『데이터 분석가(DS / DA)』

# 매력점수와 가입process 측면에서의 glam 서비스 개선 아이디어

2020. 02. 01

지원자 민경수

CUPIST

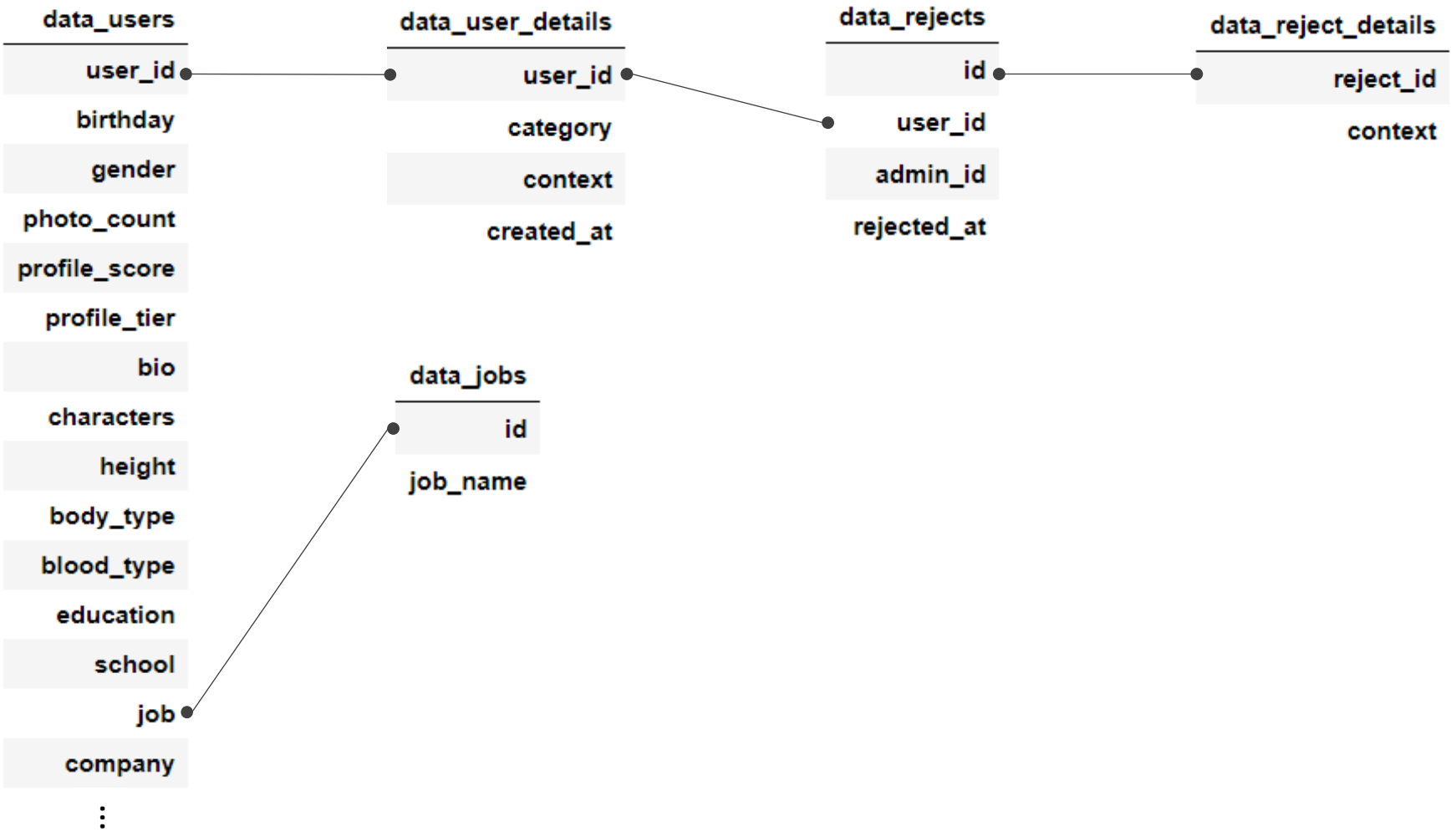
glam

# Table of Contents


1. 데이터 구조 & 선정
2. 데이터 정제
3. 탐색적 자료분석 (EDA)
4. 분석 1 : 프로필 매력점수 분석
5. 분석 2 : process 탈락 분석
6. 분석 시사점 및 개선의견

