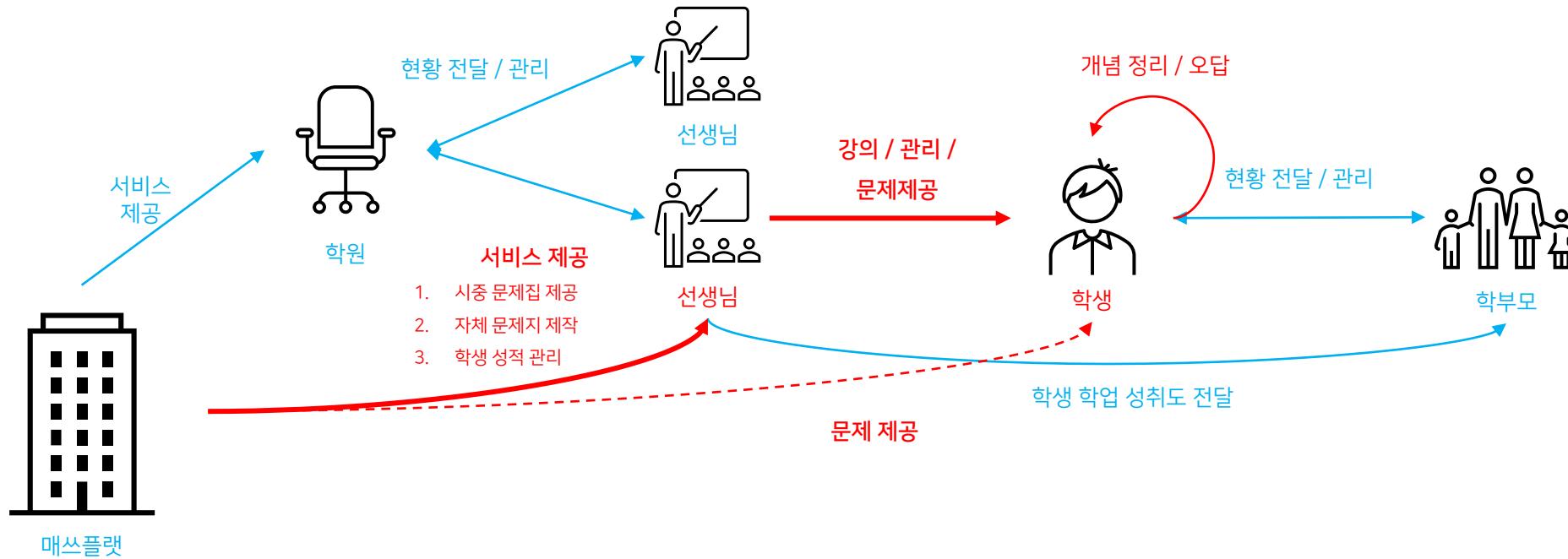
Service Analysis

Target Relationship Analysis



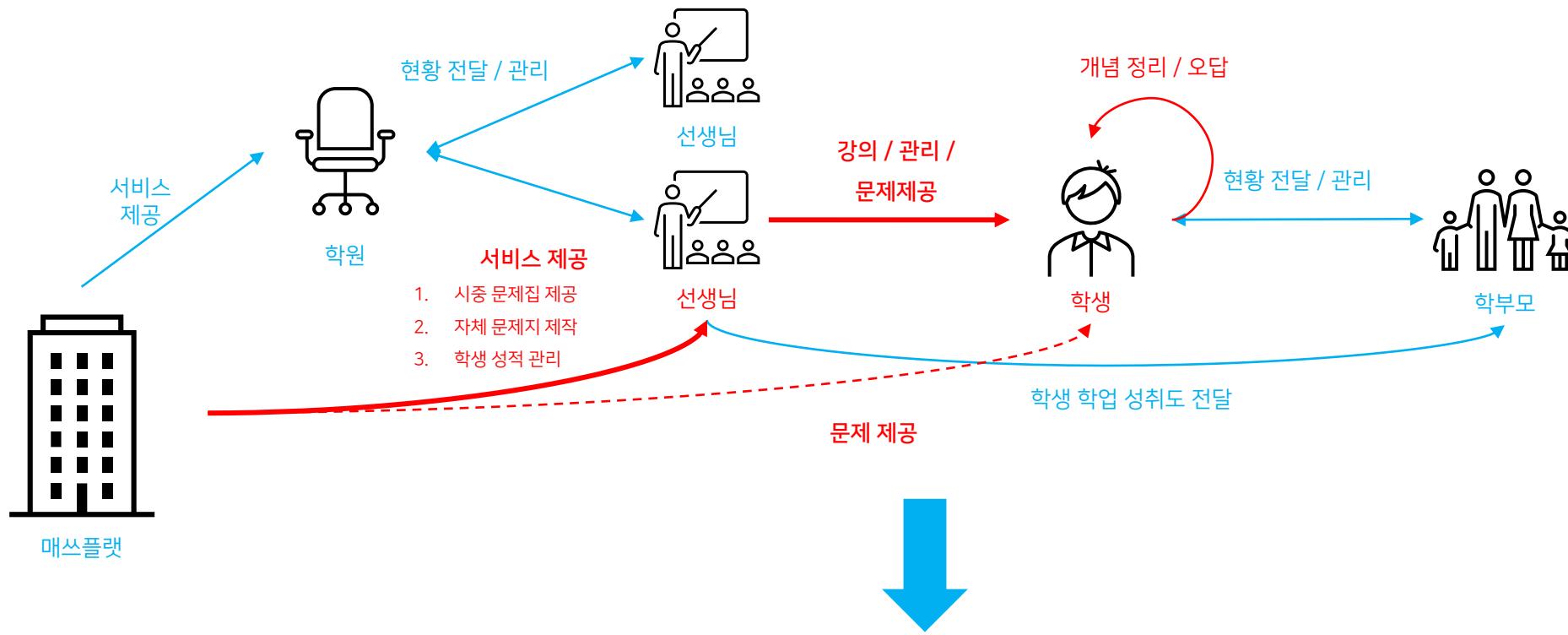
Service Analysis

Target Relationship Analysis



Service Analysis

Target Relationship Analysis



실력 향상!

1. By 적절한 문제 제공
2. 오답 관리

Service Analysis

Target Service Analysis

Workbook

"내 수업에서 쓰는 교재가 등록되어 있네? 자동채점에 유사문제도 있잖아?"

시중교재, 교과서 연동

The screenshot shows a dashboard with various sections: '교과' (Curriculum), '교재' (Textbook), '교과서' (Curriculum Textbook), and '교재 등록' (Textbook Registration). It includes a search bar and a sidebar with user information and navigation links.

학생별 교재 등록 → 시중교재 채점 진행 → 오답에 대한 유사문제 생성 → 평가보고서 학습평가

Worksheet

"오늘 나갈 단원만 골라서 간편하게 학습지를 만들어 볼까?"

빠르고 쉬운 학습지 제작

The screenshot shows a configuration interface for creating worksheets. It includes sections for '선택 단원' (Selected Chapter), '선택 단원' (Selected Chapter), and '선택 단원' (Selected Chapter). Below are buttons for '학습지 생성' (Worksheet Generation), '범위, 문항수, 난이도 설정' (Range, Question Count, Difficulty Setting), '개별 문제 확인/추가' (Individual Problem Check/Add), and '디자인 설정 및 인쇄' (Design Settings and Print).

"지난 시간에 아이들이 틀린 문제로 오답 클리닉 학습지를 만들어볼까?"

학생별 맞춤 오답관리

다양한 맞춤 학습지 제작 가능

The screenshot shows a configuration interface for creating worksheets based on student errors. It includes a list of student names and error types. Below are buttons for '기간별 학습지' (Period-based Worksheet), '단원별 학습지' (Chapter-based Worksheet), '무한 유사문제 학습지' (Infinite Similar Problem Worksheet), and '시중교재 연동 학습지' (Curriculum Textbook Integrated Worksheet).

Report

"학생별 학습량/정답률을 학부모님 상담할 때 쓰기 좋은데?"

학습관리 리포트 제공

테마별 분석 리포트 제공

The screenshot shows a report generation interface. It includes sections for '학습 대상 선택' (Select Learning Target), '학부모님 상담' (Consultation with Parents), and '보고서 제작' (Report Generation). Below are buttons for '학습내용별 리포트' (Content-based Report), '난이도/유형별 리포트' (Difficulty/Type-based Report), '학부모 앱 안내' (Introduction to Parent App), and '선생님 코멘트' (Teacher Comments).

Service Analysis

Target Service Analysis

Workbook

"내 수업에서 쓰는 교재가 등록되어 있네? 자동채점에 유사문제도 있잖아?"

시중교재, 교과서 연동

학생별 교재 등록

시중교재 채점 진행

오답에 대한 유사문제 생성

평가보고서 학습평가

Worksheet

"오늘 나갈 단원만 골라서 간편하게 학습지를 만들어 볼까?"

빠르고 쉬운 학습지 제작

학습지 생성

범위, 문항수, 난이도 설정

개별 문제 확인/추가

디자인 설정 및 인쇄

"지난 시간에 아이들이 틀린 문제로 오답 클리닉 학습지를 만들어볼까?"

학생별 맞춤 오답관리

다양한 맞춤 학습지 제작 가능

기간별 학습지

단원별 학습지

무한 유사문제 학습지

시중교재 연동 학습지

Report

"학생별 학습량/정답률을 학부모님 상담할 때 쓰기 좋은데?"

학습관리 리포트 제공

테마별 분석 리포트 제공

학습내용별 리포트

난이도/유형별 리포트

학부모 앱 안내

선생님 코멘트



1. 적절한 문제 제공
2. 오답 관리

Service Analysis

Target Service Analysis

Workbook

"내 수업에서 쓰는 교재가 등록되어 있네? 자동채점에 유사문제도 있잖아?"

시중교재, 교과서 연동

학생별
교재 등록

시중교재
채점 진행

오답에 대한
유사문제 생성

평가보고서
학습평가

Worksheet

"오늘 나갈 단원만 골라서 간편하게 학습지를 만들어 볼까?"

빠르고 쉬운 학습지 제작

학습지
생성

범위, 문항수,
난이도 설정

개별 문제
확인/추가

디자인 설정
및 인쇄

"지난 시간에 아이들이 틀린 문제로 오답 클리닉 학습지를 만들어볼까?"

학생별 맞춤 오답관리

다양한 맞춤 학습지 제작 가능

기간별
학습지

단원별
학습지

무한 유사문제
학습지

시중교재
연동 학습지

Report

"학생별 학습량/정답률을 학부모님 상담할 때 쓰기 좋은데?"

학습관리 리포트 제공

테마별 분석 리포트 제공

학습내용별
리포트

난이도/유형별
리포트

학부모 앱
안내

선생님
코멘트

- 회사 자체 서비스
- 효과성 검증 Needs ↑

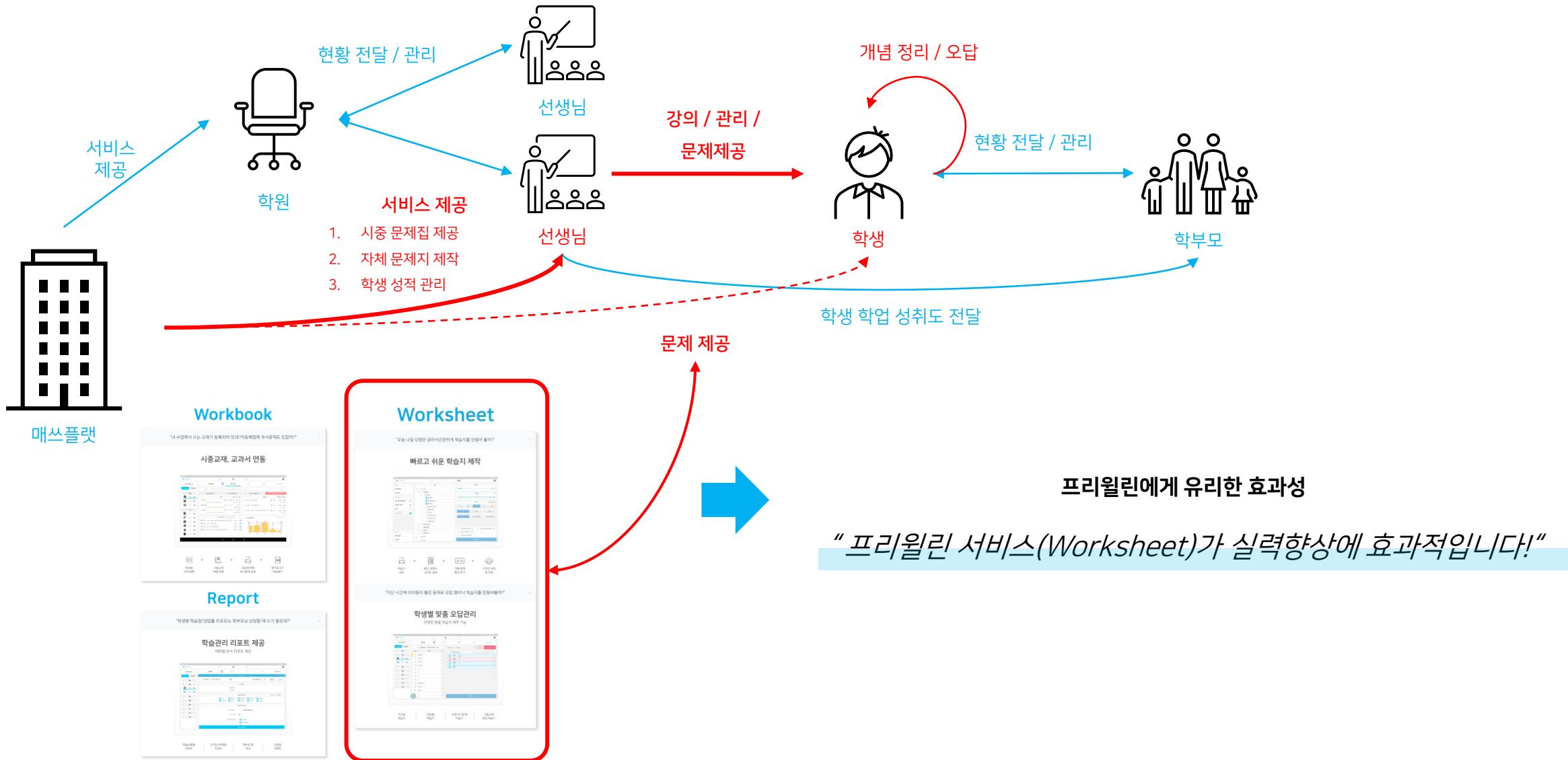


- 적절한 문제 제공
- 오답 관리



Service Analysis

Target Relationship Analysis



#3 실력 Score Metric

Service Analysis

Score Metric

기존 Score Metric problem

기존 서비스의 실력 평가 방법

summary 평가결과요약

내용 영역 성취율

89 %

행동 영역 성취율

88 %

종합 성취율

89 %

종합 등급

2 등급

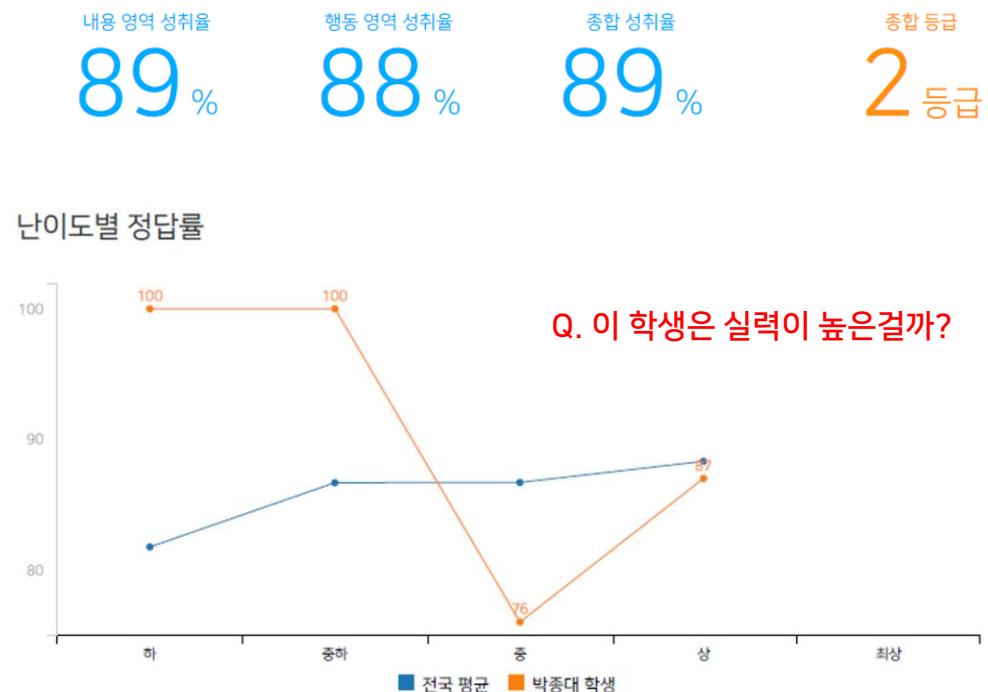
“정답률”만 사용해서 학생의 실력을 평가하고 있다!

Score Metric

기존 Score Metric problem

기존 서비스의 실력 평가 방법

summary 평가결과요약



정답률만 고려할 경우,



학생 A



학생 B

문 문제: 난이도 상 30개
정답률 : 50%

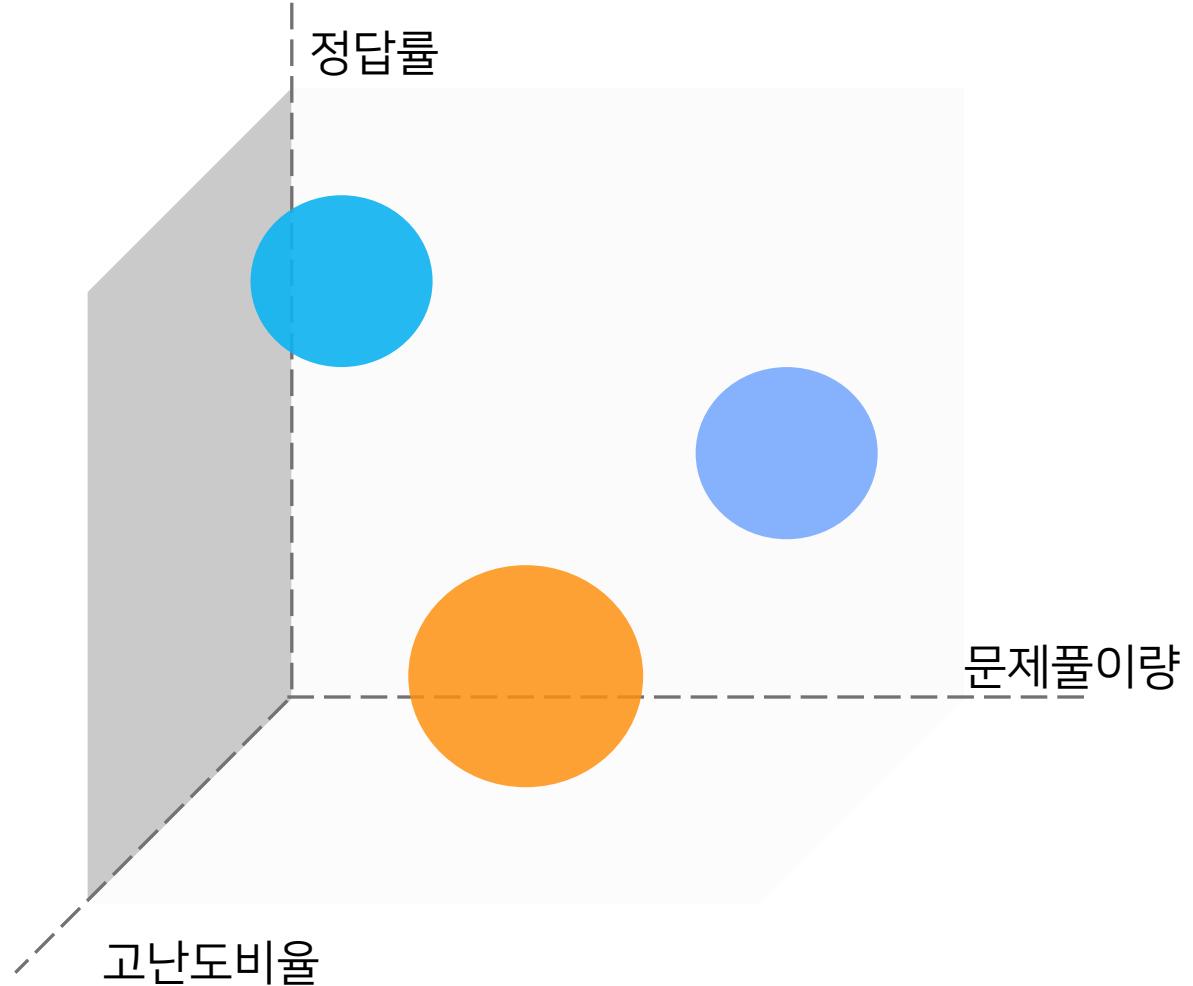
문 문제: 난이도 하 40개
정답률 : 80%

누가 더 실력이 높은지 평가 불가

새로운 실력 지표 필요

Score Metric

Score Metric Defining



실력 향상 요소

1. 성공 경험 (정답률)
2. 유형 숙지 정도 (문제 풀이량)
3. 개념 응용능력 (고난도 비율)

시각적으로 관찰하기엔 Good

학생들의 실력을 평가하기엔 Bad

3가지의 정보를 결합하여 하나의 스코어로 표현하자!

Score Metric

Score Metric Defining

난이도별 정답률의 합

Score Metric

$$score = \sum_{i=1}^5 (level i \times level i correct rate)$$

User Action	Result	Score		
난이도	문제풀이량	난이도	문제풀이량	
1	1000	1	A 90%	
2	800	2	B 73%	
3	700	3	C 75%	10.94점 (0~15점 척도)
4	400	4	D 67%	
5	100	5	E 73%	

Score Metric

Score Metric Application

학생들의 실력 향상 여부 및 정도 확인

개인 실력 상승 확인

학생 한명의 Score 기준 절대적 위치

예)

- 실력변화
- 실력상승 속도
- 단원별 실력변화 추이
- 시간이 지남에 따른 실력향상 가속도
- 시간이 지남에 따른 문제푸는 양 변화

상대적 실력 상승 확인

다른 그룹과 비교한 상대적 위치

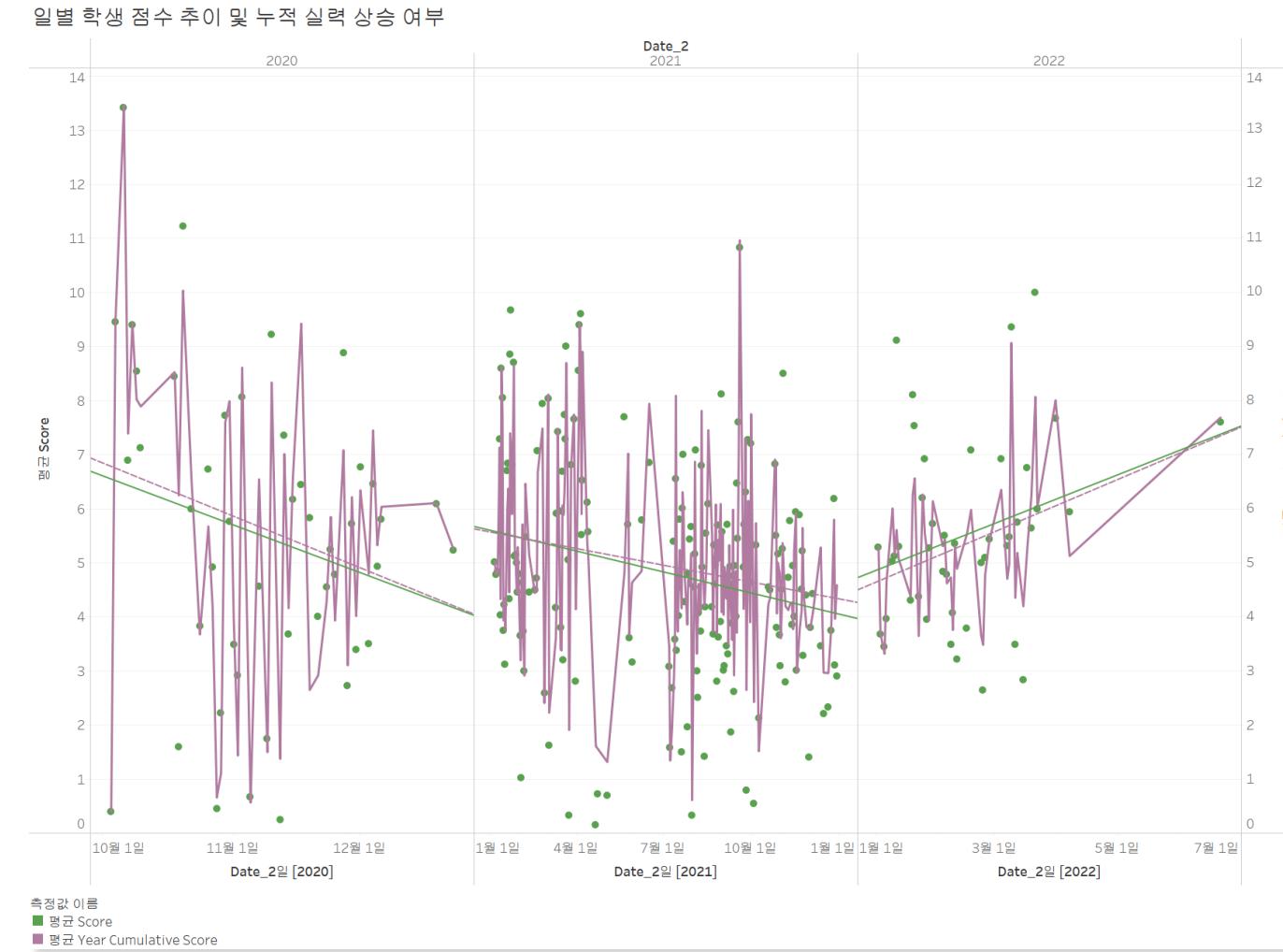
예)

- 전체 평균대비 실력변화 정도
- 학생의 시작 등급 구간 내 상승 정도
- 학생의 시작 등급 구간 내 상승 속도
- 학생의 시작 등급 구간 내 문제 푸는 양

Score Metric

Market Players Defining

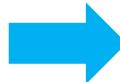
전처리 코드



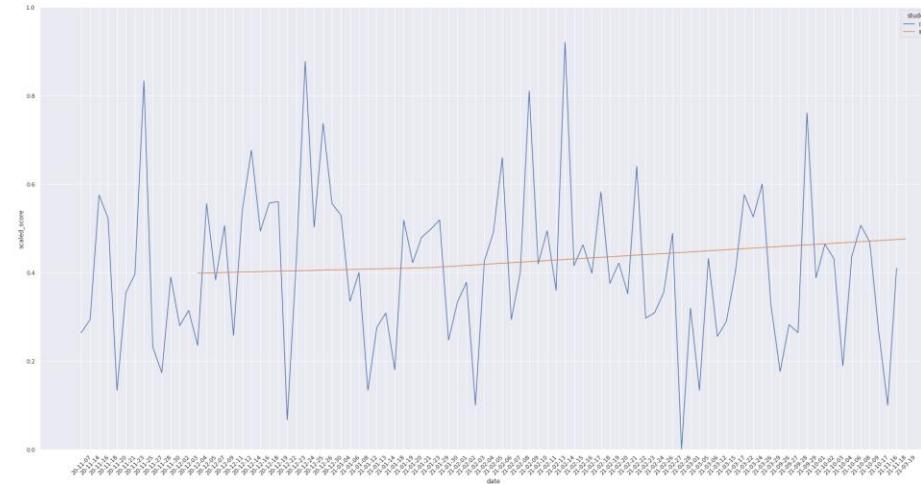
Score Metric

Market Players Defining

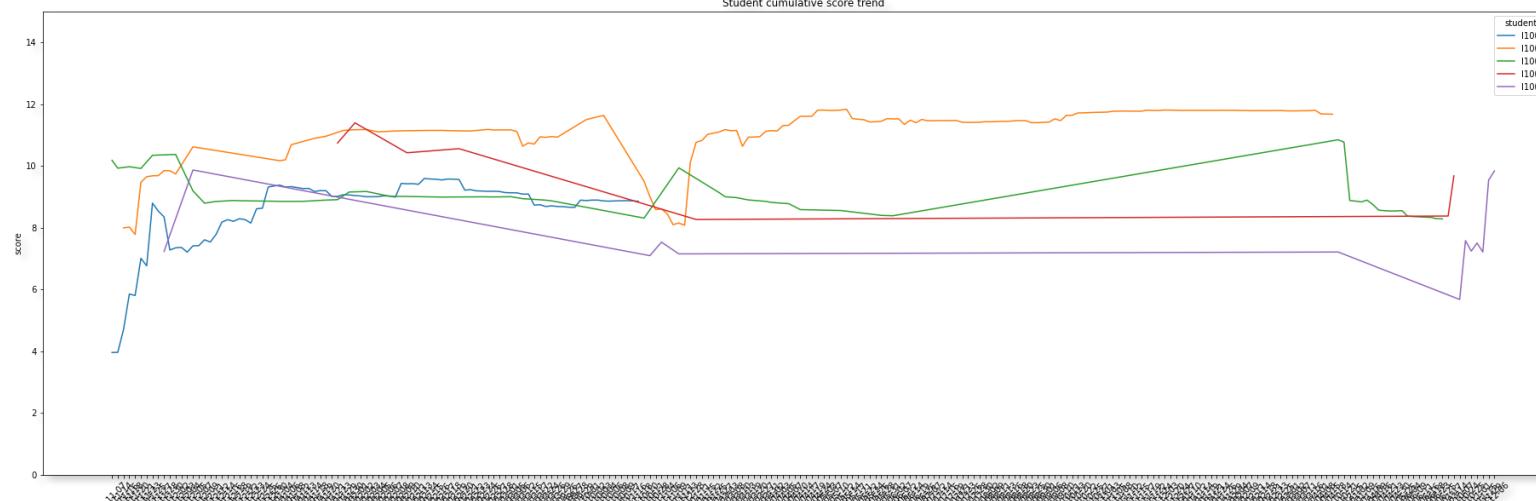
전처리 코드



상대적 성적 향상 여부 확인



특정 2학년 일별 성적 비교



동일 학원 내 학생들 누적 실력 비교

#4 서비스 효과성 검증

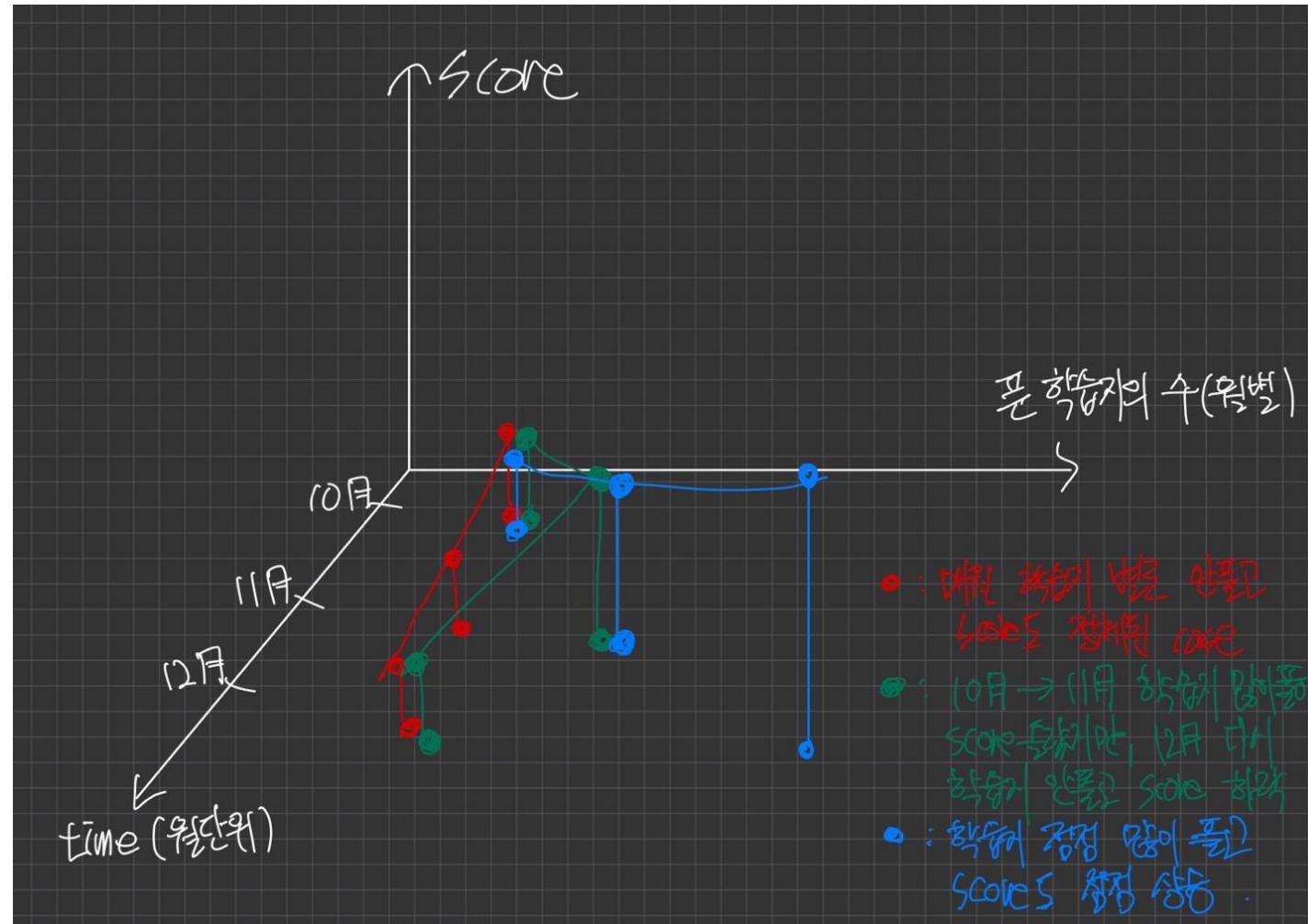
Service Effectiveness Analysis

Service Effectiveness Analysis

Case Analysis

실제로 Worksheet이 효과적일까?

