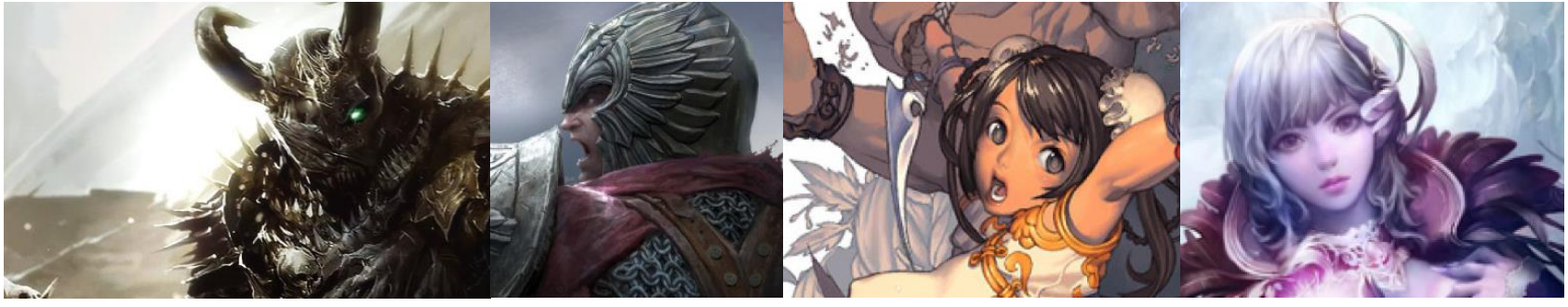
2019 빅콘테스트 대회 설명

Analysis 분야 챔피언 리그



2019. 7.17. 엔씨소프트

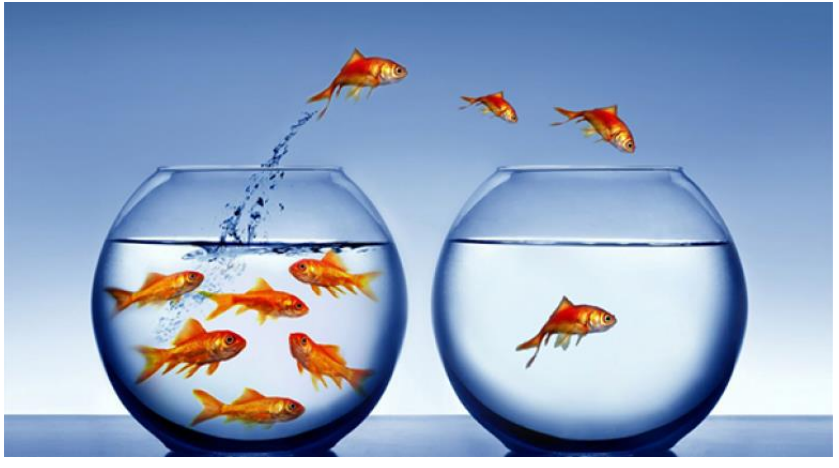
2019 빅콘테스트
2019 BIG CONTEST

개요

개요

대회 주제

➢ **리니지** 고객(유저) 활동 데이터를 활용하여 잔존 가치를 고려한 **이탈 예측** 모형 개발



개요

분석 대상

- 리니지
  - ✓ 1998년 9월 1일부터 엔씨소프트에서 서비스 중인 MMORPG
  - ✓ 2016년 기준 누적 매출 3조 2천억, 전 세계 누적 이용자수 2천만명
  - ✓ https://lineage.plaync.com/

개요

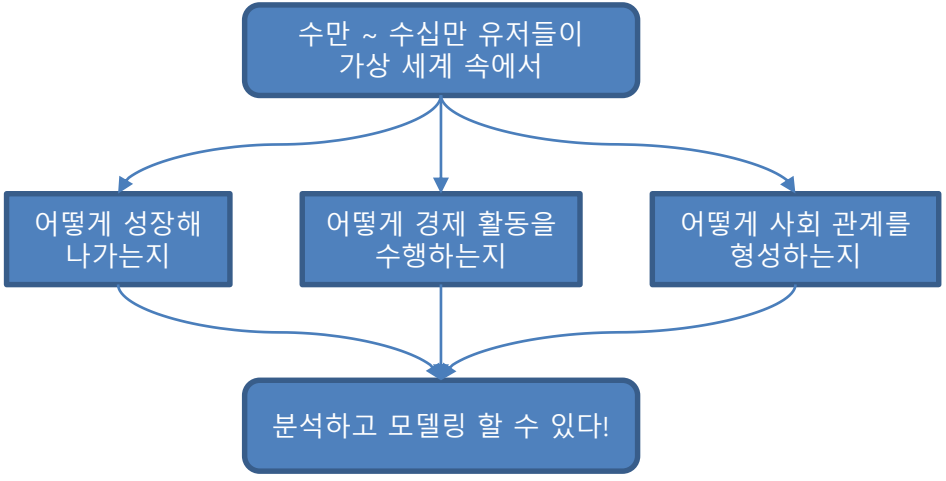
리니지의 특징

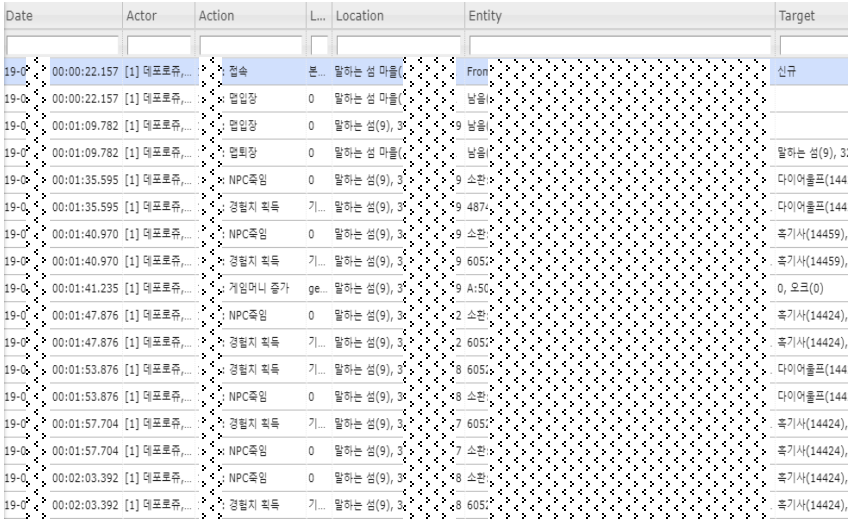
- 높은 자유도에 기반하여 현실과 유사한 다양한 활동 가능
  - ✓ 성장 및 경제 활동: 퀘스트, 레벨업, 사냥, 낚시, 물물교환 및 상업 활동
  - ✓ 사회 활동: 친구, 혈맹, 파티, 결혼, 전투, 채팅
  - ✓ 그 외 유저 스스로 상호 교류를 통해 다양한 컨텐츠 생성

개요

데이터 분석 측면에서 본 게임 데이터의 매력

- 이런 다양한 활동들이 모두 데이터로 기록됨
  - ✓ 누가, 언제, 어디서, 무엇을, 어떻게 하고 있는지 관찰 가능
  - ✓ 현실에서 접하기 힘든 고품질 데이터



