서울대학교 자연과학대학 통계학과 학사졸업논문

레이팅 시스템을 이용한 초등수학 학습 솔루션 개발 및 로지스틱 회귀분석을 통한 최적화

이 논문을 이학사 학위논문으로 제출함

2023년 08월

논문심사 대상자 소속 : 통계학과

학번 : 2017-10052

성명 : 장규찬 (인)

<학위논문의 윤리 서약>

1. 본 학위논문은 본인이 직접 연구하여 작성한 것이다.
2. 정확한 출처 제시 없이 다른 사람의 글이나 생각을 가져오지 않았다.
3. 인용한 문헌의 내용이나 자료(도표나 데이터)를 조작(위조 혹은 변조)하지 않았다.

본 학위논문은 위의 항목들을 준수하여 작성한 것임을 확인합니다.

2023년 6월 1일 작성자 : _____장규찬_____ (서명)

목 차

1. 서론

1.1 연구의 배경과 목적

(1) 연구의 배경

“지금 엄마 말 안들으면 ~ 커서 눈물 쏙 빠질 텐데~
공부 안 하니 ~ 방법 없을까? 그래 빨간 빨간 빨간펜~~“

개그우먼 박미선 씨가 광고했던 빨간펜 광고 카피 중 일부이다. 초등학생 엄마들의 고민과 걱정을 일부 과장하여 재밌게 표현한 광고로, 아직까지 잘 만들어진 광고 중 하나로 회자된다. 겨우 광고 카피 하나지만, 한국의 수많은 학부모, 사교육 시장에 지대한 영향을 끼쳤다. 구몬, 빨간펜 등 기존 연산 학습지는 문제 개수 위주의 학습법을 채택하여 아이들에게 연산 반복 학습을 유도하였고, 기존 문제집 시장에서는 각 출판사마다 모두 개념부터 시작하여 유형, 응용, 심화, 사고력까지 문제집을 분절화하여 마치 한 학기 내 모든 문제집을 다 풀어야 되는 것처럼 만들었다.

2010년대 중반 이후로 초등수학에도 변화의 바람이 불어오기 시작했다. 기존 문제집 회사들은 에듀테크(EdTech) 회사를 창립하여 오프라인 시장에서 온라인 시장으로의 진출을 희망하고 있다. 개념 학습뿐 아니라 문제까지 모두 온라인으로 전환하여 기존 ‘인강’ 수준에 그쳤던 온라인 학습 시장의 판도가 흔들리고 있다. 급기야 2020년대 초반에는 자본력을 앞세워 수많은 매스미디어 광고판을 장악하고 있다.

기존 연산 학습지의 광고부터 지금 광고를 쭉 돌이켜보면 최근 광고들에만 나타나는 흥미로운 지점이 있다. 현재 초등 에듀테크 시장을 선도하고 있는 다수의 기업들은 자신들의 소개 페이지에 ‘개인화’, ‘AI’, ‘초격차 학습’, ‘AI학습’, ‘인공지능 학습’, ‘학습 추천’, ‘실력 분석’ 등의 키워드로 자신들의 서비스를 소개하고 있다. 최근 머신러닝, 딥러닝을 통한 수많은 기술 진보가 교육 시장에서도 발생하고 있는 것을 체감할 수 있다.

그럼에도 불구, 현재 초등 학습 서비스의 ‘개인화 학습’은 아기의 걸음마 단계에 그치고 있다. 연구 진행 중 레퍼런스로 참고하였던 모든 메이저 학습 서비스들은 사실상 ‘IF-ELSE’ 수준의 출제 방식을 채택하고 있다. 예를 들면 ‘기본 학습’을 2문제 풀고, 2문제 다 맞혔으면 ‘오늘의 심화 학습’, 1문제 이상 틀렸으면 ‘복습해요’로 넘어가는 식이다. 이러한 학습을 ‘AI 학습’이라 부르는 것은 중립적인 시선에서도 상당히 어폐가 있다.

이같은 상황이 발생한 원인은 AI를 서비스화시키지 쉽지 않다는 점에 있다. 수많은 학습 데이터를 수집하는 비용은 둘째로 하더라도, 유저(아이)가 실시간으로 활용하는 서비스에서 AI 연산의 무게를 버티기 매우 어렵다. 최근 공개된 트위터 피드 추천 로직을 보아도 이같은 문제를 확인할 수 있다. 트위터는 피드를 추천하기 위한 수십 개의 기준에 점수를 부여하여 점수의 가중 합으로 정렬하는 로직을 차용하고 있다. 이것은 점수 자체를 계산하는 데 들어가는 계산 리소스는 견딜 수 있지만, 클라이언트 단에서 서버에 요청할 때의 계산은 최소화해야 하기 때문이다.

서버 단에서의 무거운 연산과 클라이언트 단에서의 가벼운 계산이 수월하게 동작하려면 대개 서버 단에서 완성적으로 구축된 데이터 파이프라인이 동반된다. 구축 비용 대비 효용을 따졌을 때 수지타산이 맞기 어려운 구조인 것이다. BM(Business Model)상 해당 부분이 추가적인 유저 이익을 발생시켜야 하는데, 대다수의 AI 서비스의 효용은 마케팅에서 발생한다. 대다수의 유저는 AI 서비스에 관심이 있기보다, AI를 사용함으로써 얻는 이익과 만족감, 혹은 더 적은 학습으로 더 큰 성장을 이끌 수 있다는 점에 주목한다. 유저의 학습 곡선은 서비스에서뿐 아니라 학습하는 유저 스스로도 측정이 어렵고, 심지어는 유저가 성장을 느끼기 전 시점에 대체로 서비스에 락인(lock-in)되어 유저에게 추가적인 과금을 만들어내기 제약이 크다.

매쓰팡(이하 서비스)에서는 이러한 초등 에듀테크 산업 구조 하에서 저렴한 비용과 함께 더 적은 데이터로 학습 추천 알고리즘 제작을 기획하게 되었다. 이에 일반적인 딥러닝(deep-learning) 모형보다는 훨씬 데이터가 적게 들어가면서 높은 수준의 모델 파워를 얻을 수 있는 로지스틱 회귀분석을 떠올리게 되었다. 로지스틱 회귀분석은 이진-모형(binary)로써 문제 정오답 모형에 가장 근접할뿐 아니라, 딥러닝 측면에서 pre-trained 모형에 가깝다고 할 수 있다. 하지만 이 접근 역시도 한계가 명확한데, 클라이언트 단에서 로지스틱 회귀분석 정도 크기의 모형을 실시간으로 적합하기에는 적절하지 않기 때문이다.

엘로 레이팅(이하 Elo Rating) 모형은 멀티 플레이 기반 실력 측정 모형이다. 서로의 사전 측정된 실력을 기반으로 대결 결과를 반영하여 각 유저의 사후 측정 실력을 구하게 된다. 나는 이 모형을 이용해보고자 하였다. 엘로 레이팅 모형은 추후 밝히겠지만 그 근간을 이진 로짓 모형(교육학에서의 IRT와 동일)에 두고 있다. Rasch 모형 하에서의 초기 문제 난이도는 사람이 아무리 정교하게 잡더라도 한계가 있다. 나는 엘로 레이팅 모형을 사용해 유저 레이팅과 문제 레이팅을 동시에 수렴시키는 방식을 착안하였다. 이미 해당 내용에 대한 사전 연구가 존재하였으며, 사전 연구를 기반으로 서비스 학습 DB를 활용해 '교과 학습 실력 측정 모형(이하 학습 실력 측정 모형)'을 우선 제작하였다. 이를 토대로 '교과 문항 출제 솔루션'을 개발하게 되었다.

(2) 연구의 목적

가. 학습 실력 측정 모형 개발

엘로 레이팅 시스템을 활용해 서비스를 사용하는 유저의 문제 제출을 기반으로 학습 실력을 측정하는 모형을 개발한다. 초등 71개 대단원에 대하여 유저_대단원 DB를 구축하는 것이 목표이다.

나. 교과 문항 출제 솔루션 개발

위의 학습 실력 측정 모형을 활용해 서비스 내에서 문제를 출제하는 솔루션을 개발한다. 유저_대단원 DB를 업데이트하면서 유저에게 적절한 문제를 출제하는 구조를 고민하였다. 매일 적절한 난이도의 15문항 학습을 담당하는 미션 기능을 개발, 배포하는 것이 목적이다.

다. 미션 기능 지표 추적

미션 기능 기획을 진행하며 세웠던 유저 가설은 아래와 같다.

- 유저들은 원하는 문제 난이도를 만나지 못하여 이탈하는 경우가 존재한다. 너무 어렵거나 너무 쉬운 문제를 주는 일(특히 너무 어려운 문제)은 지양해야 한다.
- 적당한 난이도의 문제를 계속적으로 만난다면 유저는 끝까지 문제를 풀이할 것이다.

이 가설을 기반으로 서비스에서는 미션 기능 개발을 진행하며 크게 두 가지 지표를 겨냥하여 이를 개선하고자 하였다.

- 미션 페이지 방문 기준 세 번째 미션 완료(15문항 모두 풀이 완료) 비율
- 첫 번째 미션 완료(5문항 풀이 완료) 기준 세 번째 미션 완료(15문항 풀이 완료) 비율

해당 지표들에 대하여 배포 전후 측정한 DB, Amplitude 데이터를 활용하여 검증하는 것을 목표로 두었다.

1.2 연구 절차

가. 문제 추천 솔루션이 제공해야 할 가치 탐색

- 초등수학 시장에서 문제 추천 솔루션이 제공해야 할 가치 탐색 (1.1)
- 초등 에듀테크 산업 구조 하에서 저렴한 비용, 적은 데이터로 학습 추천 알고리즘 제작 기획 (1.1)

나. 탐색적 자료 분석

- 탐색적 자료분석 진행 (2.1)
- 모든 교과 문항에 대하여 5점 척도 난이도 부여 (2.1)

다. 학습 실력 측정 모형 개발 절차

- 멀티플레이 기반 엘로 레이팅 시스템이 적합한지 확인 (2.2)
- 레이팅 설계의 결과가 IRT를 기반으로 한 Gradient Descent 방법과 동치임을 증명 (2.2)
- 학습 실력 측정 모형 구조도 작성 (2.3)
- Hyperparameter Tuning (2.3)
- 난이도별 초기 레이팅 매기는 방식 고안 (2.2 - 2.3)
- 학습 실력 측정 모형 검증 (3.1)

라. 교과 문항 출제 솔루션 개발 절차

- 문제를 찾는 과정 및 구조도 제작 (2.4)
- 유저에게 제시되는 미션 15문항 출제 / 제출 로직 (2.4)
- 대단원 내 학습량 및 학습 방식 구성 (2.4)
- 유저 컨디션 레이팅 도입 (2.4)

마. 미션 기능 지표 추적

- 미션 페이지 방문 기준 세 번째 미션 완료(15문항 모두 풀이 완료) 비율 추적
- 첫 번째 미션 완료(5문항 풀이 완료) 기준 세 번째 미션 완료(15문항 풀이 완료) 비율 추적
- 서비스 지표 추적 결과 해석

1.3. 이론적 배경

(1) 초등수학의 연산 / 교과

연산이란 교과 수학을 하기 위한 기본에 해당하는 수학 범위이다. 초등 수학 교과는 모든 초등 연산을 포함한다. 저학년 교과 수학의 경우 연산이 주를 이루며, 수학의 기본이기 때문에 꾸준한 반복이 중요하다. 아이들은 주로 눈높이, 빨간펜 등 학습지을 통하거나 엄마와 연산 문제집을 반복해서 풀게 된다.

교과 수학이란 초등 교과서에 제시된 수학이다. 교과서를 기본으로 초등 수학 전 범위를 교과라고 부른다. 연산 문제집, 사고력 문제집을 제외한 대다수의 일반적인 초등 수학 문제집을 전부 교과 문제집이라 부르며, 수록된 문항이 교과 문항이다. 이 연구의 논의 범위는 교과 문항(이하 문항)에 한정하도록 한다. 이때의 교과 문항이란 통상적으로 연산 학습지 등에서 등장하는 '연산 반복 학습'은 제외한다.[1)]

(2) 초등수학 EdTech 주요 개인화 솔루션 소개

현재 초등수학 EdTech 시장에서 상위의 점유율을 차지하고 있는 두 교육 회사의 개인화 솔루션에 대해서 소개한다.

가. 천재교과서 밀크T

크게 두 가지 서비스가 존재한다. 단원평가 서비스의 경우 적당한 분포를 그려 평가 그룹 내 내 위치를 분석해준다. 문항의 5단계 척도를 보유하여 각 수준별 몇 개 중 몇 개를 맞추었나를 기준으로 백분율로 정답률을 알려준다. 이외에도 풀이 시간 분석, 유형별 정답률 등을 제공하는 등 정보량이 다양하다.

문제 출제 방식이 매우 특이한데, 4가지 유형을 기반으로 난이도 1 문제를 우선 준다. (1번, 2번, 3번, 4번) 그 다음 5번 문항부터는 같은 유형을 맞혔으면 난이도 2 문항을, 아니라면 난이도 1 문항을 출제한다. 총 20문항을 제시하므로 16번까지 하나도 틀리지 않는다면 난이도 5 문항까지 풀이할 수 있다. 낮은 수준의 개인화 솔루션을 구현하기 위해 초기 난이도 분류 작업, 정오답에 따른 반응이 주어지는 기초적인 수준의 출제 솔루션을 제공한다.

---

1) 초등 방학 탐구 생활 시리즈, 〈연산, 교과 수학, 사고력 수학, 다른 건가요?〉. 네이버 포스트. 검색 일자. 2023.05.31.
https://m.post.naver.com/viewer/postView.naver?volumeNo=33357390&memberNo=904209

나. 웅진스마트올

AI 학습량 자동판단, AI 진단 테스트, AI 문제풀이 분석, AI 오답 원인 분석 등을 제시하는데, 모두 if-else 기반의 솔루션이다. AI 유형학습의 경우에 대표문항, 유사문항, UP 문항, DOWN 문항이 존재하는데, 실제 활용을 통해 확인한 if-else 작동 방식은 아래와 같다.

Index	경로	결과
1	대표 정답 -> UP 정답	CLEAR
2	대표 정답 -> UP 오답	GOOD
3	대표 오답 -> 유사 정답 -> UP 정답	CLEAR
4	대표 오답 -> 유사 오답 -> DOWN 정답	GOOD
5	대표 오답 -> 유사 오답 -> DOWN 오답	취약

표 1 웅진 스마트올 AI 유형학습 작동 분석

나머지 다른 AI 서비스를 활용했을 때에도 맞힌 개수를 기반으로 다음 액션을 유도하거나, 맞히는 데까지 걸린 시간을 기준으로 if-else 방식의 낮은 수준의 구현 결과를 보여준다. 이는 분석자의 역량이 높아서 발견한 것이 아닌, 서비스 내에서 충분히 파악 가능한 형태로 제작되어 애초에 대다수 유저로 하여금 AI 서비스가 아니라는 사실을 충분히 인지하는 상황에서 사용된다.