Q1. Worksheet을 많이 풀수록 실력이 오를까?



케이스 분리

1. 매월 학습지를 적게 풀고 Score도 정체된 Case
2. 10월 → 11월에 학습지를 많이 풀며 Score 올랐지만, 12월에 다시 학습지를 안 풀어서 Score 하락
3. 학습지를 점점 많이 풀며 Score도 점점 상승

분석 대상

: 매쓰플랫을 3개월 이상 이용한
2022년 고3 학생

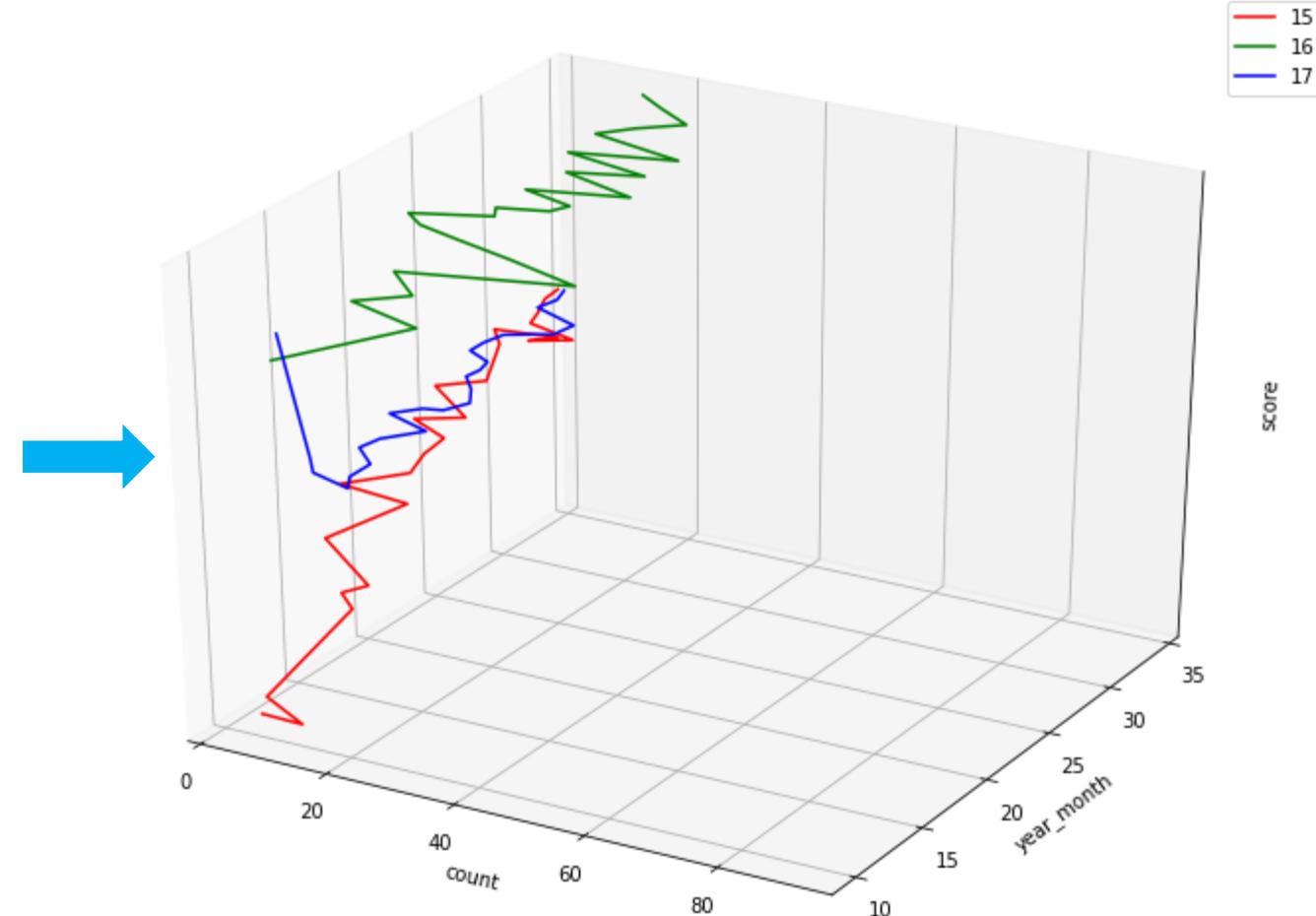
분석 기간

: 2020-10(고1 시점) ~ 2022-10(현재)

Service Effectiveness Analysis

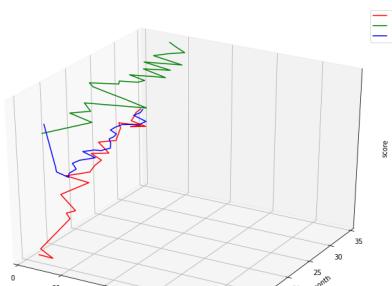
Case Analysis

Student.id (학생 ID)	Year_month (연-월)	Count_Worksheet.id (월별 학습지 개수)	Score (점수)
I100008	202010	27	8.7935
I100008	202011	62	9.3481
I100008	202012	22	9.5926
...			
I100008	202210	58	11.2263
IB113	202101	8	6.0359
IB113	202102	10	6.2926

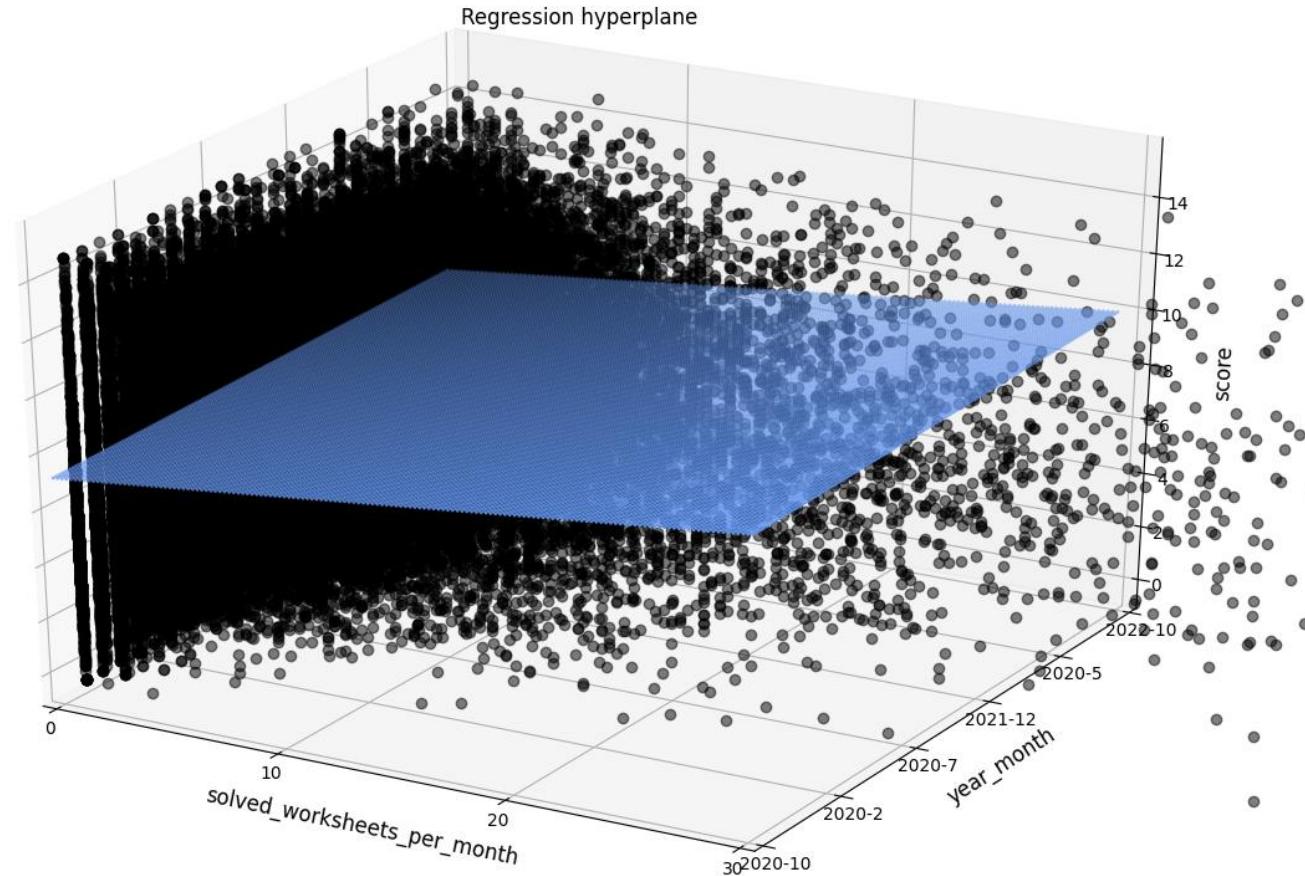


Service Effectiveness Analysis

Case Analysis



모든 학생



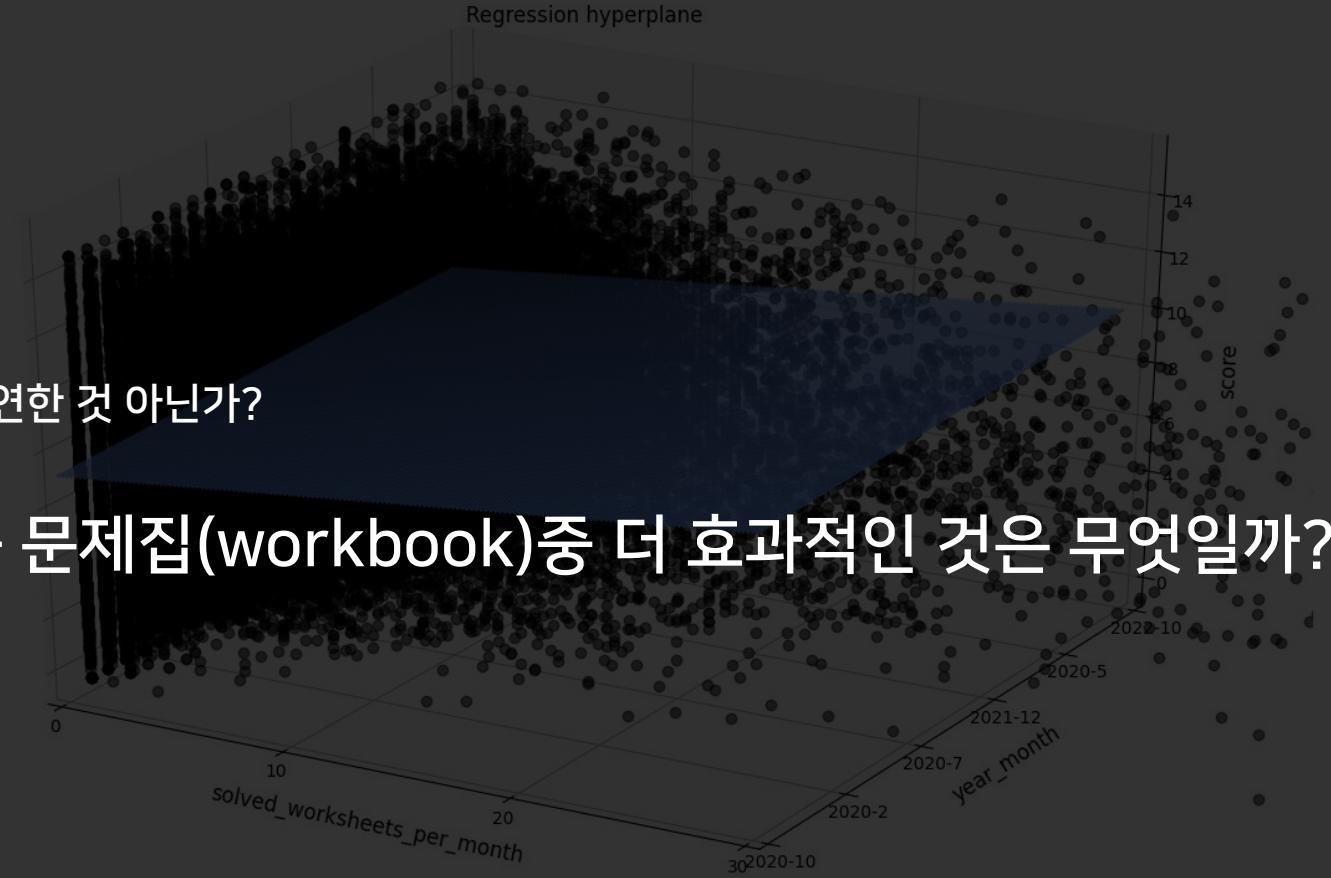
	coef	std err	t	P> t	[0.025	0.975]
const	6.9786	0.037	186.213	0.000	6.905	7.052
x1	0.0861	0.002	45.751	0.000	0.082	0.090
x2	0.0209	0.002	11.463	0.000	0.017	0.024

Service Effectiveness Analysis

Case Analysis

문제를 많이 풀면 score가 높게 나오는 것은 당연한 것 아닌가?

=> 학습지(worksheet)와 시중 문제집(workbook) 중 더 효과적인 것은 무엇일까?



	coef	std err	t	P> t [0.025 0.975]
const	6.9786	0.037	186.213	0.000 6.905 7.052
x1	0.0861	0.002	45.751	0.000 0.082 0.090
x2	0.0209	0.002	11.463	0.000 0.017 0.024

Service Effectiveness Analysis

Case Analysis

Q2. Workbook과 Worksheet 중 어떤 것이 성적 향상에 보다 효과적일까?

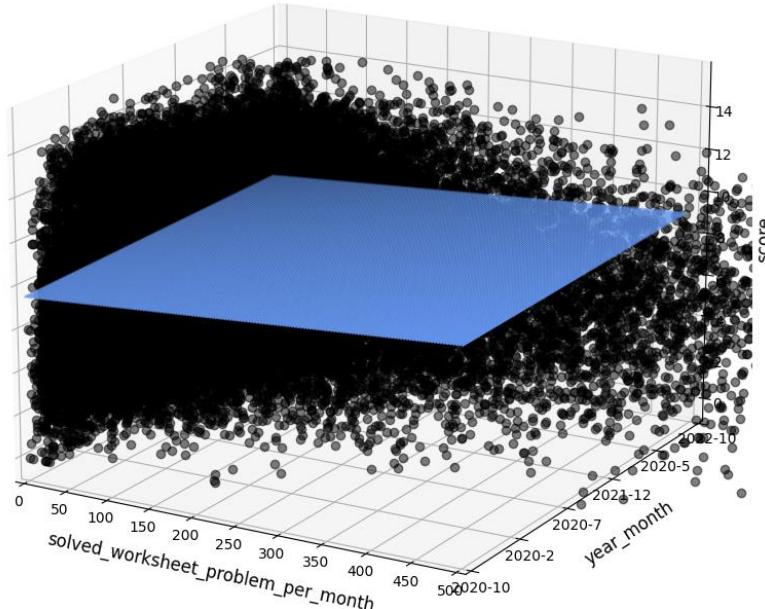
Student.id (학생 ID)	Year_month (연-월)	Count_worksheet -problem.id (월별 학습지 문제 수)	Count_workbook _problem.id (월별 시중 문제집 문제 수)	Score (점수)
I100008	202010	276	NaN	8.7935
I100008	202011	1347	NaN	9.3481
I100008	202012	330	NaN	9.5926
...				
I100008	202210	611	NaN	11.2263
I99874	202101	105	22	7.4008
I99874	202102	NaN	12	7.8442

I10008 학생 : 오랜 기간 이용했지만,
매쓰플랫에서 시중 문제집 문제는 풀지 않는 학생
→ 학습지만 끈 학생

I99874 학생 : 학습지는 적게 풀고
시중 문제집 서비스를 이용하는 학생

Service Effectiveness Analysis

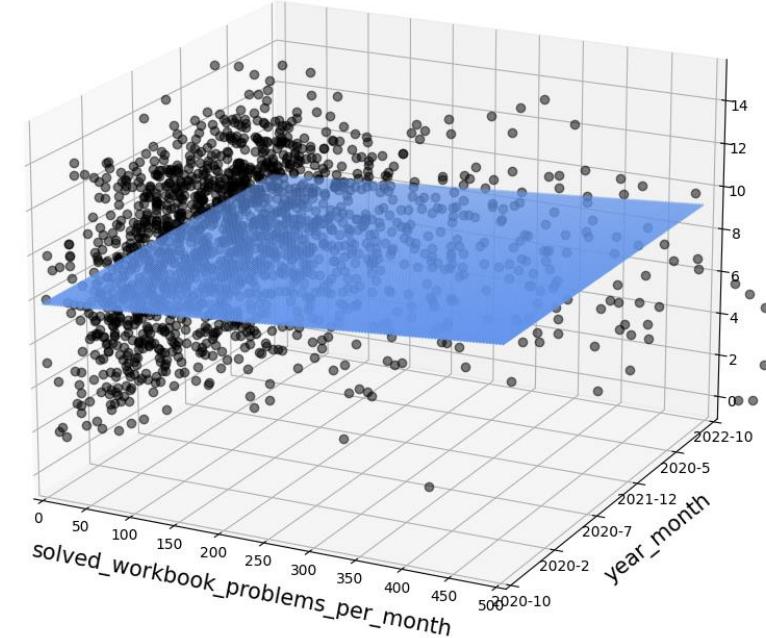
Case An



월간 푼 학습지(worksheet) 문제 수와 score

	coef	std err	t	P> t	[0.025	0.975]
const	6.8563	0.043	172.228	0.000	7.265	7.432
x1	0.0027	7.18e-05	37.205	0.000	0.003	0.003
x2	0.0164	0.002	8.120	0.000	0.012	0.020

Worksheet의 회귀계수 :
문제 수: 유의
기간별 : 유의



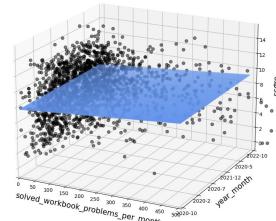
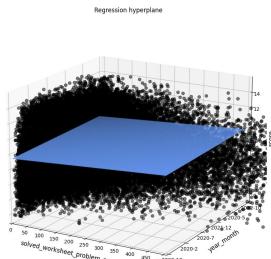
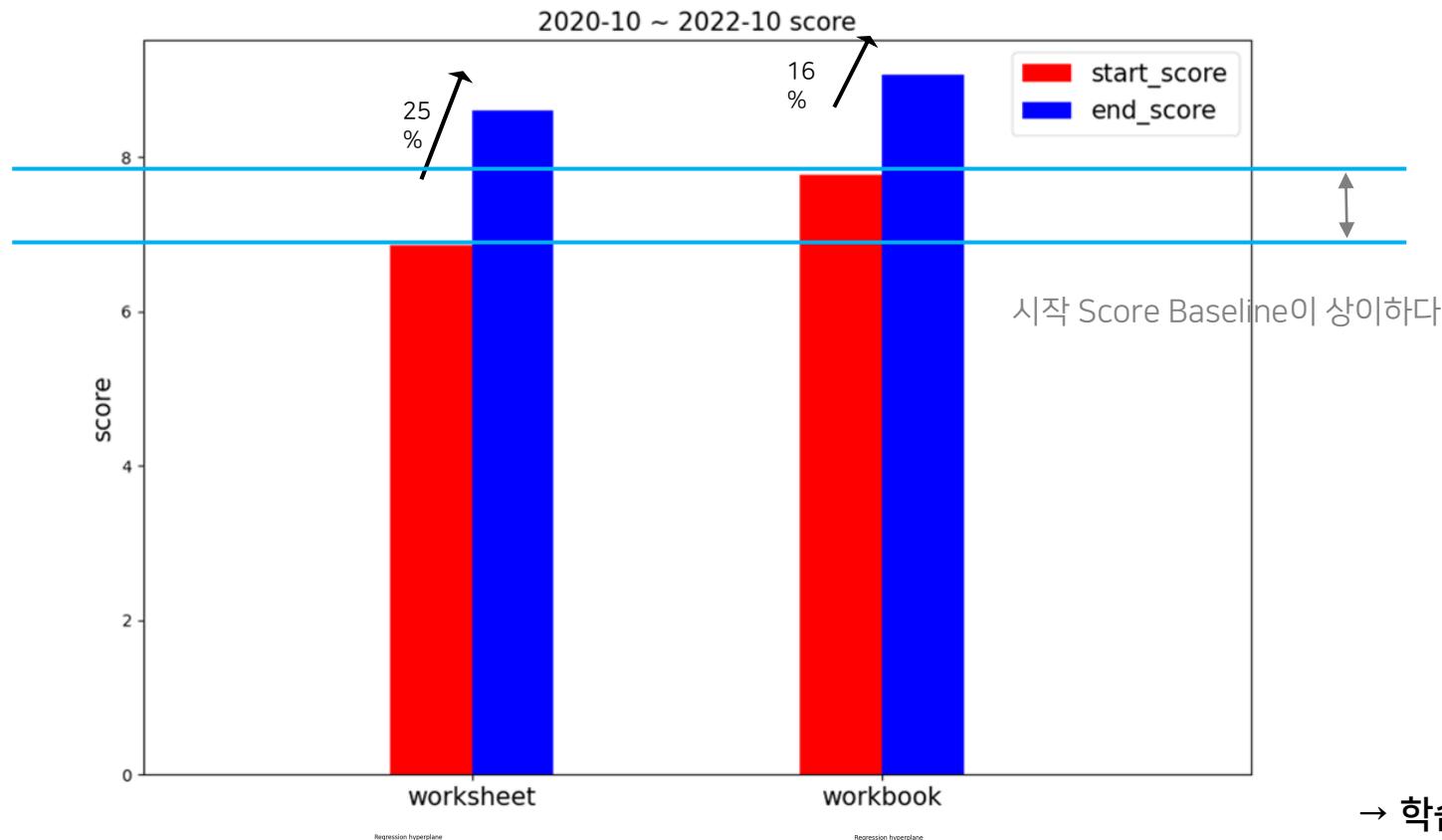
월간 푼 시중 문제집(workbook) 문제 수와 score

	coef	std err	t	P> t	[0.025	0.975]
const	7.7764	0.236	32.973	0.000	7.314	8.239
x1	0.0025	0.000	8.127	0.000	0.002	0.004
x2	0.0017	0.010	0.164	0.870	-0.018	0.022

Workbook의 회귀계수 :
문제 수: 유의
기간별 : 유의하지 않음...!!

Service Effectiveness Analysis

Case Analysis



결과 해석

1. Worksheet만을 푼 학생들이 Workbook을 위주로 푼 학생보다 9%p만큼 실력이 더 올랐다.
2. Worksheet을 위주로 푼 학생들이 기존 실력이 더 낮았지만 종료 시점에는 실력이 비슷해졌다.
3. Worksheet는 계속 풀 수록 실력 상승이 유의했지만, Workbook은 유의하지 않았다.

→ 학습지는 선생님이 선별하는 문제이므로

1. 개념 정리
2. 복습 문제
3. 오답 등 학생들을 개개인별로 관리해줄 수 있어 효과적으로 보인다.

Service Effectiveness Analysis

Case Analysis

실제로 오답 정리와 복습용으로 쓸까?

Q3. 이용자들은 실제로 Worksheet를 어떤 용도로 사용할까?

Student.id (학생 ID)	Year_month (연-월)	Worksheet_Problem.id (문제 id)	Result (채점결과)
I100008	202010	272345	CORRECT
I100008	202011	221252	WRONG
I100008	202012	221252	CORRECT
...			
I100008	202210	385512	CORRECT
IB113	202101	113432	WRONG
IB113	202102	113432	WRONG

학습지

“같은 학생이 같은 문제”
=> 2350 of 69384건 채점 데이터

Student.id (학생 ID)	Year_month (연-월)	Workbook_Problem.id (문제 id)	Result (채점결과)
I100008	202010	274211	CORRECT
I100008	202011	225124	WRONG
I100008	202012	226125	CORRECT
...			
I100008	202210	42402	CORRECT
IB113	202101	114311	WRONG
IB113	202102	105271	WRONG

시중 문제집

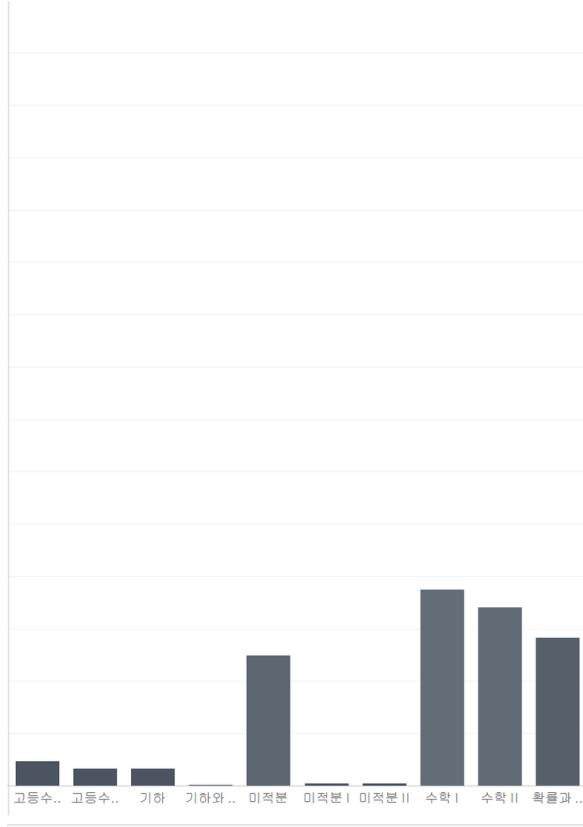
“같은 학생이 같은 문제”
=> 0 of 1808건 채점 데이터

결과 해석

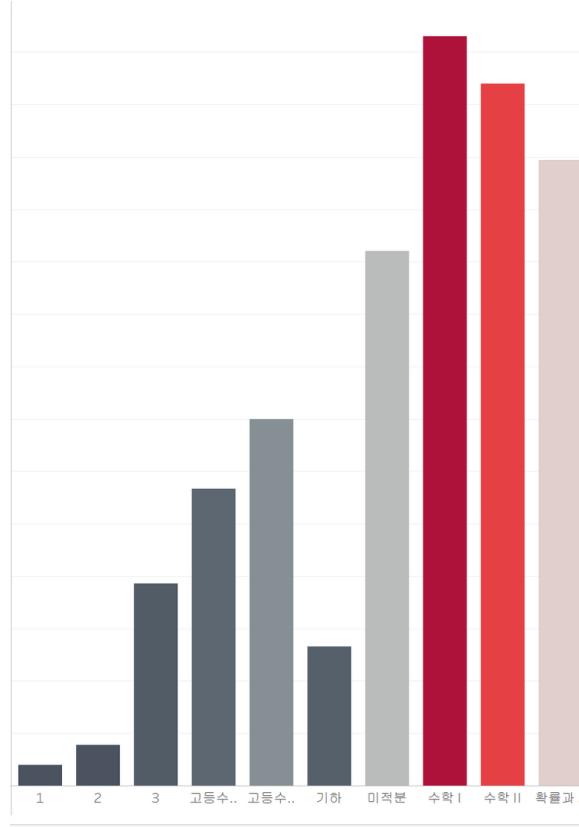
1. Workbook을 두 번 채점한 데이터는 없었다.
 2. Worksheet에서는 같은 문제를 같은 학생이 푸는 데이터들이 있다. (3.3%)
 3. 같은 유형을 푼 경우를 포함하면 더 많을 것
- 학습지는 선생님이 관리하기 때문에 채점 데이터도 많다.
→ 시중 문제집은 채점을 잘 하지 않을 뿐더러, 오답 관리는 진행되지 않는다.

Service Effectiveness Analysis

Workbook



Worksheet



X축: 과목명 Y축: Academy 수

결과 해석

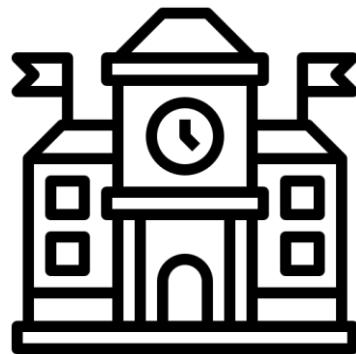
- 실제로 모든 과목에 대해서 Worksheet를 통해서 문제를 푼 수가 Workbook을 통해 문제를 푼 수보다 많았다.

→ 매쓰플랫 사용의 주된 이유: [Worksheet](#)

→ 매쓰플랫의 발전 방향을 Worksheet 위주로 진행하는 것이 바람직

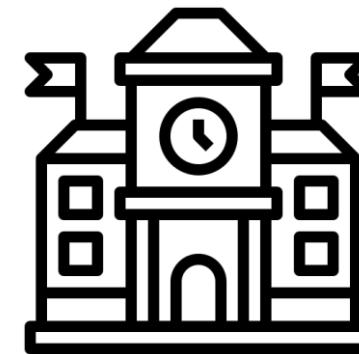
Q3. 모든 등급의 학원에 대해서도 Worksheet가 효과적일까? 동일한 추세일까?

