

‘마리화나 사용빈도 예측 모델 만들기’ 프로젝트 결과보고서

마약 중독자를 피하는 방법

이화여자대학교 데이터사이언스학과

팀명: 남채영

팀원: 2392008 남재은 2392024 이채원 2392032 정지영



목차

- 1 주제 선정
- 2 데이터 수집
- 3 데이터 전처리 & EDA
- 4 모델 생성 및 평가
- 5 결과 분석
- 6 고찰 및 느낀점
- 7 참고문헌

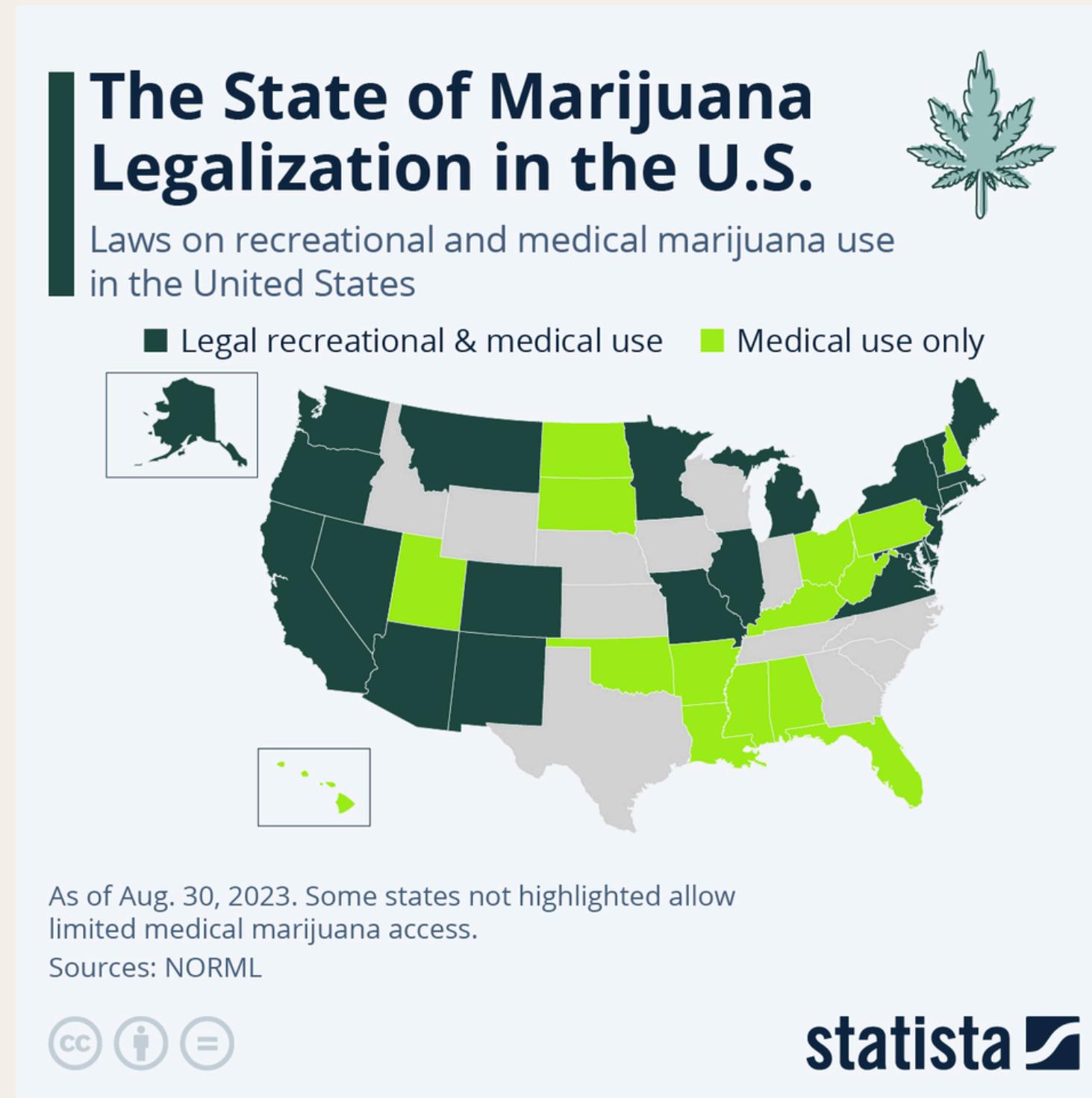
**이 프로젝트는
기존에 공개된 타 프로젝트 참고 없이
저희 팀이 자체적으로 아이디어를 내고 상의하여
주제를 선정하고
직접 데이터 수집과 가공을 거쳐
진행했음을 밝힙니다**