(3) 문항 반응 이론(Item Response Theory)

문항반응이론(item response theory)는 고전검사이론처럼 검사 총점에 의해 검사나 문항이 분석되는 이론이 아니라 문항 하나 하나의 독특한 특성을 지닌 고유한 문항특성곡선(Item Characteristic Curve)에 의해 분석된다는 이론이다.

문항 반응 이론이란 Binet과 Simon(1916)이 지능을 측정하기 위해 문항을 제작하고, 각 문항이 어느 나이의 아동들에게 적합한 문항인지 알아보기 위해 문항 특성 곡선을 그리게 된 것에서 유래한다.

덴마크의 수학자 Rasch는 확률이론에 의해 문항 반응 이론을 전개하였으며, 난이도에 순수하게 depend하는 모형을 제시하였다. 해당 곡선은 로지스틱 커브이며, 문항의 모수치를 추정하기 위해 최대우도추정법(MLE)을 활용한다. 이를 1PL Rasch 모형, 혹은 Rasch 모형이라 한다.

정답률을 추정하는 모형에 따라 주로 3개의 모형으로 나뉘며, 아래와 같다. 계수 1.7

은 고전 모형과의 연동을 위해 있는 상수로, 큰 의미를 가지지는 않는다.[2)]

모형	수식	모수
1모수 로지스틱 모형 (Rasch)	$P(\theta)=\frac{1}{1+\exp(-1.7(\theta-b))}$	θ : 능력 수준 b : 문항 난이도
2모수 로지스틱 모형	$P(\theta)=\frac{1}{1+\exp(-1.7a(\theta-b))}$	a : 문항 변별도
3모수 로지스틱 모형	$P(\theta)=c+(1-c)\times\frac{1}{1+\exp(-1.7a(\theta-b))}$	c : 문항 추측도

표 2 1PL, 2PL, 3PL 문항 반응 모형

(4) Elo Rating 시스템

엘로 레이팅 시스템이란 체스 등 2명제 게임에서 주로 사용하는 실력 측정, 평가 산출 방식이다. 엘로 레이팅 시스템을 기반으로 하는 두 대국자 A, B의 레이팅을 R_A, R_B라 하면 승률 E_A, E_B는 아래와 같다. 숫자 400의 경우 변경될 수 있다.

$$E_A=\frac{1}{1+10^{\frac{R_B-R_A}{400}}},\quad E_B=\frac{1}{1+10^{\frac{R_A-R_B}{400}}}$$

A가 승리한 경우 새로운 레이팅 R_A', R_B'은 아래와 같이 조정된다. 잃고 얻는 레이팅은 동일하며, 레이팅을 매기는 집단이 일정하다는 가정 하 제로섬 게임이 유지된다.

$R_A' = R_A + K(1-E_A)$

$R_B' = R_B + K(-E_B) = R_B - K(1-E_A)$

K는 상수로써, 레이팅의 증감 속도, 종국에는 레이팅이 실력으로 얼마나 빠르게 수렴할지를 결정한다. Gradient Descent Method로 비유하자면 learning rate에 해당한다. 다만, K 값이 너무 크면 빠르게 수렴하는 대신 한 번의 승패에 지나치게 영향을 받게 된다. 주로 K 값은 16 내지 32, 혹은 20, 40 등을 사용한다. 한 번의 승패에 지나치게 영향을 받지 않으면서 빠르게 수렴시키기 위해 Learning Rate Scheduler처럼 첫 10판, 20판 등을 K 값을 inflate시키는 방식 역시 많이 고려된다. 2014년부터

2) 김석우, 교육평가이론(문항반응이론) 강의자료, http://www.kocw.net/home/cview.do?mty=p&kemId=1266989. 부산대학교. 2017년.

국제체스연맹 FIDE는 레이팅이 2300점 미만이면서 30판 이하 혹은 18세 이하(빠른 실력 성장) 모든 플레이어에게 K=40 (일반적인 경우 대비 2배)를 적용한다.

2. 본론

2.1 탐색적 자료 분석

(1) 학습 데이터 저장 방식

서비스 내 한 아이가 문제를 푸는 상황은 크게 세 가지로 나누어볼 수 있다. 연산 자유학습, 교과 자유학습, 미션 학습이다. 이 중 '매일 학습', '진도 학습'에 해당하는 서비스 기능은 미션 학습이다. 미션 학습을 통해 아이들은 학부모가 선택한 진도에 맞추어 주간 3회 이상 학습하고, 학교 진도 속도 이상으로 교과 학습을 수행한다.

학습 방식	기능 목적	수집 데이터
연산 자유학습	15문제 내외가 담긴 스테이지 기반 형태의 연산 문제 풀이를 자유롭게 진행	2021년도 6월 ~ 2022년도 7월까지의 서비스 문제 풀이 데이터 기반. 문제 ID, 제출한 답, 정오답 여부, 풀이 시간, 문제 풀이 시각 등
교과 자유학습	4 ~ 8문제 내외가 담긴 스테이지 기반 형태의 교과 문제 풀이를 자유롭게 진행	
미션 학습	[연구 진행 전] 존재하는 교과 문항 중 일부를 적절하게 선별하여 대단원 순서대로 아이에게 잘라서 '오늘의 학습' 형태로 제공 **[출제 솔루션 제작]** 아이의 실력을 측정하여 아이에게 적절한 난이도의 교과 문항을 공급하여 아이의 학습 습관을 기른다.	

표 3 서비스 지원 학습 방식, 기능 목적 및 수집 데이터

이번 연구에서 나는 미션 학습을 개선하기 위하여 우선 학습 실력 측정 모형을 개발하고, 그를 통해 교과 문항 출제 솔루션을 서비스 가능한 형태로 제작하였다. 이 제품을 배포한 이후의 서비스 데이터를 통해 연구의 목적 달성 여부를 검증하도록 하겠다.

문제 DB에서 보유한 주요 열(column)은 아래와 같다.

Column	정의
id	문제를 구분하는 primary key이다.
created_at	DB에 생성된 시각을 밀리초(ms) 단위의 timestamp로 저장한다.
pattern_code_id	문제의 패턴을 구분하는 인자이다. RDB에서 문제 DB의 부모 테이블의 primary key이다.
question	문제 정보를 LaTeX로 저장한다.
solution	문제 정답 정보를 배열의 형태로 저장한다.

difficulty	문제 난이도 정보를 정수(int)로 보유한다. 이 스키마는 후술하겠지만 특정 기준으로 사람이 만든 데이터이다.

표 4 서비스 문제 DB 주요 열 및 그 정의

(2) 교과 문항 난이도 라벨링 및 분석

서비스 커리큘럼 팀에서 교과 문항을 제작할 때에 각 문항에 대해 5점 척도의 난이도를 라벨링하였다. 이는 통상적으로 학습 문제집과 학부모, 사교육 시장 내에서 통용되는 난이도 기준을 기반으로 한 것이다. 이 데이터를 아이의 학습 실력 측정 모형의 문제 초기 난이도로 사용하였다.[3)]

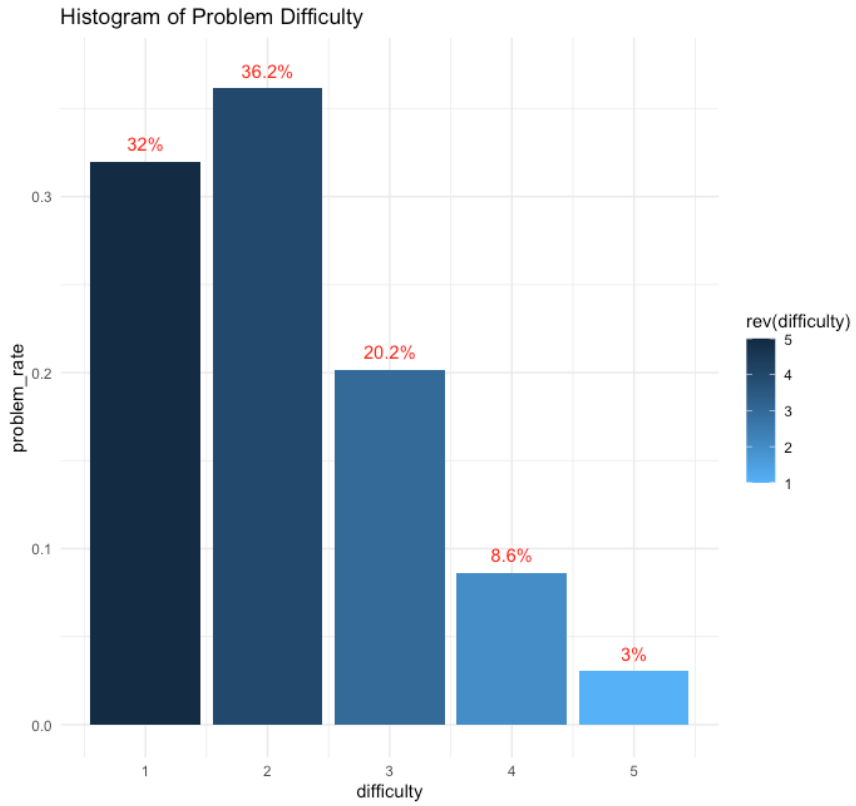
문제 제작 과정에서 모든 교과 문항에 대하여 5점 척도의 난이도를 라벨링하였다. 5점 척도의 난이도 라벨의 기준은 아래와 같다.

난이도 점수	분류 기준	비슷한 난이도의 문항
1	기초 개념 확인 문항	쎈 Step A
2	기본 유형 문항	쎈 Step B 난이도 하 / 중
3	한 가지 이상 개념을 응용한 문항	쎈 Step B 난이도 상 디딤돌 응용
4	개념을 복합적으로 응용한 문항	쎈 Step C 디딤돌 최상위S
5	심화 문항, 사고력 문항	디딤돌 최상위, 최상위쎈

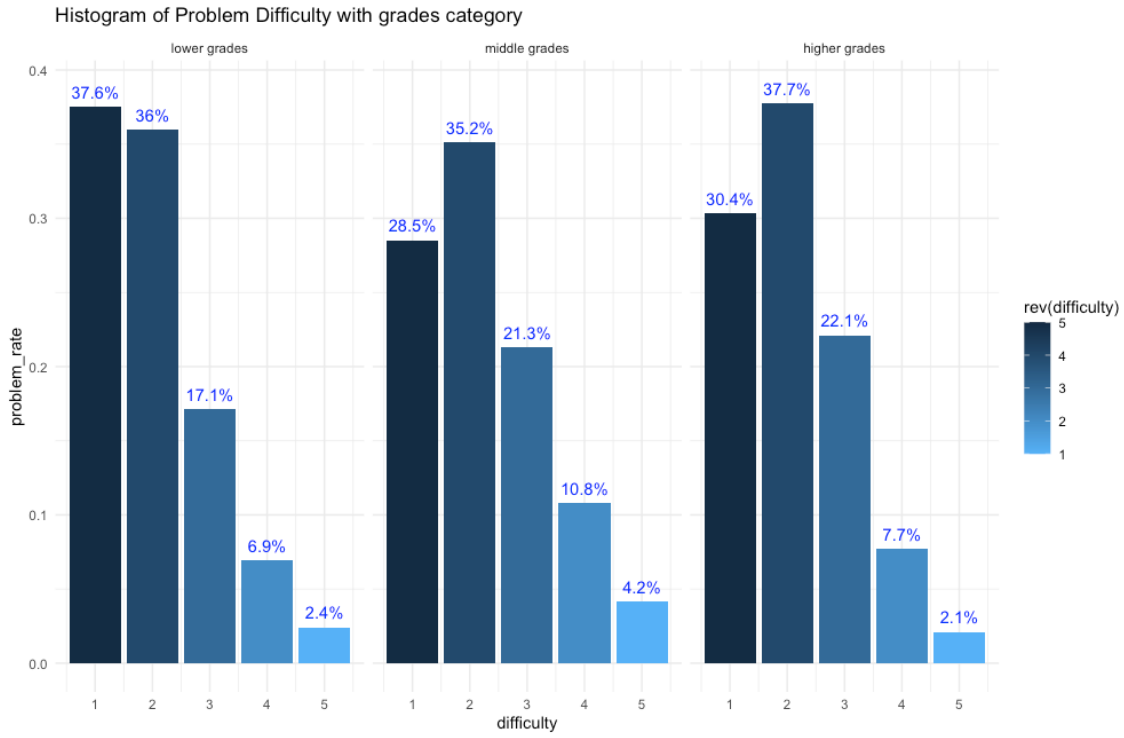
표 5 교과 문항의 난이도 점수와 분류 기준 표

해당 난이도 분류 기준 표를 활용해 보유한 교과 전 문항에 난이도 태그를 추가하였으며, 교과 문항 난이도에 대한 기초적인 정보는 아래와 같다.

3) 다만, 등간 척도임이 보장되지 않기 때문에, 후술한 분석에서는 1 ~ 5 점수를 항상 factor로 취급하였다. 점수를 rating으로 scoring하는 방식에 대하여는 아래에서 자세히 다룰 것이다.