1등급 대상 학원



• • •

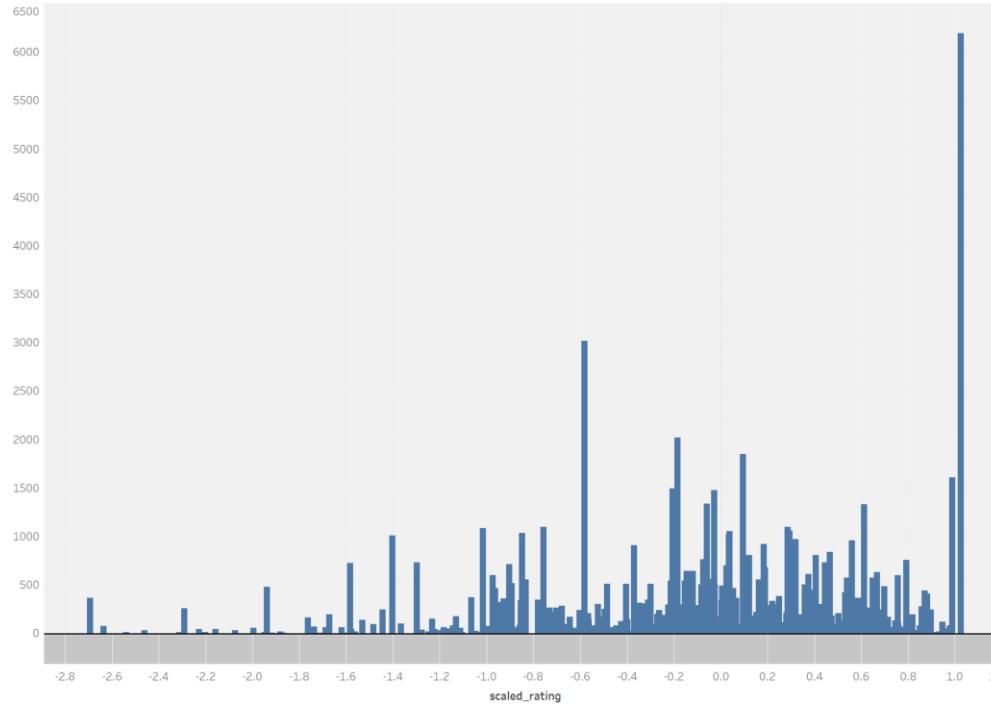
9등급 대상 학원



학원마다 매쓰플랫 서비스를 사용하는 방법이 다르지 않을까

Service Effectiveness Analysis

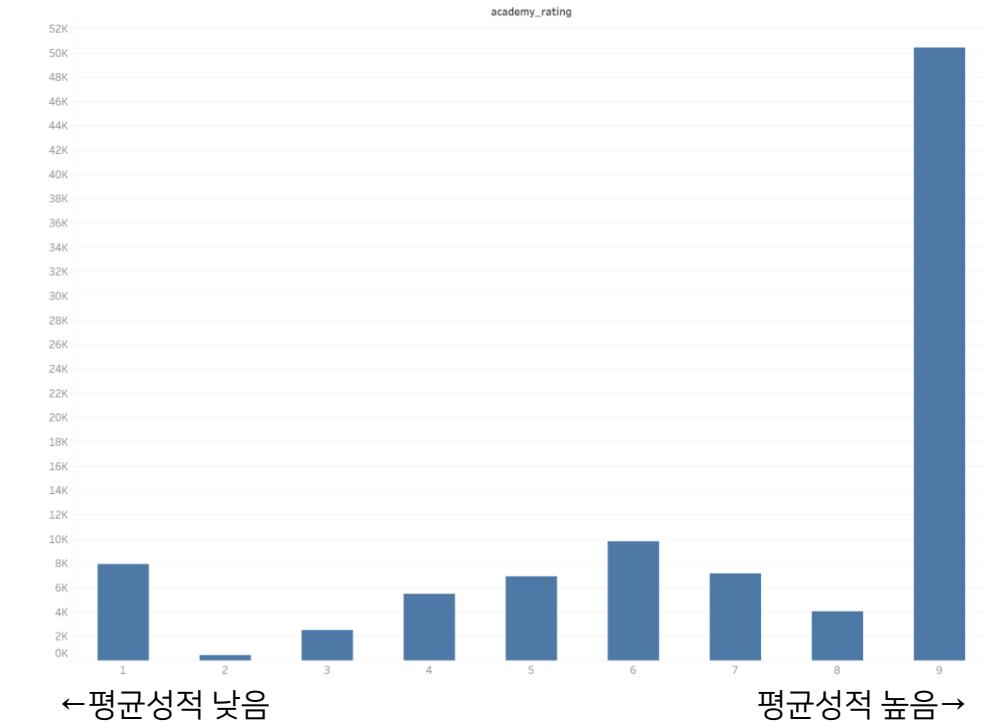
scale(학생 평균 등급)에 따른 학원 분포



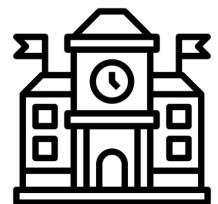
← 평균성적 낮음

평균성적 높음→

academy_rating에 따른 학원 분포



1등급 대상 학원

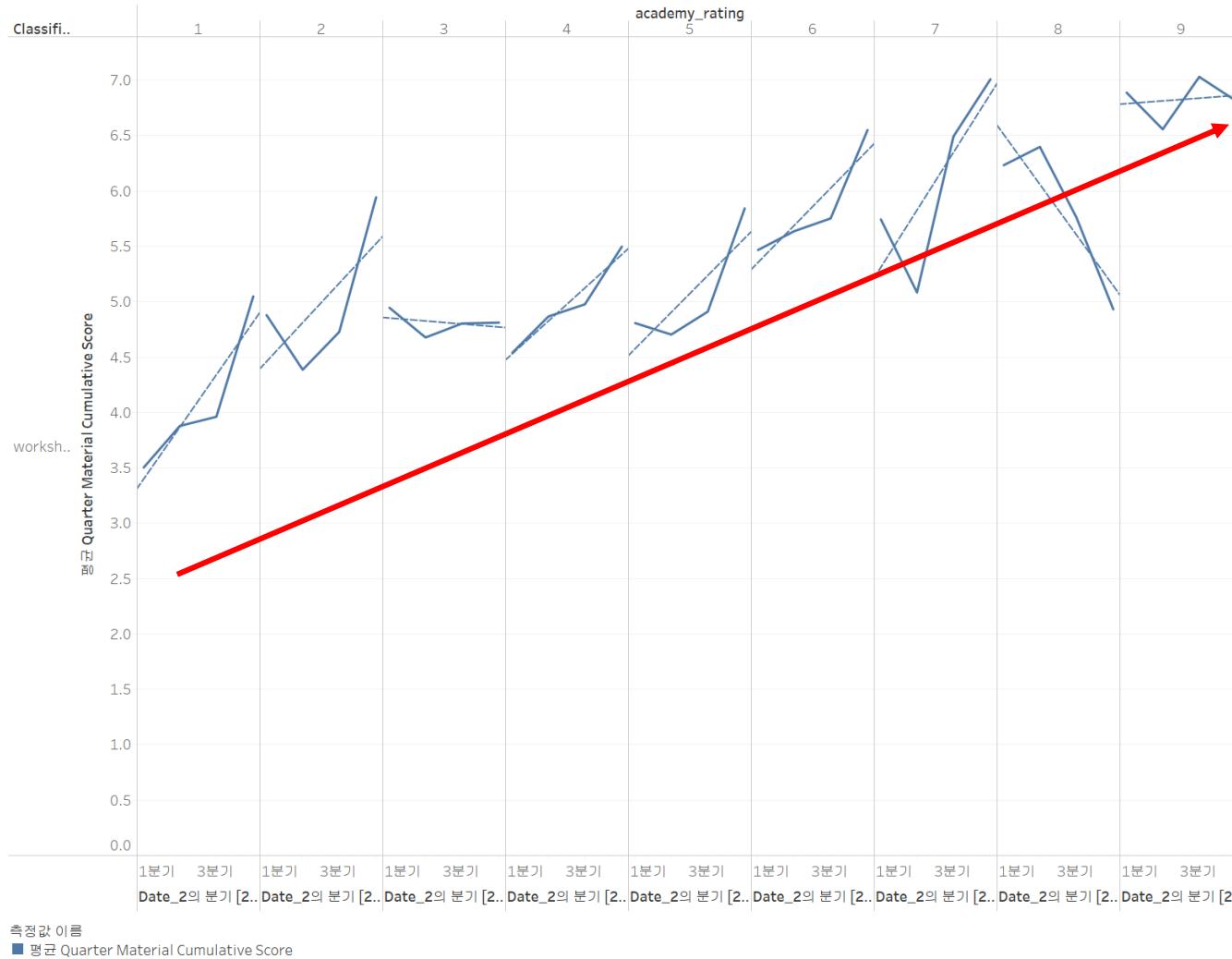


9등급 대상 학원

- 모든 학생들의 최초 분기 마지막 시점 누적 Score를 기준으로 등급 부여
 - 각 학원별로 학생들의 평균 등급을 기준으로 다시 학원 1~9등급 부여

Service Effectiveness Analysis

학원 구간별 실력 변화 추이

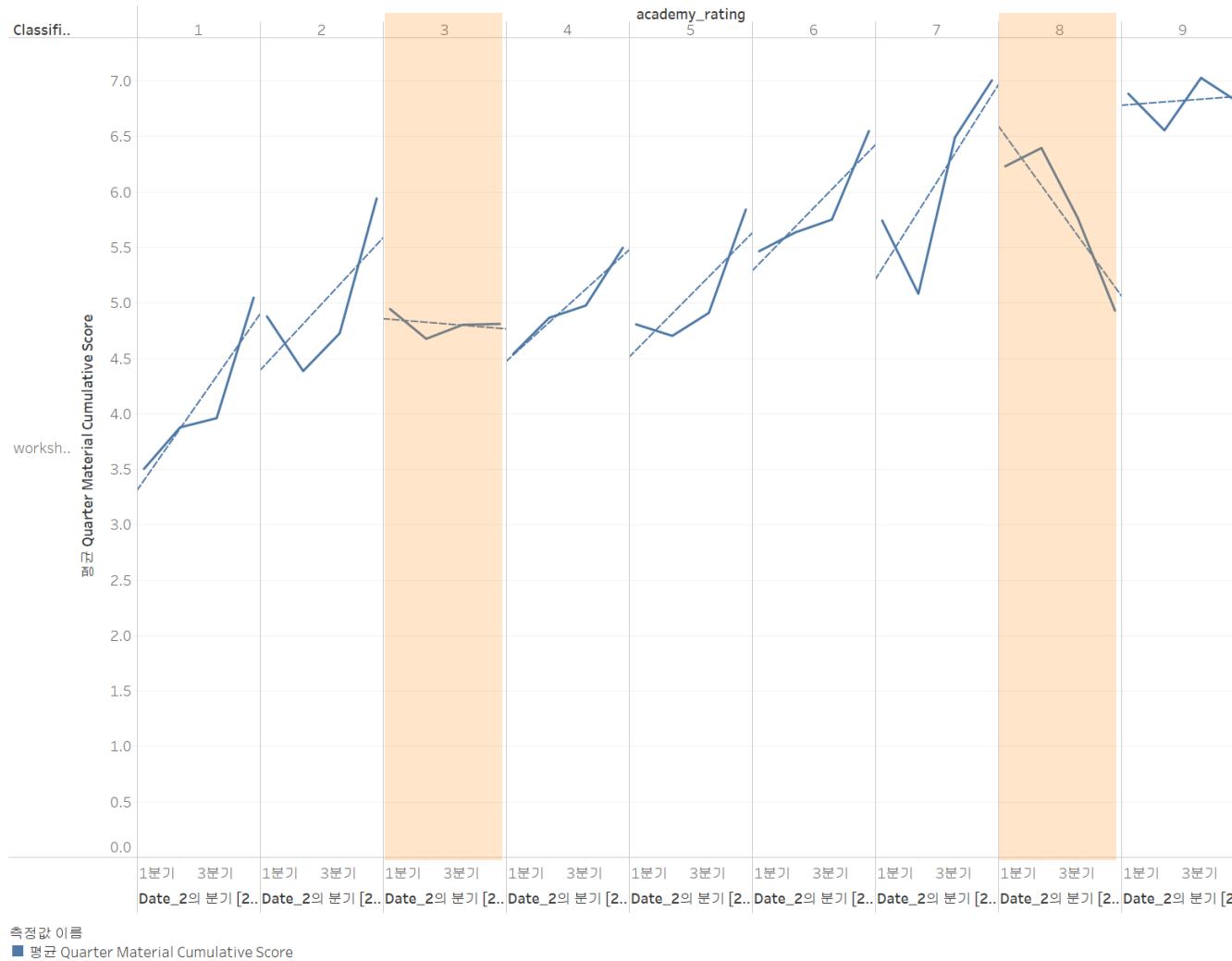


결과 해석

1. 대체적으로 모든 등급 구간 학원에 대해서 실력이 오르는 추세가 보인다.

Service Effectiveness Analysis

학원 구간별 실력 변화 추이



결과 해석

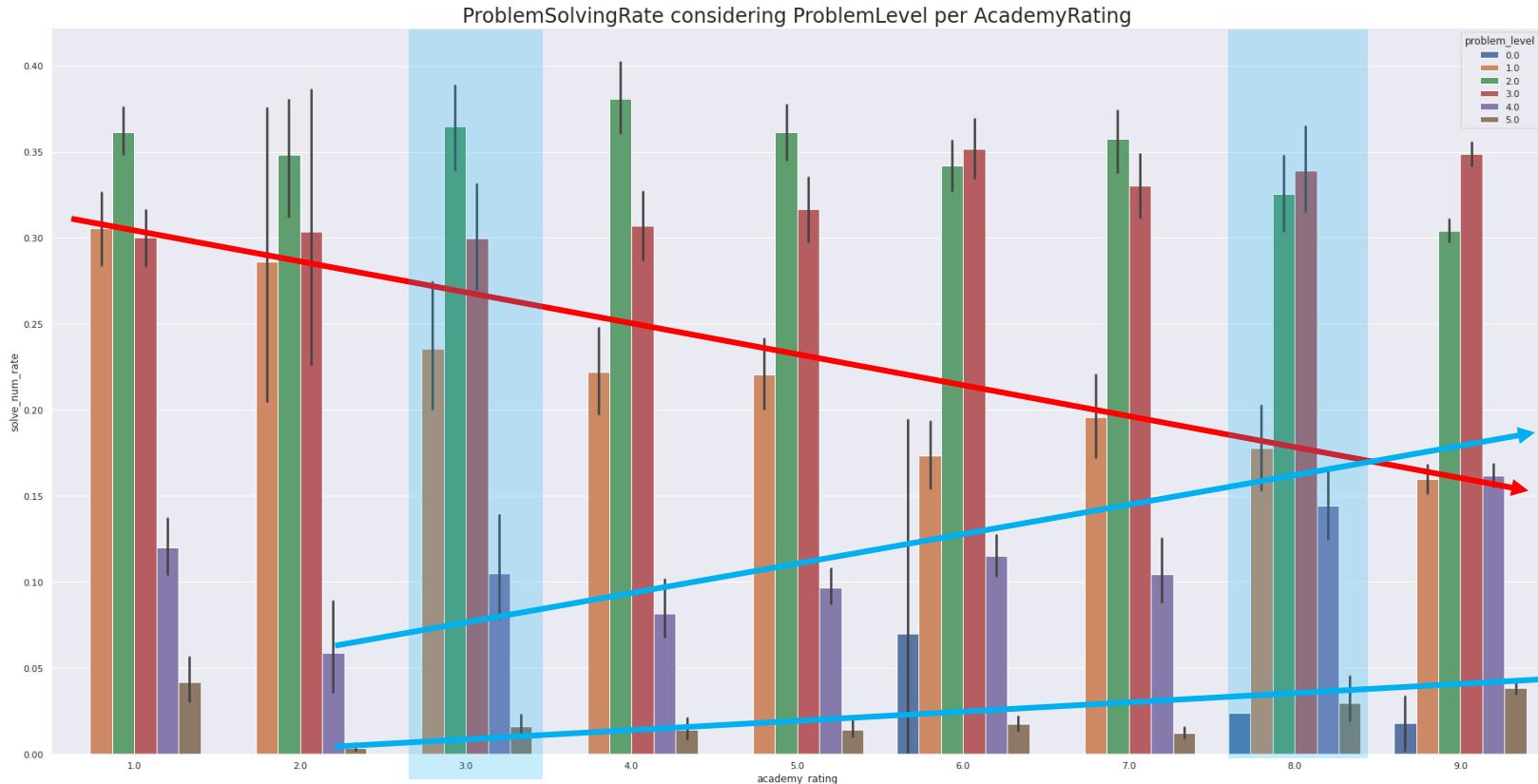
1. 대체적으로 모든 등급 구간 학원에 대해서 실력이
오르는 추세가 보인다.
 2. 특정 등급 구간의 학원에서만 다른 추세가 보인다.

(실력이 횡보하는 추세 / 떨어지는 추세)

→ 이 등급대 학생들이 푸는 문제 난이도 변화가 있지 않을까?

Service Effectiveness Analysis

학원 구간별 실력 변화 추이

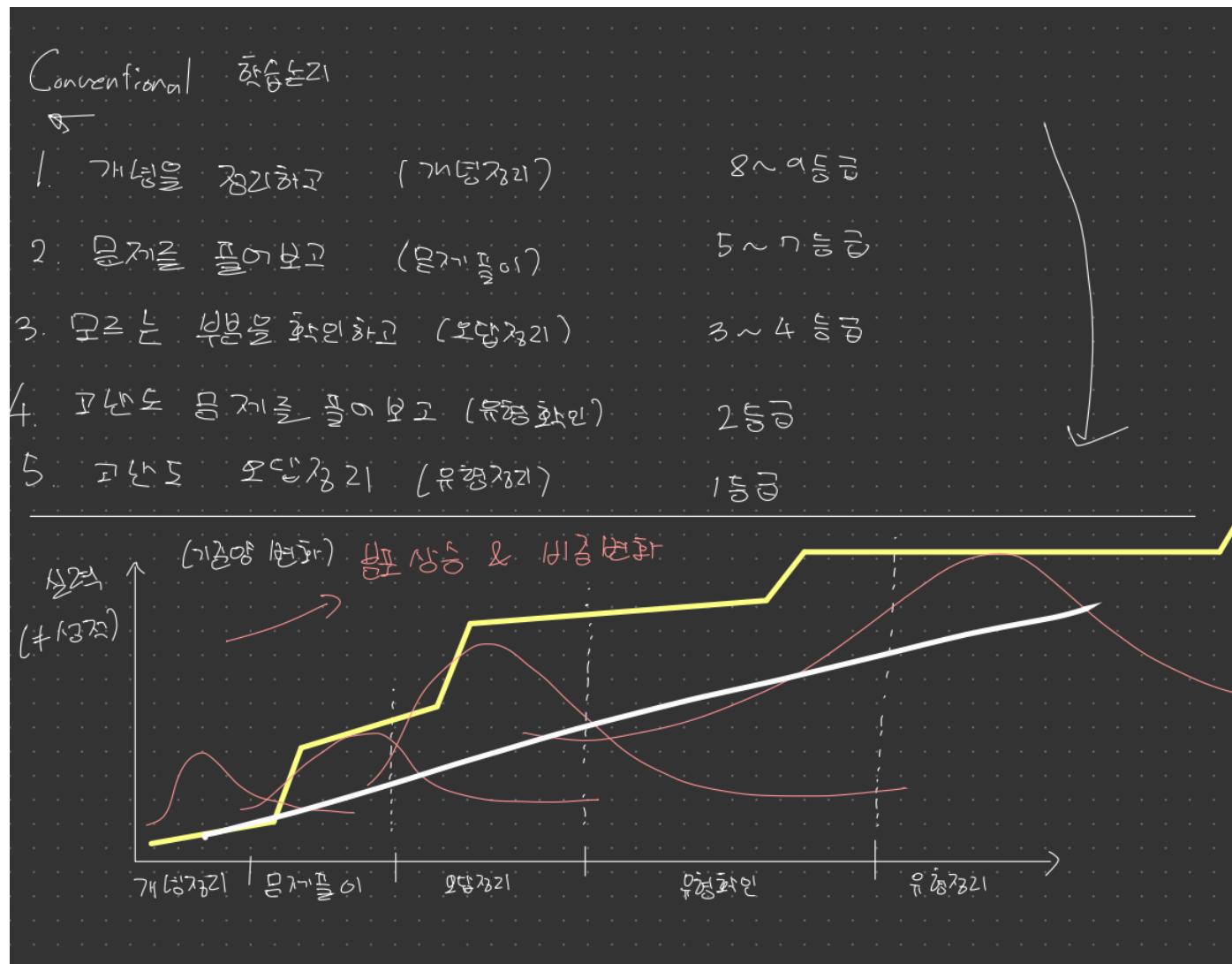


결과 해석

- 학원 level이 높아짐에 따라 1점 문제 비율은 적어지고 4점/5점 비율이 높아진다.
- 3level 학원 & 8 level 학원에서는 4점 / 5점 문제를 푸는 비율이 두드러지게 높아진다.

Service Effectiveness Analysis

학원 구간별 실력 변화 추이



설문조사를 통한 결과 학습 과정 해석

- 학생들은 (개념 정리 → 문제 풀이 → 유형확인 → 오답 정리 → 고난도 문제 정리) 순서들로 공부한다.
- 각 등급이 올라감에 따라 문제를 푸는 양이 점차 늘어나고, 등급이 오름에 따라 문제를 푸는 양상이 위의 순서대로 달라진다.

→ 때문에

- 3level 학원들의 학생들은 문제풀이/유형확인 단계에 돌입한 것
- 8level 학원들의 학생들은 고난도 문제풀이 단계에 돌입한 것으로 볼 수 있다.

Service Effectiveness Analysis

Q4. 실력을 잘 올릴 수 있는 Worksheet 구성은 어떻게 해야할까?

1.

p_num
20.0

학생의 실력변화에 영향을 주는 변수 확인

2.

p_level_mean	p_level_std	correct_rate_mean	correct_rate_std
2.65	0.875094	63.42123	14.566228

후보군

3.

unique_seq_rate	seq_std
0.6	5.292497

1. 문제 수

2. 문제의 난이도

3. 들어가는 개념 (개수 / 연관성)

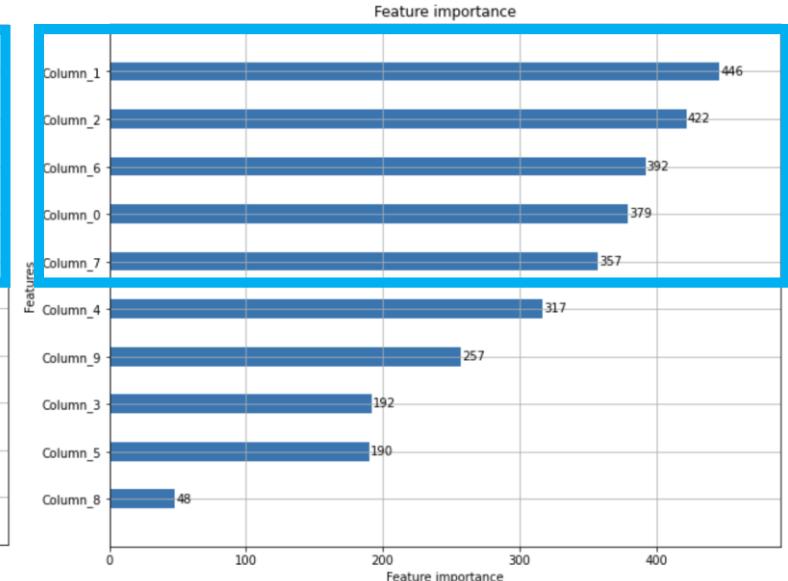
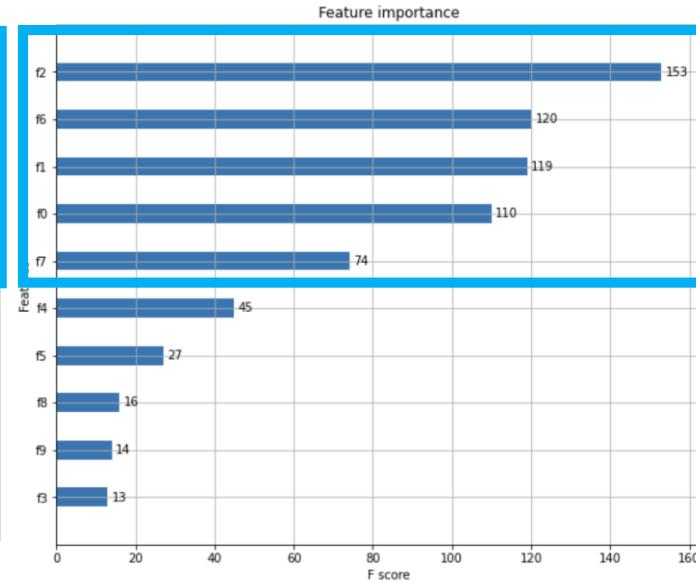
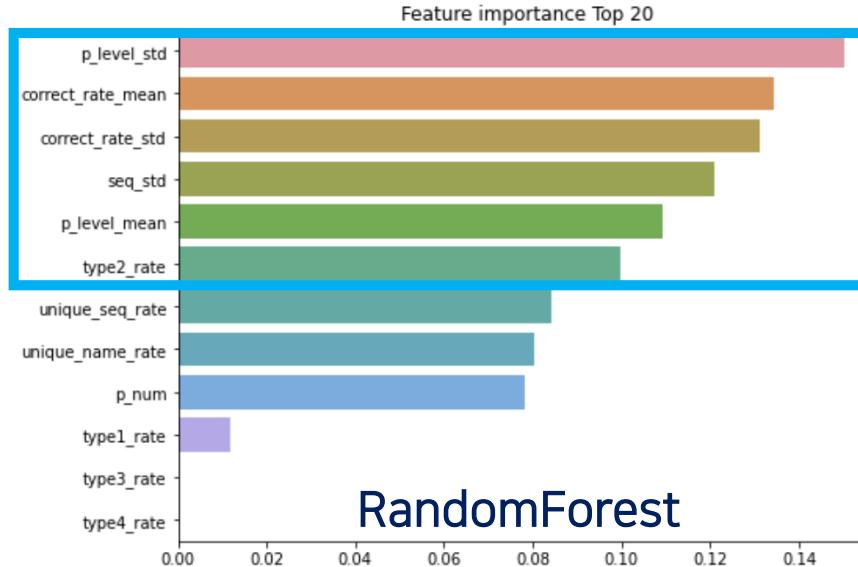
4. 문제 유형 (단답형, 주관식 등)의 비율

4.

type1_rate	type2_rate	type3_rate	type4_rate
0.0	1.0	0.0	0.0

Service Effectiveness Analysis

1. 얼마나 큰 영향을 주는가?



2. 어떻게 영향을 주는가?

중요한 Features: 다중선형회귀 회귀계수의 부호

1st. 문제 난이도

- $p_level_std = -0.0613 < 0$

2nd. 문제 정답률

- $correct_rate_mean = -0.0857 < 0$

3. 결론

난이도가 관건!

- 학습지 안에 난이도의 수준의 유사할 수록 학생 성적 향상에 도움이 된다.
- 학습지 안에 정답률이 너무 높은 문제들(쉬운 문제들)만 있으면 오히려 실력이 늘지 않는다.

#5 결론

Conclusion

Conclusion

Phases for
Studying



Metric
Score

$$score = \sum_{i=1}^5 (level\ i \times level\ i\ correct\ rate)$$

Workbook
vs
Worksheet

Worksheet!

1. 매쓰플랫 이용자들이 사용하는 서비스는 worksheet다.
2. Worksheet가 Workbook에 비해 성적 향상 속도가 빠르다.
3. Worksheet를 통해 모든 등급의 학원들의 성적향상에 기여할 수 있다.

**Good
Worksheet**

1. 정답률이 너무 높게 나오는 문제들만 제공해서는 안된다.
2. 난이도는 학생의 실력에 맞게 유사한 난이도 위주로 주는 것이 더 효과적이다.

프리월린 B

- 1) Feature extraction for clustering
- 2) K-means clustering

문제 정의

문제

- 성적 향상을 위해 본인의 강점과 약점을 파악할 필요가 있다.
- 본인과 비슷한 특성의 학생을 찾아 학습 방향을 참고할 수 없을까?

목표

- 학생 별 특성을 파악하고 학습 방향을 제시한다.

방법

- 단원별, 난이도별 정답률을 이용한 Z-score 사용하여 학생을 분류한다.
- 분류된 집단별 특성을 파악하고, 집단별 차이 유무를 판단한다.
- 집단별 차이를 보이는 경우, 차이를 유발한 객체를 추론한다.

Objectives and Analysis Methods

최종 목표

고 3 학생 별 특성 파악 및 학습 방향 제시

Step of
Analysis



실력 정의

1. 각각의 문제 & 난이도 별 평균, 표준편차 계산
2. 위에서 구한 평균, 표준편차로 학생 별 푼 문제&난이도 별 Z-Score 계산

군집화

1. Z-Score 기반으로 학생 군집화 (K-means Clustering, CNN 사용)
2. 군집이 **학업 역량이 높은 학생** vs **학업 역량이 낮은 학생** 으로 나뉘는지 여부 파악

특성 파악 및 해석

1. 군집 별 **학업 역량이 높은 학생** vs **학업 역량이 낮은 학생** 차이 파악
2. 학업 역량이 높은 학생이 되기 위한 학습 방향 제시

문제 접근

학생 분류로 특성 파악

NN 모델

- NN을 이용한 임베딩 벡터 추출, 학생 분류 및 중요 변수 추출

클러스터링

- K-MEANS 클러스터링을 사용한 학생 분류

분석 및 학습 방향 제안

- EDA / PCA 변수 별 설명력 모델을 사용한 학생 분류 및 학습 방법 제안

해결 방식

NN 모델

1D CNN을 사용한 학생분류

- 5가지 수학 과목 (수학, 미적분 등)으로 데이터 분할
- 학생 별 문제와 나이도에 따른 정답률 기반 Z-score 생성
- 1D CNN 사용한 embedding
- 반환된 vector로 학생 간 거리 계산 및 분류

변수 중요도 파악

- 과목에 따른 변수 중요도 파악
- 소결론 제시

1. Students Classification by CNN

1D-CNN 데이터 전처리

과목별 데이터 분할

Student_id	Problem_curriculum_id	level	...	Parent3_name
학생1	커리큘럼 번호	3	...	수학 I
학생2	커리큘럼 번호	4	...	수학 II
학생3	...	2	...	확률과 통계
학생4	...	5	...	미적분
학생5	...	1	...	기하
...
...

- 고3 학생 별 푼 문제의 과목에 따라 데이터 분할
- 총 5 set의 데이터 생성 (과목에 따라)

1. Students Classification by CNN

1D-CNN 데이터 전처리

학생별 문제 정답률 계산

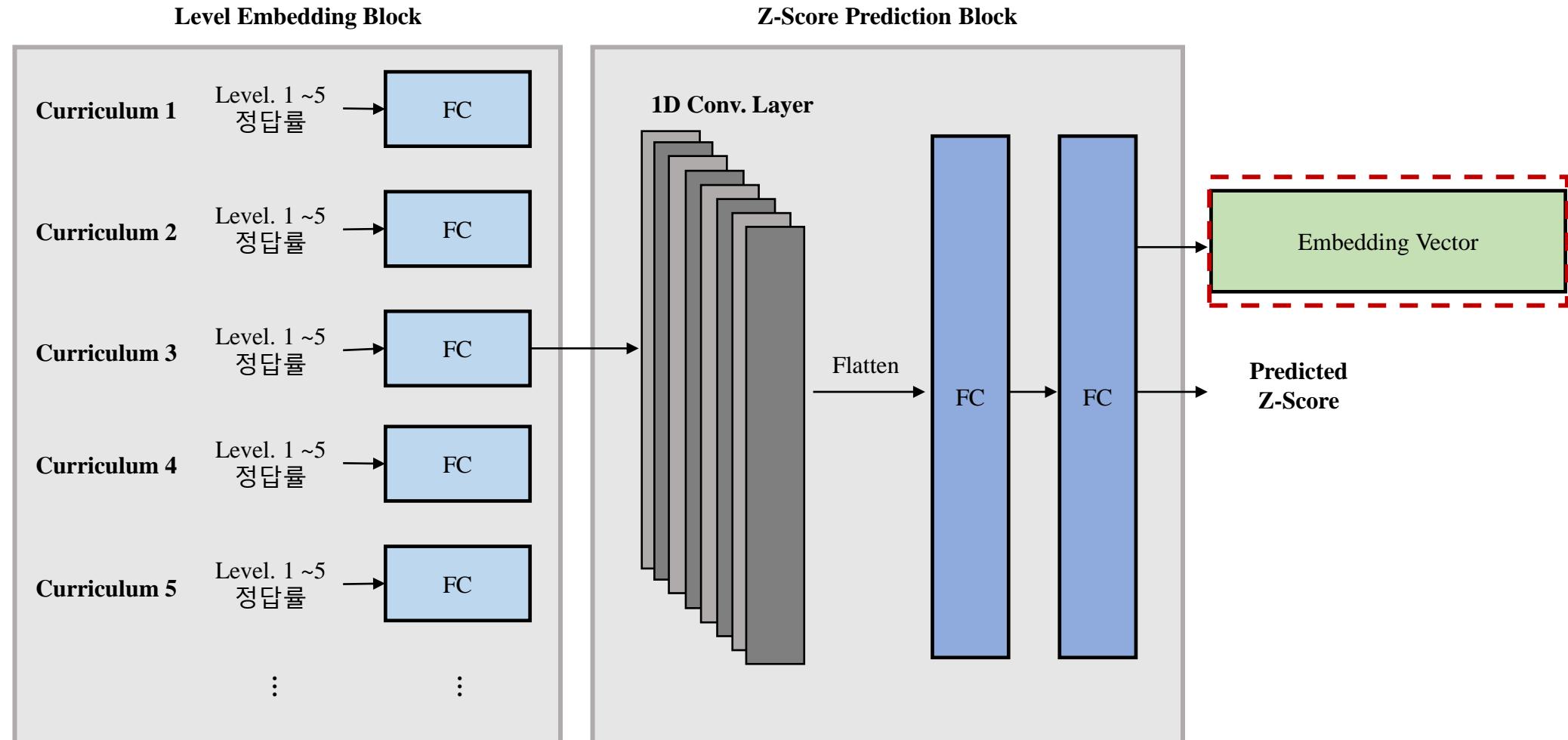
수학 정답률	문제1.난이도1	문제1.난이도2	...	문제300.난이도5	전체 정답률	Z-score
학생1	0.78	0.67	...	0.32	0.6	-0.2
학생2	0.56	0.6	...	NaN -> 0	0.5	-0.6
...			
학생7	0.87	0.77	...	0.5	0.8	0.7

- 문제+난이도별 정답률 계산
- 학생별 Z-score 계산하여 Label 값으로 설정
- 풀지 않은 문제는 NaN으로 표시 → 모델에 넣기 전 0으로 수정
- Sparse 한 칼럼 제거 (모델 성능 저하 우려 → 75% 이상의 학생이 해당 문제 풀지 않은 경우 삭제)

1. Students Classification by CNN

모식도

NN 모델



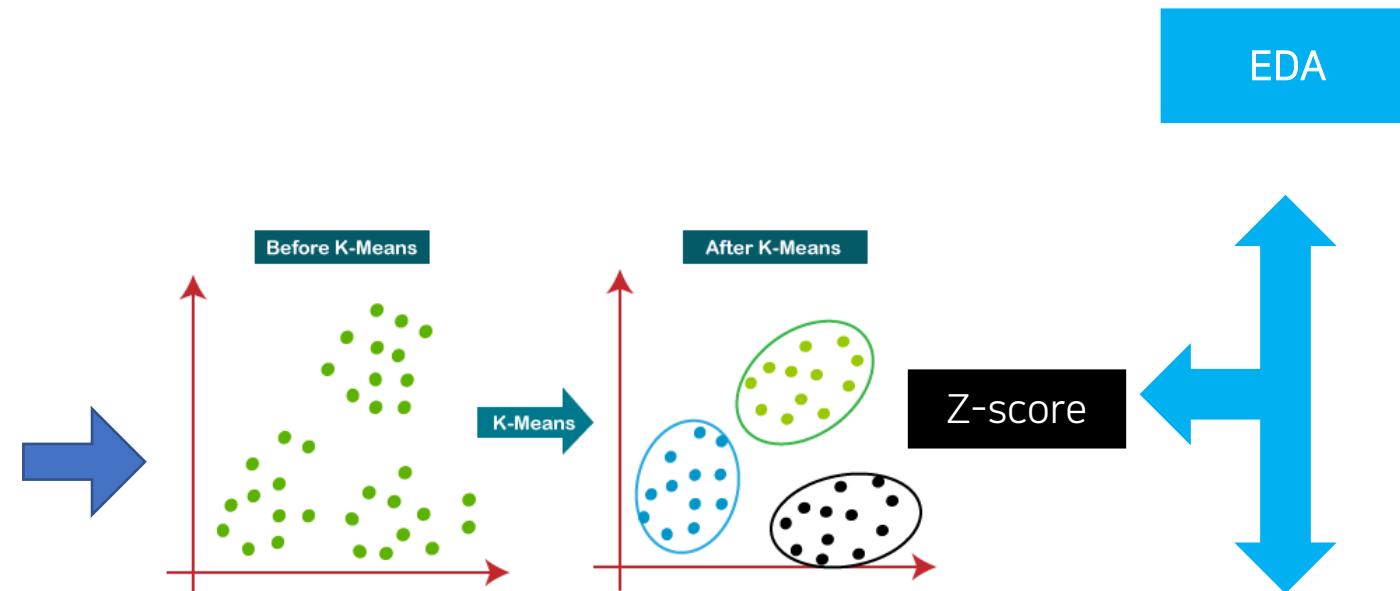
1. Students Classification by CNN

해결 방식

NN 모델

Embedding Vector

student_id	0	1	2	3	4	5	6	7	y
I427575	5.410321	5.736994	5.535293	5.644643	5.535326	5.916299	5.545562	5.567052	
I869191	3.142298	3.468971	3.267270	3.376621	3.267303	3.648276	3.277539	3.299029	
I206034	2.303088	2.629761	2.428060	2.537410	2.428093	2.809066	2.438329	2.459819	



- CNN 적용 → 7개의 칼럼으로 구성된 Embedding Vector 추출 (모델이 학생 별 특성 수치화 함)
- Embedding vector를 K-means로 군집화 (뽑힌 특성을 사용해 학생별 군집 할당)
- 군집 별 학생 문제 정답률 데이터 분석 (RF등 모델 사용해 칼럼 별 설명력 얻음)

확인 할 요소

군집 수 어떻게 정할까? & 군집화가 잘 되었을까?