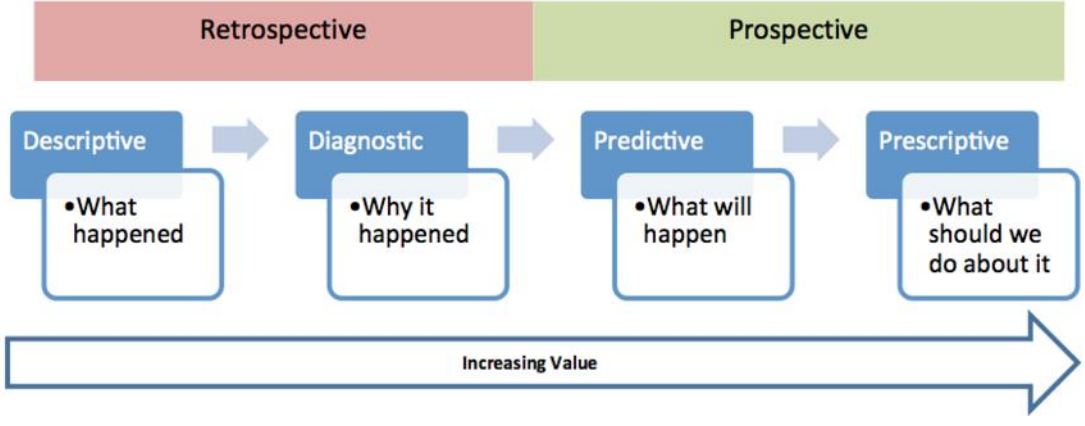


Date	Actor	Action	L...	Location	Entity	Target
19-0	00:00:22.157 [1] 데포로쥬,...	접속	본...	말하는 섬 마을(	Fron	신규
19-0	00:00:22.157 [1] 데포로쥬,...	맵입장	0	말하는 섬 마을(	남을(	
19-0	00:01:09.782 [1] 데포로쥬,...	맵입장	0	말하는 섬(9), 3	9 남을(	
19-0	00:01:09.782 [1] 데포로쥬,...	맵퇴장	0	말하는 섬 마을(	남을(	말하는 섬(9), 3
19-0	00:01:35.595 [1] 데포로쥬,...	NPC죽임	0	말하는 섬(9), 3	9 소환	다이어울프(144
19-0	00:01:35.595 [1] 데포로쥬,...	경험치 획득	기...	말하는 섬(9), 3	9 487	다이어울프(144
19-0	00:01:40.970 [1] 데포로쥬,...	NPC죽임	0	말하는 섬(9), 3	9 소환	흑기사(14459),
19-0	00:01:40.970 [1] 데포로쥬,...	경험치 획득	기...	말하는 섬(9), 3	9 605	흑기사(14459),
19-0	00:01:41.235 [1] 데포로쥬,...	게임머니 증가	ge...	말하는 섬(9), 3	9 A:50	0, 오크(0)
19-0	00:01:47.876 [1] 데포로쥬,...	NPC죽임	0	말하는 섬(9), 3	2 소환	흑기사(14424),
19-0	00:01:47.876 [1] 데포로쥬,...	경험치 획득	기...	말하는 섬(9), 3	2 605	흑기사(14424),
19-0	00:01:53.876 [1] 데포로쥬,...	경험치 획득	기...	말하는 섬(9), 3	8 605	다이어울프(144
19-0	00:01:53.876 [1] 데포로쥬,...	NPC죽임	0	말하는 섬(9), 3	8 소환	다이어울프(144
19-0	00:01:57.704 [1] 데포로쥬,...	경험치 획득	기...	말하는 섬(9), 3	7 605	흑기사(14424),
19-0	00:01:57.704 [1] 데포로쥬,...	NPC죽임	0	말하는 섬(9), 3	7 소환	흑기사(14424),
19-0	00:02:03.392 [1] 데포로쥬,...	NPC죽임	0	말하는 섬(9), 3	8 소환	흑기사(14424),
19-0	00:02:03.392 [1] 데포로쥬,...	경험치 획득	기...	말하는 섬(9), 3	8 605	흑기사(14424),

개요

이탈 분석의 목적

- 인과 분석
  - ✓ 이탈 고객 집단 분석을 통해 이탈의 원인 파악 → 원인 제거
  - ✓ 관측된 데이터를 통해 인과 관계를 파악하는 것은 매우 어려움 (**Correlation ≠ Causation**)
- **예측 분석 ← 이번 대회의 주제**
  - ✓ 이탈 징후를 보이는 고객을 사전에 선별 → 인센티브 제공을 통해 잔존 유도