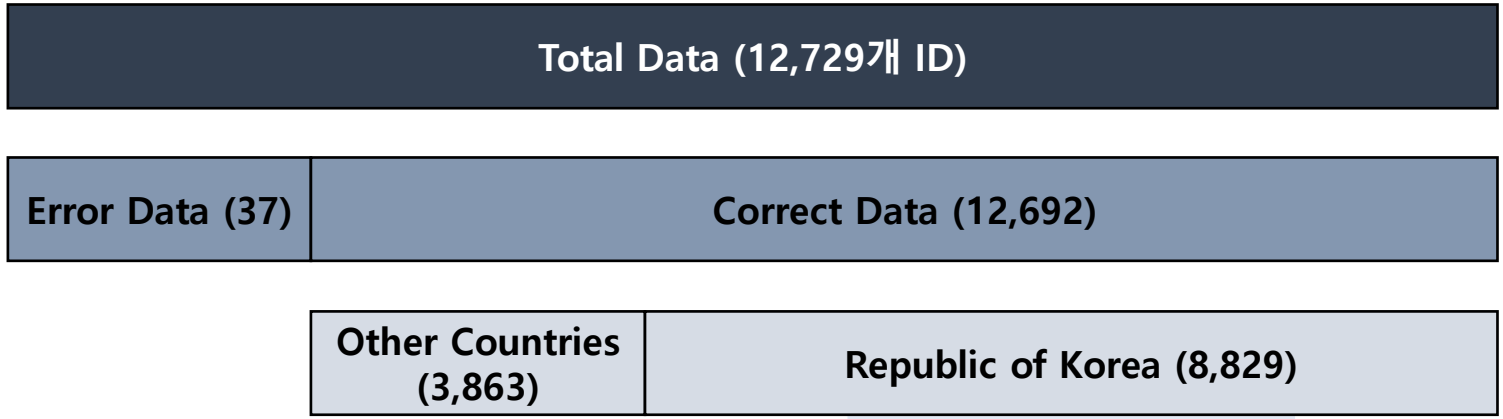
# 1. 데이터 구조 & 선정

제공받은 분석 데이터 5건은 아래와 같은 구조로 이루어졌으며, user_id 칼럼,rejects.id와 reject_id, job과 jobs_name.id 칼럼을 공유함.



# 2. 데이터 정제

다른 데이터셋과의 연결성이 가장 강한 data_users 데이터셋에서 발생한 parsing 애러 등을 제거하였음.
한국·미국 ·기타 국가중 "한국"에서 서비스를 이용하는 고객들의 데이터를 분석 대상으로 설정함.




분석 대상 데이터
(전체의 69.4%)

```
# trial and error를 통한 csv parsing error 발생 rows 제거
error_index = []
for i in data_users.index:
    if len(str(data_users.user_id[i])) != 7:   # id 형식이 아닌것 rows 제거
        #print(i, data_users.user_id[i])
        error_index.append(i)
    if i not in error_index and data_users.body_type[i] not in list(data_code_dict.keys()) and type(data_users.body_type[i]) != float:
        # body_type 코드 외의 values를 가진 rows 제거
        # nan value의 경우 제거하지 않음(아직 작업을 완료하지 않은 고객)
        #print(i, data_users.body_type[i])
        error_index.append(i)
    if i not in error_index and data_users.religion[i] not in list(data_code_dict.keys()) and type(data_users.religion[i]) != float:
        #print(i, data_users.religion[i])
        error_index.append(i)
    if i not in error_index and type(data_users
        error_index.append(i)
    if i not in error_index:
        try: float(data_users.height[i])
        except: error_index.append(i)

print('{} rows deleted due to parsing error'.form
data_users = data_users.drop(error_index)

# set data type correctly
data_users.profile_score = data_users.profile_sc
data_users.height = data_users.height.applyl
```

```
37 rows deleted due to parsing error
```