주제 선정

미국 필라델피아 켄싱턴 거리에서 마약에 취한 사람들이 기괴한 자세를 하고 있는 모습을 담은 뉴스 기사를 보며 마약 문제의 심각성에 공감하고 관심을 가지게 되었다. 그래서 미국 방문 시 마약 중독자를 피함으로써 범죄의 대상이 될 위험성과 마약 자체에 노출되는 일 또한 없애고자 마약 사용 빈도를 예측하는 모델을 제작하는 프로젝트를 진행했다. 이를 통해 마약 사용 빈도에 크게 영향을 미치는 요인이 무엇인지 알아보고자 했다.



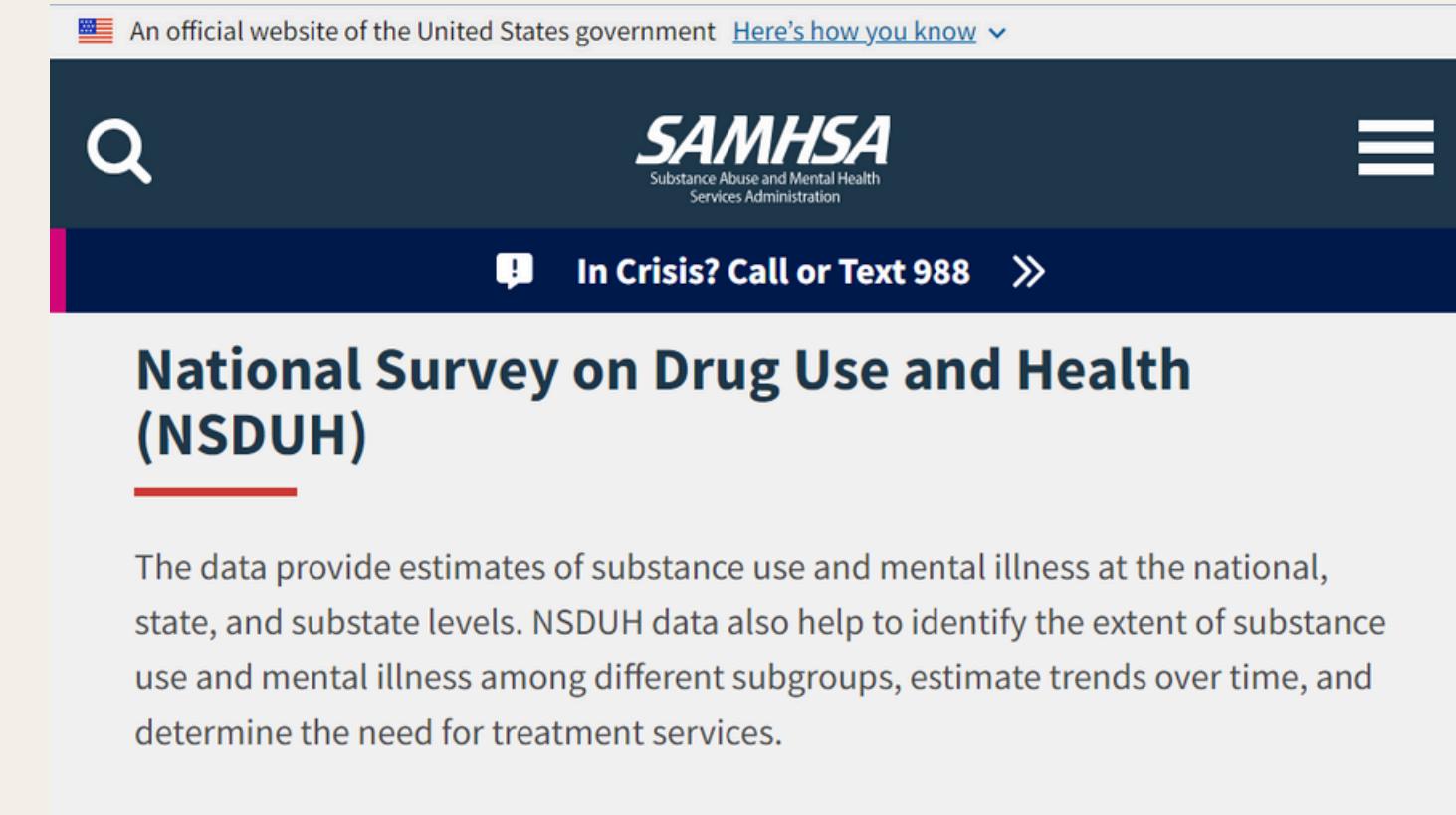
주제 선정



- 특히 미국은 여러 주에서 마약의 일종인 마리화나(대마)의 사용이 합법인 만큼, 미국을 방문하는 한국인들에게 있어 찾은 노출이 걱정됐다. 그래서 마약 종류는 **마리화나**를 중심으로 여러 종류의 불법 약물들에 대한 분석을 진행했다.
- 왼쪽의 표에서, 미국의 50개 주 중, 23개의 주에서 개인 오락용&의료용 마리화나 사용이 합법이고, 15개의 주에서 의료용 마리화나만 사용이 합법이다.

데이터 수집

SAMHSA(Substance Abuse and Mental Health Service Administration)에서 제공하는 **NSDUH**(National Survey on Drug Use and Health) 데이터를 직접 다운로드해 분석에 사용했다. 대중에 공개된 가장 최근 데이터인 2022년 데이터를 사용했다.



출처: SAMHSA 웹사이트(<https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health>)

데이터 수집

• NSDUH

- 미국 연방 정부에서 1971년부터 매년 실시한 조사
- 12세 이상 표본 약 70,000명을 대상으로 인터뷰를 진행한다
- 흡연, 음주, 약물에 대한 데이터를 포함해 약물 사용 장애, 정신질환 문제에 대한 데이터를 제공한다
- 미국 정부기관에서 제공하는 공신력 있는 자료로, 다양한 연구 및 정책 수립에 이용된다

데이터 수집

제공되는 데이터의 형태가 한정적이었기 때문에 Rdata를 저장한 후 R에서 **.csv** 파일로 저장하여 분석에 사용하였다.

The screenshot shows the 'Public Use Files' section for the year 2022. It includes a dropdown menu for the year, and sections for 'Time period' (2022), 'Collection date' (2022), 'Dataset Documentation' (with links to 'Codebook (PDF)', 'In-Person Questionnaire (PDF)', and 'Questionnaire Showcards (PDF)'), and 'Dataset Downloads' (links to 'SAS', 'SPSS', 'Stata', 'ASCII', 'Delimited', 'R', 'SAS Setup', 'SPSS Setup', and 'Stata Setup').