그래프 1 교과 문항 난이도별 문항 비율 그래프

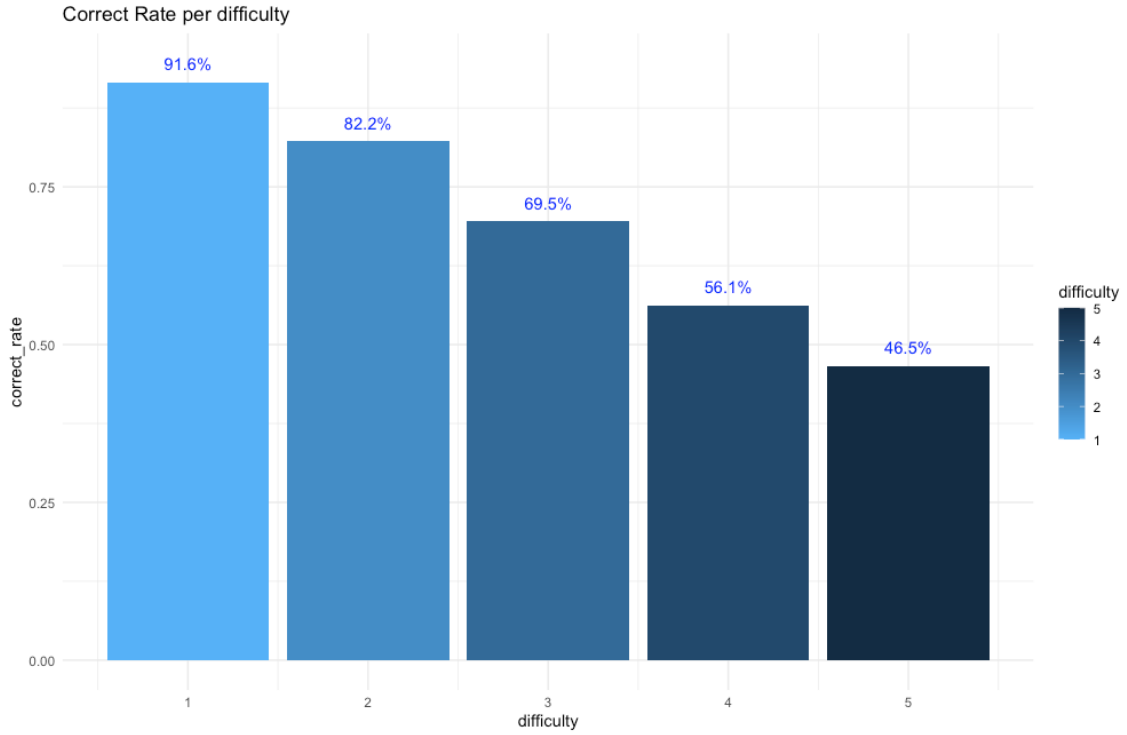


그래프 2 학년 그룹, 문항 난이도별 문항 비율 그래프

교과 문항의 경우 대체로 기본 문제와 유형 문제에 집중되어 있다. 일반적인 문제집에서도 기본, 유형 문제가 가장 많은 것을 고려하면 일견 자연스럽게 보인다. 학년에 따른 문항 난이도의 경우 대체로 고른 편이다. 다만, 저학년에서는 연산 문항에 가까운 형태의 교과 문항이 많아 난이도 1인 문제가 다른 학년에 비해 많이 존재하는 경향을 보인다.

(3) 난이도별 평균 정답률 / 아이 실력을 보정한 난이도별 정답률

위의 (2)의 난이도 분류를 기반으로 아이 실력에 대한 기초통계량을 구하고, 실제 난이도 분류의 적절성에 대해서 알아보자.

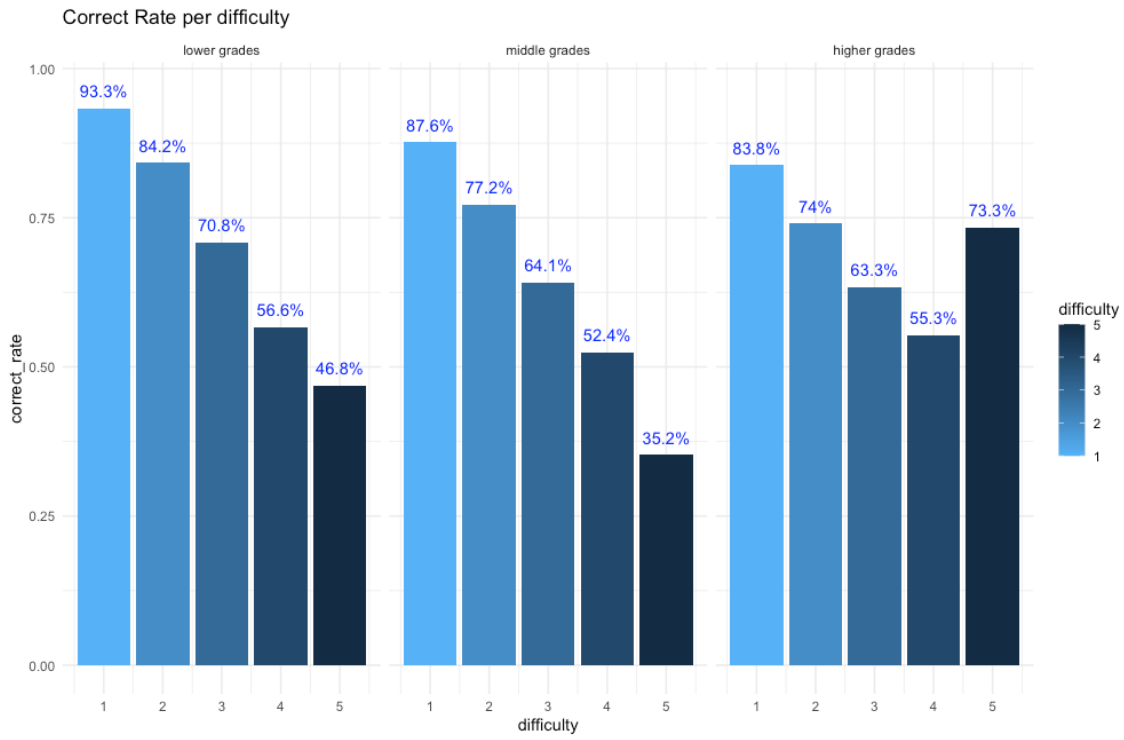


그래프 3 문항 난이도별 정답률 그래프

아이들에게 약 75 ~ 80%의 정답률 문제를 제시하는 것이 학습적으로 효과가 높다는 점을 고려할 때, 평균적인 아이들에게 난이도 2 혹은 난이도 3의 문제를 추천해주는 것이 대체로 유효할 것이다.

문제 난이도별로 대체로 선형적인 정답률 감소 추세를 보인다. 다만, 이것이 실제로 난이도 구간별 선형성을 의미하지는 않는다. 난이도의 경우 logit score 기반[4], 즉

절댓값이 커질수록 구간 밀도가 희박(sparse)해지는 경향을 띤다. 즉, 점수의 선형성이라 함은 대체로 $\log\left(\frac{p}{1-p}\right)$의 선형성을 의미한다고 보아야 할 것이다. 나는 2.3.1에서 멀티플레이 기반 레이팅 모형의 문항 초기 난이도(레이팅) 값으로 두 가지 방법을 도입해 어떤 것이 더 적합한 방식인지 확인하고, 이를 실제 서비스 런칭에 활용하였다.



그래프 4 학년 그룹, 문항 난이도별 정답률 그래프

학년 구간, 난이도별 정답률을 보면 크게 두 가지 패턴을 관찰할 수 있다. 첫째로는 저학년, 중학년, 고학년 순으로 난이도별 정답률이 높다. 수포자라는 키워드가 설명하듯 학년에 구애받지 않는 비슷한 기준(개념 확인 문제 여부, 유형 확인 문제 여부 등)으로 분류하였음에도 정답률이 감소하는 것을 통해 기초 수학 이해도가 학년에 따라 감소함을 확인할 수 있다.

둘째로 고학년의 난이도 5의 정답률이 유독 높다는 점이다. 이는 서비스 내 교과 자유학습의 구조에서 원인을 찾을 수 있다. 교과 자유학습은 Easy, Normal, Hard 스테이지 중 선택하여 문제를 풀이할 수 있는데, Hard 스테이지의 경우 난이도 5 문항

4) 문제 정오답의 경우 맞혔으면 1, 틀렸으면 0을 부여하는 binary case로, Logistic GLM을 주로 적용한다.

이 집중적으로 포함되어 있다. Hard 스테이지의 보상이 높기에[5] 상대적으로 어렵더라도 도전하는 소수의 아이들이 문제 정답률 이상치를 만들었다고 볼 수 있다.

위의 두 그래프를 토대로 볼 때 특정 유저가 문제를 많이 풀거나 특정 패턴을 가지고 풀었을 때 평균 난이도 정보의 가치가 하락함을 이해할 수 있다. 따라서 로지스틱 회귀분석을 통해 위의 두 통계량이 제거하지 못한 '유저 실력' 값을 보정하여 실제 문항 난이도 인자를 분리해보았다. 분석 상세는 아래와 같다.

조건	필터
아이별 문제 개수	50문제 이상
학년	초등 1학년 유저
문제	초등 1학년 1 ~ 2학기 교과 문제
아이 정답률	0% 초과 100% 미만

표 6 로지스틱 회귀분석 분석 상세

유저 ID 정보와 난이도를 factor로 로지스틱 회귀분석을 수행하였다. 난이도 정보만을 사용했으므로 1PL IRT (Rasch Model)을 기반으로 분석한 것이다. 결과는 아래와 같다.

인자	유저 ID	난이도 1	난이도 2	난이도 3	난이도 4	난이도 5
계수	(유저마다 가지는 값, 약 -2 ~ 2)	0 (baseline)	-1.132	-1.966	-2.643	-3.085
계수 차이		-	**1.132**	**0.834**	**0.677**	**0.442**
회귀직선으로 구한 평균 유저 정답률(%)		95.8%	88.1%	76.3%	62.1%	51.4%
평균 유저 정답률(%)		94.4%	85.6%	73.4%	59.7%	51.6%

표 7 로지스틱 회귀분석 계수 요약

위에서 서술했듯 로지스틱 회귀분석의 계수 값은 선형적이지 않고 난이도가 증가할수록 증가세가 꺾임을 알 수 있다. 계수가 난이도에 따라 작아지는 것은 난이도가 증가할수록 정답률이 감소하기 때문이다.

5) 게이미피케이션(Gamification) 형태의 학습-보상 체계를 명확히 이해하고 Hard 스테이지만 도전하는 행동 패턴을 보이는 유저는 고학년에 몰려 있다.

2.2 멀티플레이 기반 레이팅 모형 실험

(1) 레이팅 모형(Elo Rating) 실험 모식도

실제로 문제-아이 간의 멀티플레이 기반 레이팅 모형이 작동할 수 있는지를 알아보기 위하여 1학년 1학기 데이터에 대하여 실험적으로 레이팅 모형을 적용해보았다. 실험의 모식도는 아래와 같다.

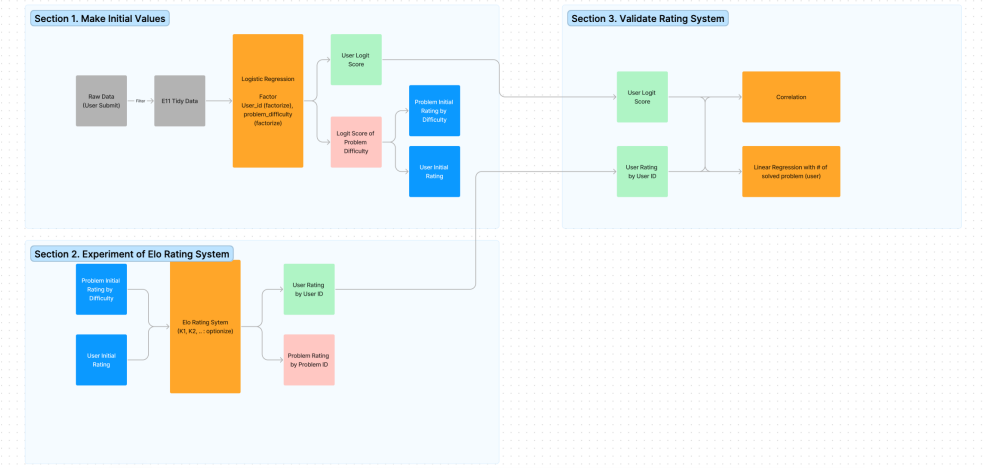


그림 5 (실험) 레이팅 모형 모식도

Section 1은 모형의 초기값을 정하는 과정이다. 유저가 제출한 문제 풀이 자료를 1학년 1학기 문제 풀이만을 포함한 정돈된 데이터로 가공한다. 이후 로지스틱 회귀분석을 통해 유저의 평균 레이팅과 문제 난이도별 문제 초기 레이팅 값을 설정한다. 로지스틱 회귀분석의 인자로는 유저 ID(팩터화), 문제 난이도(숫자 -> 팩터화)를 사용하였다. 문제 난이도를 숫자로 사용하지 않은 이유는 앞에서 서술하였듯 현재 문제 난이도를 등간 척도라 주장할 어떠한 근거도 없기 때문이다.

Section 2에서는 위 Section 1에서 구한 문제 초기 레이팅, 유저 초기 레이팅(모든 유저에 대해 동일)을 기반으로 Elo Rating 시스템을 실제 적용하여 User logit score와 문제 난이도별 logit score를 구하였다. 이 과정에서 후술할 두 가지 옵션을 적용하여 분석하였다.[6)]

Section 3에서는 Section 2에서 구한 User logit score 값의 정당성에 대해서 확인하였다. 유저의 문제 풀이 개수에 대해 측정의 inflate가 발생하는지 확인하였다. User logit score와 User Rating 간 correlation을 구하고, 실제로 User logit

6) 2.3에서 실제로 모형을 구축할 때에는 시스템 내 여러 변수(K1, K2 등)를 grid search하여 찾았다.

score에 User Rating이 (스케일을 제외하고) 수렴함을 논증하였다.

Section 2의 Elo Rating System 수식은 아래와 같다.

$$p = \frac{1}{1 + 10^{-\frac{(probrating - user\,rating)}{400}}}$$

$$(User\,Rating)_{New} = (User\,Rating)_{Old} + K_1 \times (I(correct) - p)$$

$$(Problem\,Rating)_{New} = (Problem\,Rating)_{Old} + K_2 \times (p - I(correct))$$

K값, 문제 초기 레이팅, 유저 초기 레이팅의 경우 아래 2가지를 실험해보았다.

	옵션 1	옵션 2
K value	K1 = K2 = 32	20문제 이하 : K1 = K2 = 32 * 4 21 ~ 40문제 : K1 = K2 = 32 * 2 40문제 초과 : K1 = K2 = 32
문제 초기 레이팅	650 / 850 / 1000 / 1100 / 1200	650 / 850 / 1000 / 1100 / 1200
유저 초기 레이팅	1200[7]	1200

표 8 Elo Rating 적용 실험 두 가지 옵션 요약

난이도별 문제 초기 레이팅의 경우 Elo Rating의 수식과 유저 초기 레이팅 1200점을 기반으로 로지스틱 회귀분석으로 구한 평균 유저 정답률이 되도록 역산한 뒤, 50점 단위가 되도록 자른 것이다.

$$p_{Elo} = \frac{1}{1 + 10^{-\frac{(probrating - user\,rating)}{400}}}$$ **[Elo Rating 기반 정답률 수식]**

구체적인 과정은 아래와 같다.

0. 1학년 1학기 데이터를 이용하여 진행한다.

1. logit(p) = (Intercept) + (User Land ID(Factored)) + (Land Problem Difficulty ID) + epsilon으로 로지스틱 회귀분석

2. 구한 Land Problem Difficulty Score[8]을 기반으로 logit(p_{LR}), p_{LR}를 구한다.