- 기대 군집 결과: 잘하는 학생 vs **못하는 학생** vs 보통 학생 끼리 분류
- 군집의 수는 어떻게 정할 것인가?

1. PCA후 2차원 평면 plot 확인

2. 실루엣 계수 및 그래프 확인

- 군집화가 잘 되었는가?

1. 군집 간 Z-score의 평균 비교

2. 군집 간 문제 풀이 난이도 비교 (잘하는 그룹이 고난도 문제 풀 것 기대)

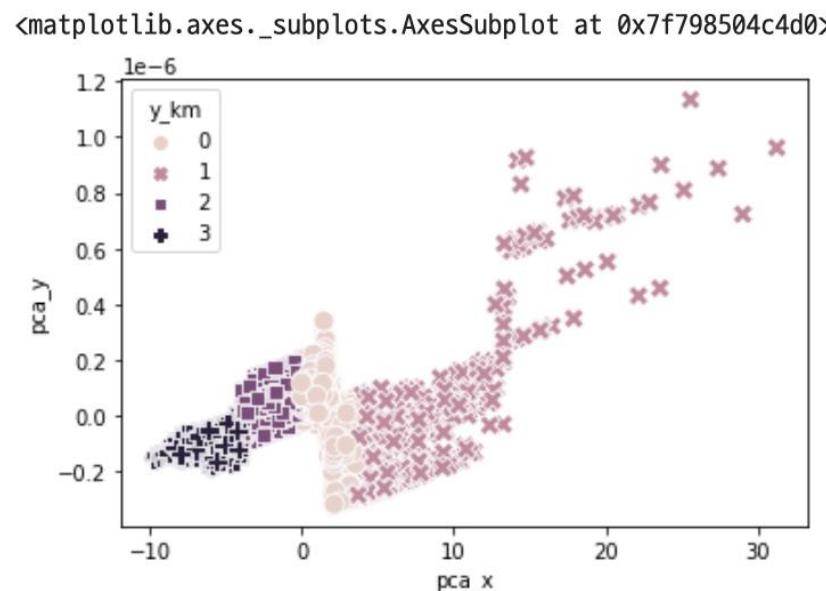
어떤 변수가 군집 별 차이를 유발했는가?

- 군집을 나누는데 큰 영향을 끼친 변수(문제 및 난이도) 선별
- 이후 잘하는 학생이 되기 위한 문제 추천

1. Students Classification by CNN

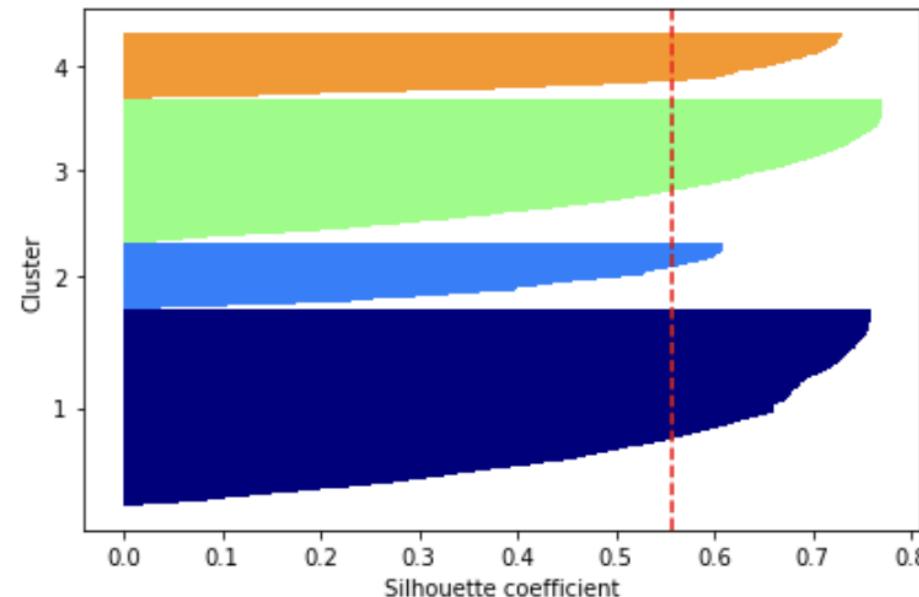
군집 수 결정 : 수학2

NN 모델



군집 수 결정 방법

- Scatterplot(PCA)로 분포 확인 (왼쪽)
- 실루엣 계수(Euclidean 거리) 기반 최적의 군집 개수 파악 (오른쪽)

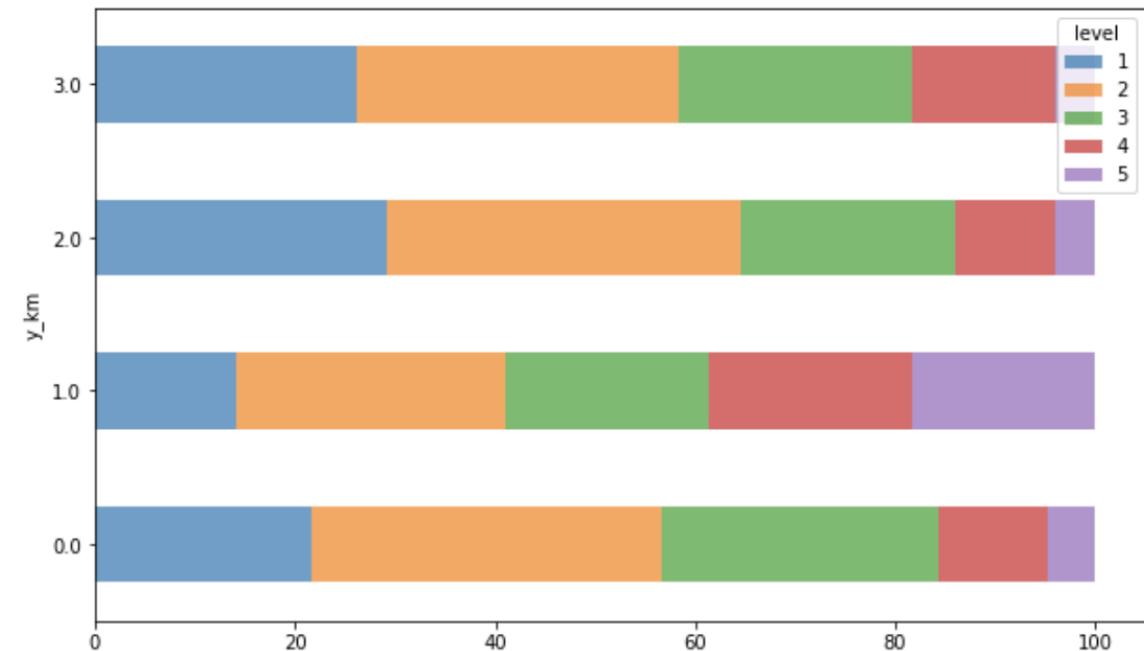
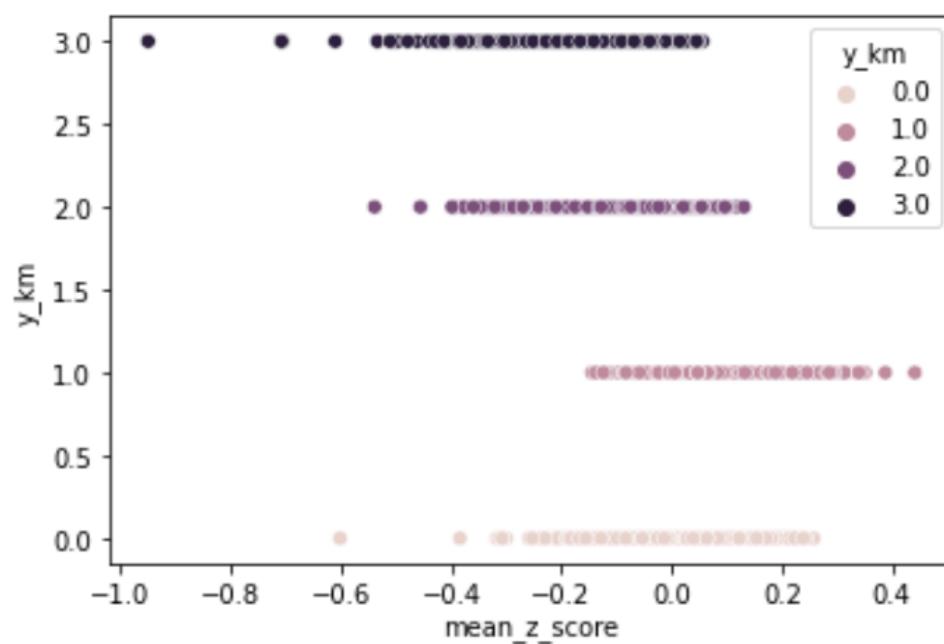


군집별 실루엣 계수

K=3 : 0.544 K=4: 0.621
 K=5 : 0.569 K=6: 0.536
 → 군집 수 =4로 진행

군집 해석 : 수학2

NN 모델

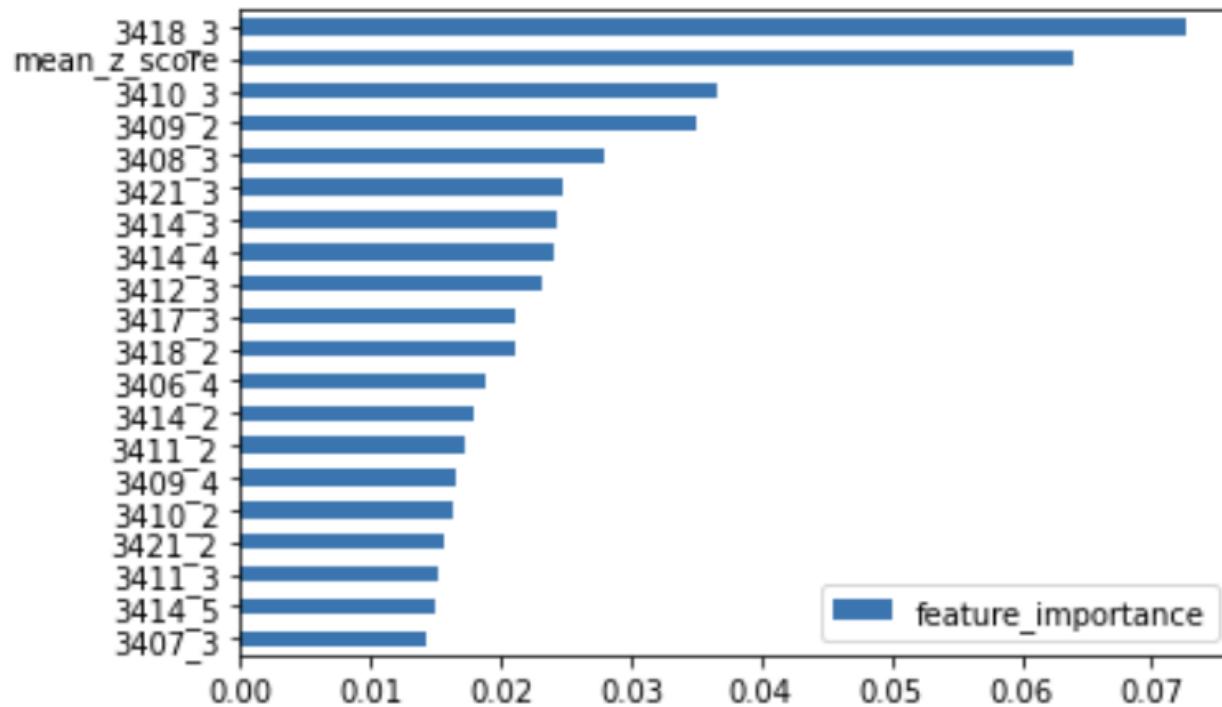


- 왼쪽 그래프 통해 군집 간 평균 Z점수에 차이가 유의미하게 나는 것 확인 가능
(1번 집단 잘함, 3번 집단 잘 못함)
 - 오른쪽 그래프 통해 잘하는 집단(1번)이 난이도 높은 문제(보라, 빨강) 많이 푸는 것 확인 & 못하는 집단(3번)이 난이도 낮은 문제(파랑, 노랑) 많이 푸는 것 확인

1. Students Classification by CNN

결과 : 수학2

결과 해석



RF를 통해 얻어진 피쳐 중요도:
(무엇이 학생들을 분류하는데 큰 영향을 주었는가?)

Mean Z-Score의 설명력 높음

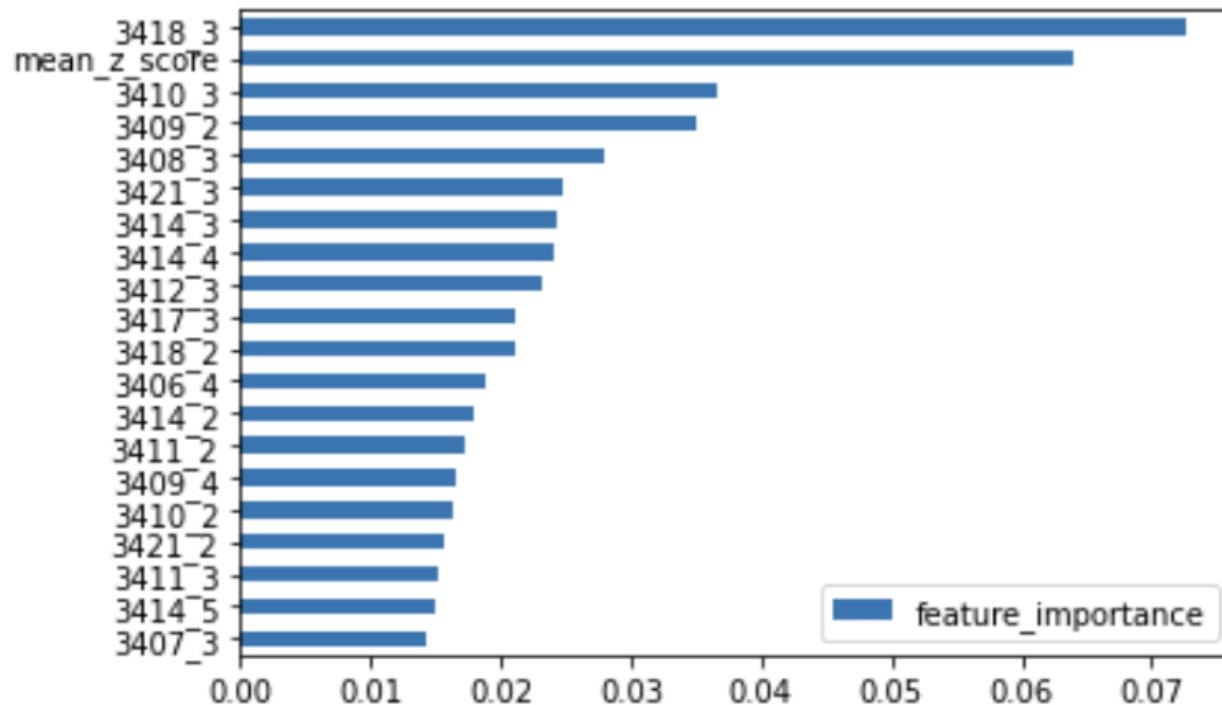
3418: 함수의 극한값의 계산

3410: 함수의 그래프

3409: 방정식과 부등식에서 활용
등의 칼럼의 설명력이 높게 나타남.

결과 : 수학2

결과 해석



소결론

Mean Z-Score의 설명력 높음

3418: 함수의 극한값의 계산

3410: 함수의 그래프

3409: 방정식과 부등식에서 활용

1. Mean Z-Score가 군집 나누는 것에 큰 기여 함
→ 예상대로 잘하는 학생 vs 못하는 학생끼리 잘 나뉨
2. 학생들이 위 3가지 문제에 의해 크게 나누어짐
→ 잘하는 학생이 되려면 위 3가지부터 탄탄히 공부해야 함

2. Students Classification by K-means

접근 방안

클러스터링

학생 분류 (by K-MEANS CLUSTERING)

- 3가지 수학 선택 과목, 2가지 필수 과목으로 데이터 분할
(미적, 확통, 기하, 수1, 수2)
- 문제/난이도별 학생들의 정답률 계산
- 정답률 자료의 설명력 제고를 위해 Z-score 활용
→ Z-score 비교를 통해 학생들 분류(cluster)

클러스터별 중요 변수 선정 (by PCA)

- 각 클러스터별로 강점과 약점이 나타나는 소단원 및 레벨 파악
- PCA 사용하여 클러스터별 차이점을 야기한 중요 변수(소단원과 난이도) 선정

접근 방안

클러스터링

CLUSTERING의 유효성 파악

- Clustering이 잘 되었는지에 대해 파악하는 작업 진행
(실루엣 계수를 통한 검정)
- 군집 별 특징과 군집을 나눠준 변수에 대한 인사이트
도출

(Ex. 각 클러스터별로 z-score 비교를 통해 강점과 약점을
가지는 소단원 및 레벨 파악)

학생 별 성적 향상 인사이트 제공
(도움이 될 소단원 및 난이도에 관한 정보 제공)

- PCA를 통해 파악한 중요변수를 활용하여 약점으로
드러나는 커리큘럼에 대한 인사이트 제공
- 클러스터링에서 나타난 군집 별 취약점 분석을 시행해
인사이트 도출, 성적 향상 도모

2. Students Classification by K-means

K-means 데이터 전처리

과목별 데이터 분할

Student_id	Problem_curriculum_id	level	...	Parent3_name
학생1	커리큘럼 번호	3	...	수학 I
학생2	커리큘럼 번호	4	...	수학 II
학생3	...	2	...	확률과 통계
학생4	...	5	...	미적분
학생5	...	1	...	기하
...
...

- 학생별 문제의 과목에 따라 데이터 분할
- 총 5 set의 데이터 생성

2. Students Classification by K-means

K-means 데이터 전처리

학생별 문제 정답률 계산

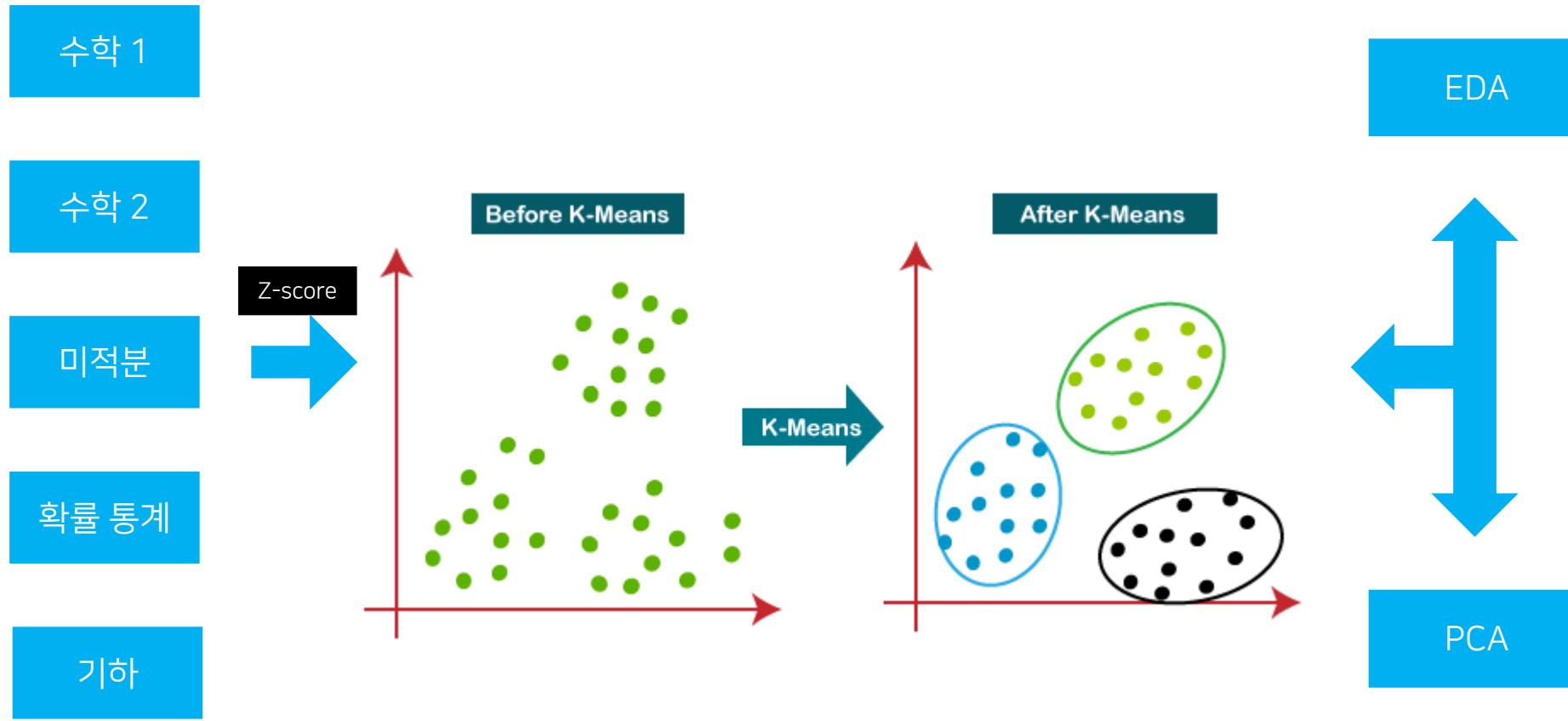
수학 정답률	문제1.난이도1	문제1.난이도2	...	문제300.난이도5	전체 정답률
학생1	1.3	1.56	...	0.32	0.6
학생2	-0.2	0.1	...	NaN -> 0	0.5
...		
학생7	-0.87	-0.77	...	-0.5	0.8

- 문제+난이도별 정답률 계산
- 학생별 Z-score를 각 Data로 사용 (NN은 정답률 사용)
- 풀지 않은 문제는 NaN으로 표시 → 모델에 넣기 전 0으로 수정
- Sparse 한 칼럼 제거 (모델 성능 저하 우려 → 75% 이상의 학생이 해당 문제 풀지 않은 경우 삭제)

2. Students Classification by K-means

모식도

클러스터링



데이터 분류

Z-score 기반 Clustering

결과 해석

2. Students Classification by K-means

확인 할 요소

군집 수 어떻게 정할까? & 군집화가 잘 되었을까?

- 기대 군집 결과: 학업 역량이 높은 학생 vs 학업 역량이 낮은 학생 vs 보통 학생끼리 분류
- 군집의 수는 어떻게 정할 것인가?

1. PCA후 2차원 평면 plot 확인
2. Elbow method 사용 (최대로 꺾이는 지점 근처가 최적의 군집 수)

- 군집화가 잘 되었는가?
 1. 군집 간 Z-score의 평균 비교
 2. 군집 간 문제 풀이 난이도 비교 (학업 역량이 높은 그룹이 고난도 문제 풀 것 기대)

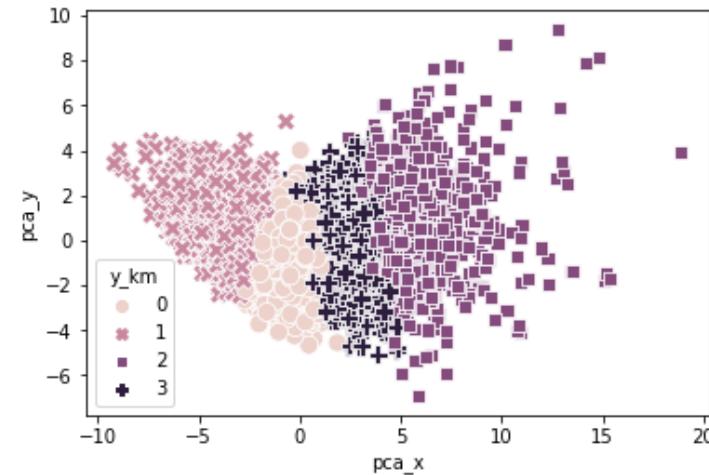
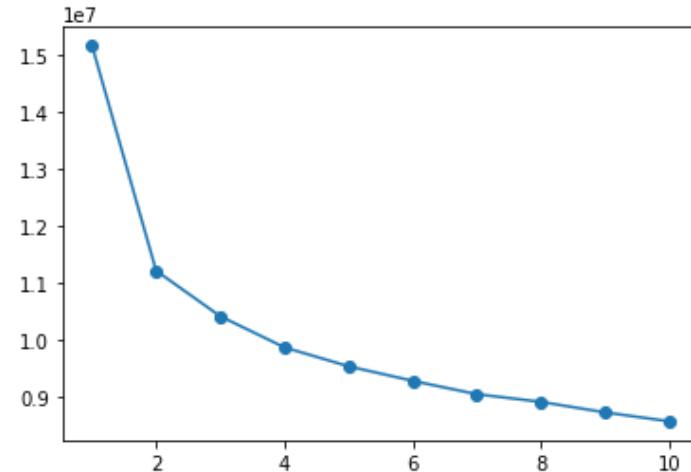
어떤 변수가 군집 별 차이를 유발했는가?

- 군집을 나누는데 큰 영향을 끼친 변수(문제 및 난이도) 선별
- 이후 학업 역량이 높은 학생이 되기 위한 문제 추천

2. Students Classification by K-means

군집 수 결정 : 수학1

클러스터링



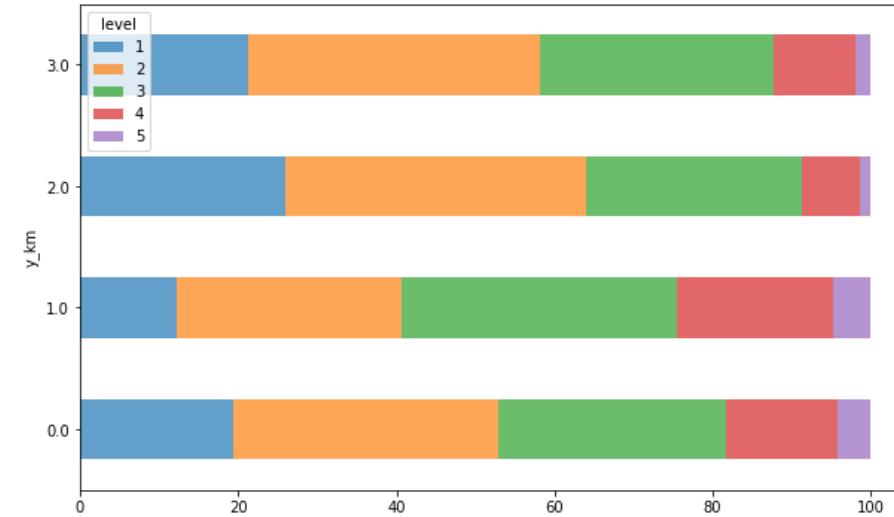
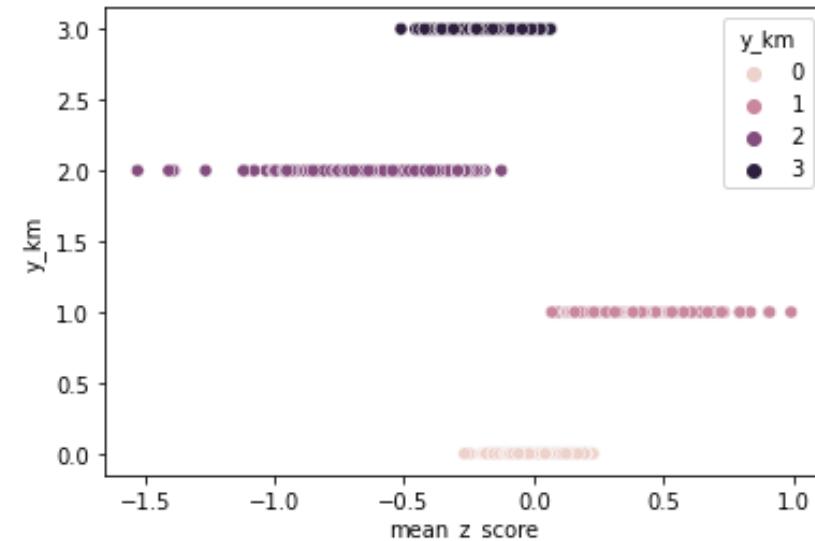
최적의 군집 수 결정

1. ELBOW METHOD 통해 [군집 수=4] 결정
2. 왼쪽 그래프 통해 [군집 수 = 4] 인 경우, 군집화 잘 되는지 판단
→ 비슷한 거리에 위치한 사람끼리 잘 묶임

2. Students Classification by K-means

군집 해석 : 수학1

클러스터링



- 왼쪽 그래프 통해 군집 간 평균 Z점수에 차이가 유의미하게 나는 것 확인 가능
(1번 집단 잘함, 2번 집단 잘 못함)
- 오른쪽 그래프 통해 잘하는 집단(1번)이 난이도 높은 문제(보라, 빨강) 많이 푸는 것 확인 & 못하는 집단(2번)이 난이도 낮은 문제(파랑, 노랑) 많이 푸는 것 확인

2. Students Classification by K-means

결과 : 수학1

군집을 결정한 주요 요인 파악

성적 높은 학생군

PC1

3402_5	0.185274
3389_5	0.172999
3386_4	0.167569
3400_4	0.164761
3391_5	0.158082
3399_4	0.144495
3403_5	0.142420
3384_5	0.122909
3386_3	0.120524
3393_4	0.112553

성적 낮은 학생군

PC1

3392_2	0.212236
3392_3	0.196346
3390_3	0.177055
3389_2	0.165409
3394_2	0.163951
3391_2	0.163247
3390_2	0.142163
3380_3	0.131029
3391_3	0.127467
3389_3	0.119642

분석 결과

- 성적 높은 학생군 내의 설명력이 가장 높은 단원
→ 상용로그, 등차수열, 일반각과 호도법 등 고난도 문제
- 성적 낮은 학생군 내의 설명력이 가장 높은 단원
→ 수열의 합, 등비수열, 등차수열 등 저+중난도 문제