```
1 # Rdata 불러 오기  
2 load("NSDUH_2022.Rdata")  
3  
4 # csv 파일로 저장하기  
5 write.csv(NSDUH_2022, file = "NSDUH_2022.csv")
```

데이터 전처리 & EDA

1. **코드북 탐구**: 우선 데이터와 함께 제공되는 코드북을 간단히 살펴보았다. raw/imputed/recoded의 총 세가지 형태의 데이터가 있었다. imputed/recoded 데이터가 존재할 경우, raw data 보다는 그것들을 사용하길 권장하고 있었다.

The screenshot shows the Table of Contents for the 2022 National Survey on Drug Use and Health Public Use File Codebook. The table lists sections and their corresponding page numbers. The sections include Introduction, NSDUH Overview, Codebook Overview, Confidentiality of Data, Organization of the Data File, History and Development, Questionnaire Changes for the 2022 NSDUH, Strengths and Limitations of NSDUH, Sample Design Overview, Stratification and Selection of First, Second, and Third Stage Sampling Units, Selection of Dwelling Units, Selection of Individuals, Sample Design Variables, Data Collection and Response Rates, Sample Weights, Public Use File Weight Calibration, Usable Interviews, Logical Editing, Editing Procedures for the Nicotine Section through the Methamphetamine Section, Editing Procedure for Prescription Drug Variables, Standard Code Conventions, Statistical Imputation, Implication Indicators, Constraints and Consistency, Variance Estimation of Estimated Numbers of Individuals, Statistical Significance of Differences, and Quality Assessment of Public Use File Estimates and Standard Errors.