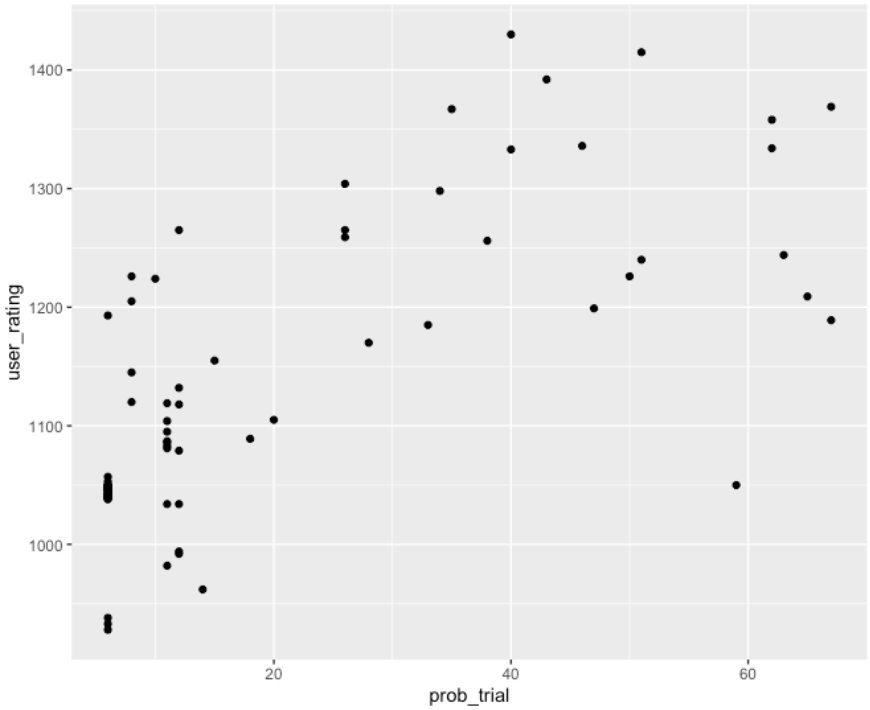
3. $p_{LR} = p_{Elo}$, User Rating 평균 1200점을 활용해 Land, Difficulty별 Intial Problem Rating을 구한다. 이후 서비스 편의를 위해 50 * round(Problem Rating / 50)으로 치환한다.

---

7) 1200점은 임의로 정해진 숫자로써, Elo Rating 시스템에서는 대체로 400의 배수로 초기 레이팅을 정한다. 레이팅 시스템 특징 상 scale은 그다지 큰 의미를 가지지 않는다.

8) 정확히는 User Land ID별 스코어의 평균이 0이 되도록 Land Problem Difficulty ID의 스코어를 조정하여야 한다.

(2) Rasch Model 기반 로지스틱 회귀분석과 비교 - 설명력 / 상관계수 / 잔차 분석



그래프 6 유저별 유저 레이팅 - 문제 풀이 횟수 산점도

위 그래프는 유저별 유저 레이팅 - 문제 풀이 횟수 산점도이다. 상대적으로 문제 풀이 횟수가 많은 유저가 레이팅이 높음을 관측할 수 있다. 문제 풀이 횟수에 따라 유저 레이팅이 highly-inflated 되는 것인지 확인할 필요가 있다.

이를 규명하기 위해 2.1과 같은 형식의 로지스틱 회귀분석(1PL Rasch Model)을 데이터를 바꾸어 진행하였고, 회귀계수 중 유저 ID에 귀속되는 정보를 User logit score이라 정의하였다.

먼저 유저별 레이팅을 User logit score, 유저별 문제 풀이 횟수로 선형 회귀분석하여 User logit score와 문제 풀이 횟수의 선형 설명력을 구하였다.

	user logit score 계수(p-value)	# of solved prob 계수(p-value)	R sq.	Adj. Rsq.
rating ~ user logit score + # of solved problem	85.74 (2.4e-14)	0.117 (0.515)	**80.38%**	79.34%

rating ~ user logit score	86.97 (2.8e-15)	-	**80.15%**	79.65%

표 9 (옵션 2) 결과를 기반으로 회귀분석한 자료 (모형 1의 VIF = 1.072)

```
Call:
lm(formula = L1_rating ~ user_score_fit1, data = user_rating_fit_E11)

Residuals:
     Min       1Q   Median       3Q      Max
-115.965  -22.653   -2.439   21.582  108.368

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1288.000      6.845  188.18  < 2e-16 ***
user_score_fit1   86.968      6.930   12.55 2.84e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.83 on 39 degrees of freedom
Multiple R-squared:  0.8015,    Adjusted R-squared:  0.7965
F-statistic: 157.5 on 1 and 39 DF,  p-value: 2.843e-15
```

그림 7 (옵션 2) 결과를 기반으로 회귀분석한 자료

두 모형을 비교해볼 때 유저의 문제 풀이 개수 팩터를 추가한 모형은 AIC, BIC, Adj. R squared 측면에서 모두 불리하여 설명력의 증가에 영향을 끼치지 못함을 알 수 있다. 즉, 문제 풀이 개수로 인해 유저 레이팅이 inflate되는 상황은 존재하지 않음을 알 수 있다.

Rating을 User logit score로 설명하는 것에 대한 설명력은 약 80% 수준으로 측정되었다. 이는 상당히 높은 수치인데, logit score와 Elo Rating System 하의 레이팅 수치가 비율을 제외하면 대체로 일치한다는 의미로 볼 수 있다. 그 이유를 아래에서 서술하겠다.

$$p_{Elo} = \frac{1}{1+10^{-\frac{(probrating - user\,rating)}{400}}}$$

$$p_{3PL} = \frac{1}{1+10^{\frac{b\times(a-u)}{400}}} \times (1-c) + c,\ \Phi = \frac{1}{1+10^{\frac{b\times(a-u)}{400}}}$$

$$\frac{\partial p}{\partial a} = \frac{(p-c)(1-p)}{1-c} \times \frac{b \times \log 10}{400}$$

log likelihood $L = -\log p \times I(correct) - \log(1-p) \times I(wrong)$

$$\frac{\partial L}{\partial p} = -\frac{1}{p} I(correct) + \frac{1}{1-p} I(wrong)$$

$$\frac{\partial L}{\partial a}|_{correct} = \frac{-b \times \log 10}{400} \times \frac{1}{p} \times \frac{(p-c)(1-p)}{1-c} = \frac{-b \times \log 10}{400} \times \frac{\Phi}{p} \times (1-p) = -\frac{\partial L}{\partial u}|_{correct}$$

$$\frac{\partial L}{\partial a}|_{wrong} = \frac{b \times \log 10}{400} \times \frac{p-c}{1-c} = \frac{b \times \log 10}{400} \times \Phi = -\frac{\partial L}{\partial u}|_{wrong}$$

p_{Elo}, p_{3PL}는 각각 Elo Rating, 3PL IRT 하에서의 정답률이다. 우리는 Rasch Model, 즉 1PL IRT를 이용하고 있으므로, $b = 1, p = \Phi$인 경우에 대해 생각하기로 한다. Rasch Model 하에 log Likelihood의 gradient 방향 $\frac{\partial L}{\partial a}$은 각각 Correct, Wrong 인 경우에 대해 각각 $1-p$, p에 비례한다. 이는 Elo Rating 시스템 하에서의 수식과 정확히 동일하다. 해당 알고리즘이 보장하는 것은 gradient descent method를 통해 Elo Rating System은 1PL IRT 하에서의 log Likelihood를 minimize하는 방향으로 움직이도록 한다는 것이다.

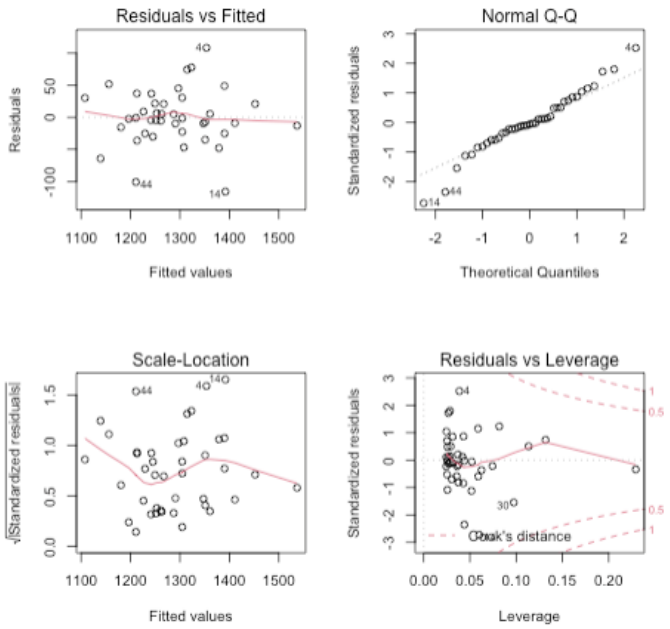
위 결과가 어쩌면 자연스러운 것은 Elo Rating System의 경우 상대를 이길 확률이 Logistic Curve를 따른다. 결과적으로 우리는 로지스틱 회귀분석 없이 머신러닝의 Gradient Descent 방식으로 각 유저의 로짓 점수로 레이팅을 수렴시키는 방식을 도입하게 된 것이다.

로지스틱 회귀분석 대신 Elo Rating System을 쓰는 것이 효율적인 이유는 Rating System은 선형 시스템이기 때문이다. 서비스를 구축하면서 매번 Complexity가 엄청난 로지스틱 회귀분석을 유저별로 돌리는 것은 계산 리소스 문제뿐 아니라 접속한 유저에 대한 동시성 문제가 발생한다.[9] 출제 알고리즘은 측정 정확성을 확보함과 동시에, 그것이 빠르게 계산될 수 있음을 보장할 수 있어야 하는 것이다.

추가로 유저별로 레이팅 움직임을 보았을 때 여러 대단원을 푼 유저의 레이팅이 크게 흔들림을 관찰할 수 있었다. 따라서 실제 기능 개발 전 2.3에서 대단원별 레이팅을 도입하는 것에 대한 개념을 추가하게 되었다. 이는 초등수학 학습 관점에서도 암시하는 점이 있는데, 도형이 강한 아이가 반드시 수와 연산 관련 대단원을 잘 푼다고 담보할 수 없기 때문이다. 만약 유저 실력 값을 하나의 수치로 관리하게 될 경우, 당연히도 유저 데이터가 흐려지는 것을 막을 수 없다. 이하 2.3에서 초등 대단원 71개에 대하여 유저별로 따로 관리하는 형태로 User Land Rating[10] 개념을 도입하게 되었다.

9) 전체 데이터에 대한 로지스틱 회귀분석은 매 유저 제출마다 유저별로 스코어가 바뀌는 결과를 초래한다. 접속하지 않은 유저 정보가 서비스 내에서 실시간으로 변하는 것은 결코 자연스럽지 않다. 심지어 유저의 학습 실력은 문제를 푸는 시간, 과정에 의해 실제로도 늘 변한다. 과거에 제출한 데이터가 현재에도 동일한 weight로 내 실력 측정에 반영되는 것은 바람직하지 않다.

10) 대단원 정보를 Land라 명칭하였다.



그래프 8 (모형 1) 잔차 회귀진단

위 표 7의 (모형 1)에 대한 잔차 회귀진단을 시행하였다. 그래프 6의 네 개의 그래프에서 크게 특이사항을 발견하지 못하였다. 특히 Residual-fitted Curve의 경우 거의 x축과 평행한데, 이는 모든 레이팅 대에서 유저 레이팅이 User logit score로 적절히 수렴함을 암시한다.

2.3 학습 실력 측정 모형 개발 및 로지스틱 회귀분석 기반 최적화

(1) 학습 실력 측정 모형 모식도

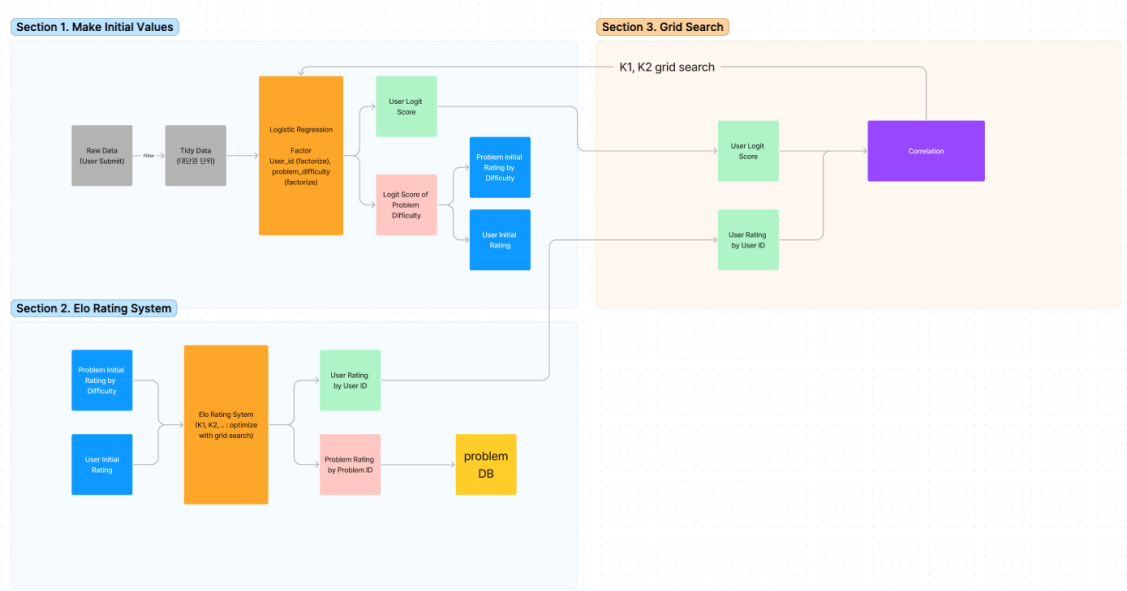


그림 9 학습 실력 측정 모형 모식도

2.2의 실험 결과를 기반으로 멀티 플레이 방식의 Elo Rating 모형의 작동 가능성을 확인하고, 실제 학습 실력 측정 모형 개발에 착수하였다.

Section 1과 Section 3은 최적의 K1, K2를 구하는 과정이다. Section 2를 거치며 현재까지 유저가 제출한 문항 기반으로 구한 유저 레이팅을 기반으로 로지스틱 회귀분석 결과와 Pearson Correlation과 Spearman Correlation을 구해 가장 상관계수가 높아지는 hyperparameter을 찾는 grid search 과정을 거쳤다. 대단원별 난이도별 초기 문항 레이팅의 경우 대단원 단위의 Tidy Data를 활용해 로지스틱 회귀분석을 통해 구하였다. 로지스틱 회귀분석의 결과를 기반으로 Elo Rating 수식의 User Rating 부분에 평균 유저 레이팅을 집어넣어 각 문제 난이도별 초기 레이팅 값을 세팅하였다.