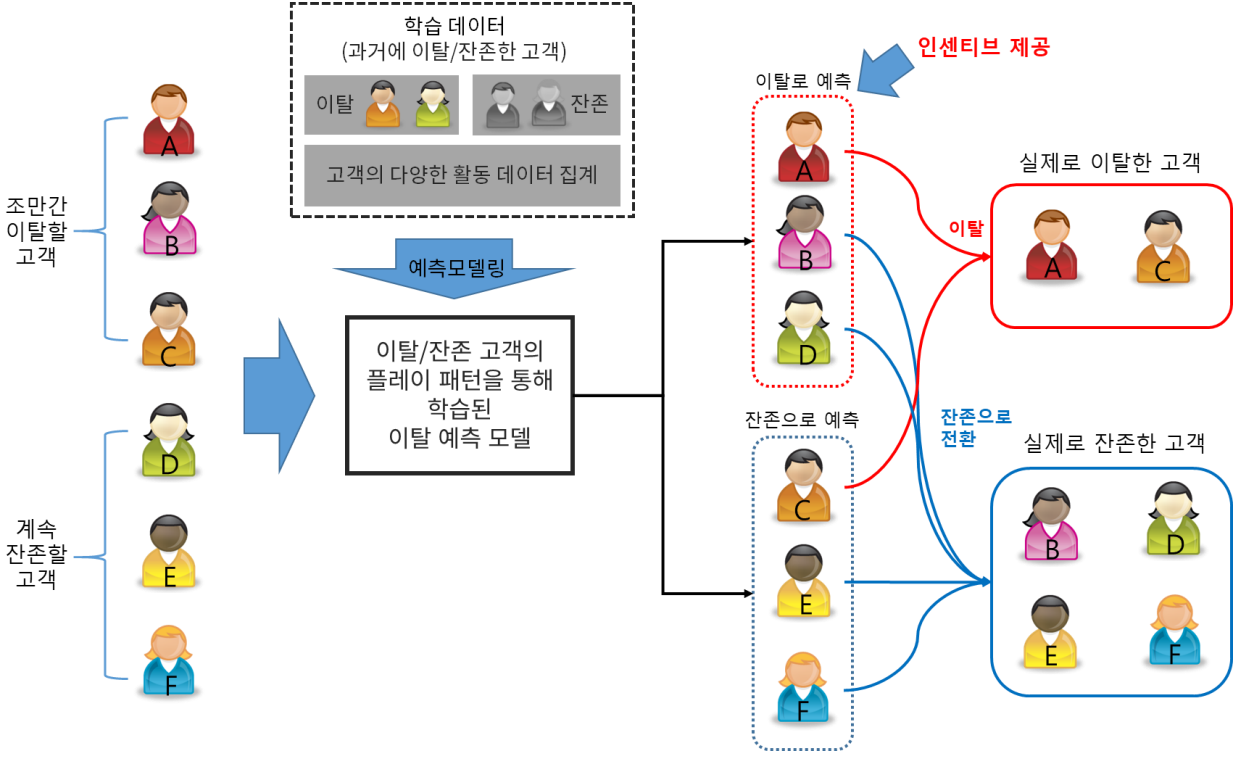


출처: https://dai-global-developments.com/articles/disruption-for-good

개요

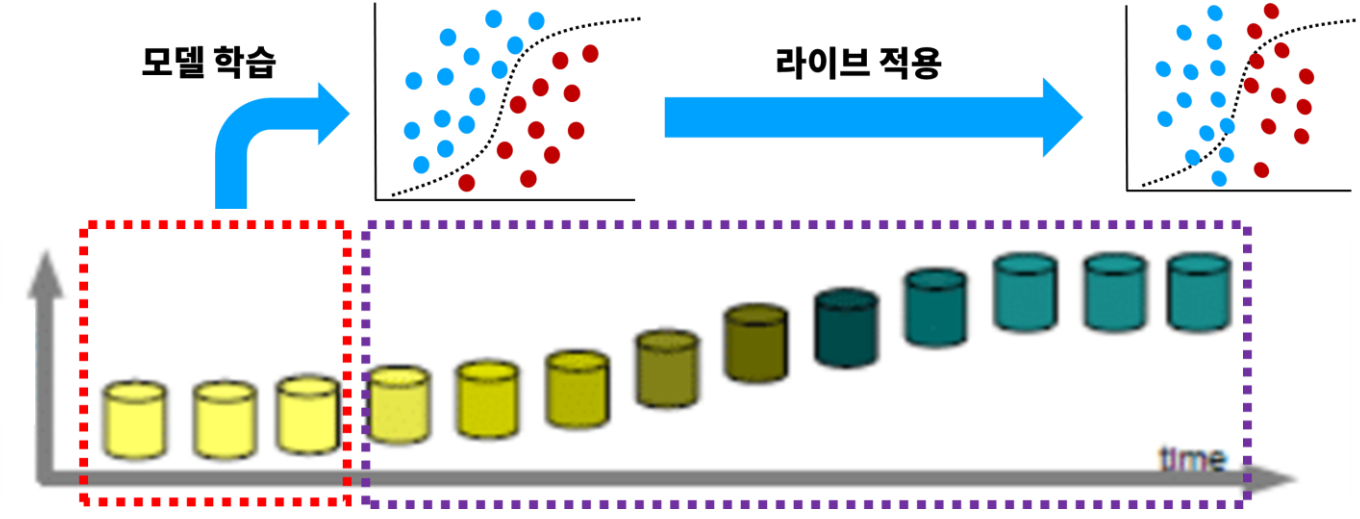
고객 이탈 예측

➢ 일반적인 고객 이탈 예측 모델 적용 시나리오



일반적인 이탈 예측 모형의 문제점 #1

- 시간에 따른 패턴의 변화를 고려하지 않은 모형
  - ✓ 데이터의 통계적 특성이 변하면 과거 데이터의 패턴을 학습한 모델의 유용성이 점차 떨어짐

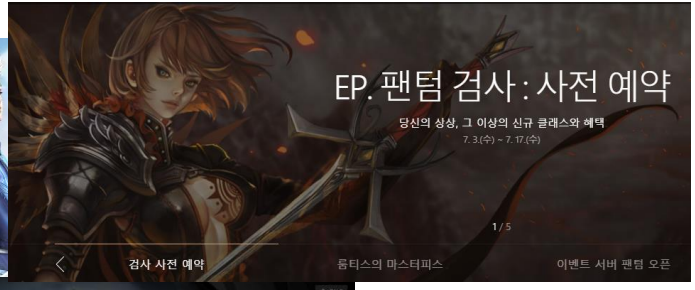


출처:https://www.researchgate.net/publication/270787580_A_Survey_on_Supervised_Classification_on_Data_Streams

개요

일반적인 이탈 예측 모형의 문제점 #1

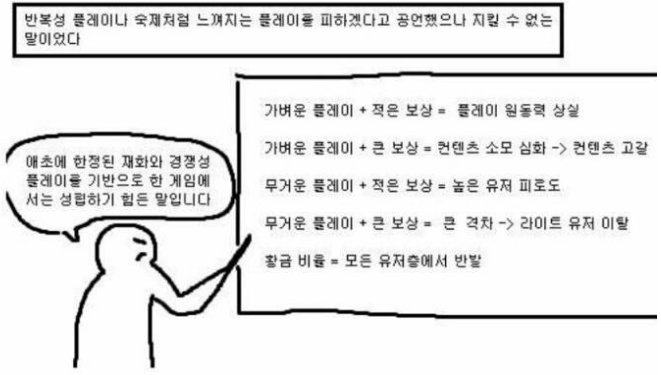
- 온라인 게임 데이터의 특징
  - ✓ 빈번한 게임 업데이트 및 이벤트로 인한
    - 게임 밸런스 변화
    - 주요 컨텐츠 삭제 및 추가
    - 비즈니스 모델 변경