Section	Page
Introduction	i-1
1. NSDUH Overview	i-1
2. Codebook Overview	i-2
3. Confidentiality of Data	i-3
4. Organization of the Data File	i-4
5. History and Development	i-6
6. Questionnaire Changes for the 2022 NSDUH	i-8
7. Strengths and Limitations of NSDUH	i-9
8. Sample Design Overview	i-10
8.1 Stratification and Selection of First, Second, and Third Stage Sampling Units (Census Tracts, Census Block Groups, and Field Enumeration Segments)	i-12
8.2 Selection of Dwelling Units	i-13
8.3 Selection of Individuals	i-14
8.4 Sample Design Variables	i-14
9. Data Collection and Response Rates	i-15
10. Sample Weights	i-16
11. Public Use File Weight Calibration	i-18
12. Usable Interviews	i-19
13. Logical Editing	i-21
13.1 Editing Procedures for the Nicotine Section through the Methamphetamine Section	i-21
13.2 Editing Procedure for Prescription Drug Variables	i-23
13.3 Standard Code Conventions	i-24
14. Statistical Imputation	i-25
14.1 Implication Indicators	i-26
14.2 Constraints and Consistency	i-28
15. Variance Estimation of Estimated Numbers of Individuals	i-29
16. Statistical Significance of Differences	i-32
17. Quality Assessment of Public Use File Estimates and Standard Errors	i-33

데이터 설명 목차

The screenshot shows the Table of Contents for the 2022 National Survey on Drug Use and Health Public Use File Codebook. The table lists sections and their corresponding page numbers. The sections include Identification, Sample Weighting and Estimation Vars, Geography, City, Segment, Block, Demographics, Demographic Variables, Education, Recoded Education, Employment, Recoded Employment, Household Composition (Roster), Prior Information, Health Insurance, Recoded Health Insurance, Income, Recoded Income, Special Topics, Recoded Special Topics, Substance Use, Nicotine Dependence, Recoded Nicotine Dependence, Alcohol, Recoded Consumption of Alcohol, Marijuana, Market Information for Marijuana, Blunts, Cocaine, Crack, Heroin, Hallucinogens, Inhalants, Methamphetamine, Pain Relievers Screen, Pain Relievers, Tranquilizers Screen, Tranquilizers, Stimulants Screen, Stimulants, Sedatives Screen, Sedatives, Other Substance Use, Recoded Substance Use, Frequency of Drug Use, Last Month Frequency of Use, Past Month Frequency of Use, Age at Date of First Drug Use, Recoded Drug Use, Other Substance Use, Special Drugs, Recoded Special Drugs, Recoded Special Drugs, Substance Use Disorder, Recoded Substance Use Disorder, Prior Substance Use, Alcohol and Drug Treatment, Recoded Alcohol and Drug Treatment, Risky Ability, Recoded Risky Ability, and Recoded Risky Ability.

Section	Page
IDENTIFICATION	3
SAMPLE WEIGHTING AND ESTIMATION VARS	4
GEOGRAPHY	4
CITY	4
SEGMENT	6
BLOCK	7
DEMOGRAPHICS	8
DEMOGRAPHIC VARIABLES	8
EDUCATION	11
RECODED EDUCATION	16
EMPLOYMENT	21
RECODED EMPLOYMENT	22
HOUSEHOLD COMPOSITION (ROSTER)	28
PRIOR INFORMATION	30
HEALTH INSURANCE	31
RECODED HEALTH INSURANCE	32
INCOME	32
RECODED INCOME	42
SPECIAL TOPICS	43
RECODED SPECIAL TOPICS	50
SUBSTANCE USE	51
NICOTINE DEPENDENCE	51
RECODED NICOTINE DEPENDENCE	62
ALCOHOL	62
CONSUMPTION OF ALCOHOL	62
RECODED CONSUMPTION OF ALCOHOL	83
MARIJUANA	83
MARKET INFORMATION FOR MARIJUANA	95
BLUNTS	101
COCAINE	105
CRACK	109
HEROIN	112
HALLUCINOGENS	116
INHALANTS	118
METHAMPHETAMINE	140
PAIN RELIEVERS SCREEN	144
PAIN RELIEVERS	145
TRANQUILIZERS SCREEN	154
TRANQUILIZERS	155
STIMULANTS SCREEN	161
STIMULANTS	162
SEDATIVES SCREEN	170
SEDATIVES	171
RECODED SUBSTANCE USE	172
FREQUENCY OF DRUG USE	177
LAST MONTH FREQUENCY OF USE	181
PAST MONTH FREQUENCY OF USE	181
AGE AT DATE OF FIRST DRUG USE	187
RECODED DRUG USE	188
OTHER SUBSTANCE USE	208
SPECIAL DRUGS	256
RECODED SPECIAL DRUGS	259
RECODED SPECIAL DRUGS	271
SUBSTANCE USE DISORDER	273
RECODED SUBSTANCE USE DISORDER	340
PRIOR SUBSTANCE USE	346
ALCOHOL AND DRUG TREATMENT	373
RECODED ALCOHOL AND DRUG TREATMENT	390
RECODED ALCOHOL AND DRUG TREATMENT	392
RISKY ABILITY	403
RECODED RISKY ABILITY	409
RECODED RISKY ABILITY	414