Hyperparameter tuning을 마치고 난 이후에는 Section 2를 다시 한 번 수행하며 problem DB에 실을 최종 문항 레이팅을 구하였다.[11] Section 2의 Elo Rating

11) 해당 과정에서 minibatch 방법을 시도하였으나, learning rate(현재에서는 K)를 줄여 동일 데이터

System 수식은 아래와 같다. (2.2와 동일)

$$p = \frac{1}{1 + 10^{-\frac{(probrating - user\,rating)}{400}}}$$

$$(User\,Rating)_{New} = (User\,Rating)_{Old} + K_1 \times (I(correct) - p)$$
$$(Problem\,Rating)_{New} = (Problem\,Rating)_{Old} + K_2 \times (p - I(correct))$$

(2) 난이도별 문항 초기 레이팅 부여

난이도별 문제 초기 레이팅의 경우 Elo Rating의 수식과 유저 초기 레이팅 1200점을 기반으로 로지스틱 회귀분석으로 구한 평균 유저 정답률이 되도록 역산한 뒤, 50점 단위가 되도록 자른 것이다.

$$p_{Elo} = \frac{1}{1 + 10^{-\frac{(probrating - user\,rating)}{400}}}$$ [Elo Rating 기반 정답률 수식]

구체적인 과정은 아래와 같다.[12)]

0. 데이터 수가 부족하여 1학년, 2학년, 3~4학년, 5~6학년 4가지로 구분하여 문항 초기 레이팅을 구한다.

1. logit(p) = (Intercept) + (User ID(Factored)) + (Problem Difficulty ID) + epsilon으로 로지스틱 회귀분석

2. 구한 Problem Difficulty Score[13)]을 기반으로 logit(p_LR), p_{LR}를 구한다.

3. $p_{LR} = p_{Elo}$, User Rating 평균 1200점을 활용해 Difficulty별 Intial Problem Rating을 구한다. 이후 서비스 편의를 위해 50 * round(Problem Rating / 50)으로 치환한다.

	난이도 1	난이도 2	난이도 3	난이도 4	난이도 5
1학년	750	900	1050	1200	1300
2학년	850	1050	1100	1250	1350

를 수 번 반복 학습시키는 방식이 적어도 현재 데이터 수준에서는 K를 늘리는 것에 비해 큰 이익이 없고, 특정 유저 데이터가 과다 학습되어 overfitting을 유도하여 제외하였음을 밝힌다.

12) 대단원별로 초기 레이팅을 구하기에는 절대적인 데이터 수가 부족하여 상대적으로 큰 단위로 쪼개어 진행하였다.

13) 정확히는 User Land ID별 스코어의 평균이 0이 되도록 Land Problem Difficulty ID의 스코어를 조정하여야 한다.

3-4학년	850	1000	1150	1300	1400
5-6학년	900	1000	1150	1350	1450

표 10 Problem Database로 이식될 Problem Initial Rating

(3) 모수 K1, K2 값 찾기 - grid search

모형의 hyperparameter를 구하기 위해 grid search를 진행하였다. 찾아야하는 hyperparameter의 경우 아래와 같다.

	특이사항
K1	유저 대단원 레이팅의 학습률
K2	문항 레이팅의 학습률
유저의 문제 풀이 개수별 K1 변화	어느정도 수렴했다 판단되면 학습률을 줄여줄 필요가 있음
유저의 문제 풀이 개수별 K2 변화	어느정도 수렴했다 판단되면 학습률을 줄여줄 필요가 있음

표 11 찾아야하는 하이퍼파라미터 내역

Grid Search의 경우 두 번에 나누어 진행하였다. 우선, K1, K2를 고정해두고 K1 변화량을 조절하였다. 1학년 1학기 1단원 데이터에 대하여 대단원별로 대단원 레이팅을 구하고, User logit score과의 상관계수를 계산하였다.

Option	**K1**	**K2**	**K1 변화**	**K2 변화**[14]
Option A	32	32	~10 : 4배, 11~20 : 2배	~5 : 0배
Option B	32	32	~10 : 6배, 11~20 : 2배	~5 : 0배
Option C	32	32	~10 : 4배, 11~20 : 3배	~5 : 0배
Option D	32	32	~5 : 4배, 6~30 : 2배	~5 : 0배
Option E	32	32	~5 : 6배, 6~30 : 3배	~5 : 0배
Option F	32	32	~5 : 4배, 6~30 : 3배	~5 : 0배

표 12 첫 번째 Grid Search 파라미터 내역

이 파라미터들로 유저 제출 데이터를 1회 학습시키고 구한 상관계수이다.

Option	Pearson Correlation	Spearman Correlation
Option A	0.869	0.915
Option B	0.828	0.891
Option C	0.87	0.914

14) K2 값이 0인 이유는 상대적으로 문제의 난이도는 DB에 이미 정확히 측정되있는 것과 다르게, 유저의 풀이 실력은 완전히 초기 상태이기 때문이다. 정보의 정확성 차이로 인해 유저의 실력이 rough하게라도 측정되기 전까지의 값은 버리는 것이 옳다고 판단하였다.

Option D	**0.929**	**0.936**
Option E	0.897	0.924
Option F	0.892	0.922

표 13 첫 번째 Grid Search 결과

Option D, 즉 K1 변화를 5문제까지 4배, 30문제까지 2배한 경우가 가장 높았다. 이 경우가 가장 잘 동작한다고 가정하고 두 번째 Grid Search를 진행하였다.

Option	K1	K2	K1 변화	K2 변화
Option A	32	32	~5 : 4배, 6~30 : 2배	~5 : 0배
Option B	40	32	~5 : 4배, 6~30 : 2배	~5 : 0배
Option C	48	32	~5 : 4배, 6~30 : 2배	~5 : 0배
Option D	64	32	~5 : 4배, 6~30 : 2배	~5 : 0배
Option E	32	10	~5 : 4배, 6~30 : 2배	~5 : 0배
Option F	40	10	~5 : 4배, 6~30 : 2배	~5 : 0배
Option G	50	10	~5 : 4배, 6~30 : 2배	~5 : 0배
Option H	100	10	~5 : 4배, 6~30 : 2배	~5 : 0배

표 14 두 번째 Grid Search 파라미터 내역

이 파라미터들로 유저 제출 데이터 1회 학습하여 구한 상관계수이다.

Option	Pearson Correlation	Spearman Correlation
Option A	0.929	0.936
Option B	0.929	0.935
Option C	0.922	0.932
Option D	0.904	0.911
Option E	0.928	0.934
Option F	**0.931**	**0.942**
Option G	0.924	0.934
Option H	0.868	0.879

표 15 두 번째 Grid Search 결과

두 번의 Grid Search를 통해 확정한 Elo Rating 내부 로직은 아래와 같다.

$$p = \frac{1}{1 + 10^{-\frac{(probrating - user rating)}{400}}}$$

$K_1 = 160, K_2 = 0$ (유저가 대단원 내 문제 풀이가 5회 이하일 때)

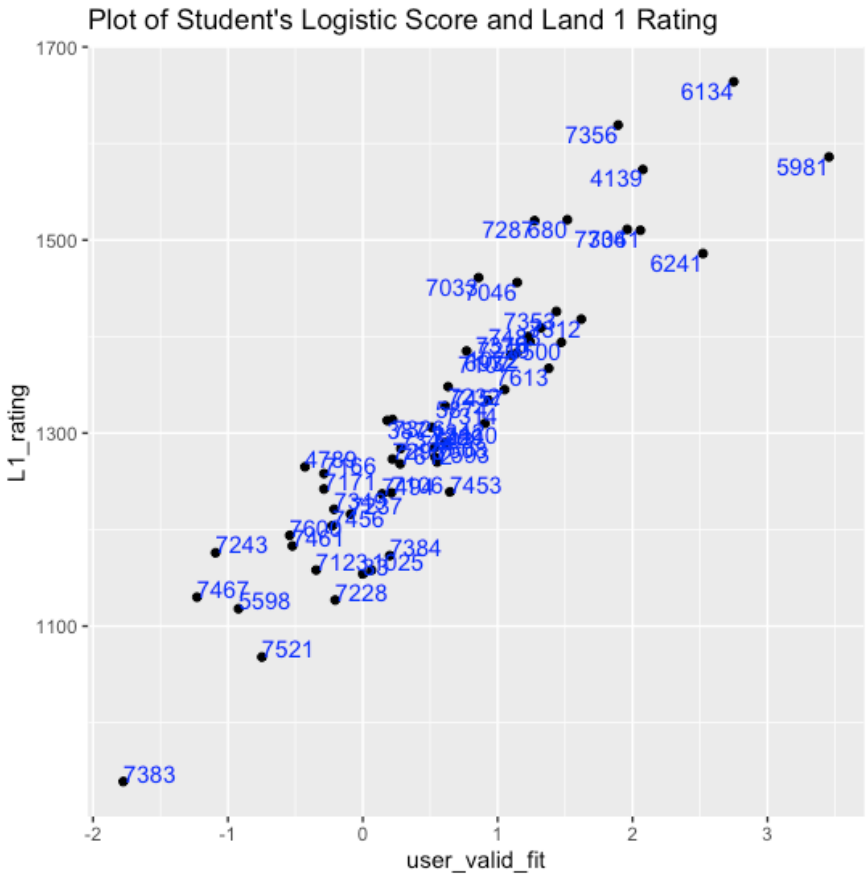
$K_1 = 80, K_2 = 10$ (유저가 대단원 내 문제 풀이가 5회 초과 30회 이하일 때)

$K_1 = 40, K_2 = 10$ (유저가 대단원 내 문제 풀이가 30회 초과일 때)

$$(UserRating)_{New} = (UserRating)_{Old} + K_1 \times (I(correct) - p)$$

$$(ProblemRating)_{New} = (ProblemRating)_{Old} + K_2 \times (p - I(correct))$$

확정된 모형 파라미터를 통해 구한 유저 레이팅과 User logit score의 산점도는 아래와 같다.[15)]



그래프 10 유저 레이팅, User Logit Score 산점도

이하 3.1 연구 결과에서는 해당 실력 측정 모형을 통해 구한 예상 정답률이 얼마나 잘 작동하였는지 cross-entropy와 잔차 분석을 통해 평가해볼 것이다.

15) 상대적으로 유저들의 대다수 점수대가 1200점을 상회하는데, 이는 1학년 전체 데이터로 문제 초기 레이팅을 잡았음에 반해, 1학년 1학기 1단원 9까지의 수 난이도가 너무 낮기 때문이다. 다만, 이 부분 역시 [교과 문항 출제 솔루션]의 제작 방식 여하에 따라 전혀 문제되지 않을 수 있다. 애초에 [학습 실력 측정 모형]이 추구하는 바 역시 대단원과 상관없이 획일화된 문제 난이도(개념 확인급, 응용 문항, 실력 문항 등) 하에서 어느 정도의 문제 풀이가 적합한지를 구한 것이기 때문이다.

2.4 학습 실력 측정 모형을 이용한 교과 문항 출제 솔루션 개발

(1) 교과 문항 출제 솔루션 모식도

학습 실력 측정 모형을 이용하여 유저들에게 어떤 학습 곡선을 제시할 것인가는 여전히 숙제이다. 유저의 정확한 학습 실력, 즉 특정 문제를 제시했을 때 유저가 맞힐 확률을 서비스에서 알고 있는 것은 분명한 이익이다. 그럼에도 사실 핵심은 이 값을 이용해 유저에게 알맞은 문제를 찾는 것이다.

교과 문항 출제 솔루션은 크게 세 가지 프레임으로 구성된다.

가. 서버가 문제를 찾는 과정

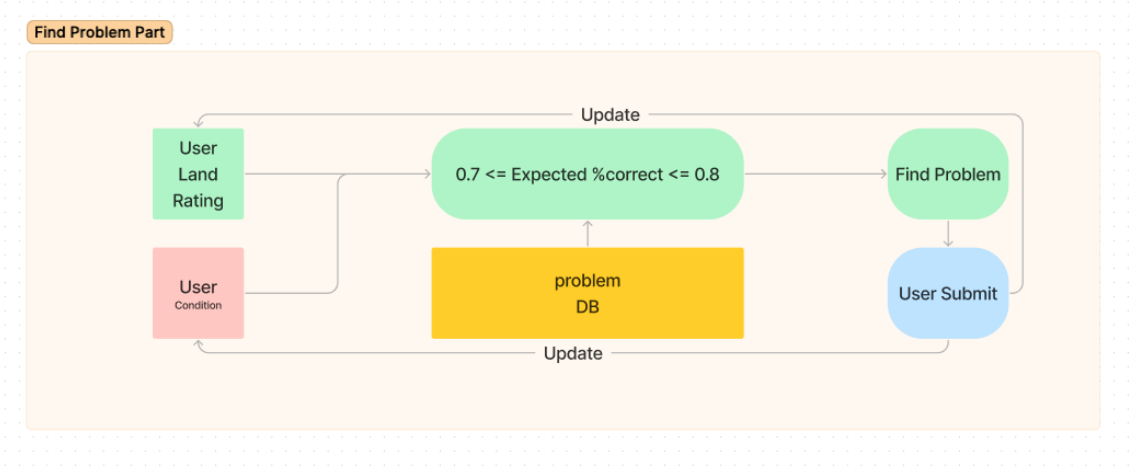


그림 11 서버 문제 탐색 로직 모식도

첫 번째 프레임은 문제를 찾는 과정이다. 서버에서는 쿼리를 통해 유저, 대단원별 레이팅 값과 특정 시간을 기점으로 0으로 초기화되는 유저 컨디션 지표, 두 가지 값을 불러와서 기대 정답률을 계산하고 70% 이상 80% 이하에 해당하는 문제를 문제 DB에서 불러온다. 해당 문제 중 적당한 weight로 한 문제를 뽑고 유저에게 문항을 제시한다. 유저의 응답에 따라 유저 컨디션 지표와 유저, 대단원별 레이팅 지표가 업데이트된다.