[일반] 업데이트 속도 가불기 공식 찾았다

ㅂㅎㅂㅎ 2018.12.27 13:30

조회 103 댓글 3 갤로그



느린 업데이트 -> 지루함 -> 헤비유저 이탈

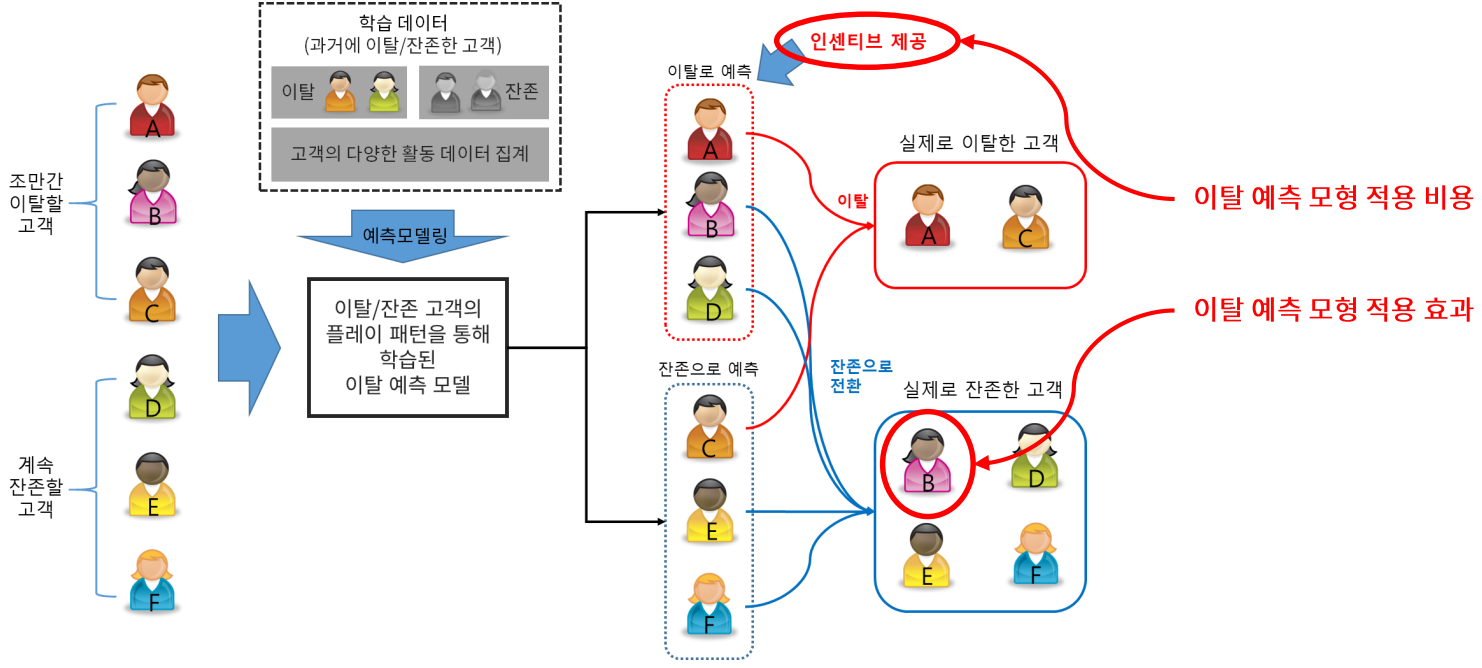
빠른 업데이트 -> 빡빡함 -> 라이트유저 이탈

적당한 속도 -> 모든 유저 이탈

출처: https://bbs.ruliweb.com/community/board/300143/read/40471624

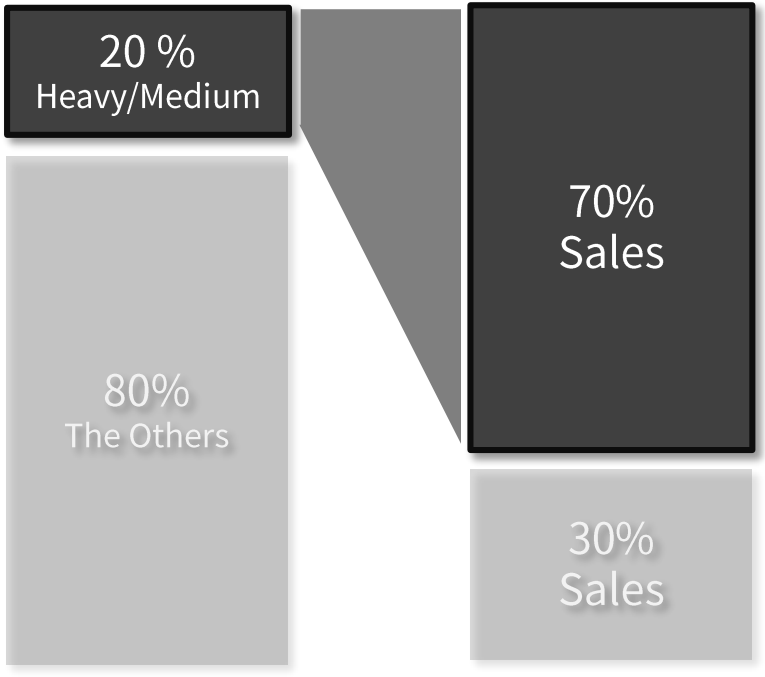
일반적인 이탈 예측 모형의 문제점 #2

- 기대 이익을 고려하지 않은 예측 모형
  - ✓ 이탈 예측의 실제 목표는 예측을 정확히 하는 것이 아니라 이탈 방지를 통한 잔존 가치 보존임
  - ✓ 기대 이익 = 이탈 예측 모형 적용 효과 – 비용 (**예측 정확도 ≠ 기대 이익**)



일반적인 이탈 예측 모형의 문제점 #2

- **잔존 가치**가 높은 고객에 대한 이탈 방지가 중요함
  - ✓ 악성 유저에 대한 이탈 예측이 필요할까?
  - ✓ 잔존 가치는 어떻게 추정해야 할까?
- 적절한 **인센티브**를 산정해야 함
  - ✓ 인센티브가 너무 낮으면 → 고객의 관심 얻을 수 없음
  - ✓ 인센티브가 너무 높으면 → 이익보다 손실이 더 큼
- 적절한 **인센티브 제공 시기** 선정도 중요함
  - ✓ 너무 빠르거나 늦으면 효과가 없음

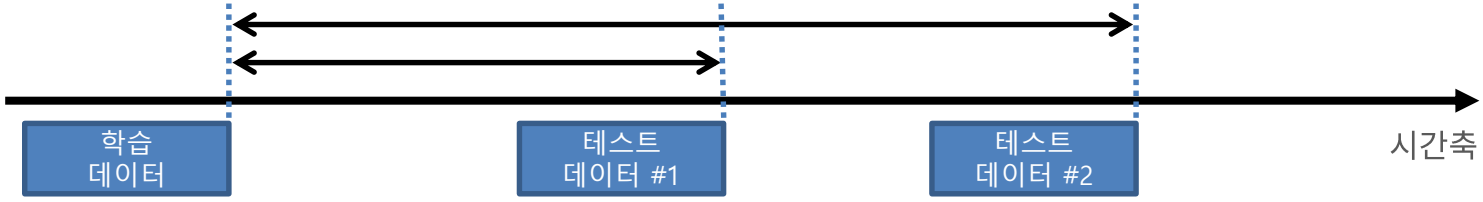


20%의 고객이 70%의 매출 기여

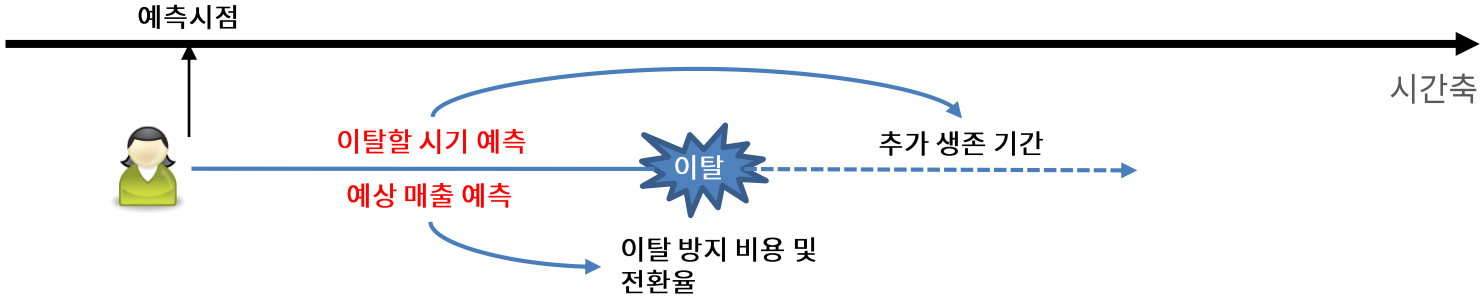
개요

문제의 의도 및 목표

- 시간의 변화에 강건한 모델 구축
  - ✓ 학습 데이터와 **시점이 다른 두 개의 테스트 데이터** 제공



- 고객의 잔존 가치를 고려한 모델 구축
  - ✓ 고객별 **예상 매출**과 **이탈 시기**를 예측하고 이를 이용한 이탈 방지 정책 적용 시 예상되는 기대 이익 평가
    - 기대 이익 = 전환율 ×(추가 생존 기간 × 단위 기간당 예상 매출) – 이탈 방지 비용