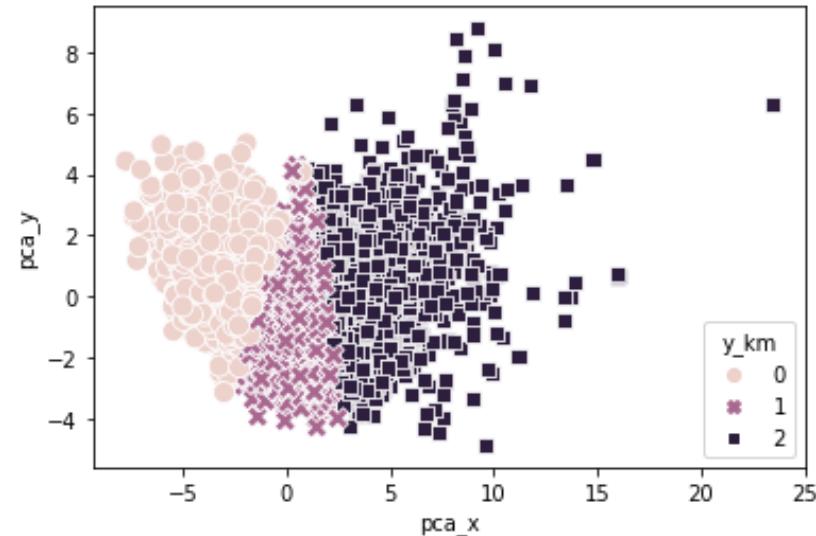
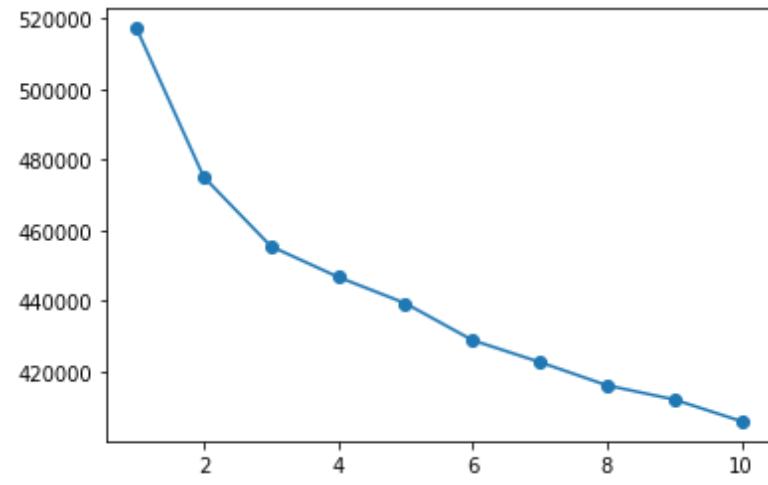
※ 결론

- 수학1에서 잘/못하는 집단 나누는 것은 위 문제들
- 못하는 학생들이 수열의 합부터 차근차근 풀어야 나가야 함
- 잘하는 학생들은 상용로그, 등차수열 등을 학습해야 BEST 될 수 있음

2. Students Classification by K-means

군집 수 결정 : 수학2

클러스터링



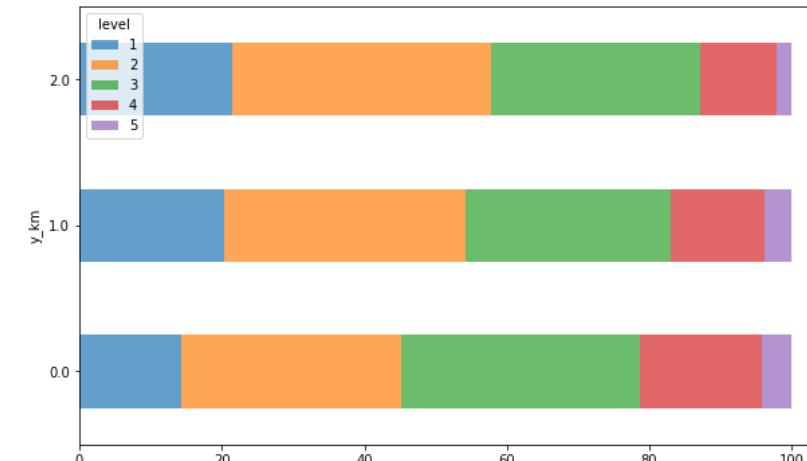
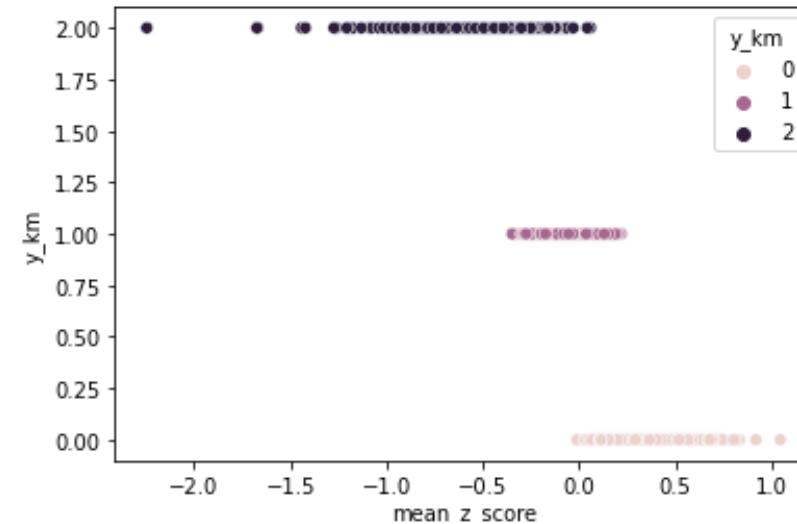
최적의 군집 수 결정

1. ELBOW METHOD 통해 [군집 수=3] 결정
2. 왼쪽 그래프 통해 [군집 수 = 3] 인 경우, 군집화 잘 되는지 판단
→ 비슷한 거리에 위치한 사람끼리 잘 묶임

2. Students Classification by K-means

군집 해석 : 수학2

클러스터링



- 왼쪽 그래프 통해 군집 간 평균 Z점수에 차이가 유의미하게 나는 것 확인 가능
(0번 집단 잘함, 2번 집단 잘 못함)
- 오른쪽 그래프 통해 잘하는 집단(0번)이 난이도 높은 문제(보라, 빨강) 많이 푸는 것 확인 & 못하는 집단(2번)이 난이도 낮은 문제(파랑, 노랑) 많이 푸는 것 확인

2. Students Classification by K-means

결과 : 수학2

군집을 결정한 주요 요인 파악

성적 높은 학생군

	PC1
3414_5	0.380038
3421_5	0.309684
3409_4	0.296956
3414_4	0.290782
3411_5	0.287714
3409_5	0.242267
3408_5	0.232648
3416_4	0.222030
3408_4	0.190042
3417_4	0.187680

성적 낮은 학생군

	PC1
3409_2	0.202572
3417_3	0.194199
3414_3	0.193661
3417_2	0.188851
3414_2	0.186018
3415_2	0.183530
3415_3	0.157682
3410_2	0.154526
3409_3	0.153849
3407_3	0.153022

분석 결과

- 성적 높은 학생군 내의 설명력이 가장 높은 단원
→ 정적분, 연속함수 성질, 방·부등식의 활용 등 고난도 문제
- 성적 낮은 학생군 내의 설명력이 가장 높은 단원
→ 방·부등식의 활용, 넓이, 정적분 등 저+중난도 문제

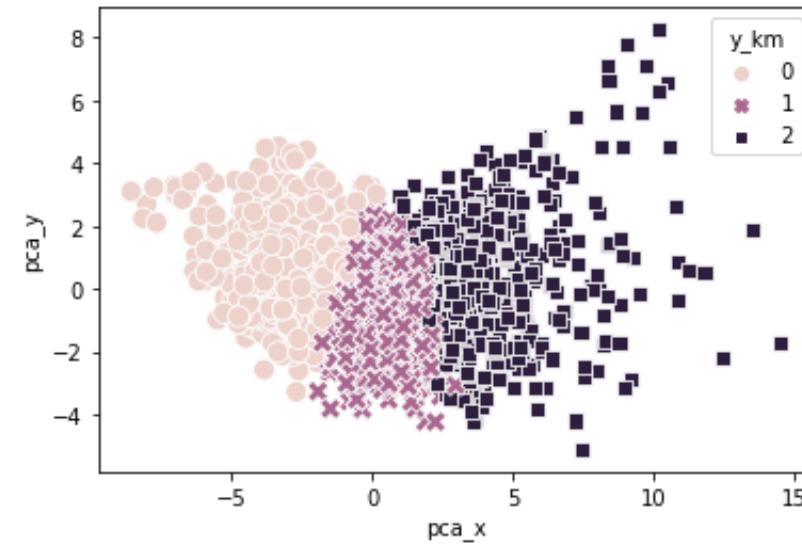
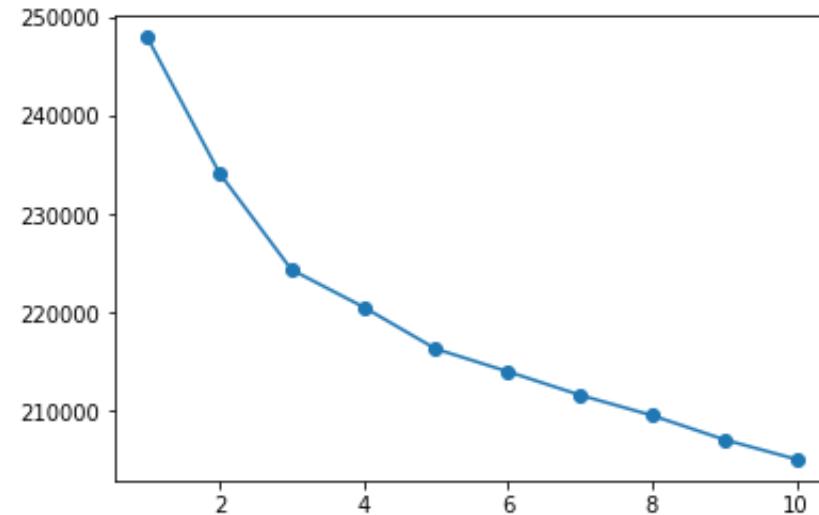
※ 결론

- 수학2에서 잘/못하는 집단 나누는 것은 위 문제들
- 잘하는 학생은 고난도위주 문제 풀이, 못하는 학생은 저난도 문제부터 풀어야 함
- **못하는 학생**들이 방정식의 활용부터 차근차근 풀어야 해 나가야 함
- 잘하는 학생들은 정적분 등을 학습해야 BEST 될 수 있음

2. Students Classification by K-means

군집 수 결정 : 미적분

클러스터링



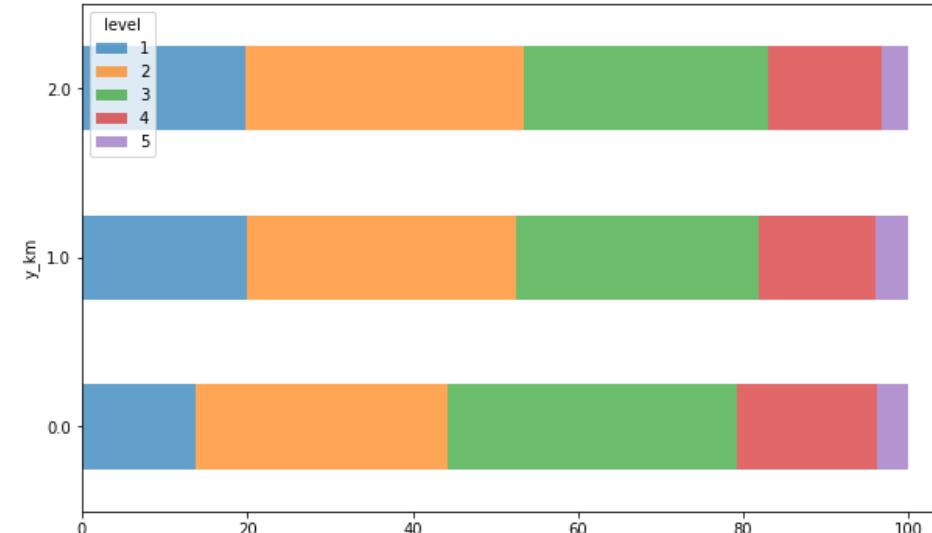
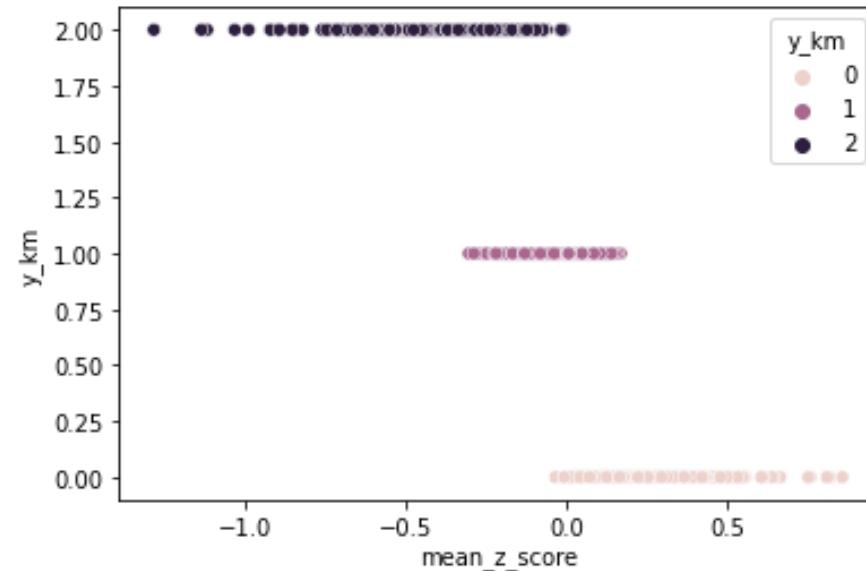
최적의 군집 수 결정

1. ELBOW METHOD 통해 [군집 수=3] 결정
2. 왼쪽 그래프 통해 [군집 수 = 3] 인 경우, 군집화 잘 되는지 판단
→ 비슷한 거리에 위치한 사람끼리 잘 묶임

2. Students Classification by K-means

군집 해석 : 미적분

클러스터링



- 왼쪽 그래프 통해 군집 간 평균 Z점수에 차이가 유의미하게 나는 것 확인 가능
(0번 집단 잘함, 2번 집단 잘 못함)
- 오른쪽 그래프 통해 잘하는 집단(0번)이 나이도 높은 문제(보라, 빨강) 많이 푸는 것 확인 & 못하는 집단(2번)이 나이도 낮은 문제(파랑, 노랑) 많이 푸는 것 확인

2. Students Classification by K-means

결과 : 미적분

군집을 결정한 주요 요인 파악

성적 높은 학생군

PC1

3446_5	0.384601
3460_5	0.323164
3460_4	0.293653
3446_4	0.268747
3454_4	0.210544
3445_5	0.176476
3450_4	0.165661
3448_4	0.163358
3463_4	0.159534
3454_5	0.155769

성적 낮은 학생군

PC1

3455_1	0.229684
3455_2	0.178646
3449_1	0.018264
3465_5	0.002394

분석 결과

- 성적 높은 학생군 내의 설명력이 가장 높은 단원
→ 함수의 그래프, 함수의 정적분, 함수의 그래프 등 고난도 문제
- 성적 낮은 학생군 내의 설명력이 가장 높은 단원
→ 수열의 극한, 매개변수의 미분, 도형의 부피 저+고난도 문제

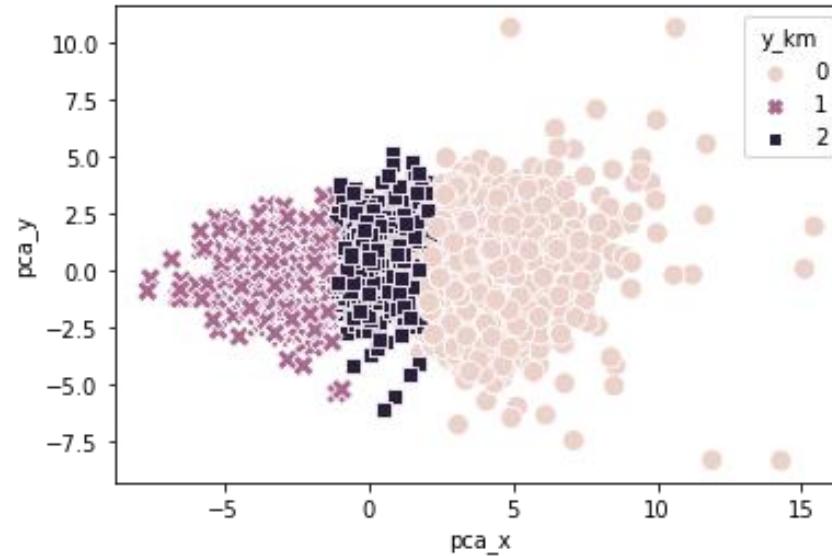
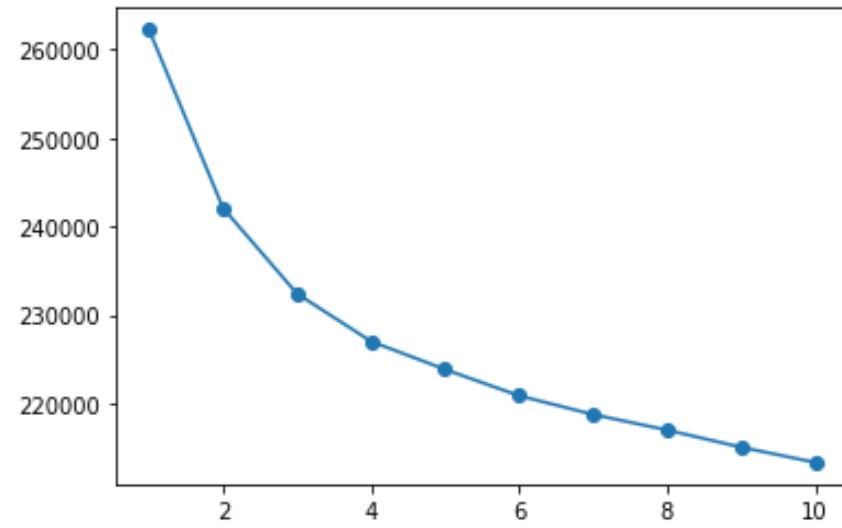
※ 결론

- 미적분에서 잘/못하는 집단 나누는 것은 위 문제들
- 잘하는 학생은 고난도위주 문제 풀이, 못하는 학생은 저난도 문제부터 풀어야 함
- 못하는 학생들이 수열의 극한부터 차근차근 풀이해 나가야 함
- 잘하는 학생들은 함수에 대해 철저히 학습해야 BEST 될 수 있음

2. Students Classification by K-means

군집 수 결정 : 확률 통계

클러스터링



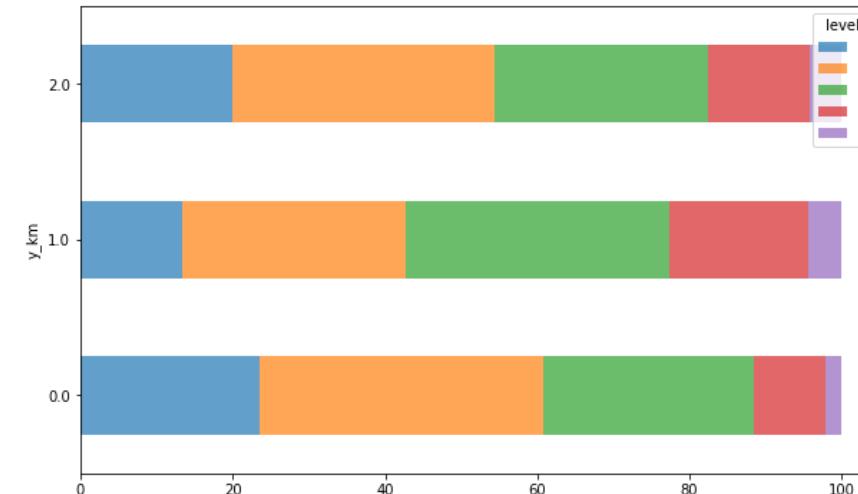
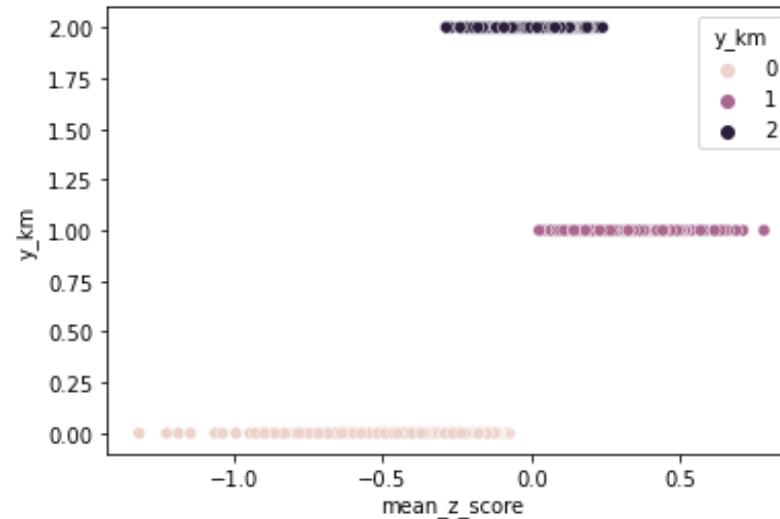
최적의 군집 수 결정

1. ELBOW METHOD 통해 [군집 수=3] 결정
2. 왼쪽 그래프 통해 [군집 수 = 3] 인 경우, 군집화 잘 되는지 판단
→ 비슷한 거리에 위치한 사람끼리 잘 묶임

2. Students Classification by K-means

군집 해석 : 확률 통계

클러스터링



- 왼쪽 그래프 통해 군집 간 평균 Z점수에 차이가 유의미하게 나는 것 확인 가능
(1번 집단 잘함, 0번 집단 잘 못함)
- 오른쪽 그래프 통해 잘하는 집단(1번)이 난이도 높은 문제(보라, 빨강) 많이 푸는 것 확인 & 못하는 집단(0번)이 난이도 낮은 문제(파랑, 노랑) 많이 푸는 것 확인

2. Students Classification by K-means

결과 : 확률 통계

군집을 결정한 주요 요인 파악

성적 높은 학생군

PC1

3427_4 0.227714

3438_4 0.223460

3429_3 0.217684

3439_4 0.208860

3433_3 0.205380

3436_3 0.204423

3426_4 0.199165

3432_3 0.197005

3436_4 0.193376

3433_4 0.185683

성적 낮은 학생군

PC1

3422_2 0.175424

3423_1 0.161807

3424_2 0.160689

3422_1 0.160230

3423_2 0.155225

3424_1 0.116941

3422_3 0.106193

3423_3 0.083680

3424_3 0.068591

3441_1 0.039139

분석 결과

- 성적 높은 학생군 내의 설명력이 가장 높은 단원
→ 평면벡터의 연속확률변수, 조건부확률, 정규분포 등 고+중난도 문제
- 성적 낮은 학생군 내의 설명력이 가장 높은 단원
→ 순열, 조합, 이항정리 등 저난도 문제

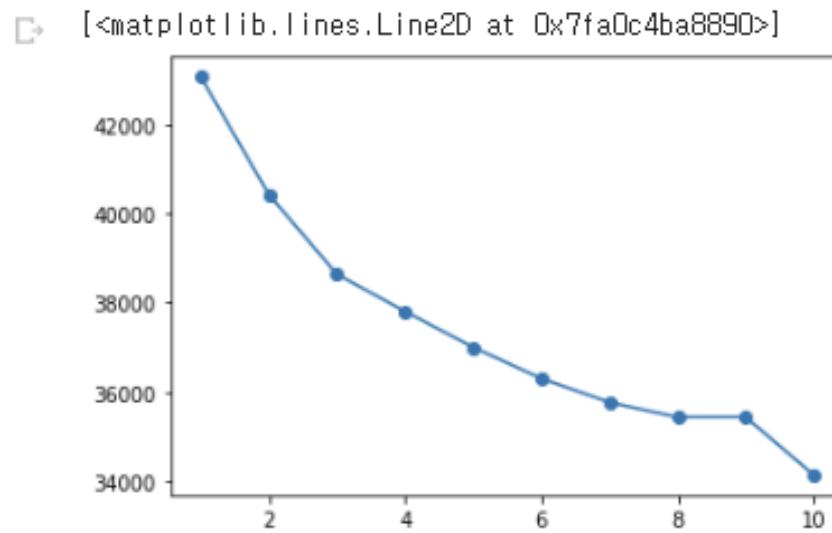
※ 결론

- 확률 통계에서 잘/못하는 집단 나누는 것은 위 문제들
- 잘하는 학생은 고난도위주 문제 풀이, 못하는 학생은 저난도 문제부터 풀어야 함
- **못하는 학생**들이 순열, 조합부터 차근차근 풀어야 나가야 함
- 잘하는 학생들은 확률, 분포에 대해 학습해야 BEST 될 수 있음

2. Students Classification by K-means

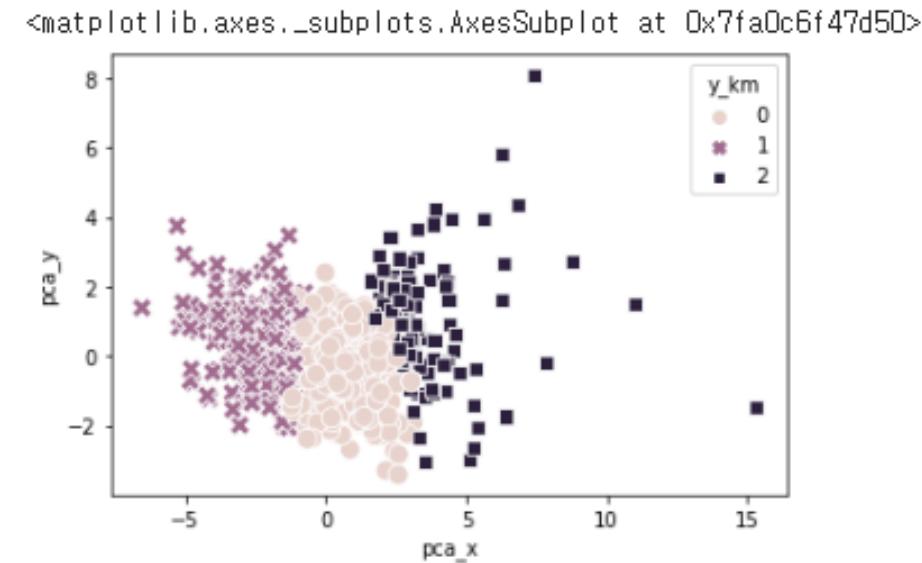
군집 수 결정 : 기하

클러스터링



최적의 군집 수 결정

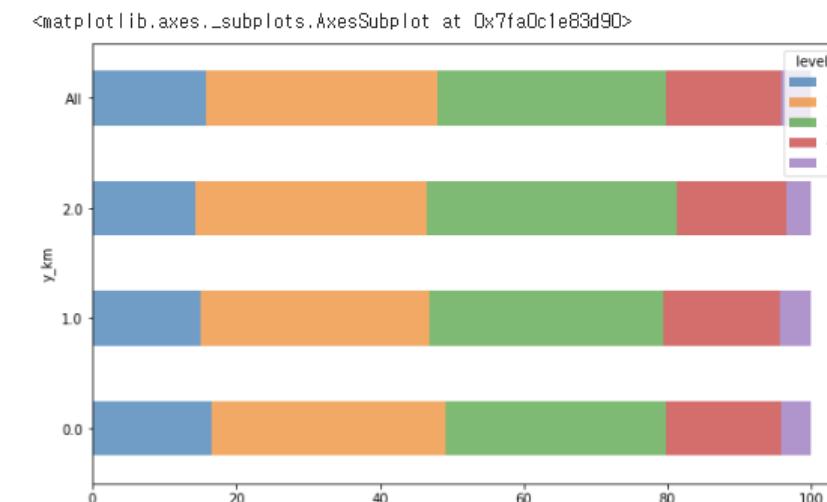
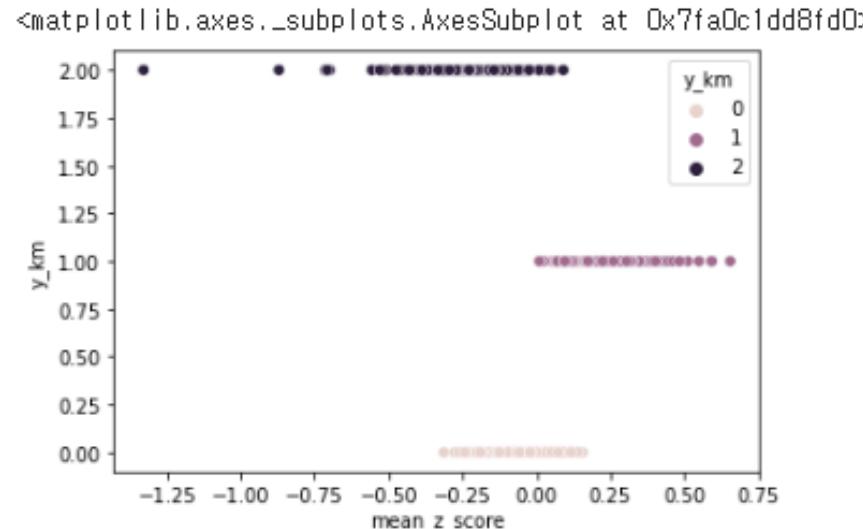
1. ELBOW METHOD 통해 [군집 수=3] 결정
2. 왼쪽 그래프 통해 [군집 수 = 3] 인 경우, 군집화 잘 되는지 판단
→ 비슷한 거리에 위치한 사람끼리 잘 묶임



2. Students Classification by K-means

군집 해석 : 기하

클러스터링



- 왼쪽 그래프 통해 군집 간 평균 Z점수에 차이가 유의미하게 나는 것 확인 가능
(1번 집단 잘함, 2번 집단 잘 못함)
- 그러나 오른쪽 그래프의 집단 별 품 문제 난이도에 유의미한 차이 없음

2. Students Classification by K-means

결과 : 기하

군집을 결정한 주요 요인 파악

성적 높은 학생군

PC1	
3479_5	0.274222
3482_4	0.270812
3483_5	0.257212
3473_4	0.217337
3485_4	0.217263
3483_4	0.214142
3475_4	0.188547
3475_5	0.183000
3488_4	0.174746
3486_4	0.173742

성적 낮은 학생군

PC1	
3482_4	0.178875
3483_4	0.069811
3488_4	0.052146
3484_4	0.037898
3485_4	0.036366
3486_4	0.028930
3479_5	0.020626
3483_5	0.018534
3482_5	0.017308
3484_3	0.014376

분석 결과

- 성적 높은 학생군 내의 설명력이 가장 높은 단원
→ 평면벡터의 내적, 포물선, 타원의 고난도 문제
- 성적 낮은 학생군 내의 설명력이 가장 높은 단원
→ 포물선, 타원, 타원과 직선의 고난도 문제

※ 결론

- 기하는 잘하는/못하는 학생들이 잘 분류되지 못함
- 기하가 가장 어려운 과목이기 때문에 예상 됨
- 못하는 학생들 잘하는 학생 간의 문제 풀이 방향 제시하기 힘듦

최종 결론 도출

학생 간 실력에 따른 분류

- 설명력이 높은 변수의 난이도를 통해 Clustering의 유효성 판단 가능
(잘하는 집단 고난도 / 못하는 집단 저난도 위주)
→ 우리의 클러스터링이 학생을 잘 분류해 주었다고 판단
- 교사들에게 학생 실력에 따른 군집 제시 가능
- 학생이 속한 군집의 강, 약점 파악으로 학생 별 지도 방향 설정에 참고

군집 형성에 영향 주는 중요 문제 파악

- 군집 형성에 영향 많이 끼치는 문제, 난이도 파악
- 따라서 학생 별 실력 향상 위한 추천 문제군 제시 가능

보완점

군집 형성 이후 추가적 분석 부재

- 군집의 특징을 분석하여 추천할 단원을 제시했으나 추가적인 분석의 부재
- 군집 별 약점, 성적 이외 변수를 Feature에 추가한다면 개선된 인사이트 도출할 수 있을 것

NN 모델의 설명력 부재

- 피쳐에 성적을 넣고 나온 임베딩을 다시 성적 분석에 사용함
- NN 모델을 사용해 임베딩하는 과정에서 칼럼 (객/주관식 여부, 성적과 연관이 있을 만한 이외 변수)을 추가해 분석했다면 의외의 인사이트가 도출될 가능성 존재

프리월린 A

- 1) K-means(DTW) Time series Clustering
- 2) Embedding 기반 정답률 예측 모델

#1 K-means(DTW) Time series Clustering

Conclusion

- ✓ 학생이 푼 최근 문제 결과값을 소단원별, Level(고난이도/저중난이도)별 시계열흐름으로 클러스터링 진행
- ✓ 소단원 별로 어느 집단에 속하는지에 따라, 집단 내 비교 가능
- ✓ 속한 집단의 상위집단 특성을 바탕으로 학생에게 공부 방식을 맞춤화하여 제안 가능

데이터(input) 설명

- 지수함수 / 정규분포 / 함수의 극한 세 단원 선택
- student_id(학생) 별 update_datetime 기준으로 정렬
- 각 단원별 최신 50개 문제 풀이 결과 데이터 사용
- 문제를 맞췄으면? “Level(난이도) x 1” 문제를 틀렸으면? “0”

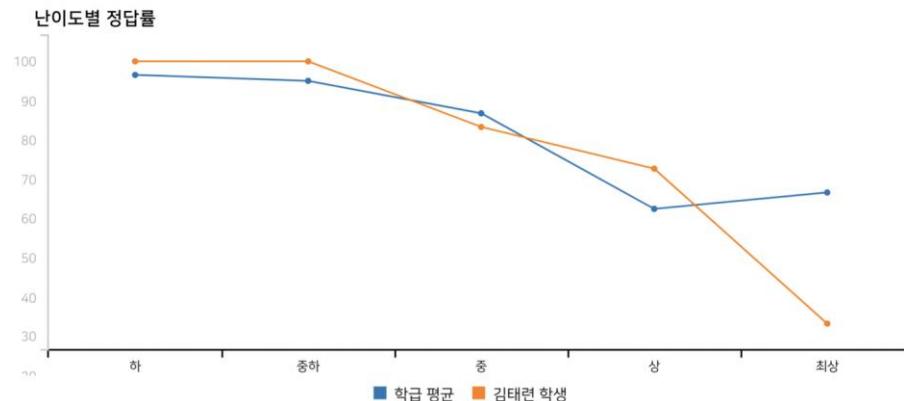


문제 인식

프리윌린 기준 보고서 방식

OVERVIEW 종합

난이도별 평가



전체 학생과 특정 학생의 상대적 비교

K-means(DTW) Time series Clustering

- ✓ 학생이 푼 최근 문제 결과값을 소단원별, Level(고난이도/저중난이도)별 시계열흐름으로 클러스터링 진행
-
- ✓ 소단원 별로 어느 집단에 속하는지에 따라, 집단 내 비교 가능
- ✓ 속한 집단의 상위집단 특성을 바탕으로 학생에게 공부 방식을 맞춤화하여 제안 가능
- ✓ 더욱 구체적으로 본인의 실력 확인 가능

K-means(DTW) Time series Clustering

K-means(DTW) Time series Clustering이란?