나. 유저에게 제시되는 미션 문항 출제 / 제출 로직

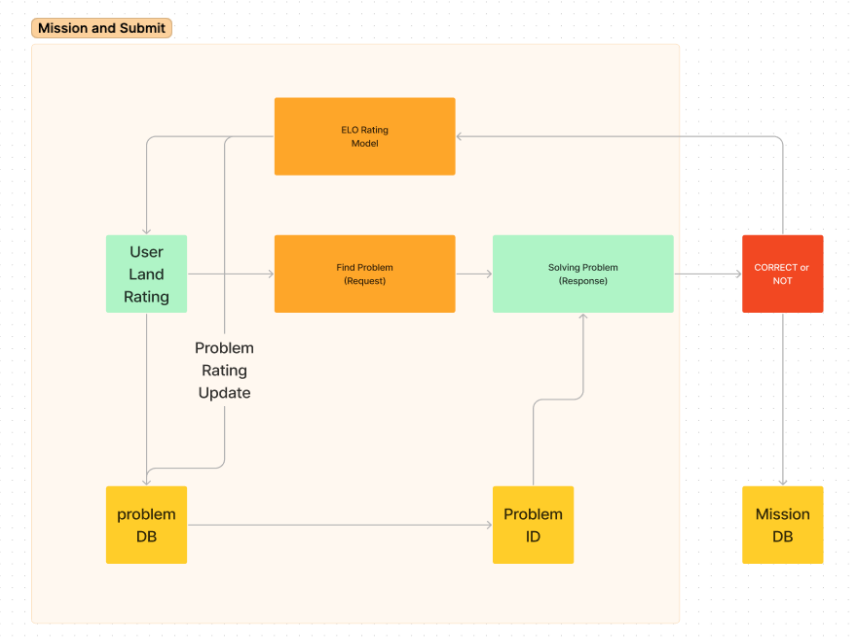


그림 12 미션 문항 출제 / 제출 로직 모식도

두 번째 프레임은 유저에게 제시되는 미션 문항의 출제 / 제출 로직이다. 위 첫 번째 프레임으로 선발된 문제가 제출되면 미션 DB에 제출, 저장된다. 해당 정보는 2.3에서 구성한 Elo Rating 모형에 집어넣어 새로운 유저 레이팅, 문제 레이팅을 구하여 DB에 저장된다.

다. 대단원 내 학습량 및 학습 방식 구성

세 번째 프레임은 대단원별 학습 구성이다. 대단원이 여러 개의 소단원으로 구성되어 있고, 초등수학 학습 구성에 따라 소단원별로 적당한 양만큼 학습시켜야 한다. 이는 다소 난해한 문제이다. 개념적으로 볼 때 소단원별로 난이도에 차이가 존재하여 특정 난이도를 학습해야 하는 유저에게 특정 소단원의 출제 비율이 높아져야 할 필요가 있는 경우가 많기 때문이다.

나는 유저별로 계산되는 소단원 비율 리스트를 제작하여 해당 대단원 내 소단원별 미션 출제 비율을 조절하는 알고리즘을 구현하여 이를 해결하였다. 소단원 비율 리스트 알고리즘의 개요는 아래와 같다.

그림 13 대단원 내 학습량 / 학습 방식 구성 모식도

List A = rep(1 / (소단원 개수) , (소단원 개수)) 형태의 uniform 비율 리스트
List B = 현재 유저 예상 정답률 75%에 가장 가까운 50문항의 소단원을 찾아 각 소단원별 (소단원 내 50문항 중 몇 문항이 포함되었는지 비율)을 구한 비율 리스트
List C = List A와 List B의 m : n 가중평균

해당 List C를 매 문제 제출마다 업데이트하여 소단원별 출제 비율로 이용하였다. 이 방식을 통하면 robust하면서도 적당히 유저가 필요로 하는 소단원을 추천하는데 충분함을 모든 71개 대단원에 대하여 실험적으로 입증하였다.

(2) 개발 과정

가. 유저 컨디션 레이팅 소개

유저 컨디션 레이팅은 유저의 다이나믹한 풀이 경험을 위해 설계된 값이다. 특정 트리거(날짜, 소단원 변경 등)로 초기화되며, 아이가 연속으로 틀리는 경험을 했을 때 절대적인 실력 값은 건드리지 않으면서도 잠시 출제용 레이팅을 하락시켜 쉬운 문제를 만나도록 하는 역할을 수행한다. (즉, 문제를 찾을 때 유저 레이팅 대신 유저 레이팅 + 컨디션 레이팅을 사용한다.)

구체적인 로직은 아래와 같다.

$$K_{Condition} = K_C = 50$$

$$p_{user} = \frac{1}{1 + 10^{\frac{(probrating) - (user\ rating)}{400}}}$$

$$p_{user,C} = \frac{1}{1 + 10^{\frac{(probrating) - (user\ rating) - (Condition\ rating)}{400}}}$$

$$(User\,Rating)_{New} = (User\,Rating)_{Old} + K_1 \times (I(correct) - p_{user})$$
$$(Condition\,Rating)_{New} = (Condition\,Rating)_{Old} + K_C \times (I(correct) - p_{user,C})$$

나. 미션에서 저장하는 데이터

유저가 미션에서 문제 정답을 제출하면 서비스는 아래 데이터를 mission DB에 저장한다.

변수	의미
정오답	정답이면 correct, 오답이면 wrong으로 저장
DB 기록 시각	DB 로그 남긴 시간 저장
DB 업데이트 시각	DB 로그 수정한 시간 저장
문제 ID	문제 DB 내 primary key인 ID를 기록
유저 ID	유저 DB 내 primary key인 ID를 기록
커리큘럼 정보	현재 유저가 학습 중인 커리큘럼 정보 저장
문제 찾는 시점 유저 레이팅	로그 데이터
문제 제출 시점 유저 레이팅	로그 데이터
문제 찾는 시점 문제 레이팅	로그 데이터
문제 제출 시점 문제 레이팅	로그 데이터
문제 찾는 시점 컨디션	로그 데이터
문제 제출 시점 컨디션	로그 데이터

표 16 미션에서 저장하는 데이터 일람

해당 데이터를 통해 위 모형이 실제로 정답를 잘 예측하는지 사후 검증을 거치게 되었다. 실제로 유저의 문항 정오답을 잘 예측하는지 여부를 확인하는 과정은 3.1.(3)에서 다루겠다.

3. 결론

3.1 연구 결과

(1) 학습 실력 측정 모형 개발

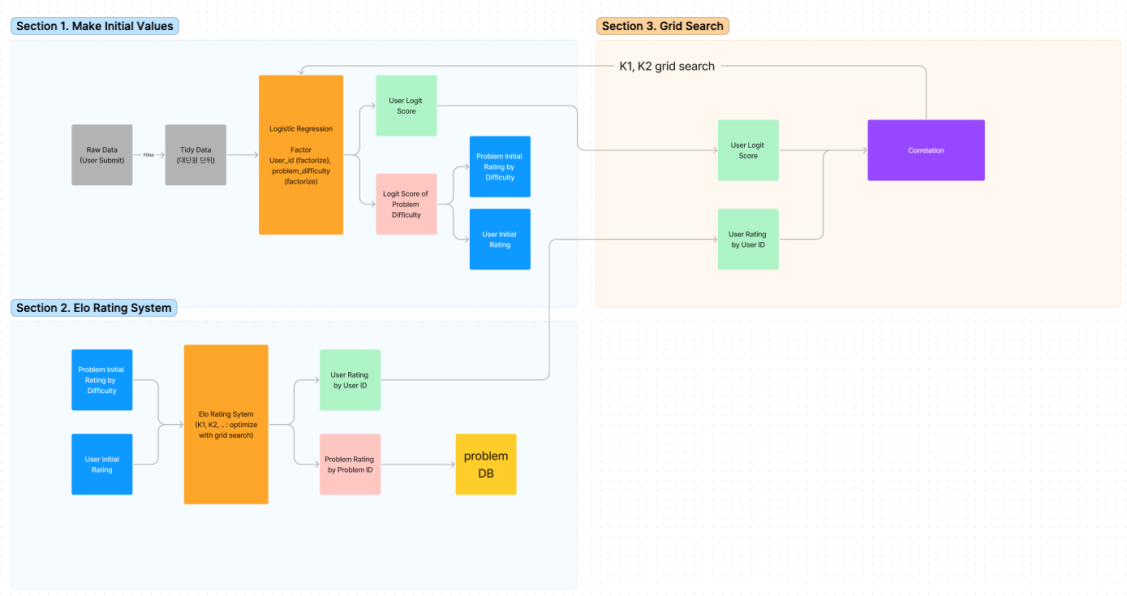


그림 14 학습 실력 측정 모형 모식도

Elo Rating System을 응용하여 문제-아이 반응 모형, 즉 학습 실력 측정 모형을 개발하였다. 완성된 학습 실력 측정 모형의 스펙(Specification)은 아래와 같다.

$$p = \frac{1}{1 + 10^{-\frac{(prob\,rating - user\,rating)}{400}}}$$

$K_1 = 160, K_2 = 0$ (유저가 대단원 내 문제 풀이가 5회 이하일 때)
$K_1 = 80, K_2 = 10$ (유저가 대단원 내 문제 풀이가 5회 초과 30회 이하일 때)
$K_1 = 40, K_2 = 10$ (유저가 대단원 내 문제 풀이가 30회 초과일 때)

$$(User\,Rating)_{New} = (User\,Rating)_{Old} + K_1 \times (I(correct) - p)$$
$$(Problem\,Rating)_{New} = (Problem\,Rating)_{Old} + K_2 \times (p - I(correct))$$

해당 모형을 통해 서비스 DB에는 실시간으로 유저의 실력과 문제의 난이도 정보가 업데이트된다. K1을 40 미만으로 줄이지 않은 것이 더 지표가 좋았던 이유는 유저의 실력은 평균적으로 우상향하기 때문이다. 만약 일반적인 ML(Machine Learning) 학습 상황이라면 descending하는 learning rate scheduler를 사용하여 유저의 문항 풀이 횟수에 따라 지속적으로 K1 값을 줄여주는 것이 효과적이었을지 모른다. 하지만, 학습의 경우 유저의 실력은 매번 변하는 값이므로 K1을 일정 이하로 줄이는 게 효과가 크지 않다.

(2) 교과 문항 출제 솔루션 개발

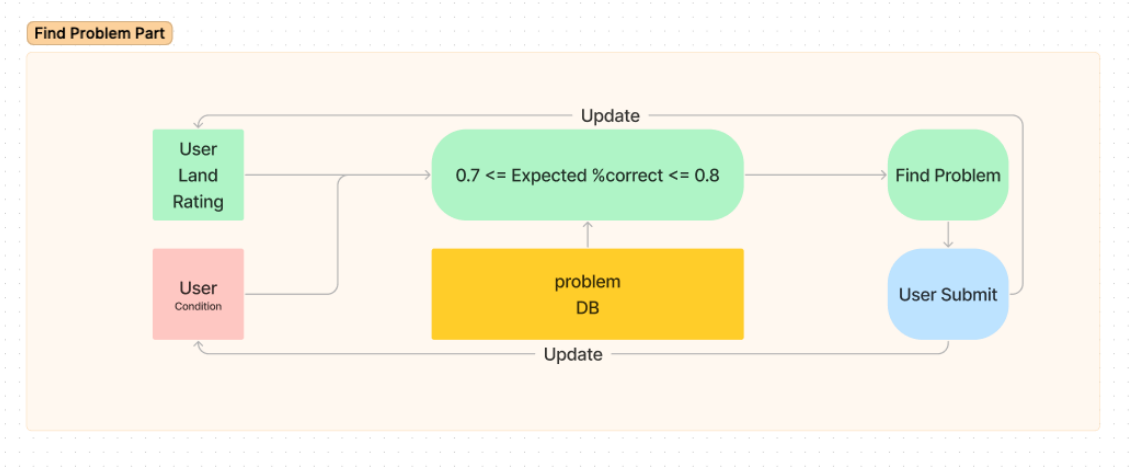


그림 15 서버 문제 탐색 로직 모식도

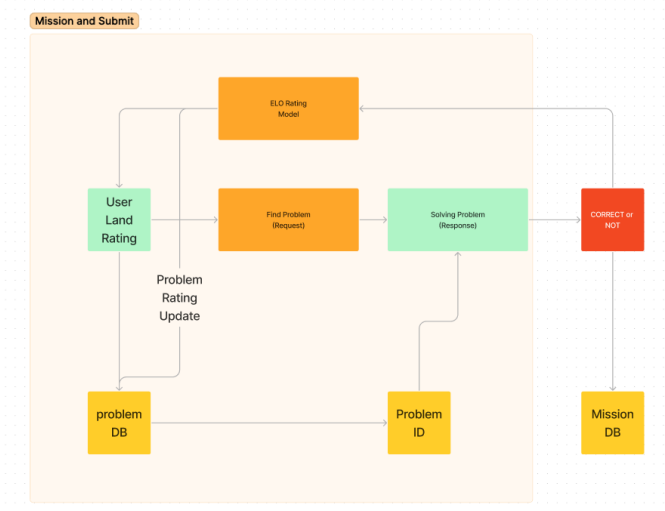


그림 16 미션 문항 출제 / 제출 로직 모식도

그림 17 대단원 내 학습량 / 학습 방식 구성 모식도

위의 학습 실력 예측, 즉 문제 단위로 유저가 맞힐 확률 정보를 기반으로 교과 문항을 어떻게 배치할 것인가에 관해 서비스 출제 솔루션을 개발하였다. 대단원 단위로 소단원들의 index를 배열하고, 소단원별로 유저에게 필요한 수준의 문항 비율을 계산하여 공급하는 방식을 통해, 특정 수준대의 아이들이 풀지 못하거나 이해할 수 없는 문제가 나오지 않도록 설계하였다.

이 방식으로 서비스에서는 실제 개발에 착수하였다. 오늘의 미션(이하 미션) 기능을 개발하여 모든 유저들이 하루에 약 40문항 정도를 풀 수 있도록 구성되었다. 그 중 <도망치는 악동들을 잡아라> 3개의 시리즈가 위의 출제 솔루션을 기반으로 출제된다.