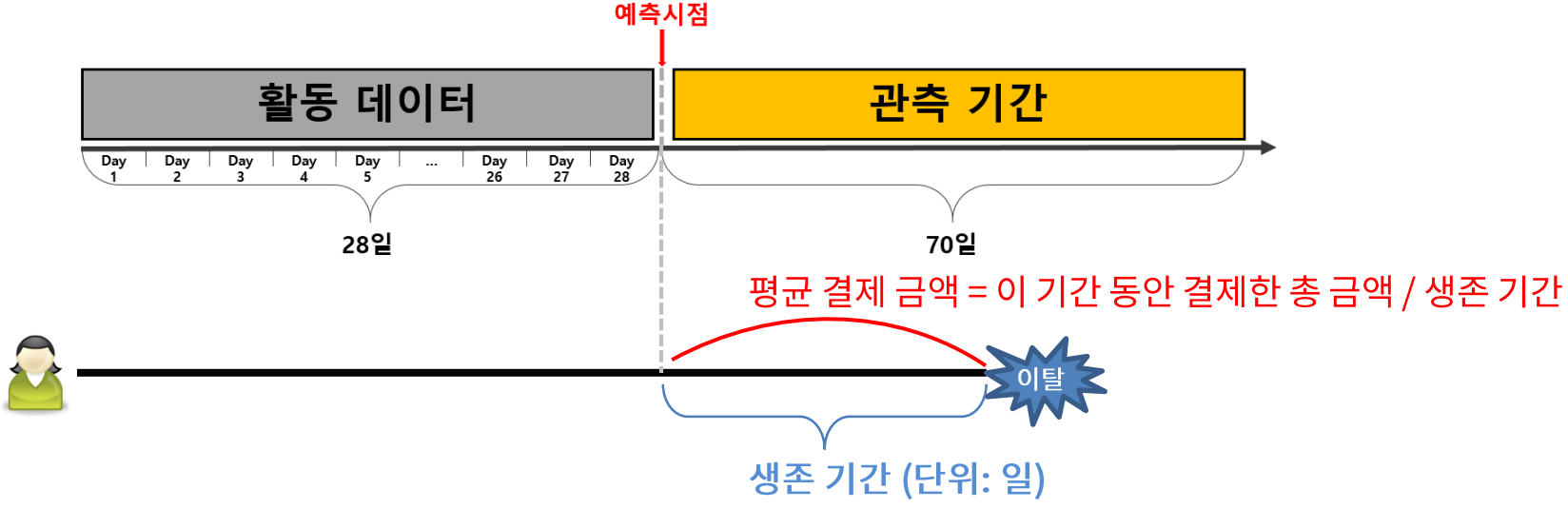


문제 설명

문제 설명

데이터 구성

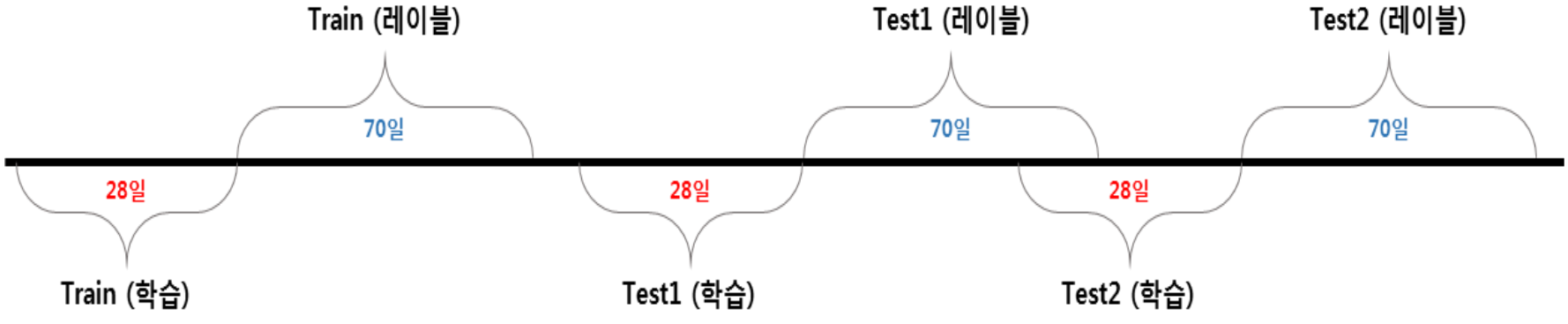
- 예측 시점에서 과거 28일간의 활동 데이터를 이용하여 모델 학습
- 예측 시점 이후 70일 간의 관측을 통해 집계된 실제 고객별 **이탈 시점(생존 기간)** 및 **평균 결제 금액** 예측
  - ✓ 64일 동안 이탈하지 않은 유저는 잔존으로 처리 (이탈 여부 판단 기간 7일 감안)
  - ✓ 예측 시점 이후 발생한 고객별 총 결제 금액을 활동 일수로 나눠 일 평균 결제 금액 집계



문제 설명

데이터 구성

- 학습 및 평가 데이터 구성 방식 및 규모
  - ✓ Train 데이터: 4만 계정
  - ✓ Test 데이터 1 & 2: 각각 2만 계정



문제 설명

데이터 구성

- 데이터의 종류 별 CSV 파일 제공 (총 16종)
  - ✓ 예측 대상은 유저 아이디 기준
  - ✓ 피처 데이터는 유저 아이디 기준과 캐릭터 아이디 기준이 혼재되어 있음
    - **하나의 유저는 다수의 캐릭터 보유 가능**

데이터 세트			데이터 내용	
Train	Test1	Test2		
train_label.csv	-	-	대상 유저들의 생존 기간 및 평균 결제 금액	Labels
train_activity.csv	test1_activity.csv	test2_activity.csv	대상 유저의 **캐릭터별** 활동 이력	Features
train_combat.csv	test1_combat.csv	test2_combat.csv	대상 유저의 **캐릭터별** 전투 이력	
train_pledge.csv	test1_pledge.csv	test2_pledge.csv	대상 유저 **캐릭터별** 소속 혈맹 전투 활동 정보	
train_trade.csv	test1_trade.csv	test2_trade.csv	대상 유저의 **캐릭터별** 거래 이력	
train_payment.csv	test1_payment.csv	test2_payment.csv	대상 유저의 일별 결제 금액	

문제 설명

데이터 구성 – 레이블 데이터

- train_label.csv
  - ✓ 각 **유저**의 생존 기간과 일별 평균 결제 금액 제공
  - ✓ 생존 기간은 1 ~ 64 값을 가지며, **64 는 잔존**을 의미함

변수	설명
acc_id	유저 아이디
survival_time	생존 기간(일)
amount_spent	일별 평균 결제 금액

문제 설명

데이터 구성 – 기본 활동 데이터

- train_activity.csv, test1_activity.csv, test2_activity.csv
  - ✓ 각 **캐릭터**의 일일 주요 활동 집계

변수	설명
day	날짜
acc_id	유저 아이디
char_id	캐릭터 아이디
server	캐릭터 서버
playtime	일일 플레이시간
npc_kill	NPC를 죽인 횟수
solo_exp	솔로 사냥 획득 경험치
party_exp	파티 사냥 획득 경험치
quest_exp	퀘스트 획득 경험치
boss_monster	보스 몬스터 타격 여부 (0=미타격 ,1= 타격)
death	캐릭터 사망 횟수
revive	부활 횟수
exp_recovery	경험치 복구 횟수(성당)
fishing	일일 낚시 시간
private_shop	일일 개인상점 운영 시간
game_money_change	일일 아데나 변동량
enchant_count	7레벨 이상 아이템 인첸트 시도 횟수

문제 설명

데이터 구성 – 거래 데이터

- train_trade.csv, test1_trade.csv, test2_trade.csv
  - ✓ **캐릭터** 간 일별 거래 (교환, 개인 상점) 이력

변수	설명
day	거래 발생 일
time	거래 발생 시간 (00:00:00 ~ 23:59:59)
type	거래 구분 (교환창 = 1, 개인상점 = 0)
server	거래 발생 서버
source_acc_id	주는/판매 유저 아이디
source_char_id	주는/판매 캐릭터 아이디
target_acc_id	받는/구매 유저 아이디
target_char_id	받는/구매 캐릭터 아이디
item_type	아이템 종류 weapon (무기), armor (방어구), accessory(액세서리), adena (아데나), spell (스킬북), enchant_scroll (강화주문서), etc (기타)
item_amount	거래 아이템 수량
item_price	거래 가격 - 교환창 거래 (Type=1)일 경우 NA

문제 설명

데이터 구성 – 전투(PvP) 데이터

- train_combat.csv, test1_combat.csv, test2_combat.csv
  - ✓ **캐릭터** 전투 활동 일일 집계

변수	설명
day	날짜
acc_id	유저 아이디
char_id	캐릭터 아이디
server	캐릭터 서버
class	**직업 (오른쪽 표 참조)**
level	**레벨 (오른쪽 표 참조)**
pledge_cnt	혈맹간 전투에 참여한 횟수
random_attacker_cnt	본인이 막피 공격을 행한 횟수
random_defender_cnt	막피 공격자로부터 공격을 받은 횟수
temp_cnt	단발성 전투 횟수
same_pledge_cnt	동일 혈맹원 간의 전투 횟수
etc_cnt	기타 전투 횟수
num_opponent	전투 상대 캐릭터 수