변수 설명 목차

데이터 전처리 & EDA

2. **사용할 변수 선택**: 자료조사를 통해 마약 사용 빈도와 연관이 있는 요소를 선정하고 관련된 변수명을 모두 골라 정리했다. 총 484개의 변수가 골라졌으며, 변수명을 모두 텍스트로 직접 작성했다. 아래 사진은 그 일부이다.

데이터프레임의 변수명: ['irmjfm', 'inhlmolst', 'irlsdage', 'irherfy', 'ysdsovrl', 'grskcigpkd', 'sevrstmany', 'alcyrbfr', 'cigdlmlu', 'ecstmoagl', 'mhntprobs', 'yhltmde', 'ilimfothfg', 'cocnedever', 'argupar', 'ircrkfy', 'irsalviarec', 'suntpriv', 'mhntmeds', 'stndalc', 'mhtnseekpy', 'irmethamyfu', 'grskbngwk', 'yflmjmo', 'yhbchmde', 'irstmnmyfu', 'stndsmj', 'irpcpyfu', 'iwrkstat', 'irfentnmyr', 'aomdmde', 'mhtoutppy', 'arxmdeyr', 'trimest', 'drvindettag', 'suntcare', 'irsednmyfu', 'yaltmde', 'suntcost', 'irpcpage', 'synstmyr', 'mhntpthnk', 'asocmde', 'kratomflag', 'smkyrlast', 'hlcallfg', 'Immipy', 'adsuitpacom', 'imfyr', 'grsklsdwk', 'ymdetrxrx', 'mhtskthpy', 'sutoutinhpy', 'rlgdcsn', 'amipy', 'crkylu', 'sevyrtrqany', 'difobther', 'ahbchmde', 'ghbyr', 'lsdaglst', 'coclfinanc', 'irdamtfxrec', 'coclalcuse', 'sunthndl', 'talkprob', 'whodastotsc', 'irwrkstat', 'sevyrpnrany', 'pcpaglst', 'irsednmminit', 'cigarmlu', 'mhntnohlp', 'grskcocwk', 'flvvapmon', 'cocaglst', 'smipppy', 'cigaglst', 'cocyrbfr', 'rcvsutnmht', 'mhtoutndoc', 'anursmde', 'synstmmmon', 'mhntfflke', 'mthnedever', 'mdeimpy', 'irhallucyfu', 'ydocmde', 'prmjmo', 'sutrtalcpy', 'mmipy', 'suntconsq', 'sutrdralpy', 'crkaglst', 'irpnrmminit', 'mhntconsq', 'pcpylu', 'rcvysubprb', 'cigylu', 'mhtinppy', 'yomdmde', 'kratommmon', 'irherrc', 'lsdylu', 'ghbflag', 'svyrsudany', 'ymdeyr', 'ifather', 'svyropiany', 'imother', 'sevyrsedany', 'alcaglst', 'sutoutppy', 'schfelt', 'yusuithkyr', 'grskmrjwk', 'othins', 'prob', 'snrldcasn', 'alcylu', 'spdpstyr', 'mhtoutotpy', 'sutoutstmpy', 'sutoutdrgpy', 'sutoutopipy', 'suntwher', 'opiimfnmmn', 'snystole', 'imfmon', 'stnddnk', 'irmjfy', 'catage', 'hallyrlst', 'irecstmorec', 'grsklsdtry', 'hersmoever', 'cigarylu', 'ymdeharx', 'mhntnoopn', 'snfamjev', 'grskherry', 'irlsdyfu', 'parchkhw', 'cocmlu', 'irecstmooyfu', 'cigmlu', 'alcmu', 'cigyrbfr', 'ytxmdeyr', 'snysell', 'income', 'acounmde', 'adocmde', 'kratomyr', 'ysdswrk', 'irtrqnmrec', 'irinhalyfq', 'sutneedpy', 'inhlaqlst', 'vnurasmde', 'imfnedflaq']

데이터 전처리 & EDA

3. **변수명 확인**: 2단계에서 선정한 변수들만을 골라낸 데이터프레임을 만들었다. 이 과정에서, 모든 변수명이 대문자로 제시된 코드북과 달리 원본 데이터의 변수명에는 **대문자와 소문자가 섞여있어 오류가 발생**했다. 데이터프레임에 해당 변수명이 존재하는지 여부를 확인하는 코드를 이용해 오류가 발생한 변수명들을 수정했다. 아래는 오류 발생 변수들 중 일부에 대한 수정 전후이다.