**2_1. divide data by country (Korea, US, ETC)**

```
id_kr = data_users[data_users.region == 'KR'].user_id
id_us = data_users[data_users.region == 'US'].user_id
id_etc = data_users[(data_users.region != 'KR') & (data_users.region != 'US')].user_id
```

```
data_users_kr = data_users[data_users.user_id.isin(id_kr)]
data_users_us = data_users[data_users.user_id.isin(id_us)]
data_users_etc = data_users[data_users.user_id.isin(id_etc)]
#print(len(data_users_kr), len(data_users_us), len(data_users_etc))
print('한국 이용자 수 : ', len(data_users_kr))
```

```
한국 이용자 수 :  8829
```

# 3. 탐색적 자료분석 (EDA, Exploratory Data Analysis)

**glam 서비스 이용자는 차단 여부/휴면 여부/탈퇴 여부에 따라 아래 표와 같이 MECE가 충족되는 8개 그룹으로 분류됨.**

*MECE(mutually exclusive and collectively exhaustive): 중복되거나 탈락되는 인자 없이 완벽하게 분류되는 것

| 차단 여부 (is_out) | 휴면 여부 (is_deactivated) | 탈퇴 여부 (is_blocked) | 설명 | 이용자 수 |
|---|---|---|---|---|
| X | X | X | 최근 5일 내 활동 이력이 있는 실사용 이용자 | 3,067 |
| X | X | ○ | 최근 5일 내 탈퇴한 이용자 | 1,250 |
| X | ○ | X | 접속한지 5일 이상 경과된 휴면 이용자 | 1,430 |
| X | ○ | ○ | 탈퇴한지 5일 이상 지난 이용자 | 2,555 |
| ○ | X | X | 최근 5일 내 활동 중 차단된 이용자, 탈퇴 x | 89 |
| ○ | X | ○ | 최근 5일 내 활동 중 차단된 이용자, 탈퇴 o | 15 |
| ○ | ○ | X | 차단된 지 5일 이상 지난 차단 이용자, 탈퇴 x | 391 |
| ○ | ○ | ○ | 차단된 지 5일 이상 지난 차단 이용자, 탈퇴 o | 32 |
| | | | | 8,829 |

## 3. 탐색적 자료분석 (EDA, Exploratory Data Analysis)

**차단 여부/휴면 여부/탈퇴 여부로 구분되는 8개의 그룹은 각각의 속성에 따라 다시 4개의 세그먼트로 통합할 수 있음.**
(e.g. 차단된 이용자들은 해당 이용자의 접속 시점이나 탈퇴 여부 등은 중요하지 않다고 판단되기 때문에 하나의 세그먼트로 통합)

| 차단 여부 (is_out) | 휴면 여부 (is_deactivated) | 탈퇴 여부 (is_blocked) | 설명 | 이용자 수 |
|---|---|---|---|---|
| X | X | X | 최근 5일 내 활동 이력이 있는 실사용 이용자 | 3,067 |
| X | X | ○ | 최근 5일 내 탈퇴한 이용자 | 1,250 |
| X | ○ | X | 접속한지 5일 이상 경과된 휴면 이용자 | 1,430 |
| X | ○ | ○ | 탈퇴한지 5일 이상 지난 이용자 | 2,555 |
| ○ | X | X | 최근 5일 내 활동 중 차단된 이용자, 탈퇴 x | 89 |
| ○ | X | ○ | 최근 5일 내 활동 중 차단된 이용자, 탈퇴 o | 15 |
| ○ | ○ | X | 차단된 지 5일 이상 지난 차단 이용자, 탈퇴 x | 391 |
| ○ | ○ | ○ | 차단된 지 5일 이상 지난 차단 이용자, 탈퇴 o | 32 |
| | | | | 8,829 |