서비스 명칭	출제 방식
도망치는 악동들을 잡아라 1	2.4의 출제 솔루션
도망치는 악동들을 잡아라 2	2.4의 출제 솔루션
도망치는 악동들을 잡아라 3	2.4의 출제 솔루션
연구소 잠복 작전	연산 스테이지로 이동
공장 전투 작전	교과 스테이지로 이동
몬스터가 훔쳐간 팡 되찾기	틀린 문제 다시 풀기로 이동

표 17 오늘의 미션 내 서비스 명칭과 출제 방식

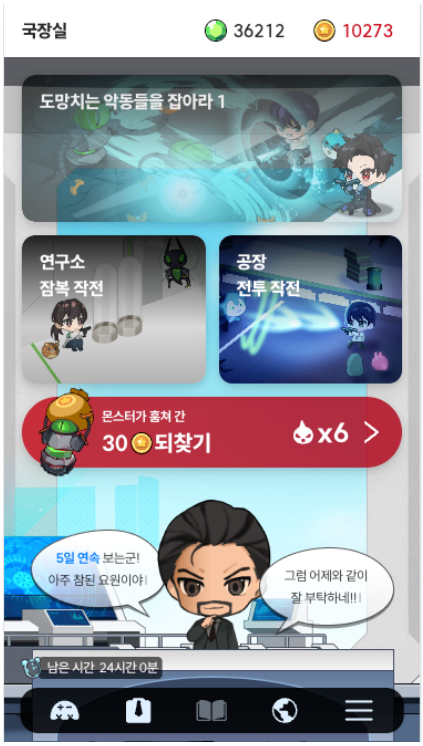
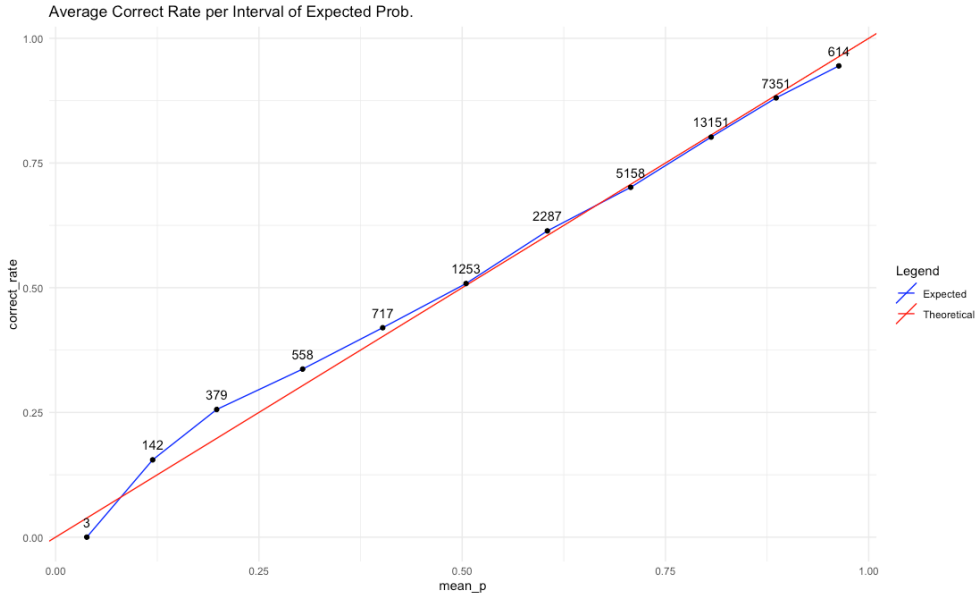


그림 18 미션 기능 배포 후 실제 화면

(3) (1)의 모형이 실제로 정답 잘 예측하는지 검증

해당 모형을 통해 특정 한 문제를 풀이하는 유저가 어느 확률로 이 문항을 맞힐지 예측할 수 있게 되었다. 문제의 대단원 정보, 문제의 현재 레이팅, 유저의 현재 레이팅으로 구성된 세 개의 변수로 얼마나 정확히 확률을 예측하는지 검증한다.

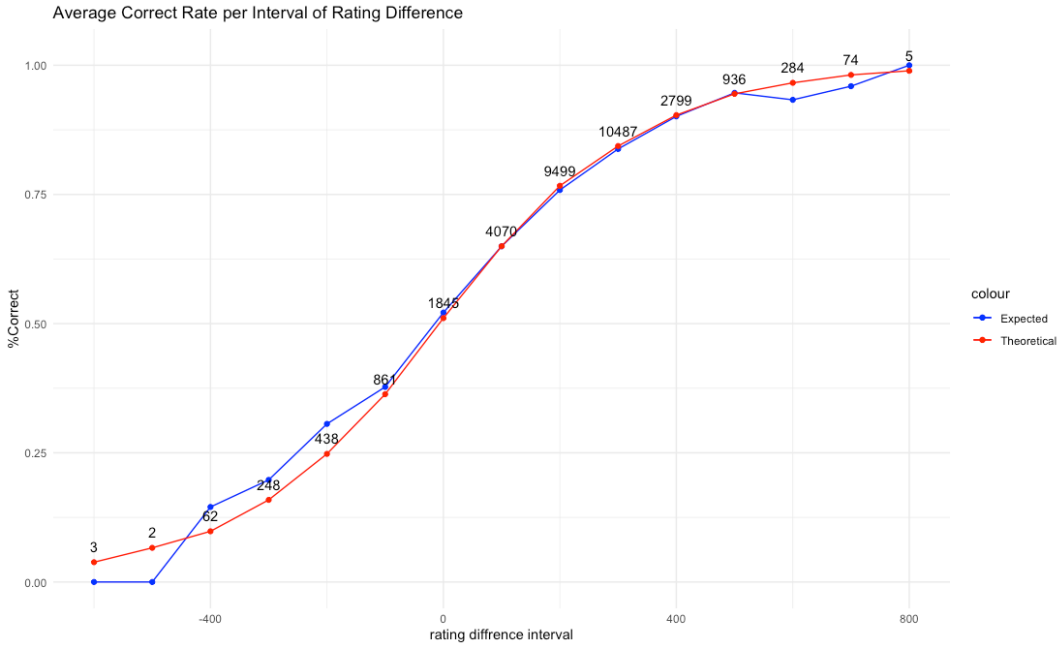
가. 예측 정답률 구간별 실제 정답률



그래프 19 기대 정답률 구간에 따른 평균 정답률 그래프

유저의 예측 정답률이 x, 정답률이 y인 그래프이다. 기대 정답률, 즉 Elo Rating의 결과물로 구해진 예상 정답률의 평균(파란색)이 y = x 그래프 (빨간색)에 거의 근접하는 결과를 확인할 수 있다. 특히 데이터가 많은 70% ~ 90% 구간에서는 거의 정확히 정답률을 예측한다.

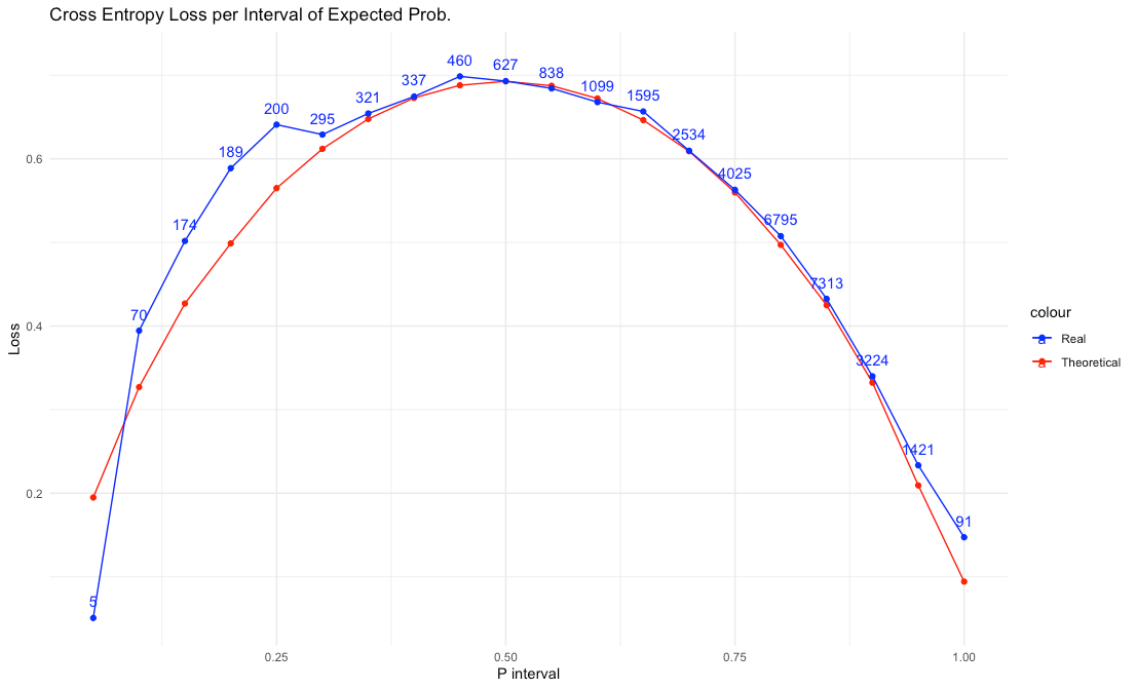
나. 유저 레이팅과 문제 레이팅 차이 구간에 따른 평균 정답률 차이



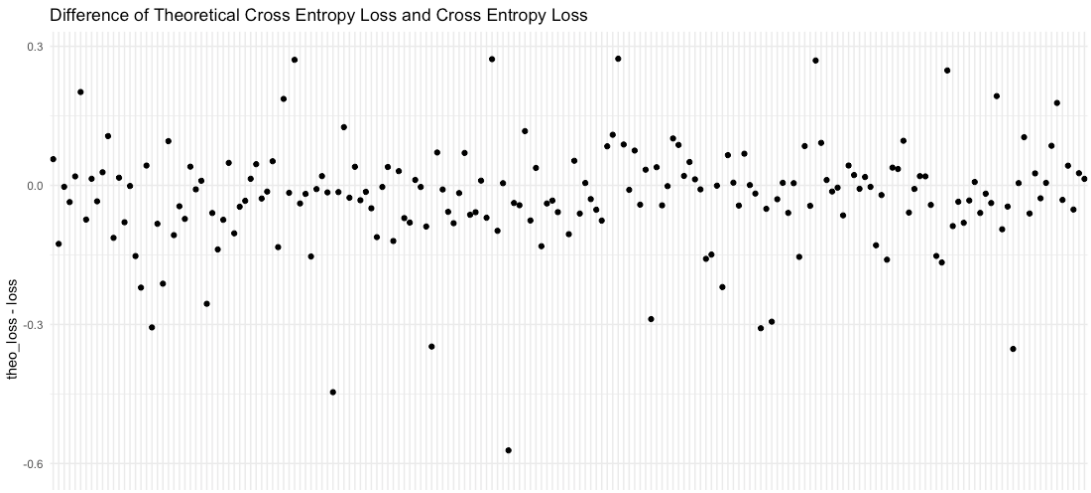
그래프 20 유저 레이팅과 문제 레이팅 차에 따른 평균 정답률 그래프

비슷하게 유저와 문제 사이 레이팅의 차이를 구하여 그 구간에 따른 평균 정답률 차이를 도식화한 것이다. Elo Rating의 결과물로 구해진 예상 정답률의 평균(파란색)이 로지스틱 곡선(빨간색)에 거의 근접함을 확인할 수 있다.

다. Cross Entropy Loss의 이론값, 측정값 비교



그래프 21 기대 정답률 구간에 따른 Cross Entropy Loss의 이론값과 측정값



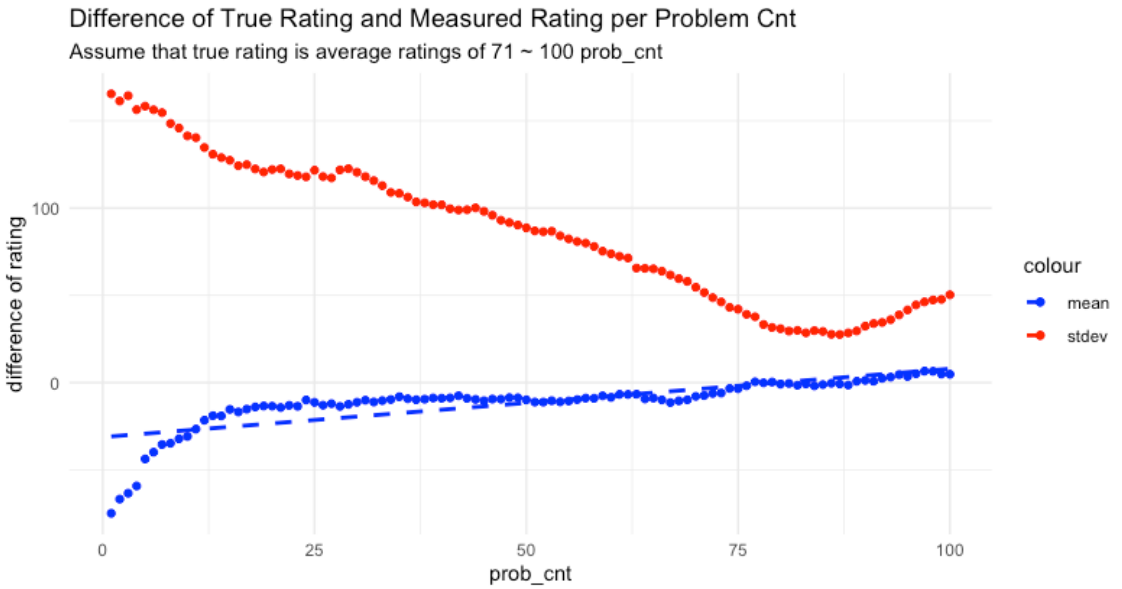
그래프 22 유저별 Cross Entropy Loss의 이론값과 측정값 차이 산점도

위 그래프는 예상 정답률 구간에 따른 Cross Entropy Loss를 비교한 그래프이다. 2.3에서 증명하였듯 Elo Rating System에 따른 레이팅 수렴은 Rasch Model 하에서의 Cross Entropy Loss를 Minimize하는 방향, 즉 Gradient Descent Method를 수행한다. 그의 결과가 입증되듯, Thoeretical Cross Entropy Loss (-plogp - (1-p)log(1-p), 빨간색)에 거의 근접하는 실제 Cross Entropy Loss (-Correct *

logp - Wrong * log(1-p), 파란색)이 도출되었다.

아래 그래프의 경우 두 Cross Entropy의 차를 유저별로 그려놓은 것이다. 역시 0에 대부분 유저가 수렴함을 확인할 수 있다. 유저 ID가 낮은 순으로 그려놓은 것인데, Loss가 증가하는 경향도 발견되지 않는다.

라. 유저의 대단원 내 문제 풀이 개수별 유저 레이팅 수렴 속도



그래프 23 유저의 문제 풀이 횟수에 따른 실제 레이팅과 측정 레이팅 차이 그래프

각 문제까지 보고 측정된 레이팅과 실제 레이팅의 차이를 도식화한 것이다. 파란색은 유저, 대단원별 레이팅 차이의 평균, 빨간색은 차이의 표준편차이다. 실제 레이팅, 즉 유저의 진짜 실력의 경우 측정할 방법이 명확하지 않으므로 71번째 측정에서 100번째 측정까지의 값을 평균내었다. 이때 두 가지 패턴을 발견할 수 있다.