<문제아들>

```
[ 'anyhlt12', => ANYHLTI2
  'anyndlred', => anyndlrec
  'aynnedflag', => anynedflag
  'cosusvhilt2', => cosusvhlt
  'crskcocmon', => 'grskcocmon'
  'crskmrjmon', => 'grskmrjmon'
  'cutoutherpy', => 'sutoutherpy'
  'drivinage', => 'drvinage'
  'frimfanyyr', => 'primfanyyr'
  'frvinaldrg', => 'drvinaldrg'
  'gbhmon', => 'ghbmon'
  'govtrog', => 'govtprog'
  'grskgerwk', => GRSKHERWK
  'hallaglist', => HALLAGLST
```

```
[ ] ## 부가코드
### 변수 실제로 존재하는지 확인하는 코드.
### 코드북에는 모든 변수가 대문자로 써있었는데 실제 변수명은 대소문자 섞여있었기 때문에 확인자 사용함.

# 데이터프레임의 모든 컬럼명을 리스트로 추출
column_names = data1.columns.tolist()

# 변수명 있는지 확인
if any(element == 'IRPNRNM30FQ' for element in column_names):
    print("리스트에 있습니다.")
else:
    print("리스트에 없습니다.")

➡️ 리스트에 있습니다.
```

데이터 전처리 & EDA

4. NULL 값 처리:

NSDUH 데이터 중 사용하려는 대부분의 변수들이 NULL값을 가지지 않았다.

다만, 특정 연령대만이 대상인 질문 등(예를 들어, 12-18세 청소년에게만 해당되는 질문)에 대해서는 대부분의 사람들은 NULL값을 갖고 주제에 대한 시사점이 적다고 판단되었다. 그래서 NULL을 많이 포함하는 변수를 제거하고 분석에 사용하지 않았다.

```
✓ for col in df_final.columns:  
    | nulls=df[col].isnull().sum()  
    | if nulls>5000:  
    |     df.drop([col],axis=1,inplace=True)  
✓ 0.0s
```

```
df_final.dropna(axis=0,inplace=True)  
✓ 0.0s
```

데이터 전처리 & EDA

5. **X변수 최종 선정:** 최종적으로 41개의 X변수가 선택되었다. 아래는 사용한 변수를 표로 모두 정리한 모습이다. 지역, 연령, 학업, 고용상태, 가족형태, 경제 상황, 범죄, 종교 등에 대한 변수들이 주를 이룬다.

41개 종속변수		
변수명	변수 설명	형태(이진/카테고리/수치) *값이 의미하는 바
PDEN10	인구 밀도 (2010 Census data, December 2009 Core Based Statistical Area classifications (OMB 제공) 기반)	1. CBSA에서의 세그먼트에서 인구 밀도가 빙만영 미안 3. 세그먼트 not in a CBSA
COUTYP4	시골 / 도시 (2013 Rural/Urban Continuum Codes 기반)	1. 대도시 2. 소도시 3. Nonmetro (시골)
AIIND102	인디언 지역 (AlAs)	1. 인디언 지역 2. 아님
CATAG2	나이	1. 12-17 2. 18-25 3. 26 이상
CATAG3	나이	1. 12-17 2. 18-25 3. 26-34 4. 35-49 5. 50 이상
CATAG6	나이	1. 12-17 2. 18-25 3. 26-34 4. 35-49 5. 50-64 6. 65 이상
CATAG7	나이	1. 12-13 2. 14-15 3. 16-17 4. 18-20 5. 21-25 6. 26-34 7. 35 이상
PREGAGE2	pregnancy age	1. 15-17 2. 18-25 3. 26-44 4. 26 또는 45 이상
NEWRACE2	인종	1. NonHisp White 2. NonHisp Black / Afr Am 3. NonHisp Native Am/AK Native 4. NonHisp Native Hi/Other