Segment 1. 실사용 이용자

Segment 2. 탈퇴 이용자

Segment 3. 휴면 이용자

Segment 4. 차단 이용자

**Segment 1 (실사용 이용자) : 최근 5일 내 사용이력이 있는 이용자들로 구성된 세그먼트**

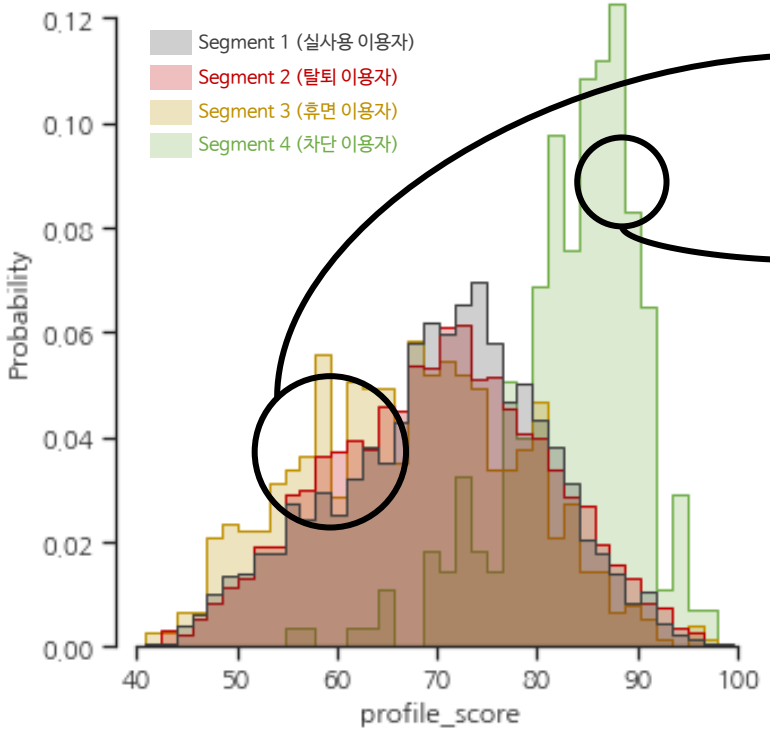
**Segment 2 (탈퇴 이용자) : 이용자 스스로 서비스를 탈퇴한 세그먼트**

**Segment 3 (휴면 이용자) : 최근 5일간 서비스를 이용치 않은 세그먼트**

**Segment 4 (차단 이용자) : 서비스 이용규칙 위빈으로 차단된 세그먼트**

# 4. 분석 1 : 프로필 매력점수 분석

**4개 segment별로 profile score(프로필 매력평가 점수)를 시각화한 결과는 아래와 같음.**
(추이 비교를 위해 분포확률로 정규화)




일반적인 이용자(실사용, 휴면, 탈퇴) segment의 경우 평균 점수가 67.4~70.4 사이인 정규분포 양상을 보임

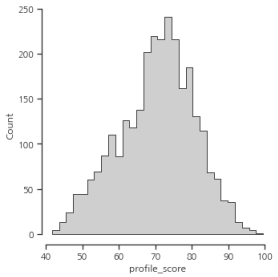
운영자에게 차단당한 이용자 segment의 경우, 평균 점수가 83.6으로 치중되어 있음