학생이 푼 문제 패턴 추세로 군집을 나누는 시계열 클러스터링



같은 시점끼리 유클리디안 거리로 비교를 하면 비슷한 패턴을 가진 시계열들을 같은 군집으로 묶을 수 없음



K-means DTW Time series Clustering



DTW 거리계산법으로 시점 차이가 나더라도 유사한 패턴을 가진다면 같은 군집으로 묶을 수 있게 됨

DTW 거리 수식

$$W = w_1, w_2, \dots, w_K, \max(m, n) \leq K < m + n - 1$$

$$DTW(Q, C) = \min \left\{ \frac{1}{K} \sqrt{\sum_{k=1}^K w_k} \right\}$$

→ 비슷한 패턴을 가지는 집단끼리 군집을 나누기 위해 후자인 DTW 거리계산법을 사용한 K-means clustering 사용

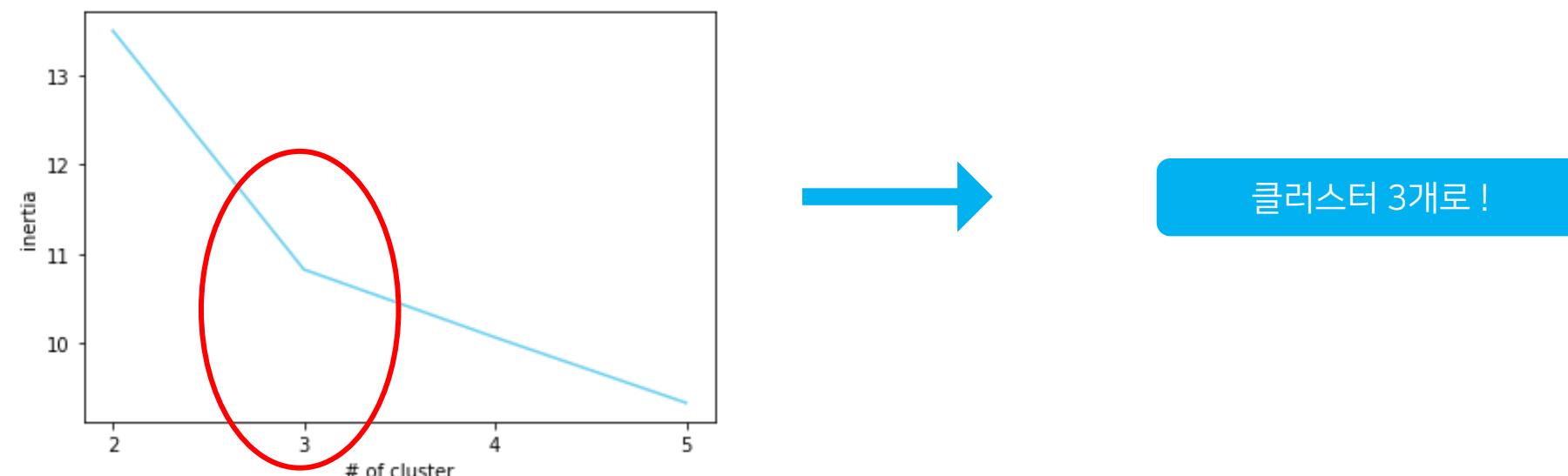
K-means(DTW) Time series Clustering

Cluster 수 설정

Inertia value

“군집 내 데이터들과 중심 데이터의 간의 거리의 합으로 군집의 응집도를 나타내는 값”

- ✓ 값이 작을수록 군집 내 응집도가 높게 군집화가 잘 되었다.
- ✓ 군집 단위 별로 inertia 값을 조회한 후 값이 급격히 떨어지는 지점이 적정 군집 수라고 판단할 수 있다.



K-means(DTW) Time series Clustering

표본 집단 및 데이터 설정

- ✓ worksheet만 사용
- ✓ student
 - 고3만 사용
 - school = 3
 - grade = 3
- ✓ student_worksheet_scoring
- ✓ result = correct / wrong 만 사용
- ✓ easy -> problem.'level' = 1, 2
- ✓ hard -> problem.'level' = 3, 4, 5



K-means(DTW) Time series Clustering

대표 단원 선택 및 클러스터링 대상 집단

[]	little_chapter[0].sort_values(ascending = False)
little_chapter_name	
지수함수 $y=a^x$ ($a>0$, $a\neq 1$)의 성질	1473168
지수부등식	1348818
정규분포	1027964
보정근과 표준법	895487
모비율의 추정	793191
지수방정식	786072
로그부등식	763498
모평균의 추정	619418
표본비율의 분포	586213
이산화률변수	475318
함수의 극한값의 계산	455937
순열	454105
로그함수 $y=\log_a x$ ($a>0$, $a\neq 1$)의 성질	406026
지수의 확장과 지수법칙	401193
지수부등식	394956
로그의 뜻과 성질	372372
로그방정식	371086
조합	370093
정적분	342685
확률의 계산과 활용	329067
지수함수	327007
등비수열	308568
함수의 그래프	307889
연속함수의 성질	306130
접선의 방정식	287682
연속확률변수	279901
이항분포	266473
미분계수	265081
등차수열	260348
도함수	249203

Easy

[]	group1['little_chapter_name'].sort_values(ascending = False)
little_chapter_name	
지수함수 $y=a^x$ ($a>0$, $a\neq 1$)의 성질	147435
지수부등식	1360356
정규분포	238712
보정근과 표준법	285730
로그부등식	242404
지수방정식	202720
모평균의 추정	196899
모비율의 추정	188648
표본비율의 분포	152531
이산화률변수	128237
순열	122998
함수의 그래프	122448
지스브드시	119967
함수의 극한값의 계산	116170
연속함수의 성질	111617
로그함수 $y=\log_a x$ ($a>0$, $a\neq 1$)의 성질	104453
조합	103529
정적분	99323
등비수열	98707
접선의 방정식	94764
지수의 확장과 지수법칙	93147
연속확률변수	92757
지수함수	89383
확률의 계산과 활용	89013
사인법칙과 코사인법칙	83573
로그방정식	80131
로그의 뜻과 성질	79964
여러 가지 수열의 합	76110
미분계수	73959
등차수열	73253
도함수	72134

Hard

```
[ ] a = set(hard_norm_50['student_id'].unique())
b = set(hard_fx_50['student_id'].unique())
c = set(hard_exp_50['student_id'].unique())
d = set(easy_norm_50['student_id'].unique())
e = set(easy_fx_50['student_id'].unique())
f = set(easy_exp_50['student_id'].unique())

[ ] st = a&b&c&d&e&f

[ ] fc = hard_fx_real.groupby('student_id').agg({'student_id':'count'})
fc['student_id'].sort_values(ascending = False)

student_id
I172112    779
I137768    555
I339347    468
I228867    446
I332869    342
...
I199744     50
I157283     50
I312681     50
I325963     50
I194214     50
Name: student_id, Length: 383, dtype: int64
```

K-means(DTW) Time series Clustering

대표 단원 선택 및 클러스터링 대상 집단

[]	little_chapter[0].sort_values(ascending = False)
little_chapter_name	
지수함수 $y=a^x$ ($a>0$, $a\neq 1$)의 성질	1473168
정규분포	1348818
로그부등식	1027964
모비우스 추정	895487
지수방정식	793191
로그부등식	786072
모비우스 추정	763498
이산화를변수	586213
함수의 극한값의 계산	455937
순열	401193
로그함수 $y=\log_a x$ ($a>0$, $a\neq 1$)의 성질	372372
지수의 확장과 지수법칙	371086
조합	370093
정적분	342685
확률의 계산과 활용	329067
지수함수	327007
등비수열	308568
함수의 그래프	307889
연속함수의 성질	306130
접선의 방정식	287682
연속확률변수	279901
이항분포	266473
미분계수	265081
등차수열	260348
도함수	249203

Easy

[]	group1['little_chapter_name'].sort_values
little_chapter_name	
지수함수 $y=a^x$ ($a>0$, $a\neq 1$)의 성질	147435
정규분포	360356
로그부등식	28712
지수방정식	285730
모평균의 추정	242404
모평균의 추정	202720
함수의 그래프	196899
이산화를변수	128237
순열	122998
함수의 극한값의 계산	122448
조합	111707
정적분	103529
등비수열	99323
접선의 방정식	98707
지수의 확장과 지수법칙	94764
연속확률변수	93147
지수함수	92757
확률의 계산과 활용	89383
사인법칙과 코사인법칙	89013
로그부등식	83573
로그의 뜻과 성질	80131
여러 가지 수열의 합	79964
미분계수	76110
등차수열	73959
도함수	73253

Hard

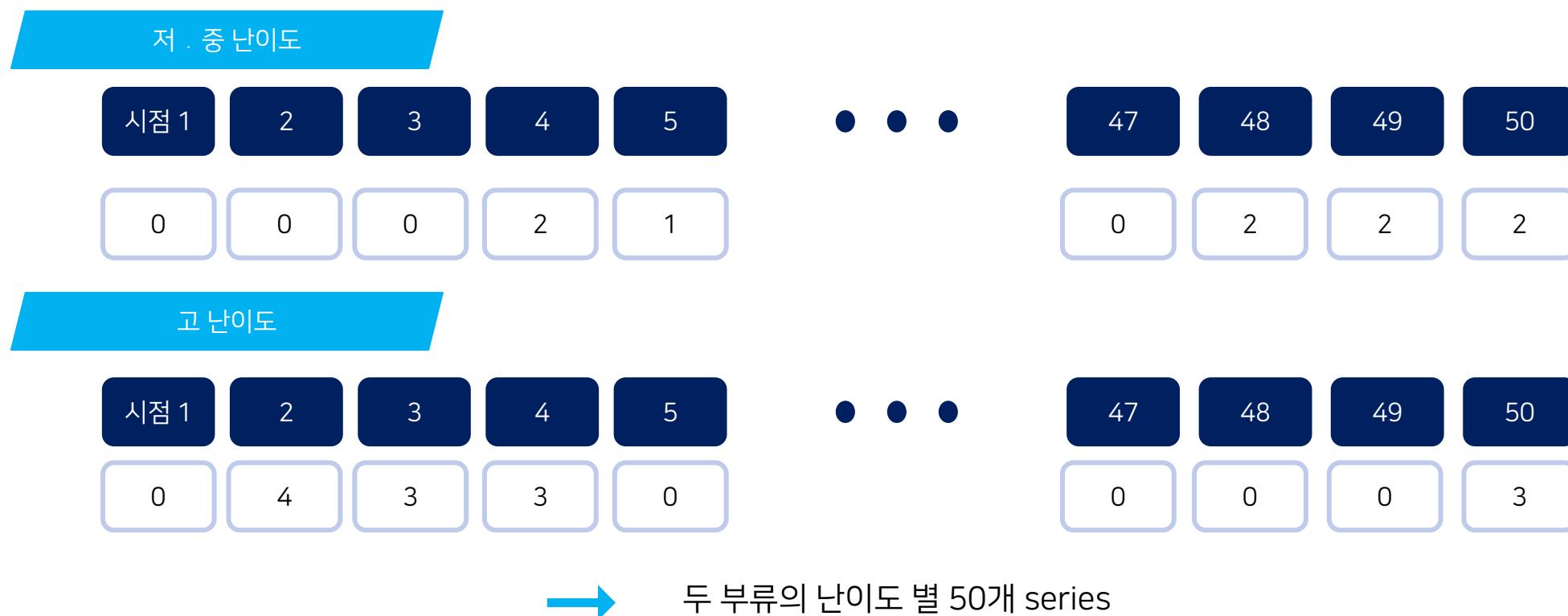
[]	a = set(hard_norm_50['student_id'].unique()) b = set(hard_fx_50['student_id'].unique()) c = set(hard_exp_50['student_id'].unique()) d = set(easy_norm_50['student_id'].unique()) e = set(easy_fx_50['student_id'].unique()) f = set(easy_exp_50['student_id'].unique())
[]	fc = hard_fx_real.groupby('student_id').agg({'student_id': count})
[]	fc['student_id'].sort_values(ascending = False)
student_id	
I172112	779
I137768	555
I339347	468
I228867	446
I332869	342
	...
I199744	50
I157283	50
I312681	50
I325963	50
I194214	50

Name: student_id, Length: 383, dtype: int64

데이터(input) 설명

student_id(학생) 별 update_datetime 기준으로 정렬,
최신 50개 문제 풀이 결과 데이터 사용
예시) student_id = '197414'

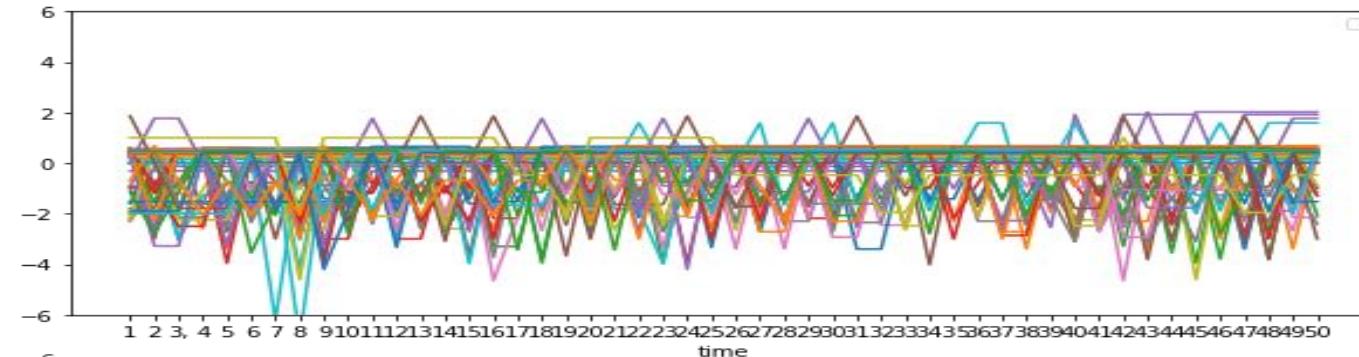
문제를 맞췄으면? "Level(난이도) x 1"
문제를 틀렸으면? "0"



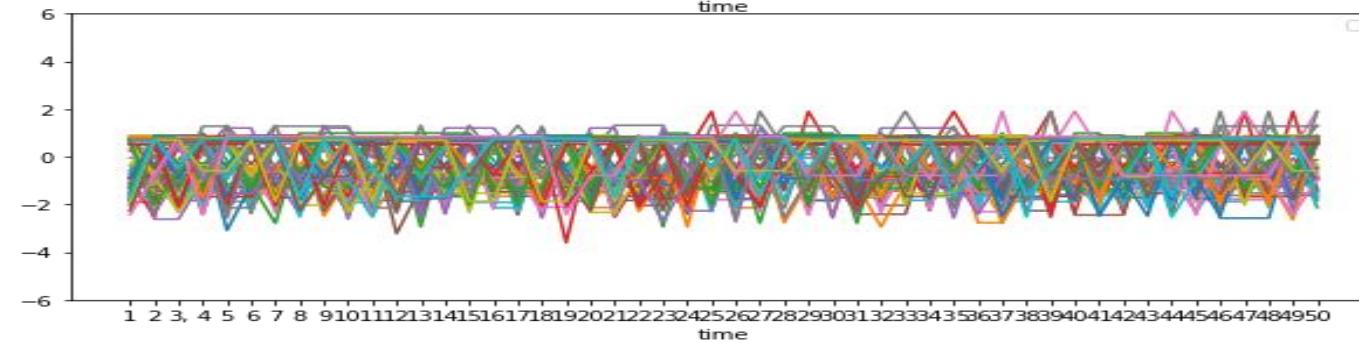
지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질

저 . 중 난이도

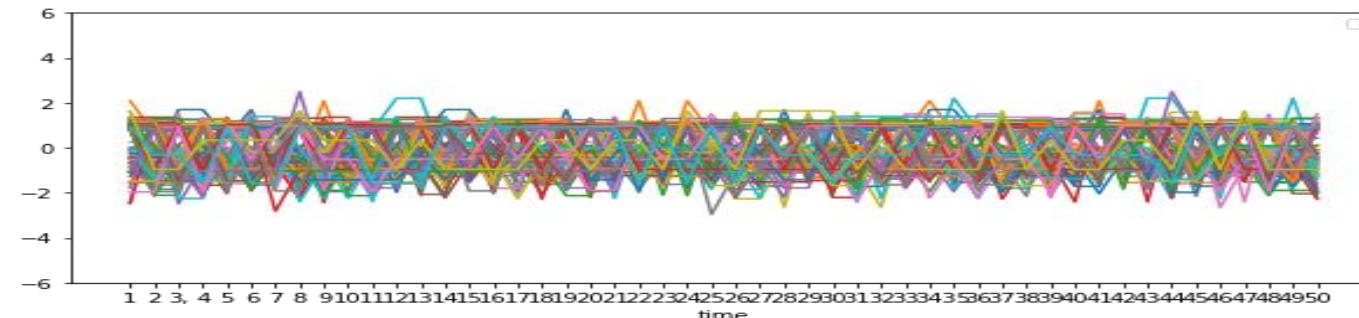
클러스터 0



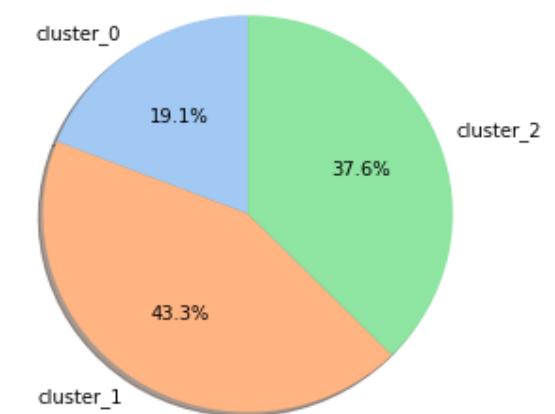
클러스터 1



클러스터 2



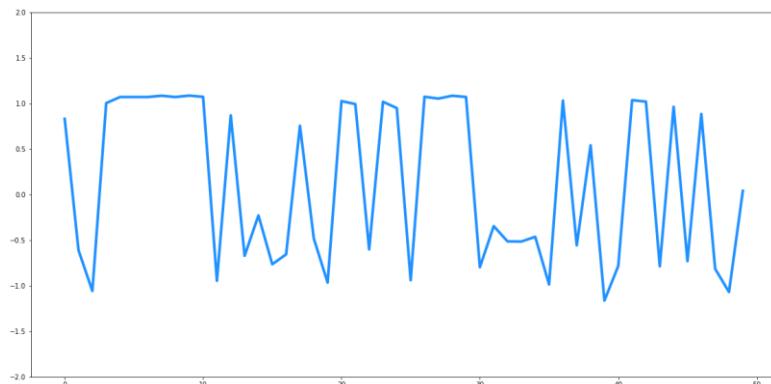
Cluster Distribution



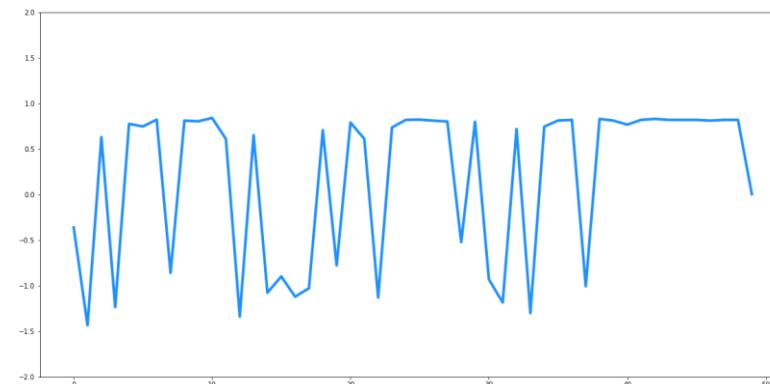
K-means(DTW) Time series Clustering

지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질

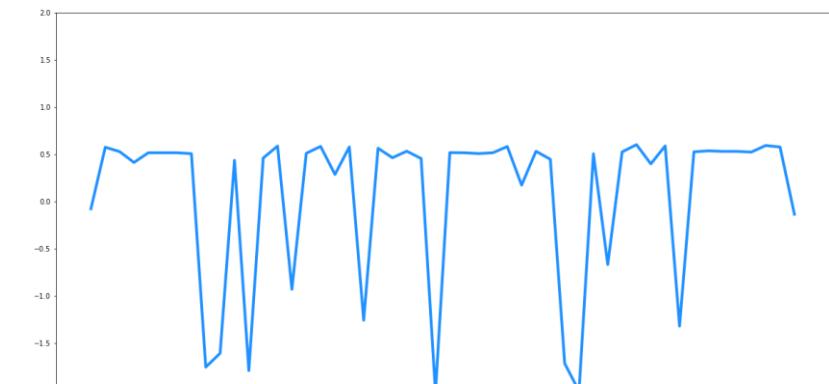
저 . 중 난이도



클러스터 0



클러스터 1



클러스터 2

→ 군집 중심 데이터만을 시계열로 나타냄 BUT 큰 차이점 발견 어려움

50개의 series 데이터가 시각적으로 성적변화(추이)를 관찰하기 힘든 짧은 기간인 점 때문에 나타난 한계로 추측

K-means(DTW) Time series Clustering

지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질

저. 중. 난이도

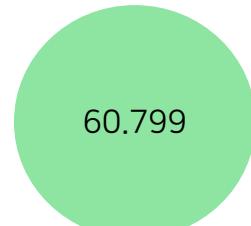
학생 별 50개 시점의 점수합 평균



클러스터 0



클러스터 1



클러스터 2

학생 별 50개 시점의 표준 편차 요약 통계량

Q1: 0.47
Q2: 0.58
Q3: 0.66
mean: 0.56

Q1: 0.64
Q2: 0.73
Q3: 0.81
mean: 0.73

Q1: 0.66
Q2: 0.75
Q3: 0.84
mean: 0.74

→ 요약통계량을 통해 집단들의 점수, 문제 풀이 결과의 움직임
(결과가 많이 왔다갔다하는지, 안정적으로 점수 받는지)으로 분류되었음을 추측 가능

'지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질' 클러스터 정리

저 . 중 난이도

클러스터 0

지수함수 $y = a^x (a>0, a\neq 1)$ 저 . 중 난이도 문제의 정답률이 높은 집단에 속하고,
문제를 안정적으로 맞히고 있는 집단

클러스터 1

지수함수 $y = a^x (a>0, a\neq 1)$ 저 . 중 난이도 문제의 정답률이 중간 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

클러스터 2

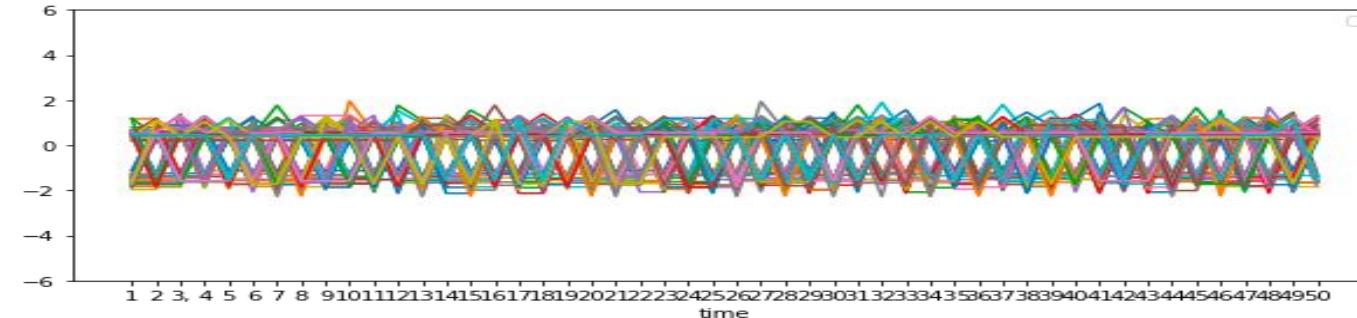
지수함수 $y = a^x (a>0, a\neq 1)$ 저 . 중 난이도 문제의 정답률이 낮은 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

K-means(DTW) Time series Clustering

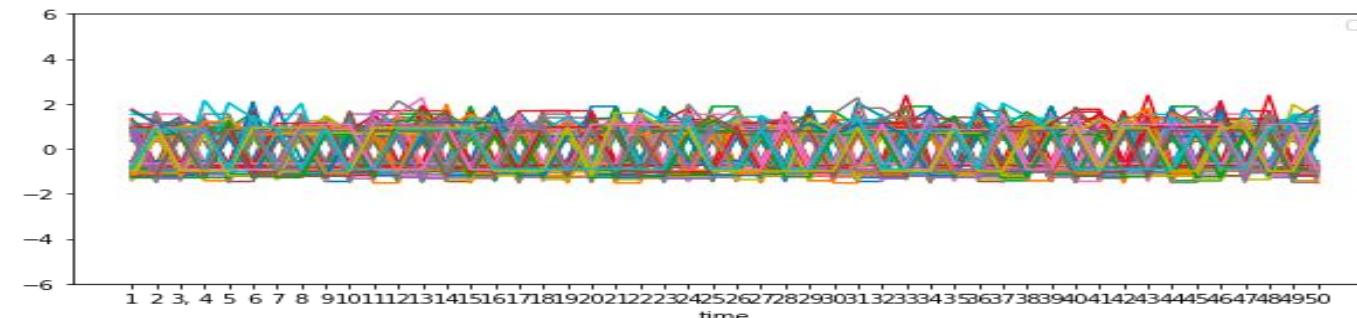
지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질

고 나이도

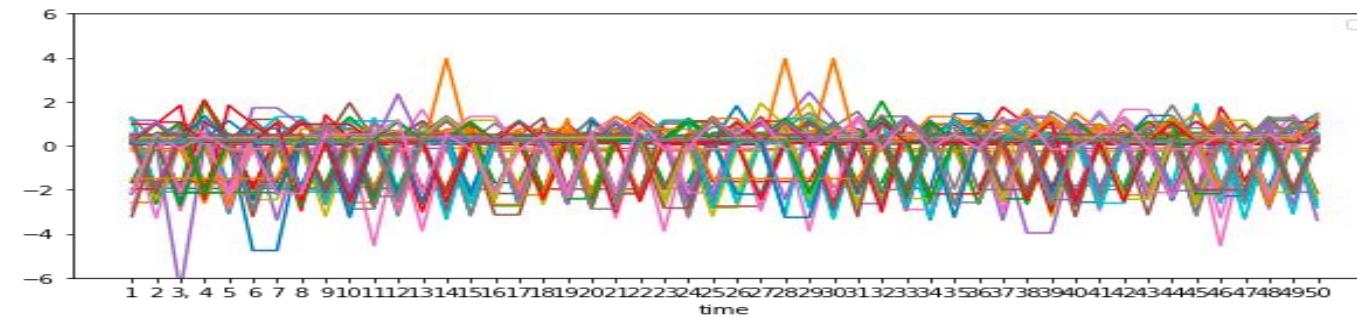
클러스터 0



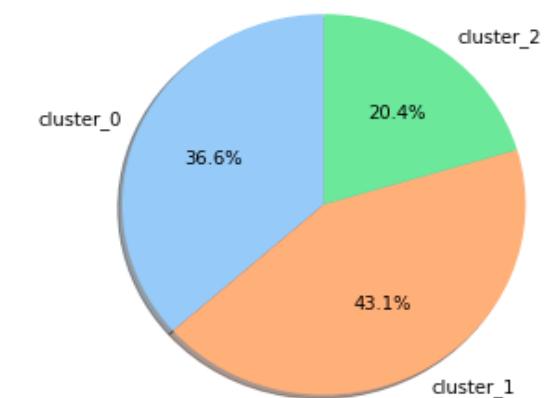
클러스터 1



클러스터 2



Cluster Distribution



K-means(DTW) Time series Clustering

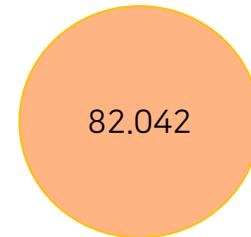
지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질

고 나이도

학생 별 50개 시점의 점수합 평균



클러스터 0



클러스터 1



클러스터 2

학생 별 50개 시점의 표준 편차 요약 통계량

Q1: 1.349
Q2: 1.411
Q3: 1.48
mean: 1.415

Q1: 1.504
Q2: 1.542
Q3: 1.509
mean: 1.545

Q1: 0.985
Q2: 1.097
Q3: 1.239
mean: 1.099

'지수함수 $y = a^x (a>0, a\neq 1)$ 의 성질' 클러스터 정리

고 나이도

클러스터 0

지수함수 $y = a^x (a>0, a\neq 1)$ 고 나이도 문제의 정답률이 중간 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

클러스터 1

지수함수 $y = a^x (a>0, a\neq 1)$ 고 나이도 문제의 정답률이 낮은 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

클러스터 2

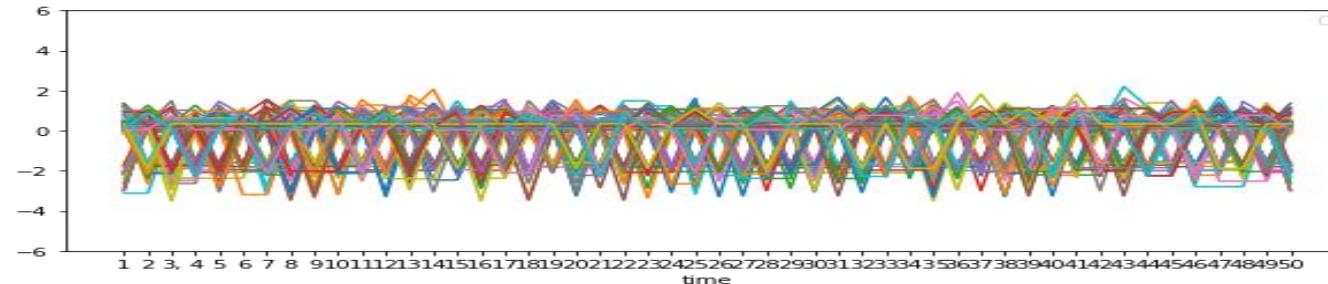
지수함수 $y = a^x (a>0, a\neq 1)$ 고 나이도 문제의 정답률이 높은 집단에 속하고,
문제를 안정적으로 맞히고 있는 집단