		Pac Isl 5. NonHisp Asian 6. NonHisp more than one race 7. 히스페닉
ENRLCOLLFT2	대학 학업	1. Full Time College Student Aged 18-22 2. Other Persons Aged 18-22 3. Unknown or not College Aged → 일의 변수 이용하면 되니까 필요 없음
ENRLCOLLST2	대학 학업 (세분화)	1. Full Time College Student Aged 18-22 2. Part Time College Student Aged 18-22 3. Not Enrolled Aged 18-22 4. Other Persons Aged 18-22 5. Unknown or not College Aged
II2WRKSTAT	15-17세 포함한 노동자 [표]의 detailed imputation indicator	1. 질문으로부터 얻음 3. 통계적으로 보정 4. Emp. stat. imp. (restr. to full-time or part-time) 9. 12-14세 → 필요 없음
IRWRKSTAT18	18세 이상의 노동자	1. full time 노동자 2. part time 노동자 3. unemployed 4. 기타 (incl. not in labor force) 99. 12-17세
IIWRKSTAT18	IRWRKSTAT18의 imputation indicator	1. 질문으로부터 얻음 3. 통계적으로 보정 9. 12-17세 → 필요 없음
II2WRKST18	IRWRKSTAT18의 detailed imputation indicator	1. 질문으로부터 얻음 3. 통계적으로 보정 4. Emp. stat. imp. (restr. to full-time or part-time) 9. 12-17세 → 필요 없음
EDFAM18	18세 이상인 가족 구성원 수 (28p)	0. 있음 1. 없음

		98 UNKNOWN
IRHHSIZ2	세대구성원 수	(categorical) 1. 고용 미하 2. 고용 3. 대학 재학 4. 대학생 5. 12-17세
IIHHSIZ2	IRHHSIZ2 imputation indicator	(categorical) 1. 설문 데이터 기반 3. 통계적 imputation
IRKI17_2	세대구성원 중 연령 18세 미만인 아이들 수	(categorical) 1. 18세 미만 어린이 없음 2. 18세 미만 어린이 1명 3. 18세 미만 어린이 2명 4. 18세 미만 어린이 3명 이상 99. 12-14세
IIKI17_2	IRKI17_2 imputation indicator	(categorical) 1. questionnaire data 3. statistically imputed data
IRHH65_2	세대구성원 중 연령 65 이상 구성원 수	(categorical) 1. 65세 이상 구성원 없음 2. 65세 이상 구성원 1명 3. 65세 이상 구성원 2명 이상
IIHHS65_2	IRHH65_2 imputation indicator	(categorical) 1. 설문 데이터 기반 3. 통계적 imputation
catage	RC-AGE 카테고리	(categorical) 1. 12-17세 2. 18-25세 3. 26-34세 4. 35세 이상
drivnage	RC-DUI age category recode	(categorical) 1. 16-25세 2. 26세 이상 3. 그 외(12-15세)
drivndetag	RC-DUI detailed age category recode	(categorical) 1. 16-20세 2. 21-25세 3. 26세 이상 4. 그 외(12-15세)
sexage	성별-연령 통합	(categorical) 1. 12-17세 남성 2. 12-17세 여성 3. 18-25세 남성 4. 18-25세 여성 5. 그외

<u>eduhighcat</u>	RC-교육 카테고리	(categorical) 1. 고용 미하 2. 고용 3. 대학 재학 4. 대학생 5. 12-17세
iwrkstat	고용 상태	(categorical) 1. 설문 데이터 기반 3. 통계적 bad data 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략
iwrkstat	고용상태 imputation indicator	(categorical) 1. 설문 데이터 기반 3. 통계적 imputation 9. 12-14세
imother	모친 유무	1. 12-17세, 모친 유 2. 12-17세, 모친 유 모름 3. 12-17세, 모친 유무 4. 18세 이상
ifather	부친 유무	(categorical) 1. 12-17세, 부친 유 2. 12-17세, 부친 유 모름 3. 12-17세, 부친 유무 4. 18세 이상
govtprog	정부보조 프로그램 1개 이상 참여(금전, 음식)	(categorical) 1. 예 2. 아니오
income	가족전체 수입 recode	(categorical) 1. \$20,000 이하 2. \$20,000-\$49,999 3. \$50,000-\$74,999 4. \$75,000 이상
snyseill	과거 12달 동안, 불법 약물 판매 횟수	(categorical) 1. 0회 2. 1-2회 3. 3-5회 4. 6-9회 5. 10회 이상 85. 논리적 bad data 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략
snrlqimp	종교적 믿음이 본인 삶에서 중요 한 부분이다	(categorical) 1. 강한 부동의 2. 부동의 3. 동의 4. 강한 동의 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략