범주	직업
0	군주
1	기사
2	요정
3	마법사
4	다크엘프
5	용기사
6	환술사
7	전사

범주	레벨	범주	레벨
0	1~4	9	45~49
1	5~9	10	50~54
2	10~14	11	55~59
3	15~19	12	60~64
4	20~24	13	65~69
5	25~29	14	70~74
6	30~34	15	75~79
7	35~39	16	80~84
8	40~44	17	85 이상

문제 설명

데이터 구성 – 혈맹 데이터

- train_pledge.csv, test1_pledge.csv, test2_pledge.csv
  - ✓ **캐릭터** 소속 혈맹 구성원들의 전투 정보 일일 집계

변수	설명
day	날짜
acc_id	유저 아이디
char_id	캐릭터 아이디
server	캐릭터 서버
pledge_id	혈맹 아이디
play_char_cnt	게임에 접속한 혈맹원 수
combat_char_cnt	전투에 참여한 혈맹원 수
pledge_combat_cnt	혈맹 간 전투 횟수의 합
random_attacker_cnt	혈맹원 중 막피 전투를 행한 횟수의 합
random_defender_cnt	혈맹원 중 막피로부터 피해를 받은 횟수의 합
same_pledge_cnt	동일 혈맹원 간 전투 횟수의 합

문제 설명

데이터 구성 – 결제 데이터

- train_payment.csv, test1_payment.csv, test2_payment.csv
  - ✓ 각 **유저**의 일별 결제 금액

변수	설명
day	날짜
acc_id	유저 아이디
amount_spent	결제 금액

문제 설명

데이터 관련 사전 정보

- 데이터 익명화
  - ✓ 민감 정보 노출을 막기 위해 주요 식별 정보 마스킹 수행
    - 마스킹 대상: 계정/캐릭터 아이디, 서버 번호
  - ✓ 수치형 데이터는 원본 값을 표준 편차로 나누어 변환
  - ✓ 예시)

데이터 종류	변환 전	표준편차	변환 후
파티 사냥 경험치	2,235,212	16.4723	135695.1973919853
일일 낚시 시간	21	864.2	0.0242999305716269

문제 설명

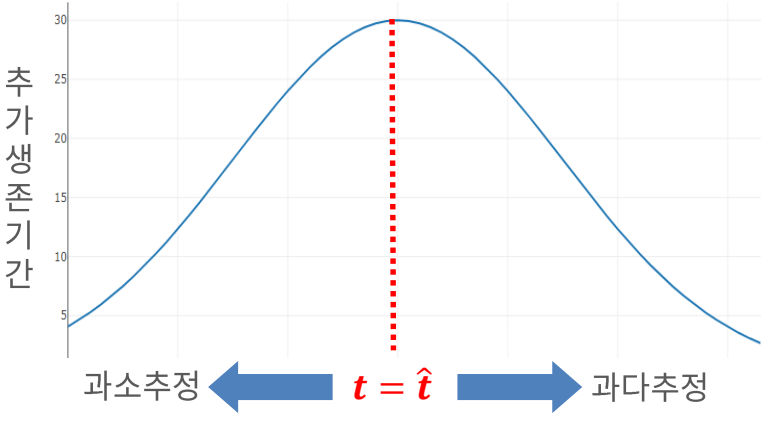
평가 방법

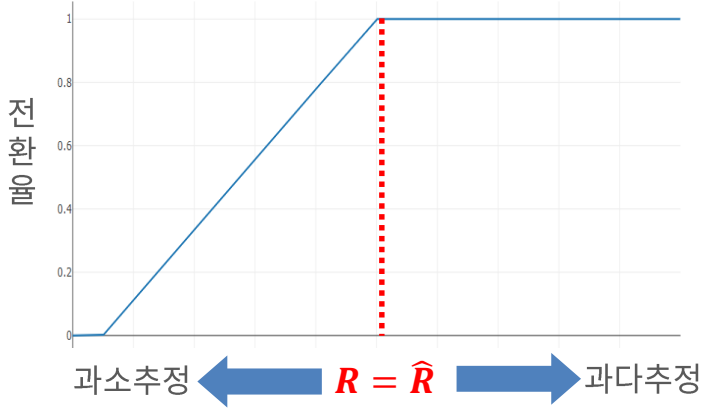
- 예측 성능 + 재현성 테스트 + 서류 심사
- 예측 성능
  - ✓ 참가팀이 제공한 '생존 기간 예측치($\hat{t}$)'와 '일별 평균 결제 금액 예측치($\hat{R}$)' 를 이용하여 계산한 기대 이익
  - ✓ 고객별 기대 이익의 총합이 높을수록 높은 점수 획득
- 재현성 테스트
  - ✓ 모델링 단계별 소스 코드 및 관련 자료 제출
  - ✓ 제출한 코드를 이용해 최종 예측 결과가 얼마나 쉽고 정확하게 재현되는지 측정
- 서류 심사
  - ✓ 탐사 분석, 전처리, 모델링 및 튜닝 등 전체 분석 과정에 대한 설명 문서
  - ✓ 체계적이고 논리적인 접근, 적절한 시각화

문제 설명

평가 방법 상세 – 예측 성능

- 기대 이익 = 잔존 가치 × 전환율 – 비용
  - ✓ 잔존 가치 = 추가 생존 기간($\boldsymbol{T}$) × 일 평균 결제 금액($\boldsymbol{R}$)
    - 추가 생존 기간은 생존 기간 예측치($\hat{\boldsymbol{t}}$)의 정확도로 결정 ($\hat{t} \geq 64$ *or* $t = 64$ 인 경우 0)
  - ✓ 전환율 = 이탈 예정 고객이 인센티브에 반응하여 잔존하게 되는 비율
    - 전환율은 일 평균 결제 금액 예측치($\widehat{\boldsymbol{R}}$)의 정확도로 결정
  - ✓ 비용 = 이탈 예정 고객에게 제공하는 인센티브
    - 비용은 추정된 잔존 가치의 1% 로 책정





문제 설명

평가 방법 상세 – 예측 성능 평가 모듈

- 참가팀이 직접 잔존 가치를 측정할 수 있도록 Python과 R로 구현된 평가 모듈 제공
  - ✓ 파일명 : score_function.py / score_function.r
- 사용방법
  - ✓ 예측 답안과 실제 답안 파일을 함수의 인자로 사용
  - ✓ 예측 답안 파일 스키마

컬럼명	컬럼설명
acc_id	유저 아이디
survival_time	생존기간 예측치
amount_spent	일별 평균 결제 금액 예측치

  - ✓ score_function.py 사용 예시

```
In [2]: from score_function import score_function
   ...: score_function('predict.csv','true.csv')
56319.66765172657
```