K-means(DTW) Time series Clustering

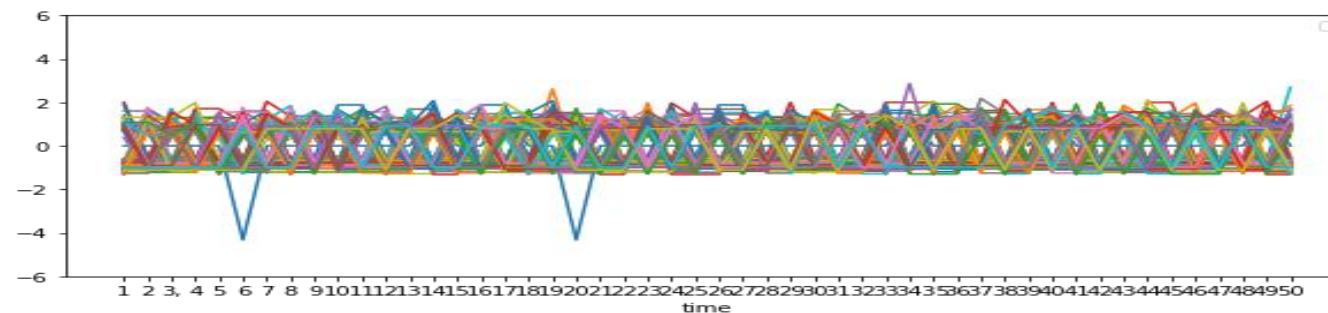
정규 분포

저. 중 난이도

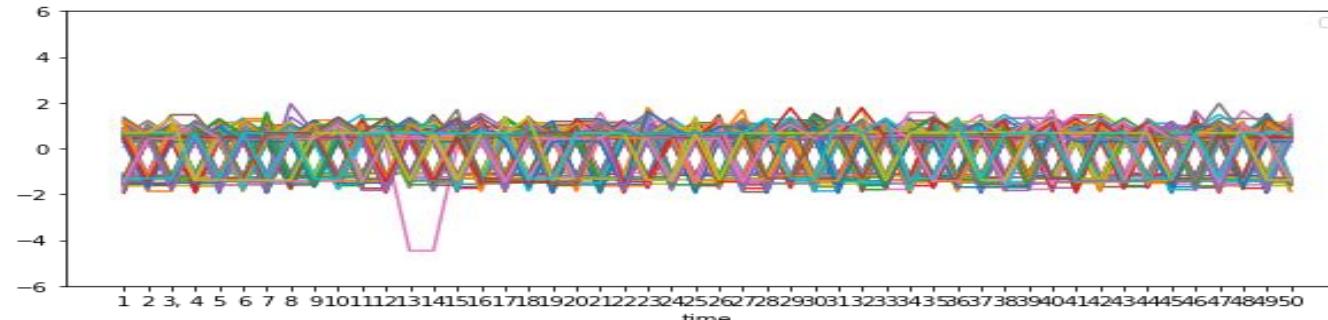
클러스터 0



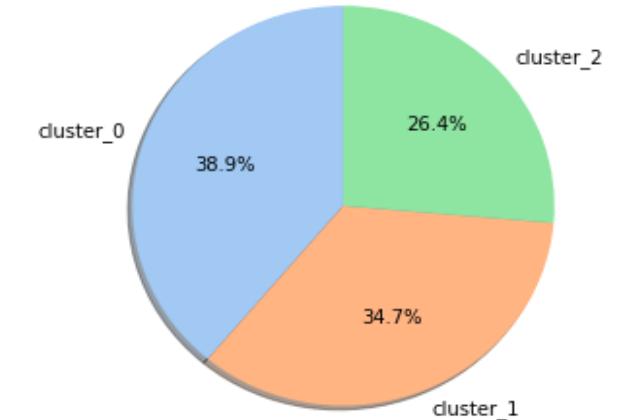
클러스터 1



클러스터 2



Cluster Distribution



정규 분포

저. 중 난이도

학생 별 50개 시점의 점수합 평균



학생 별 50개 시점의 표준 편차 요약 통계량

Q1: 0.676
Q2: 0.771
Q3: 0.858
mean: 0.764

Q1: 0.631
Q2: 0.707
Q3: 0.810
mean: 0.705

Q1: 0.618
Q2: 0.727
Q3: 0.798
mean: 0.707

K-means(DTW) Time series Clustering

정규 분포 클러스터 정리

저. 중 난이도

클러스터 0

정규분포 저 . 중 난이도 문제의 정답률이 중간 집단

클러스터 1

정규분포 저 . 중 난이도 문제의 정답률이 높은 집단

클러스터 2

정규분포 저 . 중 난이도 문제의 정답률이 낮은 집단

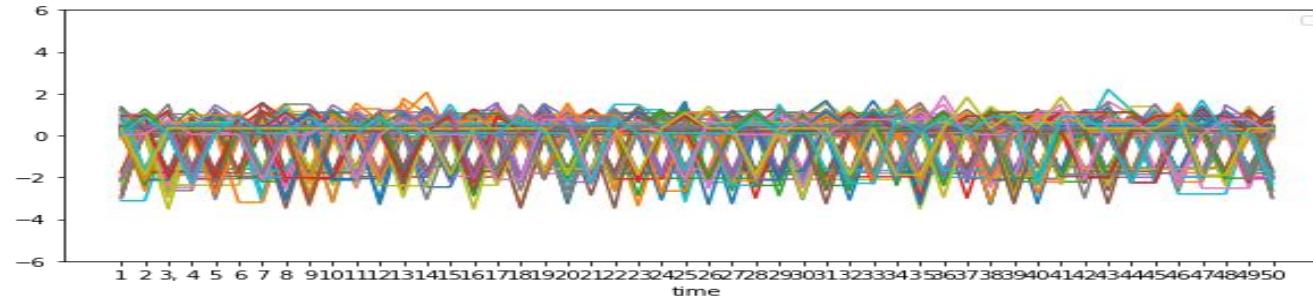
다른 클러스터링과 다르게 표준편차 간에는 큰 차이 발견하기 어려움

K-means(DTW) Time series Clustering

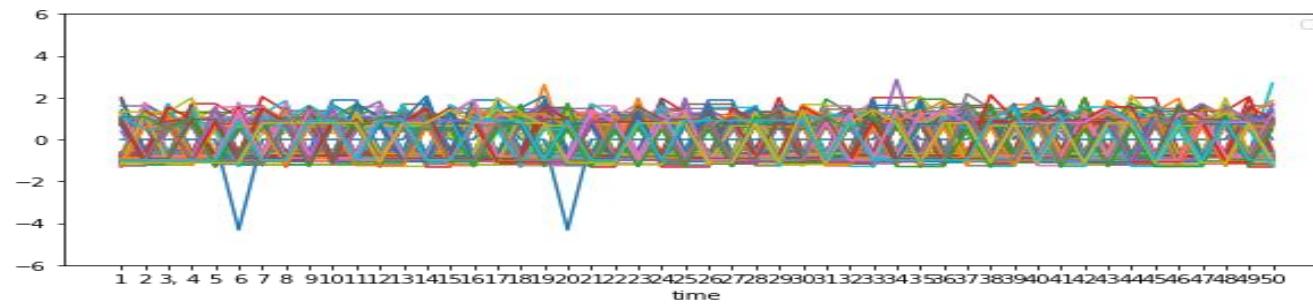
정규 분포

고 나이도

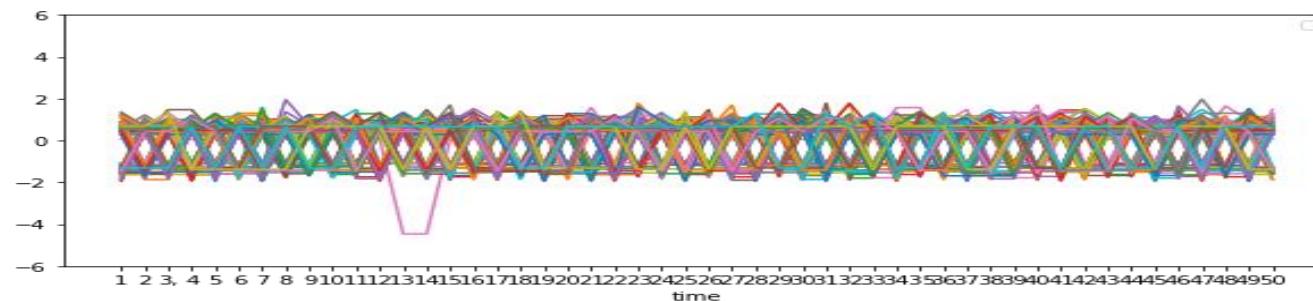
클러스터 0



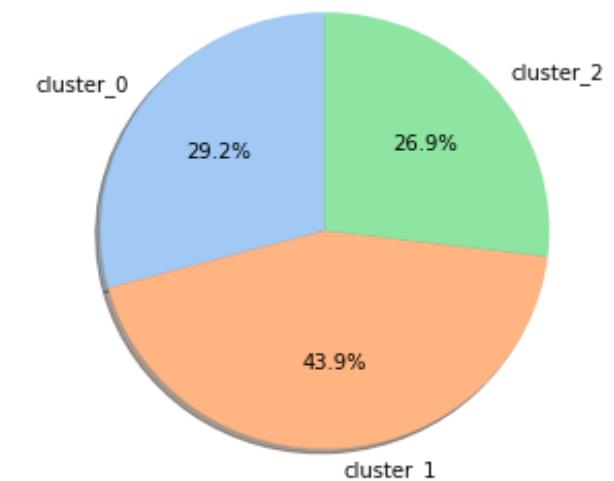
클러스터 1



클러스터 2



Cluster Distribution



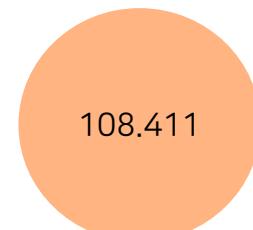
정규 분포

고 나이도

학생 별 50개 시점의 점수합 평균



클러스터 0



클러스터 1



클러스터 2

학생 별 50개 시점의 표준 편차 요약 통계량

Q1: 1.527
Q2: 1.589
Q3: 1.65
mean: 1.571

Q1: 1.455
Q2: 1.528
Q3: 1.584
mean: 1.526

Q1: 1.139
Q2: 1.263
Q3: 1.357
mean: 1.226

K-means(DTW) Time series Clustering

정규 분포 클러스터 정리

고 나이도

클러스터 0

정규분포 고 나이도 문제의 정답률이 낮은 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

클러스터 1

정규분포 고 나이도 문제의 정답률이 중간 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

클러스터 2

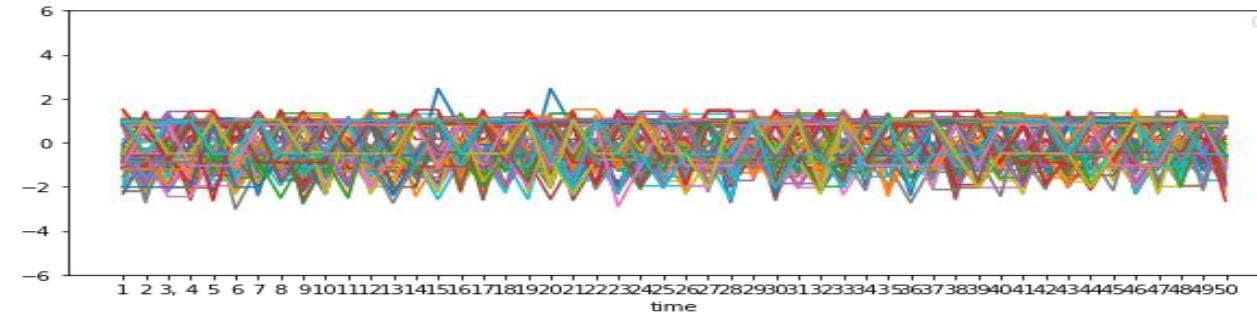
정규분포 고 나이도 문제의 정답률이 높은 집단에 속하고,
문제를 안정적으로 맞히고 있는 집단

K-means(DTW) Time series Clustering

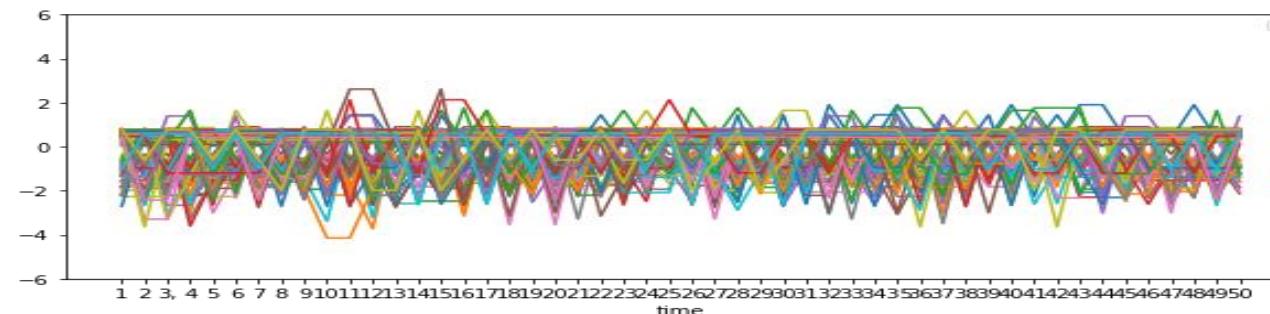
함수의 극한

저. 중 난이도

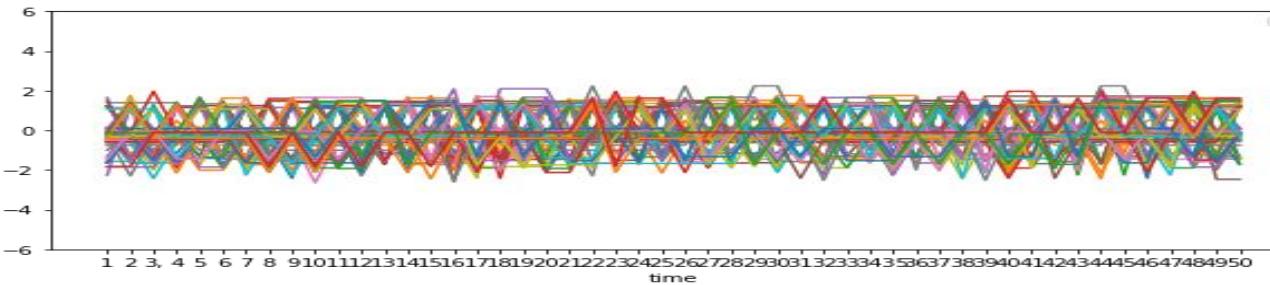
클러스터 0



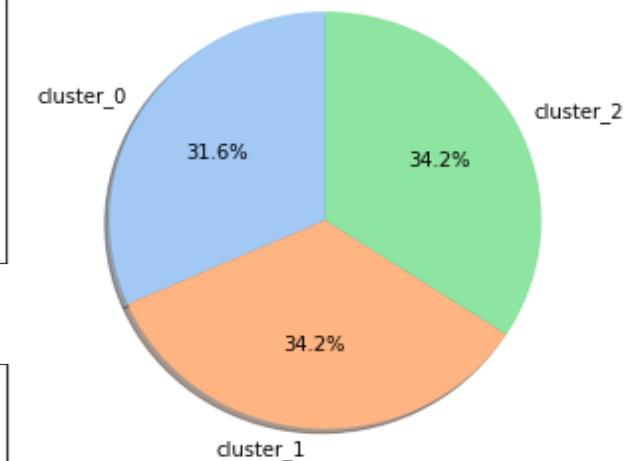
클러스터 1



클러스터 2



Cluster Distribution



K-means(DTW) Time series Clustering

함수의 극한

저. 중 난이도

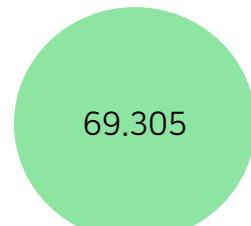
학생 별 50개 시점의 점수합 평균



클러스터 0



클러스터 1



클러스터 2

학생 별 50개 시점의 표준 편차 요약 통계량

Q1: 0.580
Q2: 0.685
Q3: 0.749
mean: 0.662

Q1: 0.600
Q2: 0.697
Q3: 0.769
mean: 0.687

Q1: 1.45
Q2: 1.52
Q3: 1.56
mean: 1.48

K-means(DTW) Time series Clustering

함수의 극한

저. 중 난이도

클러스터 0

함수의 극한 저. 중 난이도 문제의 정답률이 높은 집단에 속하고,
문제를 안정적으로 맞히고 있는 집단

클러스터 1

함수의 극한 저. 중 난이도 문제의 정답률이 낮은 집단에 속하고,
문제를 안정적으로 맞히고 있는 집단

클러스터 2

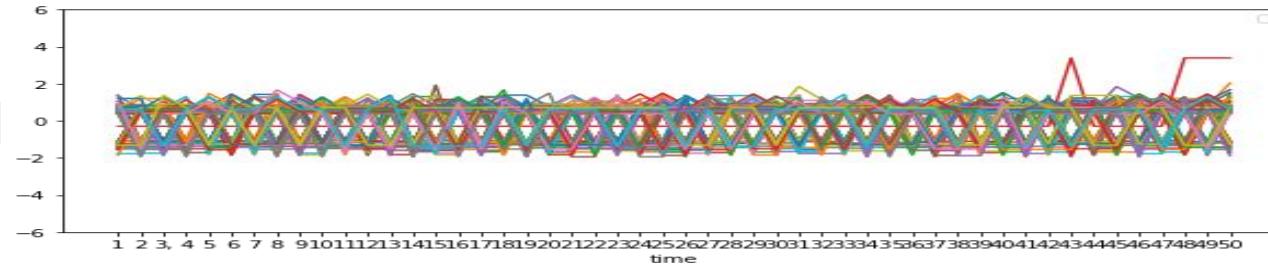
함수의 극한 저. 중 난이도 문제의 정답률이 중간 집단에 속하고,
문제를 매우 안정적으로 맞히지 못하는 집단

K-means(DTW) Time series Clustering

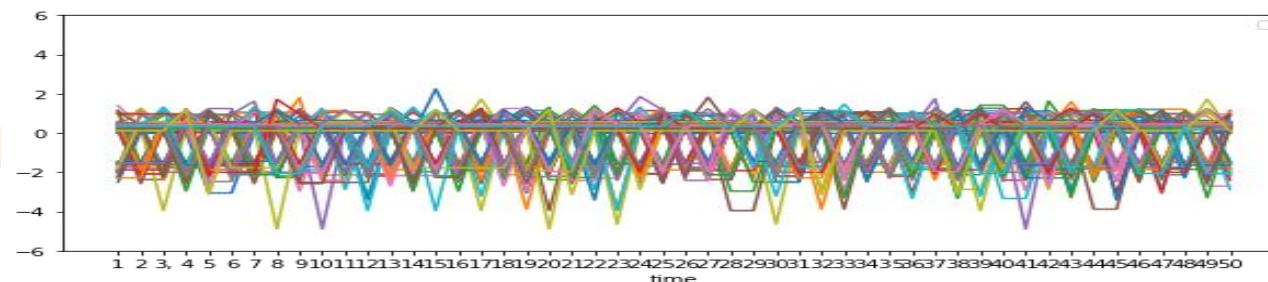
함수의 극한

고 나이도

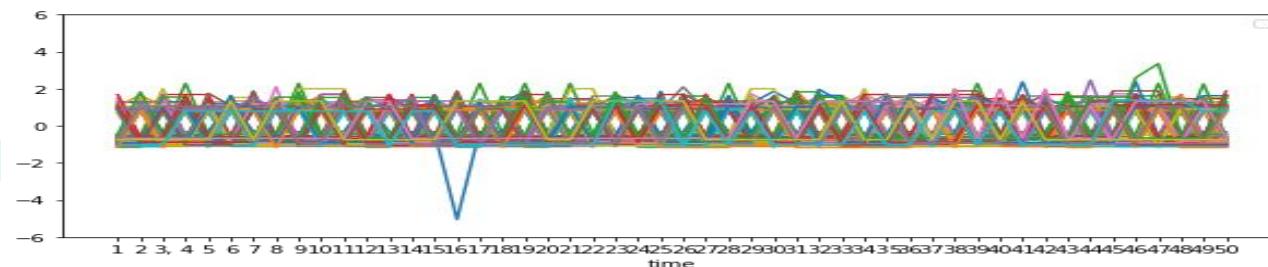
클러스터 0



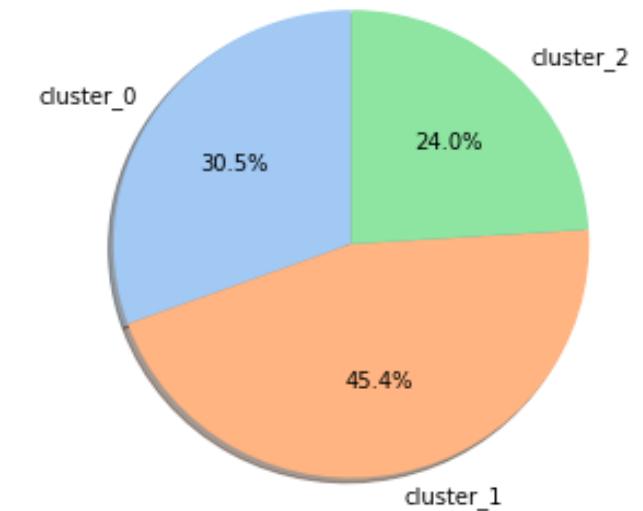
클러스터 1



클러스터 2



Cluster Distribution



K-means(DTW) Time series Clustering

함수의 극한

고 나이도

학생 별 50개 시점의 점수합 평균



클러스터 0



클러스터 1



클러스터 2

학생 별 50개 시점의 표준 편차 요약 통계량

Q1: 1.102
Q2: 1.222
Q3: 1.349
mean: 1.199

Q1: 1.434
Q2: 1.498
Q3: 1.548
mean: 1.488

Q1: 1.455
Q2: 1.526
Q3: 1.568
mean: 1.482

K-means(DTW) Time series Clustering

함수의 극한

고 나이도

클러스터 0

함수의 극한 고 나이도 문제의 정답률이 높은 집단에 속하고,
문제를 안정적으로 맞히고 있는 집단

클러스터 1

함수의 극한 고 나이도 문제의 정답률이 중간 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

클러스터 2

함수의 극한 고 나이도 문제의 정답률이 낮은 집단에 속하고,
문제를 안정적으로 맞히지 못하는 집단

K-means(DTW) Time series Clustering

단원 별 차이를 보이지 않은 학생 예시



쉬운 문제들: 정답률 높음 VS 어려운 문제들: 점수 평균이 낮고 분산이 높은 집단에 속함
 → 기초적인 개념을 다루는 쉬운 문제들에는 강하지만, 어려운 응용문제들에 약한 것으로 판단됨



쉬운 문제들: 정답률 가장 낮은 집단에 속함 VS 어려운 문제들: 정답률 높고 분산 낮은, 실력이 높은 집단에 속함

→ 두 학생 모두 단원에 상관없이 문제 난이도에 따라서 다른 양상을 보임!

K-means(DTW) Time series Clustering

단원 별 차이를 보이지 않은 학생 예시



수학 교육과정 상 지수함수, 함수의 극한, 정규분포 순으로 배우게 되기 때문에 상위 교육과정에 속하는 단원들로 갈수록 정답률이 높아지는 것으로 보여짐



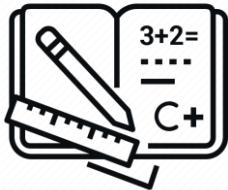
정규분포: 기초는 이미 탄탄하니, 어려운 문제집의 다양한 유형들 공략해보는 것이 바람직함

함수의 극한: 정규분포보단 우선순위가 낮지만 기초 보완

→easy 정답률 높이고 이를 응용하여 hard 문제에서도 정답률을 높여 높은 집단에 속하게 하는 것이 목표

K-means(DTW) Time series Clustering

한계점



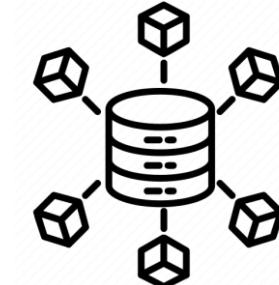
단원수 3개로 제한

학생수 확보를 위해서 문제수 50개로 제한



RAM 용량 제한으로 고3 학생들만 분석함

RAM 용량 제한으로 데이터셋 merge하는 방식에서
어려움을 겪음



클러스터링 결과 시각화의 어려움

나머지 clustering 방법은 성능이 안좋아서
kmeans만 사용

#2 Embedding 기반 정답률 예측 모델

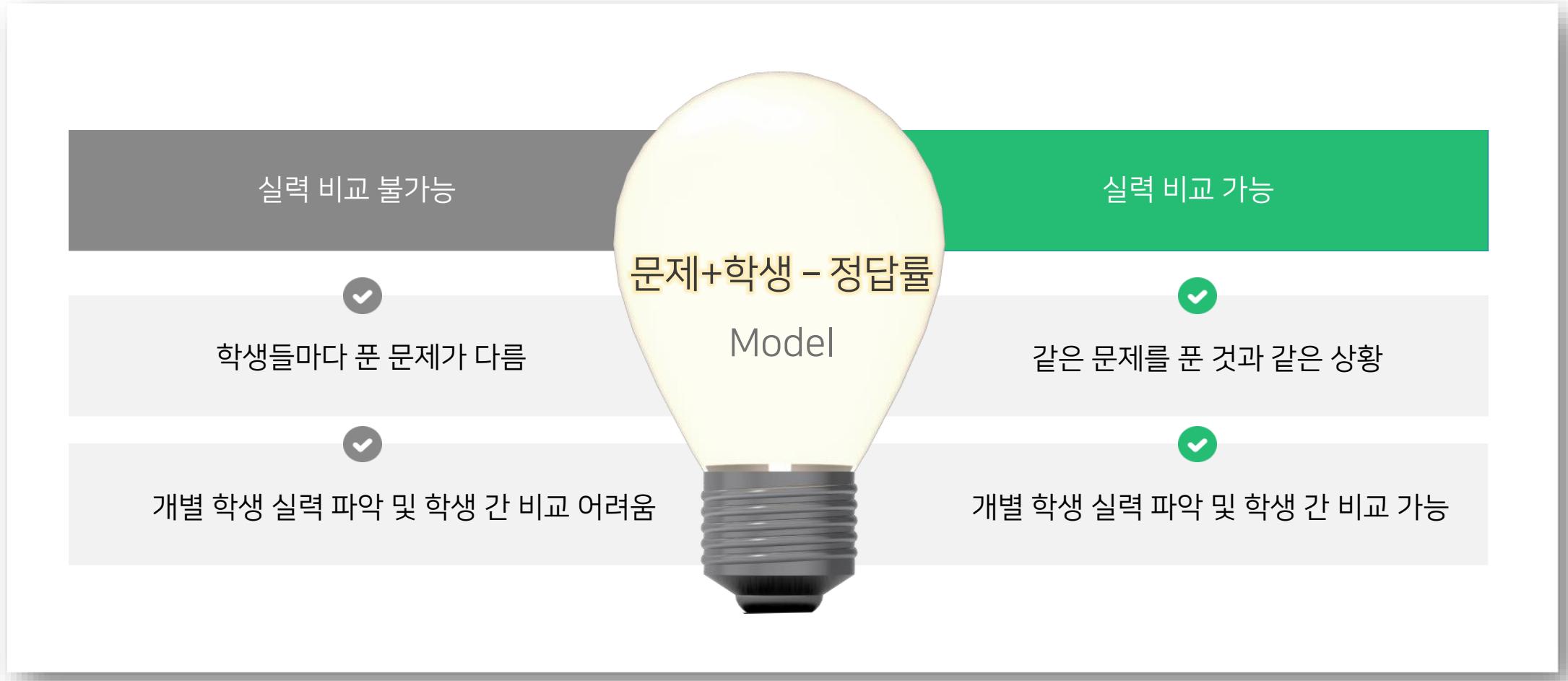
Conclusion

- ✓ 학생과 문제의 특성을 바탕으로 해당 학생이 특정 문제를 맞출 확률값을 도출하는 임베딩 모델을 구상
- ✓ 동일한 문제를 풀지 않아도 학생 간 실력 비교가 가능해 실력 지표로 활용 가능
- ✓ 개인 맞춤형 학습 및 지도가 용이해질 것이라고 기대

데이터(input) 설명

- 고3 학생 및 범위, worksheet 만 추출
- 6개의 feature 선택 - 커리큘럼 ID, 문제 정답률, 문제 유형, avg_cr(학생의 평균 정답률), o_cr(학생이 맞춘 문제들의 평균 정답률), x_cr(학생이 틀린 문제들의 평균 정답률)
- correct와 wrong의 비율 일정하게 맞춰줌
- row 수: 326555

문제 인식



문제 인식

실력 비교 불가능

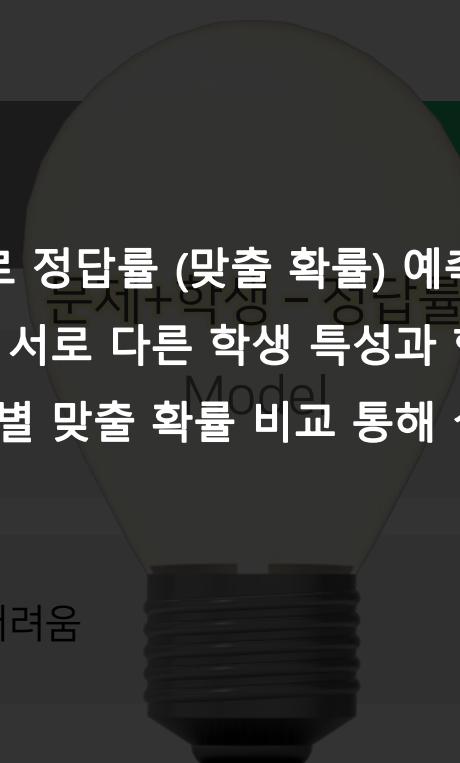
실력 비교 가능

문제, 학생을 기반으로 정답률 (맞출 확률) 예측하는 모델 학습된다면,
동일한 문제 특성을 서로 다른 학생 특성과 함께 해당 모델에 넣어

학생들마다 푼 문제마다를 학생별 맞출 확률 비교 통해 실력 비교 가능

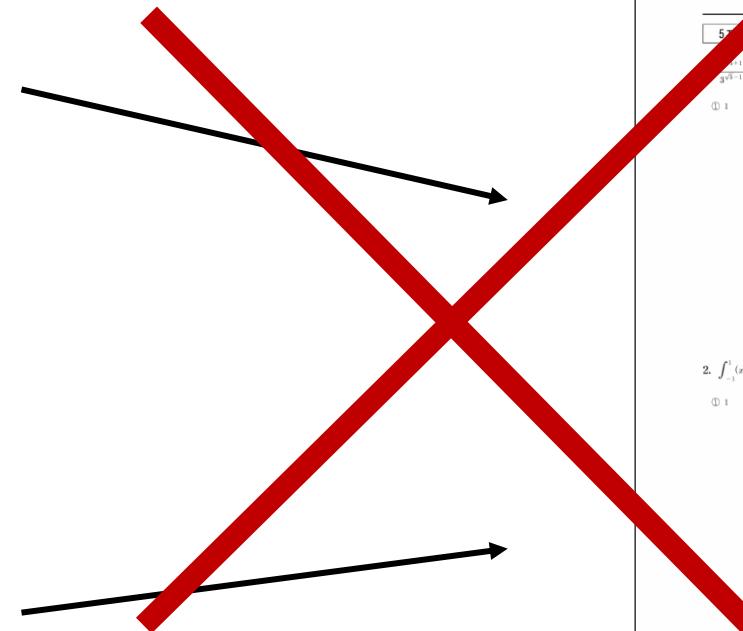
개별 학생 실력 파악 및 학생 간 비교 어려움

개별 학생 실력 파악 및 학생 간 비교 가능





모든 학생에게 같은 문제를 풀게 할 수 없는 상황



2022학년도 대학수학능력시험 예시문항 문제지
수학 영역 1
출수형

제 2 호지

57. $\frac{2x+1}{3^{x+1}-1}$ 의 값은? [2점]

① 1 ② $\sqrt{3}$ ③ 3 ④ $3\sqrt{3}$ ⑤ 9

3. 함수 $y=2^x$ 의 그래프를 y 축의 방향으로 m 만큼 평행이동한
그래프가 점 $(-1, 2)$ 를 지난 때, 상수 m 의 값은? [3점]

① $\frac{1}{2}$ ② 1 ③ $\frac{3}{2}$ ④ 2 ⑤ $\frac{5}{2}$

2. $\int_{-1}^1 (x^3 + a) dx = 4$ 일 때, 상수 a 의 값은? [2점]

① 1 ② 2 ③ 3 ④ 4 ⑤ 5

4. 함수 $y=f(x)$ 의 그래프가 그림과 같다.

$\lim_{x \rightarrow -1^-} f(x) - \lim_{x \rightarrow 1^+} f(x)$ 의 값은? [3점]

① -2 ② -1 ③ 0 ④ 1 ⑤ 2

1 20

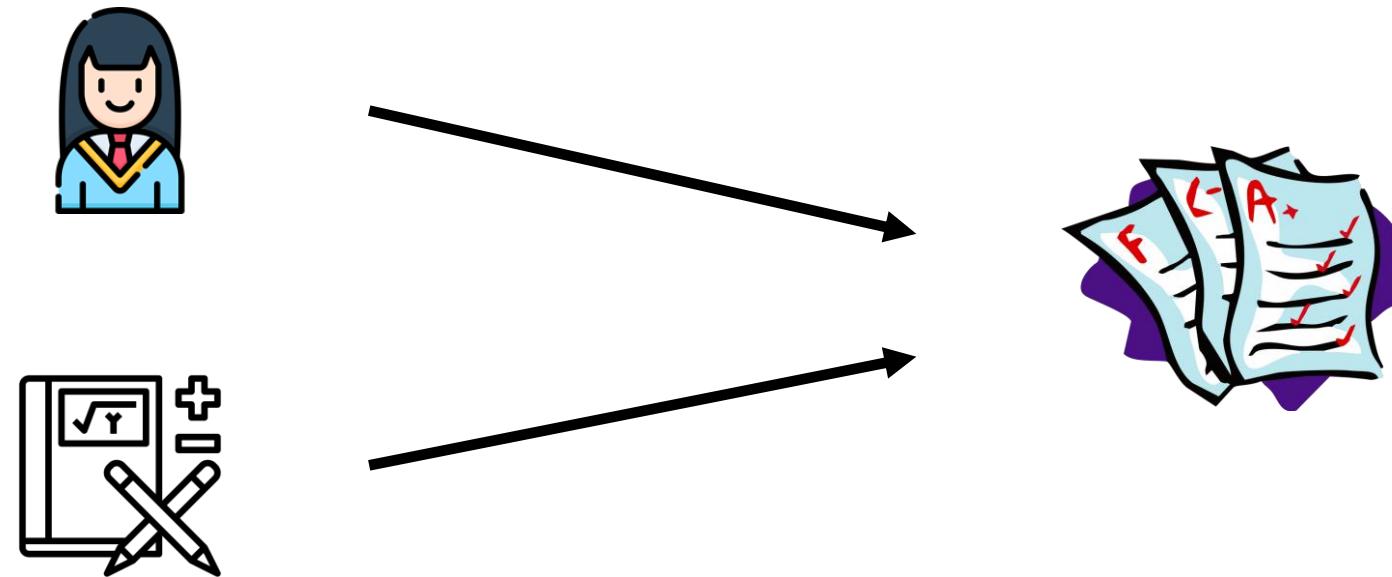
이 문제지에 관한 자작권은 한국교육과정평가원에 있습니다.