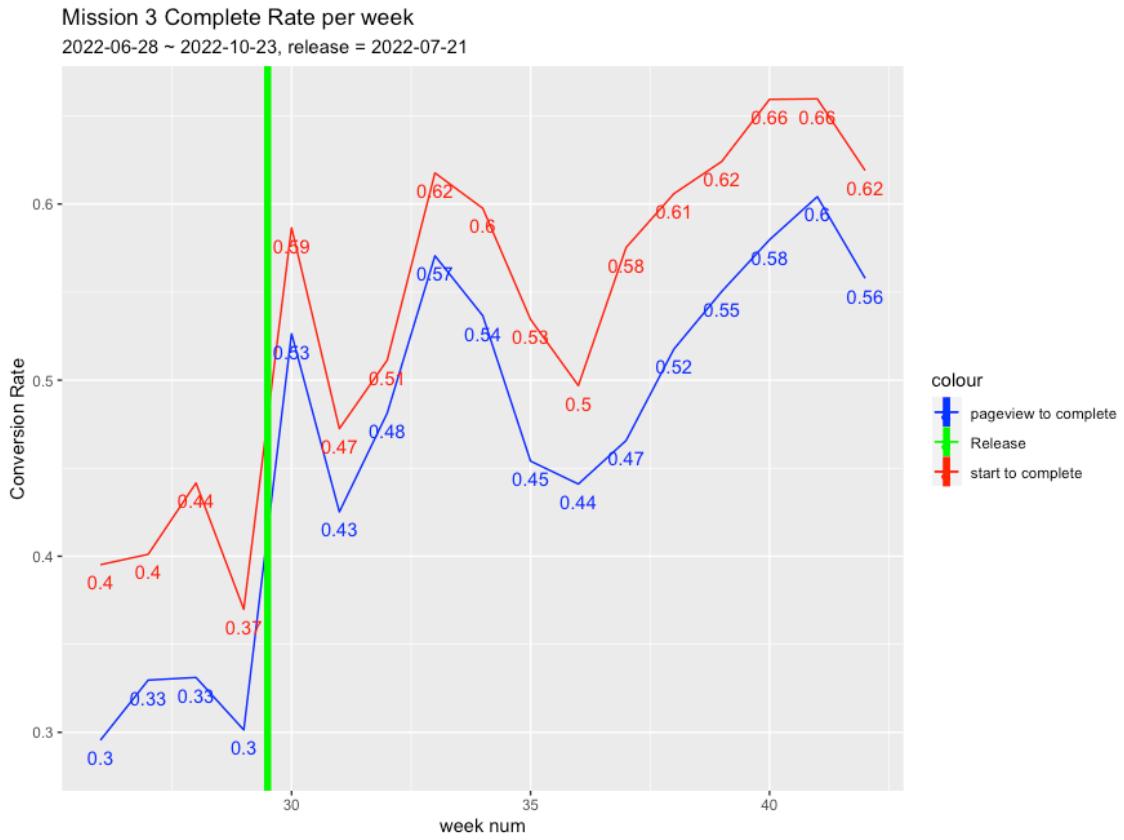
첫째로는 표준편차가 뚜렷하게 감소한다는 점이다. 실제 레이팅을 뒤쪽 값을 평균내어 썼기 때문에 85 즈음부터는 표준편차가 증가하는 경향을 보이긴 하지만, 종속성이 매우 낮은 수준의 1 ~ 20번째 데이터를 보더라도 레이팅의 측정 정확도는 뚜렷히 상승함을 알 수 있다.

둘째로는 레이팅의 평균이 꾸준히 상승한다는 점이다. 측정 표준편차가 줄어드는 것을 포함하여 고려해볼 때 유저 레이팅 평균이 문제 풀이 횟수에 비례하여 계속 상승

하는 것은 유저의 실력 상승을 어느정도 이야기한다고 볼 수 있다.

(4) 배포 이후 서비스 데이터

가. 유저의 과제 수행력



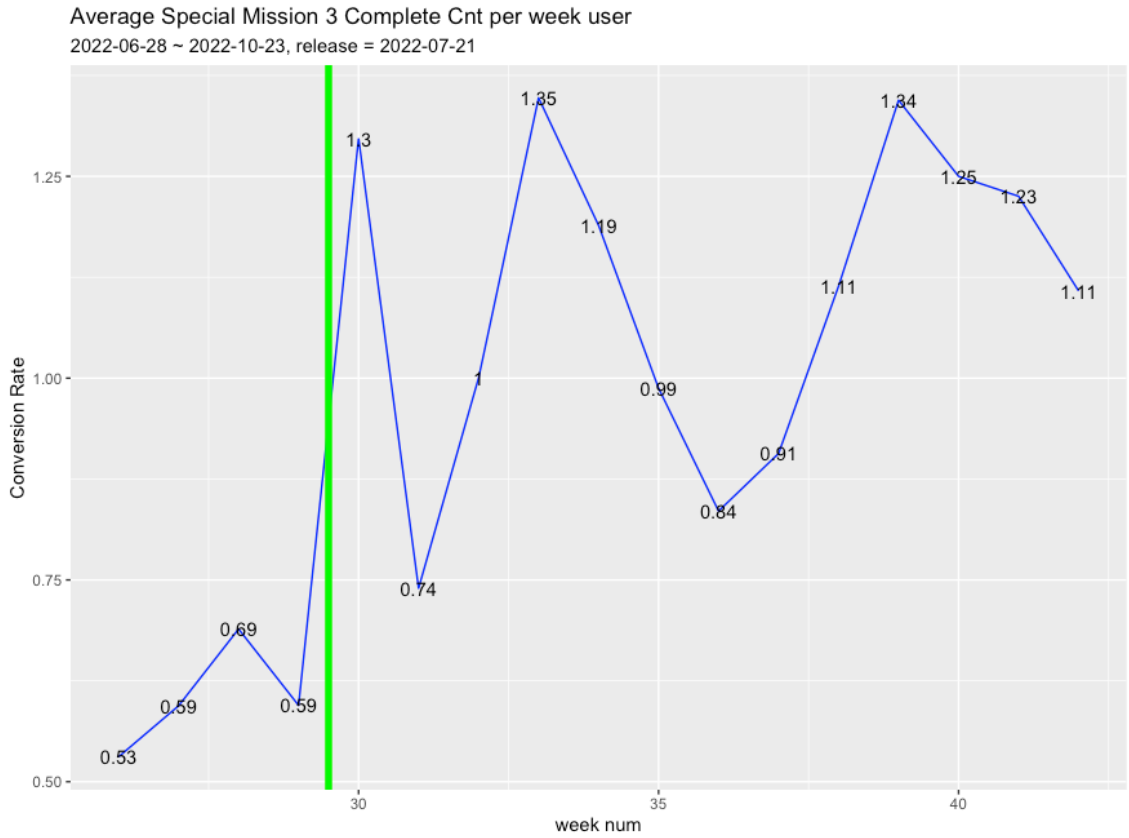
그래프 24 주별 미션 3 완료율 그래프

유저가 미션 페이지를 방문한 이후 미션 3 완료율을 나타내는 그래프가 파란색, 미션 1 시작 이후 미션 3 완료율을 나타내는 그래프가 빨간색이다. 서비스에 멤버십을 런칭한 6/27 이후 시점으로 계산하였다.

파란색 그래프의 경우 유저의 과제 수행력을 의미한다. 적당한 난이도의 문제가 출제되었을 때 유저의 과제 수행력을 높임을 어느정도 입증했다고 볼 수 있다. 빨간색 그래프의 경우 유저가 문제를 풀려고 마음 먹었을 때, 문제가 얼마나 이를 방해하였는지 나타내는 지표이다. 실제 해당 기간에 유저가 풀이한 문제 평균 난이도는 오히려 증가하였는데, 그에 반해 유저의 완료율은 크게 증가하였다. 그 원인 중 하나로는 유

저 컨디션 레이팅 도입이 있다. 실제 아이의 실력 측정과 별개로 한 문제를 틀렸을 때 - 유저가 이탈할 요인이 가장 클 때 그 다음 문제로 쉬운 문제를 줌으로써 유저 밀착을 높인다는 가설이 성공적으로 작동한 것이다.

나. 유저의 일일 학습



그래프 25 주별 미션 3 평균 완료 횟수 그래프

유저의 일일 학습을 나타내는 지표인 주당 미션 3 평균 제출 횟수이다. 1회를 한참 밑돌았던 기존 데이터와 다르게, 배포 이후 1점 초반대를 기록하는 등 고무적인 재방문(retention)을 기록하였다.

3.2 고찰 및 제언

(1) 고찰

이번 연구를 통해 엘로 레이팅 시스템을 활용한 문제-아이 반응 모형을 제작, 최적화하고, 이를 통해 문제 출제 솔루션을 개발, 배포하였다. 문제-아이 반응 모형의 경우 수만의 데이터셋을 활용해 pre-trained된 문제 초기화 데이터를 활용해 방금 서비스에 들어온 유저의 null data를 빠르게 실제 유저의 실력으로 수렴시키는 로직을 제작하였다.

개발된 모형은 실제 서비스 데이터를 활용해 검증되었다. Cross Entropy Loss를 포함해 여러 검증을 통과하여 실제로 유저의 실력을 잘 tracking함을 실험적으로 확인할 수 있었다. 또한, 연구 과정 중 1PL Rasch Model의 Gradient Descent Method와 동치임을 증명함으로써 이론적으로도 완성적인 모델임을 입증하였다.

서비스 기획, 개발, 배포 과정을 통해 '오늘의 미션' 컨셉의 문제 출제 솔루션을 런칭하였으며, 배포 후 지속적으로 데이터를 트래킹하였다. 서비스 재방문, 밀착도 등의 지속 개선을 통해 유저 맞춤형 솔루션의 활용 가능성을 보여주었다. 유저 실력을 측정하는 동시에 실력 기반 맞춤형 솔루션을 제시하여 문제 이탈 원인을 줄이는 방식이 실제로 동작함을 보여줌으로써 앞으로 Edtech에서 개인화 솔루션의 방향성을 일부 찾았다 할 수 있다.

(2) 제언

가. 3PL IRT 모형 도입

이번 연구에서는 1PL Rasch 모형을 도입함으로써 문항 난이도만 반영하였다. 다만 문제의 하위 데이터로써 c (문항 추측도 : 문항을 실력 -inf인 상태에서 맞힐 확률)를 하나하나 마킹하여 구한 상태였으므로 c를 각 문제에 대해 업데이트하지 않는 상수로 놓고 모형 개발을 진행하고자 하였다.

$$p_{3PL} = \frac{1}{1+10^{\frac{b\times(a-u)}{400}}}\times(1-c)+c,\ \Phi = \frac{1}{1+10^{\frac{b\times(a-u)}{400}}}$$

$$\frac{\partial p}{\partial a} = \frac{(p-c)(1-p)}{1-c}\times\frac{b\times\log 10}{400}$$

위 두 번째 식처럼 b에 대해서도 모수 업데이트를 진행하면 충분하다. 다만 특정 문항의 경우 데이터 개수에 따라 문항 변별도가 음수가 나오는 등 모형의 안정성에 일부 문제가 있어 이번 연구에서는 폐기하였다.

나. 풀이 시간 데이터 미활용

아이 학습 데이터에서 정보량이 많이 담긴 데이터 중 하나가 풀이 시간 데이터이다. 아이가 어느 문제, 어느 유형에 더 시간을 많이 썼냐는 단순히 아이가 '잘 푼다, 못 푼다'를 떠나 또 다른 단위에서의 실력 측정이 가능하다. 예를 들어 풀이 시간 레이팅을 만들 수 있을 것이다. 실제로 푸아송 GLM을 활용해 log(time_taken + 1)을 예측하는 모형을 개발하기도 하였다. 생각보다 잘 동작하나, 클라이언트 단에서 활용 계획을 찾지 못해 폐기되었다. 후속 연구에서는 풀이 시간 데이터를 적극 활용하여 문제 추천 알고리즘에 반영할 수 있는 아이디어가 필요하다.

다. 대단원 간 개념 연결, 단원 메타데이터 미활용

이번 연구에서는 각 대단원을 시작할 때 항상 레이팅을 1200점으로 고정하였다. 이는 데이터 품질 면에서 상당히 손해인데, 전 단원에서 1500점으로 마무리했던 아이라면 이번 단원에서 적어도 1350점은 받아도 무방하기 때문이다. 그럼에도 불구하고 시도하지 않았던 이유는 일괄적으로 값을 매기는 방식이 정량적이지 않을뿐더러, 단원이 도형, 대수, 경우의 수 등 다양하여 연결의 강약을 정확히 측정할 수 없었기 때문이다. 단원 메타데이터를 활용하지 못한 점이 더욱 빠른 수렴을 시키지 못한 것에 일부 방해가 되었음을 제언으로 남기며, 후속 연구에서는 네트워크를 활용한 인과 추론 등도 진행할 것이다.

참고문헌

정희경, 《엄마가 만드는 초등 수학 자신감》. 한빛라이프.

박태성, 이승연, 《범주형 자료분석 개론 제3판》. WILEY. 자유아카데미.

Judea Pearl, Madelyn Glymour, Nicholas P.Jewell. 김미정 역. 《의학 및 사회과학 연구를 위한 통계적 인과 추론 개정판》. 교우출판사.

Matthieu J.S. Brinkhuis, Gunter Maris. 『Dynamic Parameter Estimation in Student Monitoring Systems』. Measurement and Research Department Reports 2009-1. January 2009.

Neumann, C et al. 『Assessing dominance hierarchies: validation and advantages of progressive evaluation with Elo-rating』. LJMU research online. 2011.

Mark E.Glickman. 『A Comprehesive Guide to Chess Ratings』. Department of Mathematics. Boston University.

Margaret Martinez. 『What is Personalized Learning?』. The eLearning Developers' Journal. May 2002.

Soude Fazeli, Hendrik Drachsler, Peter Sloep. 『Applying Recommender Systems for Learning Analytics: A Tutorial』. Welten Institute, Research Centre for Learning, Teaching and Technology.

Sergio Miranda, Giovannina Albano. 『Personalized Learning in Mathematic s』. Journal of E-learning and Knowledge Society. March 2015.

Thoufeeq Ahmed Syed, Vasile Palade, Rahat Iqbal, Smitha Sunil Kumaran Nair. 『A Personalized Learning Recommedation System Architecture for Learning Management System』. Middle East College, Mustcat, Omam. Coventry University, Coventry, U.K.

Hee Jing Bang, Linlin Li, Kylie Flynn. 『Efficacy of an Adaptive Game-Based Math Learning App to Support Personalized Learning and Improve Early

Elementary School Students' Learning』. Early Childhood Education Journal. February 2022.

Xueying Tang, et al. 『A reinforcement learning approach to personalized learning recommendation systems』. British Psychological Society. 2019.

초등 방학 탐구 생활 시리즈, 〈연산, 교과 수학, 사고력 수학, 다른 건가요?〉. 네이버 포스트. 검색 일자. 2023.05.31.
https://m.post.naver.com/viewer/postView.naver?volumeNo=33357390&memberNo=904209

김석우, 교육평가이론(문항반응이론) 강의자료,
http://www.kocw.net/home/cview.do?mty=p&kemId=1266989.
부산대학교. 2017년.