문제 설명

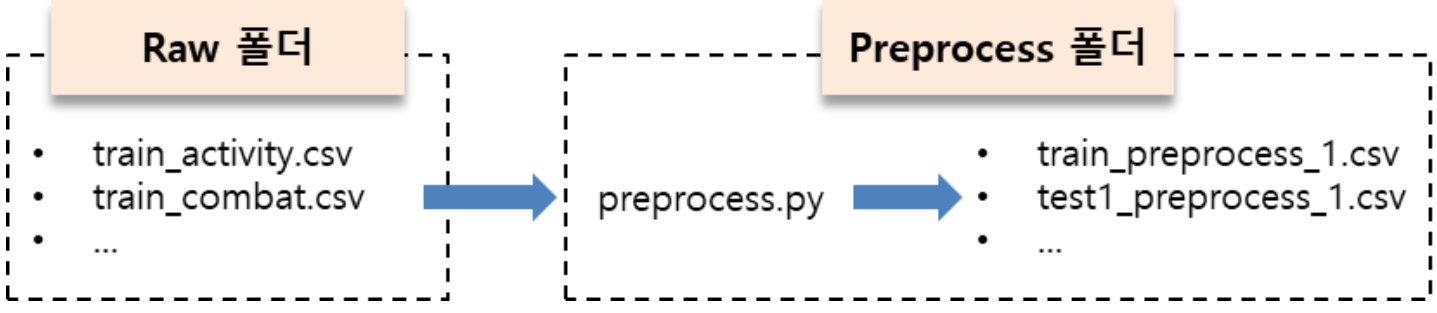
평가 방법 상세 – 재현성 테스트

- 모델링 단계별로 폴더를 구분하여 소스 코드 및 관련 자료 저장
  - ✓ 아래 폴더들을 '팀이름.zip' 으로 압축하여 제출

폴더 이름	폴더 내용
raw	원본 데이터가 적재될 폴더 (**폴더만 생성하고 데이터는 넣지 않음**)
preprocess	원본 데이터 전처리 코드 및 전처리 결과 데이터
model	최종 모델 학습용 코드 및 모델 객체
predict	테스트 데이터와 모델을 이용하여 최종 답안지를 생성하는 코드 및 최종 답안지
etc	코드 실행 방법에 대한 설명 문서 (readme.txt 혹은 readme.pdf) 및 서류 심사용 문서

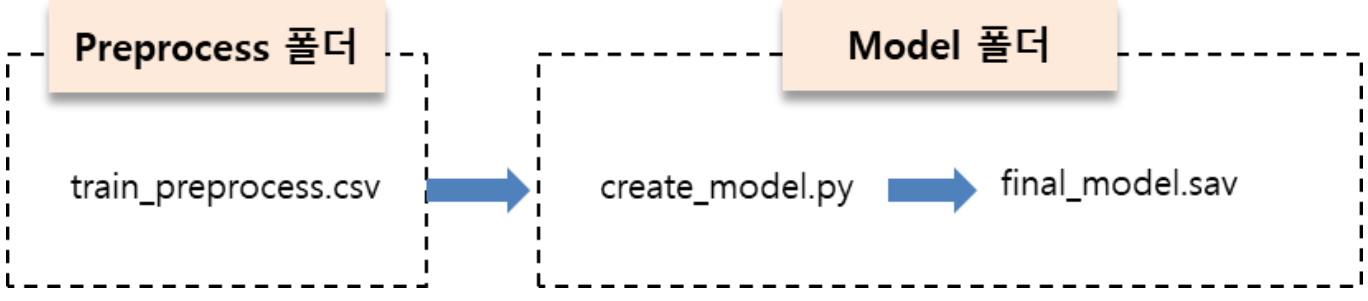
평가 방법 상세 – 재현성 테스트

- Preprocess 폴더 구성
  - ✓ 전처리 코드: raw 폴더에 위치한 원본 데이터 파일들을 불러들여, preprocess 폴더에 최종 모델의 input이 되는 데이터 파일을 저장하는 코드
    - Input : 팀이름/raw 폴더에 위치한 원본 데이터 파일
    - Output : 팀이름/preprocess 폴더에 최종 모델의 input이 되는 데이터 파일
  - ✓ 데이터 파일 명명 규칙 : 팀이름/preprocess/데이터세트_preprocess_숫자.확장자



평가 방법 상세 – 재현성 테스트

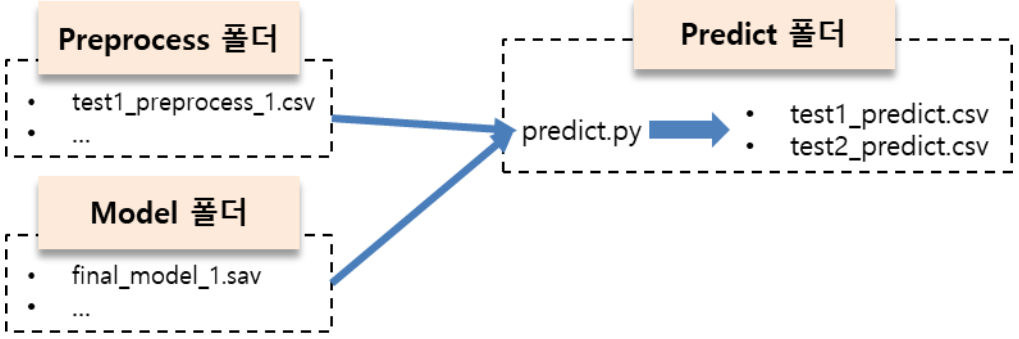
- Model 폴더 구성
  - ✓ 모델링 코드: preprocess 폴더에 위치한 데이터를 이용하여, model 폴더에 최종 답안지 생성에 사용되는 모델 객체를 저장하는 코드
    - Input : preprocess에 위치한 최종 모델의 input이 되는 데이터
    - Output : model 폴더에 최종 답안지 생성을 위해 사용되는 모델 객체
  - ✓ 모델 객체: 모델링 코드를 통해 생성되는 최종 예측 모델



문제 설명

평가 방법 상세 – 재현성 테스트

- Predict 폴더 구성
  - ✓ 예측 코드: preprocess 폴더에 위치한 데이터와,
    model 폴더에 위치한 최종 모델 객체를 불러들여 최종 답안지를 생성하는 코드
    - Input :
      1. preprocess에 위치한 최종 모델의 input이 되는 데이터
      2. model에 위치한 최종 모델 객체
    - Output :
      1. predict/test1_predict.csv
      2. predict/test2_predict.csv



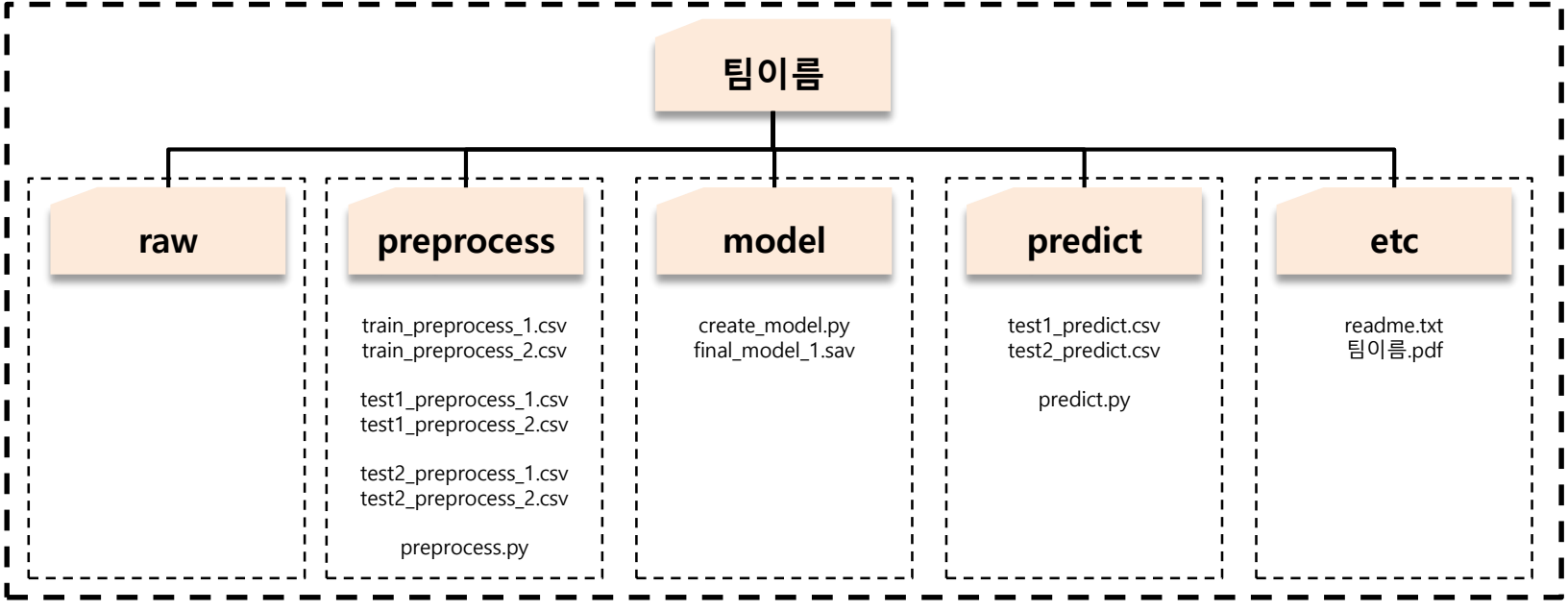
문제 설명

평가 방법 상세 – 재현성 테스트

- Etc 폴더 구성
  - ✓ **readme.txt**
    - 코드 실행에 필요한 패키지/라이브러리/모듈 리스트 및 실행 환경
    - 코드 실행 순서 및 방법
  - ✓ **팀이름.pdf**
    - 서류 심사용 자료