같은 문제를 푼 것과 같은 효과를 구현할 수 있는

“문제+학생 – 정답률 모델”을 만들면 어떨까?

목표

(학생, 문제) 바탕 정답률 예측 모델을 만들어보자.



Embedding 기반 정답률 예측 모델

Dataset

ABC student_id ↕	123 problem_id ↕	123 curriculum_id ↕	ABC name ↕	123 problem_cr ↕	ABC result ↕	123 level ↕	123 type ↕	⌚ update_datetime ↕
I100008	508,171	21	확률의 뜻	0.632	CORRECT	2	0	2021-01-04 11:40:50
I100008	399,004	21	확률의 뜻	-0.558	CORRECT	4	2	2021-01-04 11:40:50
I100008	508,131	17	확률의 곱셈정리	0.171	CORRECT	3	0	2021-01-04 11:40:50
I100008	399,631	15	사건의 독립과 종속	1.455	CORRECT	2	0	2021-01-04 11:40:50
I100008	399,781	7	연속확률변수	-2.633	WRONG	4	2	2021-01-04 11:40:50
I100008	496,123	6	이산확률변수	-1.111	WRONG	3	0	2021-01-04 11:40:50
I100008	400,337	10	이항분포	-0.833	CORRECT	3	0	2021-01-08 10:51:37

- 고3 학생 및 범위
- Worksheet only
- Row 수 : 3,625,555

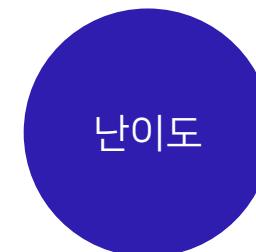
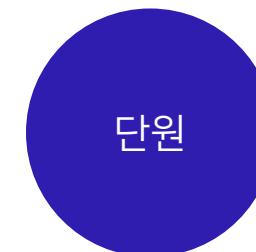
Custom Dataset

1. Feature 선택

1) 문제 feature

학생 ID	커리큘럼 ID	문제 정답률	문제 유형
I100008	33	67%	0
I100067	38	73%	2

- 문제의 주요 feature로 단원, 난이도, 유형을 선택
- 단원 : curriculum table의 id 값을 사용
- 난이도 : correct rate (정답률) 값을 사용
- 유형 : 객관식/주관식 등의 type 값을 사용



Custom Dataset

1. Feature 선택

2) 학생 feature

학생 ID	커리큘럼 ID	문제 정답률	문제 유형
I100008	33	67%	0
I100067	38	73%	2



... ... 학생의 특징을 나타내기에는 부족하지 않을까?

학생을 나타내는 데에 있어서 어떤 feature를 활용해야 할까에 대한 고민

Brainstorming

학생 지표

100문제 가정

$$\frac{1}{100} \left(\underbrace{\sum}_{\text{correct}} (100 - cr) + \underbrace{\sum}_{\text{wrong}} (-cr) \right)$$

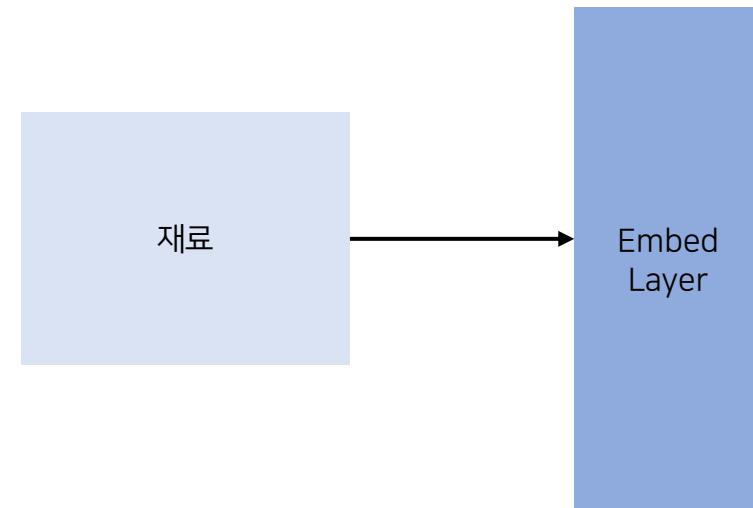
+ 가중치 부여

$$[(1 - 0.9) \times 0.1]^2$$

등등 고민해봤지만...

Brainstorming

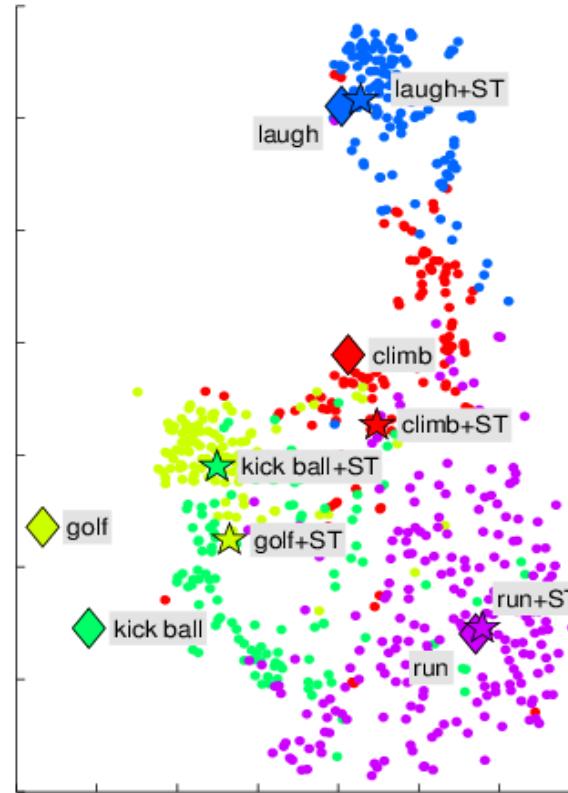
논리적인 수식보다…



(모델에게 자료를 주면 embedding layer를 거쳐서)

feature 값을 **학습** 해내도록 하는 것이
더욱 **효과적**일 것 같다!

Embedding



Effect

학생들이 embedding space로 적절히 투영되도록 학습

-> '비슷한 학생끼리 모일 것'이라는 가정



(문제) curriculum_id, type 또한 embedding

Custom Dataset

2. 학생 feature

커리큘럼 ID	문제 정답률	문제 유형	avg_cr	o_cr	x_cr
33	-0.221	0	63%	65%	57%
38	0.354	2	62%	64%	51%

새로 만든 3개의 학생 feature (재료)!

avg_cr: 학생의 평균 정답률
o_cr: 학생이 맞춘 문제들의 평균 정답률
x_cr: 학생이 틀린 문제들의 평균 정답률

왜 학생을 설명하기에 적절한 feature인가?

- 학생의 평균 정답률 (avg_cr) : 학생이 주로 푸는 문제들의 난이도 파악 가능
- 맞춘/틀린 문제들의 평균 정답률 (o_cr, x_cr) : 학생이 난이도가 높은 문제를 잘 푸는지 파악 가능

이 세 가지의 feature (재료)를 embedding layer를 거치게 하여 적절한 학생 feature를 학습하도록 함

Custom Dataset

3. 학습 활용 dataset 정리

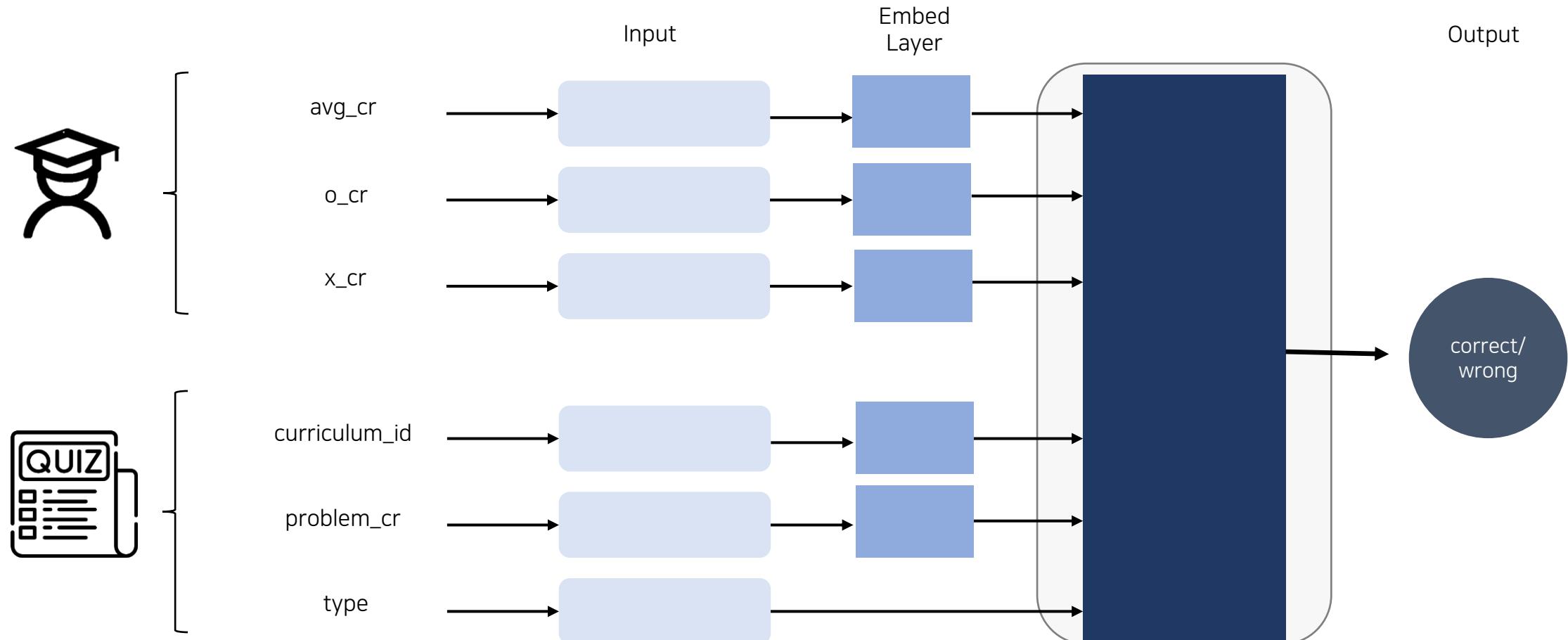
```
[ ] len(df_use[df_use['result'] == 'CORRECT'])  
465570  
  
[ ] len(df_use[df_use['result'] == 'WRONG'])  
493095
```

correct가 wrong보다 3배 가까이 많아 학습 상에 불균형



반반 정도로 비슷하게 맞춰줌

모델 구조

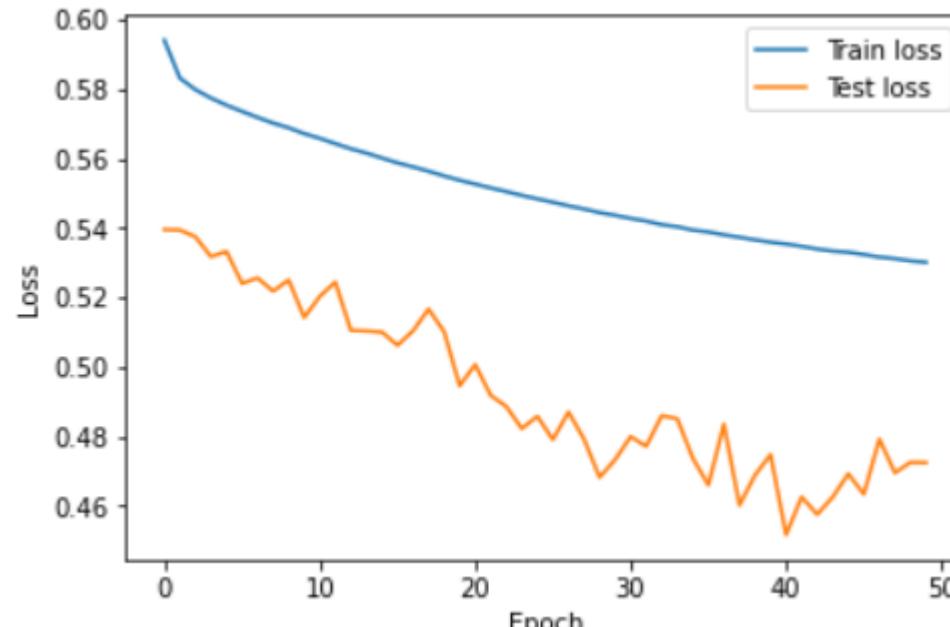


embedding layer를 각각 거치고 하나의 모델에서 학습이 이루어져 correct/wrong을 맞추도록 하는
binary classification 문제로 변환 (모델에서 출력되는 값은 softmax를 거쳐 나오는 맞출 확률)

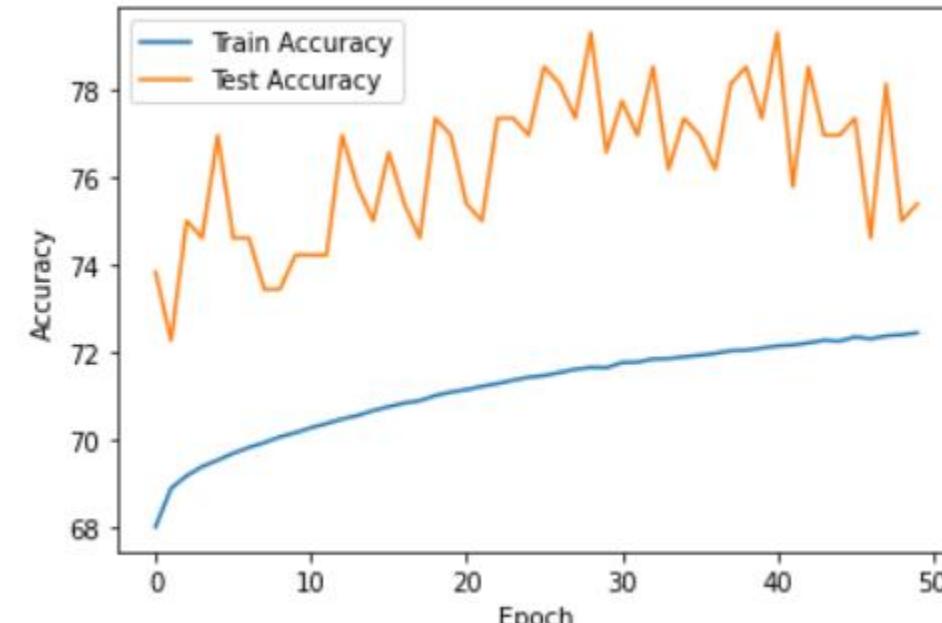
모델 구조

```
MyModel(  
    (embedding1): Embedding(100, 30)  
    (embedding2): Embedding(100, 30)  
    (embedding3): Embedding(100, 30)  
    (embedding4): Embedding(321, 50)  
    (embedding5): Embedding(3, 10)  
    (layers1): Sequential(  
        (0): Linear(in_features=180, out_features=256, bias=True)  
        (1): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)  
        (2): ReLU()  
    )  
    (layers2): Sequential(  
        (0): Linear(in_features=256, out_features=256, bias=True)  
        (1): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)  
    )  
    (layers3): Sequential(  
        (0): ReLU()  
        (1): Linear(in_features=256, out_features=32, bias=True)  
        (2): ReLU()  
        (3): Linear(in_features=32, out_features=2, bias=True)  
    )  
)
```

결과



loss

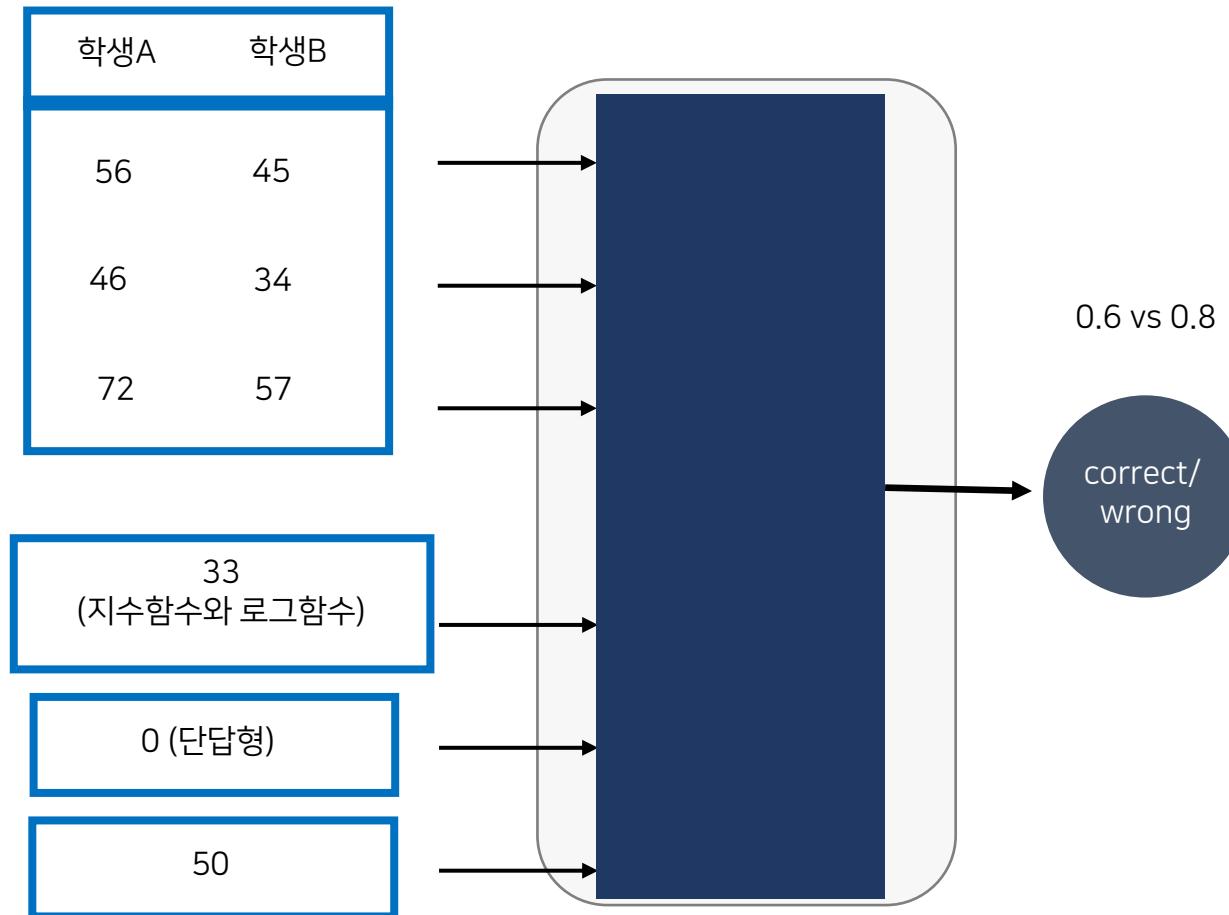


accuracy

50 epoch까지 갔을 때 embedding layer가 효과적으로 학습이 이루어졌음을 유추 가능

그렇다면 이 모델로 어떻게 **학생의 실력 지표**를 나타낼 것인가?

실력 지표



- 해당 단원, 유형, 난이도에서는 학생 A의 실력이 뛰어나다고 유추해볼 수 있음
- 단원, 유형, 난이도 다양하게 바꿔보며 간접적으로 실력 파악 가능
- 학생 A, B가 해당되는 문제 풀어본 적이 없음에도 정답률 예측 통해 실력 파악 가능!

Embedding 기반 정답률 예측 모델

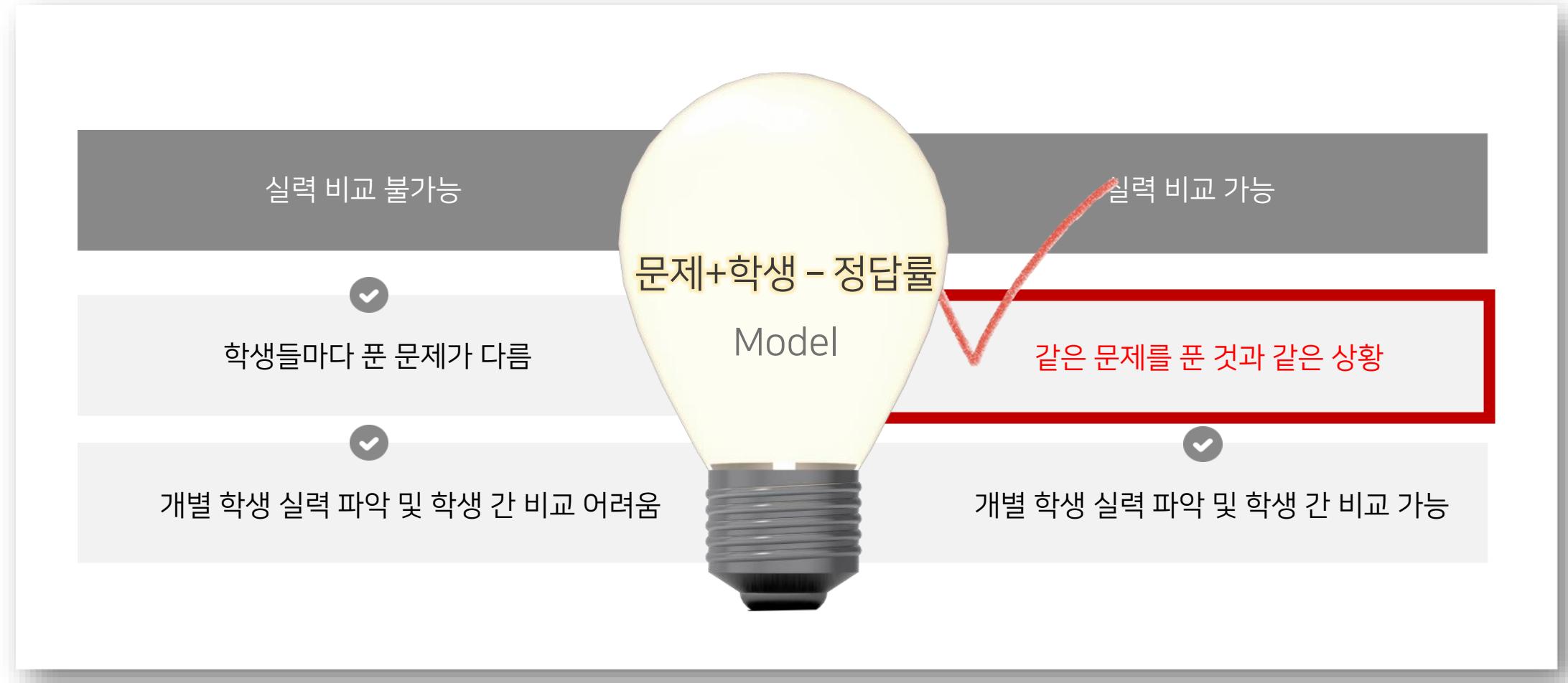
```
[ ] x1 = torch.Tensor([[70],[70]]).int().to(device)
x2 = torch.Tensor([[85],[90]]).int().to(device)
x3 = torch.Tensor([[65],[60]]).int().to(device)
x4 = torch.Tensor([[17],[17]]).int().to(device)
x5 = torch.Tensor([[0],[0]]).int().to(device)
x6 = torch.Tensor([[0.1], [0.1]]).to(device)
x6 = x6.expand(x6.shape[0], 30)

[ ] outputs = test_model(x1, x2, x3, x4, x5, x6)
outputs
tensor([[ 1.0892, -0.5989],
        [ 0.2084,  0.0356]], device='cuda:0', grad_fn=<AddmmBackward0>

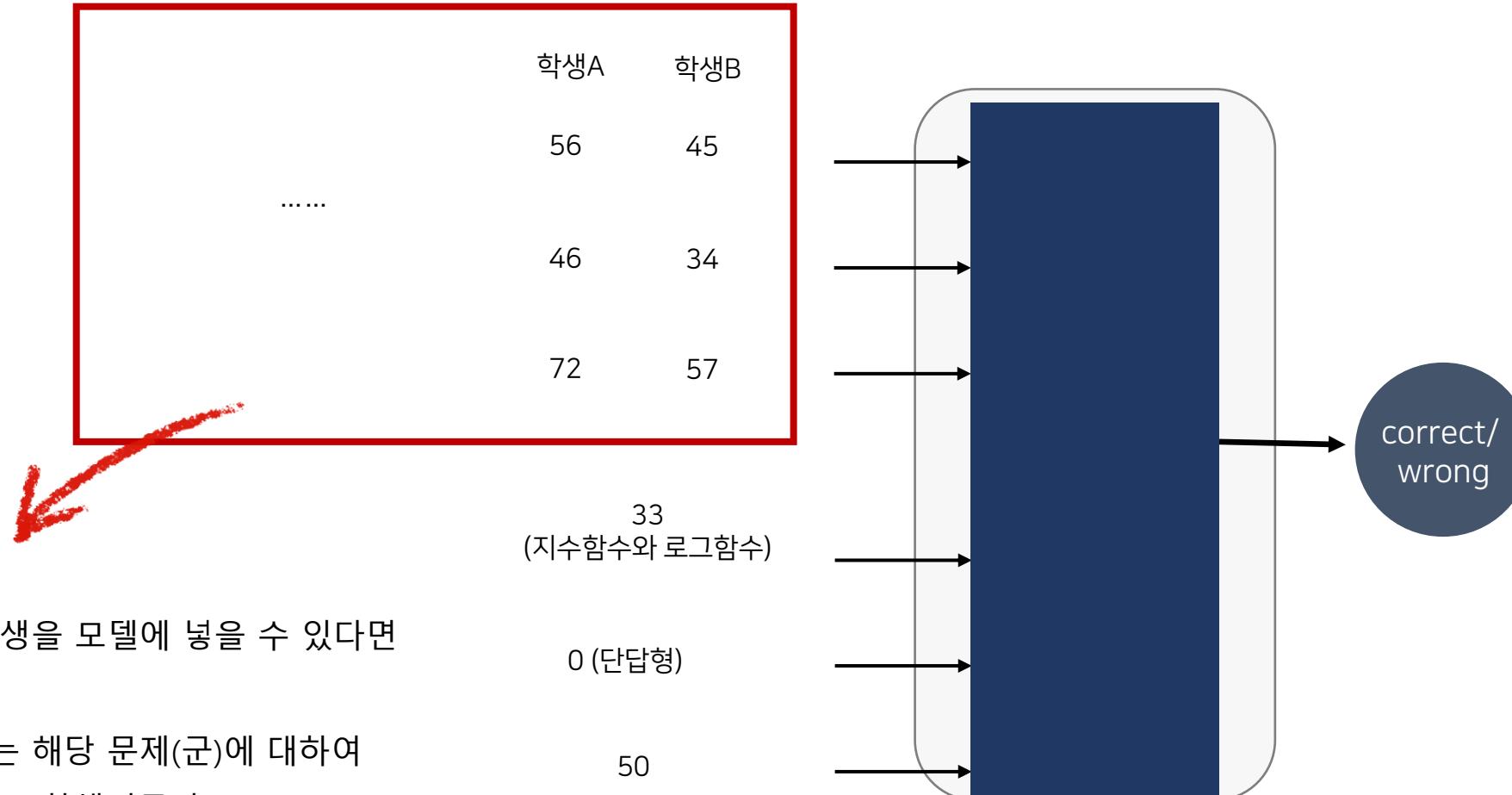
[ ] torch.nn.functional.softmax(outputs, dim=1)
tensor([[0.8440, 0.1560],
        [0.5431, 0.4569]], device='cuda:0', grad_fn=<SoftmaxBackward0>)
```

Softmax를 거쳐 확률값으로 output을 도출함

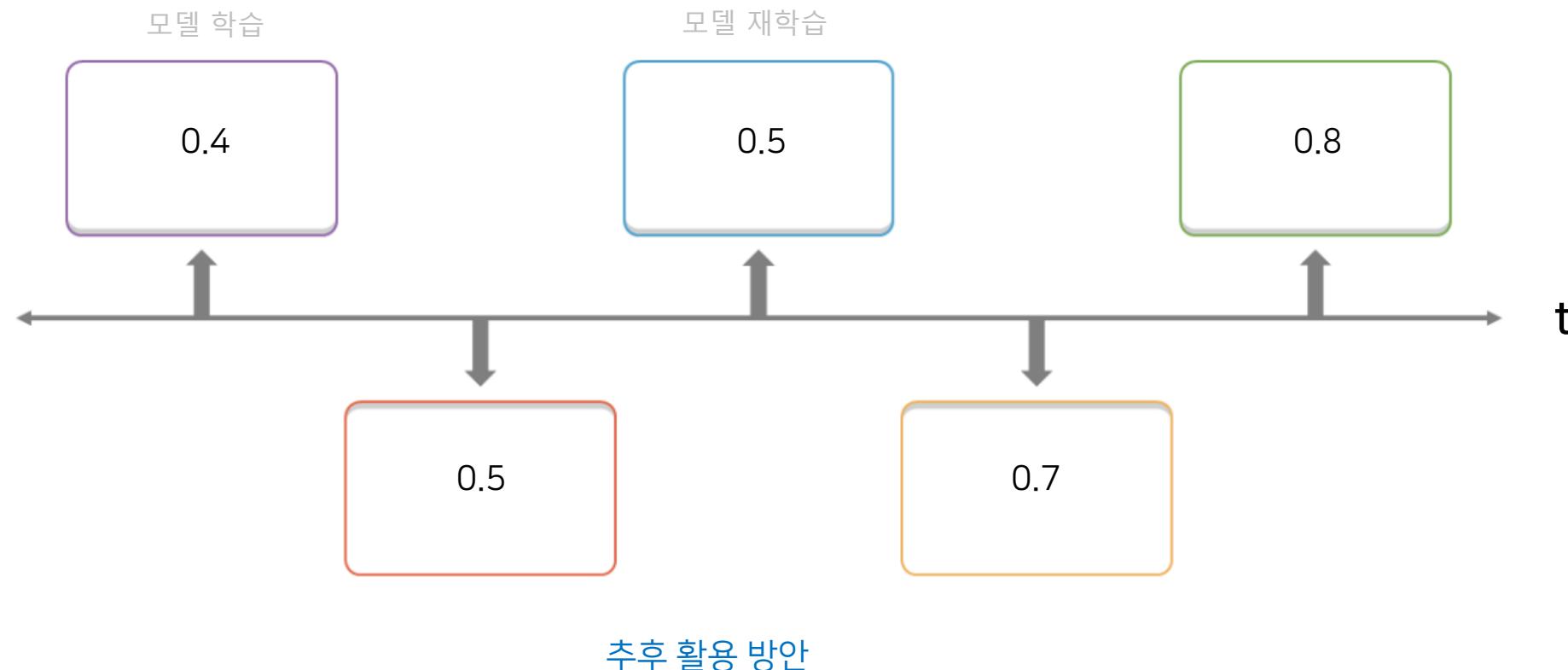
문제 인식 단계



실력 지표



실력 지표



특정 시간 구간마다 연속적인 모델 재학습이 가능하다면?

위 그림과 같이 정답률이 예측될 경우…

→ 실력 상승한다고 판단 가능

기대 효과 및 예상 적용 방안

기대 효과

- 학생이 접하지 못한 문제에 대한 정답률 예측
- 같은 문제를 푼 것과 같은 상황... 학생 간 비교 가능

-> 개별 학생 실력 지표 만들 수 있다

적용 방안

개인 맞춤형 학습 또는 지도

- 개인 맞춤형 학습지 제작 및 문제 구성이 용이해짐
- 학생 자신이 부족한 단원/유형/난이도 등을 파악함으로써 학습 방향 설정에 도움

한계점



학생을 대표하는 feature의 실효성



학생을 특징 짓기 위한 데이터 부족 (ex. 개인 정보)

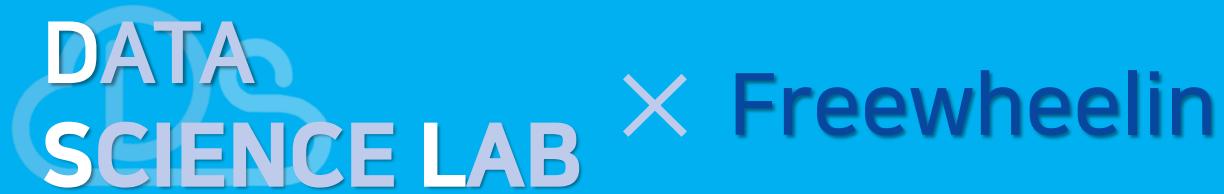


문제를 특징 짓기 위한 데이터 부족 (ex. 문제 자연어 데이터 활용 불가능)



모델 고도화

Q & A



× Freewheelin