[분석 주제 1]
비차단&차단 이용자 segment간 매력점수 차이분석
프로필 매력점수의 분포 분석

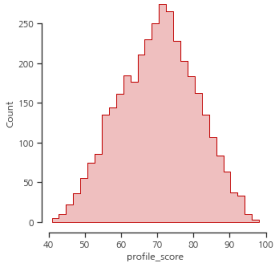
# 4. 분석 1 : 프로필 매력점수 분석

**차단당한 이용자(segment 4)의 경우, 다른 segment의 이용자들보다 상당히 높은 매력점수를 부여 받음.**

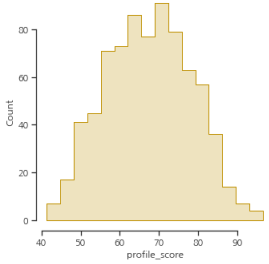
Segment 1. 실사용 이용자



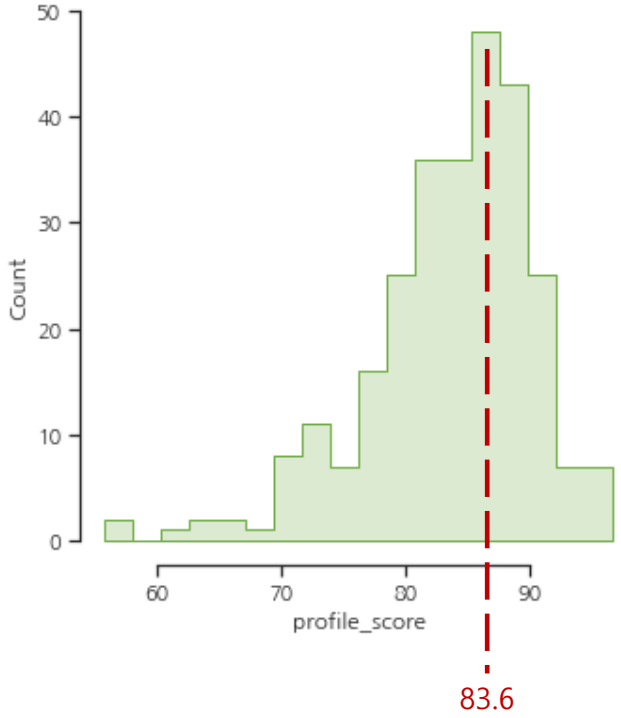

Segment 2. 탈퇴 이용자




Segment 3. 휴면 이용자



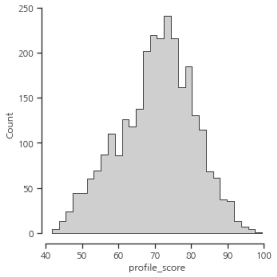

Segment 4. 차단 이용자

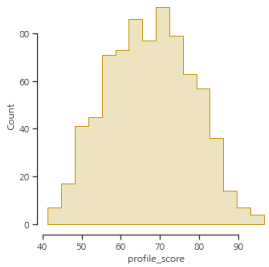


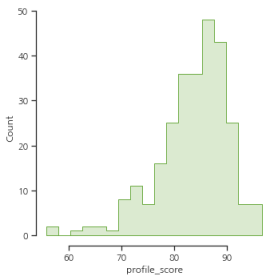

차단당한 이용자들은 차단당하지 않은 이용자들보다 더욱 매력적이라고 평가됨.
실제 이용자들의 평가이므로, 일반적인 기준으로 보았을 때 차단당한 이용자들의 프로필 사진은 실제 매력적인 여성의 사진이라고 볼 수 있음.