평가 방법 상세 – 재현성 테스트

- 최종 자료 구성 예시
  - ✓ Python 기준으로 작성된 예시이며 사용 가능한 언어 및 도구 제한은 없음 (단, 상용툴 이용 시 평가 불가)



문제 설명

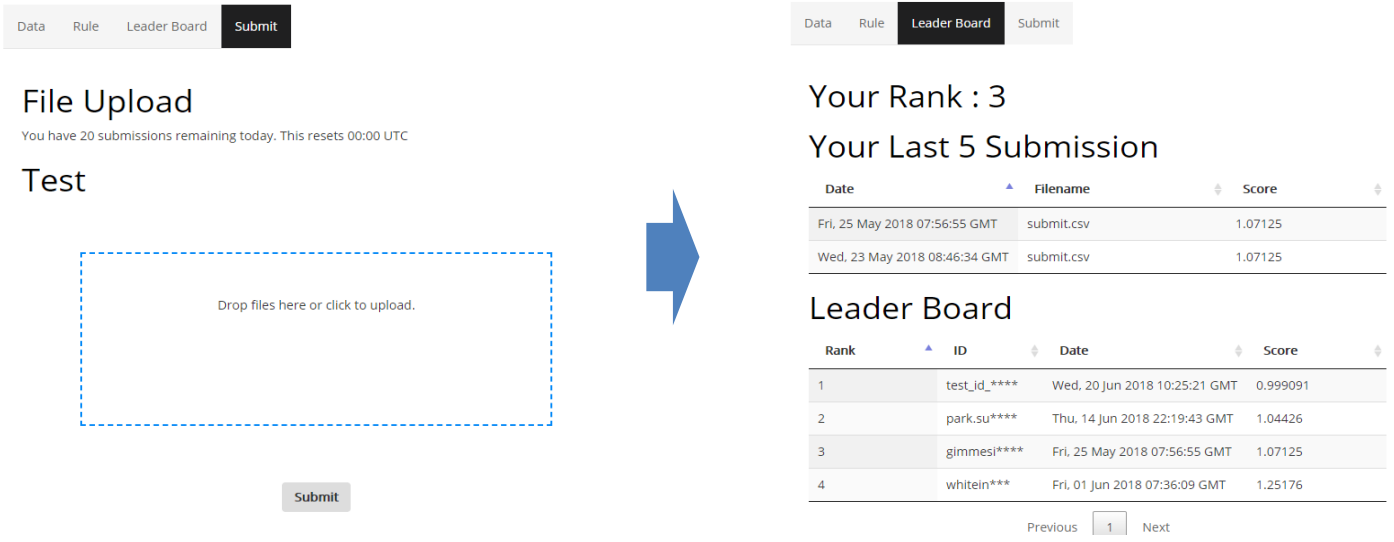
평가 방법 상세 – 서류 심사

- 본선 심사 발표 용도
  - ✓ 탐사 분석, 전처리, 모델링 및 튜닝 등 전체 분석 과정에 대한 설명 문서
  - ✓ 학습 알고리즘에 대한 설명 및 모델 해석
  - ✓ 체계적이고 논리적인 접근, 적절한 시각화

문제 설명

자율 평가

- 최종 결과 제출에 앞서 평가 데이터에 대한 예측 성능 확인 및 벤치마킹을 위한 Leader board 제공
- 어뷰징 방지를 위해 중간 평가는 **전체 평가 데이터의 20%만을 측정**한 결과 제공
- 점수 해킹 및 과도한 트래픽 부하를 막기 위해 지원자 별 **1일 5회로 횟수 제한**
- 자율 평가는 성능 확인 및 벤치마킹을 위해서만 제공되며, 최종 평가에 영향을 끼치지 않음



문제 설명

최종 평가

- 최종 예측 성능은 마지막에 제출한 예측 결과를 기준으로 전체 성능 측정하여 평가
- 예측 성능 및 재현성 테스트 결과 기준 상위 30개 팀에 대하여 서류 심사 진행
- 최종 후보 10개팀 선별하여 본선 심사 진행
  - ✓ 참가 인원 및 순위에 따라 서류 심사 대상자 및 본선 심사 대상자 수 변경 가능

문제 설명

수상자 혜택

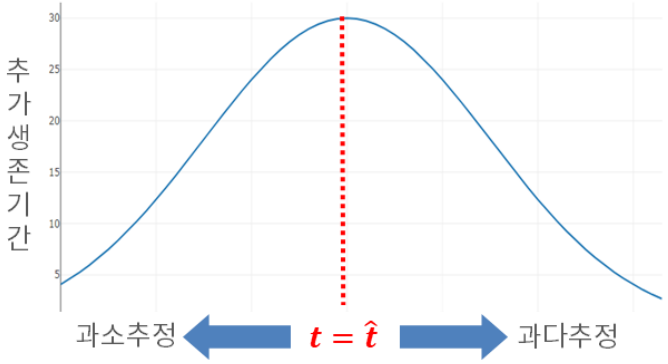
- 수상자는 Data Science 직무 채용 시 서류 전형 및 과제 테스트 우대
- 우수 평가자에 대해서 별도의 회사 세미나 및 면담 기회 제공

Q&A

Appendix

기대 이익 측정 공식

- 기대 이익 = 잔존 가치 × 전환율 – 비용
- 잔존 가치 = 추가 생존 기간($\boldsymbol{T}$) × 일 평균 결제 금액($\boldsymbol{R}$)
  - ✓ $T = \begin{cases} 0, & if\ \hat{t} \geq 64\ or\ t = 64 \\ 30 \times e^{-\frac{(t-\hat{t})^2}{2*15^2}}, & otherwise \end{cases}$
  - ✓ $\hat{t}$: 생존 기간 예측치, t: 생존 기간 실측치



- 비용 ($\boldsymbol{C}$) = 이탈 예정 고객에게 제공하는 인센티브
  - ✓ 잔존 유저로 예측되거나 예상 일 평균 결제 금액이 0인 경우 0,
    이탈 유저로 예측된 경우 $0.01 \times 30 \times \hat{R}$
  - ✓ $\hat{R}$: 일평균 결제 금액 예측치
- 전환율 ($\boldsymbol{\gamma}$) = 이탈 예정 고객이 인센티브에 반응하여 잔존되는 비율
  - ✓ $\gamma = \begin{cases} 0, & if\ \hat{C} < \frac{C_{opt}}{10}\ or\ C_{opt} = 0 \\ 1, & if\ \hat{C} \geq C_{opt} \\ \frac{10}{9}\left(\frac{\hat{C}}{C_{opt}} - 0.1\right), & otherwise \end{cases}$
  - ✓ $\hat{C}$: 예측 비용, C_{opt}: 적정 비용 ($R = \hat{R}$)