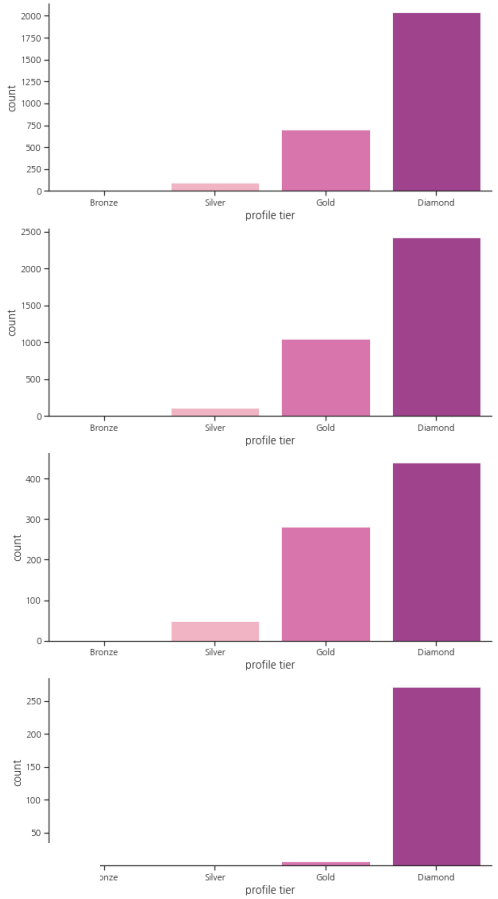
# 4. 분석 1 : 프로필 매력점수 분석

## 매력점수를 tier로 환산한 결과, Segment와 무관하게 가장 높은 tier인 Diamond tier의 분포가 압도적으로 높음

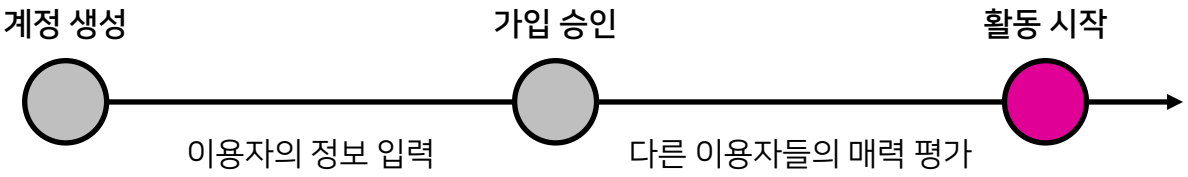




실사용 목적(수익적인 고객)을 위해서는 티어조정 필요. 10 10 50 30 이정도

- 가입 승인이 완료된 유저는 즉시 **프로필 매력 평가**를 시작하게 됩니다.
  - 매력 평가가 시작되면 글램에 접속 중인 수십 명~백여 명의 이성 유저들에게 프로필이 즉시 노출되며, 유저들은 이 프로필을 바탕으로 벌점을 부여합니다.
  - 다수의 벌점 평가 기록을 바탕으로 유저의 **매력지수**가 산출되며, **프로필 티어**가 결정됩니다.
  - 즉, 매력지수와 프로필 티어는 유저의 프로필 사진을 포함한 프로필의 각 항목에 대해 이성의 유저들이 느끼는 매력도로 인해 주로 결정됩니다.
  - 프로필 매력지수는 0 이상 100 이하의 값이며, 매력지수에 따라 프로필 티어가 Diamond(65점 이상)/Gold(50점-65점) /Silver(35점-50점)/Bronze(35점 미만)로 구분됩니다.
  - 당시 실시간 접속 유저의 수에 따라 매력 평가가 완료되는 시간이 달라질 수 있습니다.

## 5. 분석 2 : 중도탈락요인 분석

**서비스 이용을 위해 진행하는 일련의 process 과정에서, 이용자의 이탈이 발생하는 지점은 아래와 같음.**
(segment 4 차단이용자는 가입승인단계 이전에 차단되기 때문에 분석 의미가 없음)




| | | | | |
|---|---|---|---|---|
| Segment 1. 실사용 이용자 | 3,067 (100%) | -198 | -32 | 2,837 (93%) |
| Segment 2. 탈퇴 이용자 | 3,805 (100%) | -210 | -29 | 3,566 (94%) |
| Segment 3. 휴면 이용자 | 1,430 (100%) | -573 | -89 | 768 (54%) |

Process 도중 탈락하는 대부분의 이용자가 정보입력 단계에서 탈락

휴면 이용자들의 중도 탈락률이 유의미하게 큼

[분석 주제 2] 휴면 이용자를 중심으로 중도탈락요인 분석

## 3. 탐색적 자료분석 (EDA, Exploratory Data Analysis)

다른 데이터셋과의 연결성이 가장 강한 data_users 데이터셋에서 발생한 parsing 애러 등을 제거하였음.
한국·미국 ·기타 국가중 "한국"에서 서비스를 이용하는 고객들의 데이터를 분석 대상으로 설정함.

users의 admin_confirmed_at 칼럼(승인
승인일 빠름과 티어/활동성 등의 관계
제시해보자.
브론즈 ~ 다이아 점수체계 변환 -> 근
맞은 사람은 기분 안좋을테니 이런거
본인이 입력한 특징과, 이성에 대한 선
차이를 보는 것도 좋을 듯.

## 시사점이나 향후 추가 : message text를 학습시켜서
악성유저 탐지 머신러닝 알고리즘 구축
강력반려사유