<u>snytoste</u>	과거 12달 동안 \$50 가치 이상의 물건 절도/절도시도 횟수	(categorical) 1. 0회 2. 1-2회 3. 3-5회 4. 6-9회 5. 10회 이상 85. 논리적 bad data 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략
snyattak	과거 12달 동안, 상해를 입힐 목적으로 누군가를 공격한 횟수	(categorical) 1. 0회 2. 1-2회 3. 3-5회 4. 6-9회 5. 10회 이상 85. 논리적 bad data 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략
snfamjev	대마 관련 어떤 제품이라도 1-2회 사용해본 성인에게 느끼는 감정	(categorical) 1. 대마, 불용도 아님 2. 대마, 불용 3. 강한 불용 85. 논리적 bad data 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략
snrlfrnd	친구가 나와 종교적 믿음을 공유하는 것이 중요하다	(categorical) 1. 강한 부동의 2. 부동의 3. 동의 4. 강한 동의 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략

<u>snrlcsd50</u>	종교적 믿음이 내 삶에서 결정을 내리는데 영향을 미친다	(categorical) 1. 강한 부동의 2. 부동의 3. 동의 4. 강한 동의 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략
snrlfrnd	친구가 나와 종교적 믿음을 공유하는 것이 중요하다	(categorical) 1. 강한 부동의 2. 부동의 3. 동의 4. 강한 동의 94. 모름 97. 단번 거절 98. 빙진(단번 안함) 99. 정당한 생략

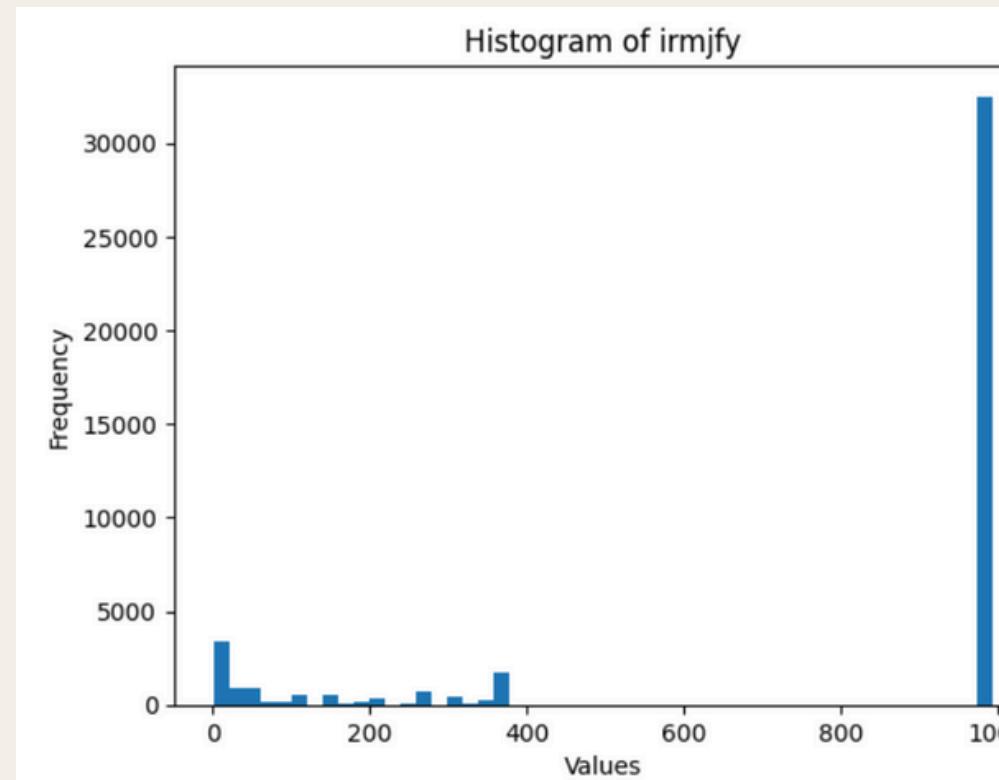
데이터 전처리 & EDA

6. **Y변수 선정**: 약물 사용에 대한 변수도 여러가지 존재했다. 살면서 지금까지 특정 약물 사용 경험 유무, 지난달 특정 약물 사용 일수, 지난해 특정 약물 사용 일수 등이 있었다. 한 가지 변수에 국한되어 사용하기 보다는 변수들 각각에 대해 모델을 여러가지 만들어 보았다. 아래는 마리화나에 대한 변수들 예시이다.

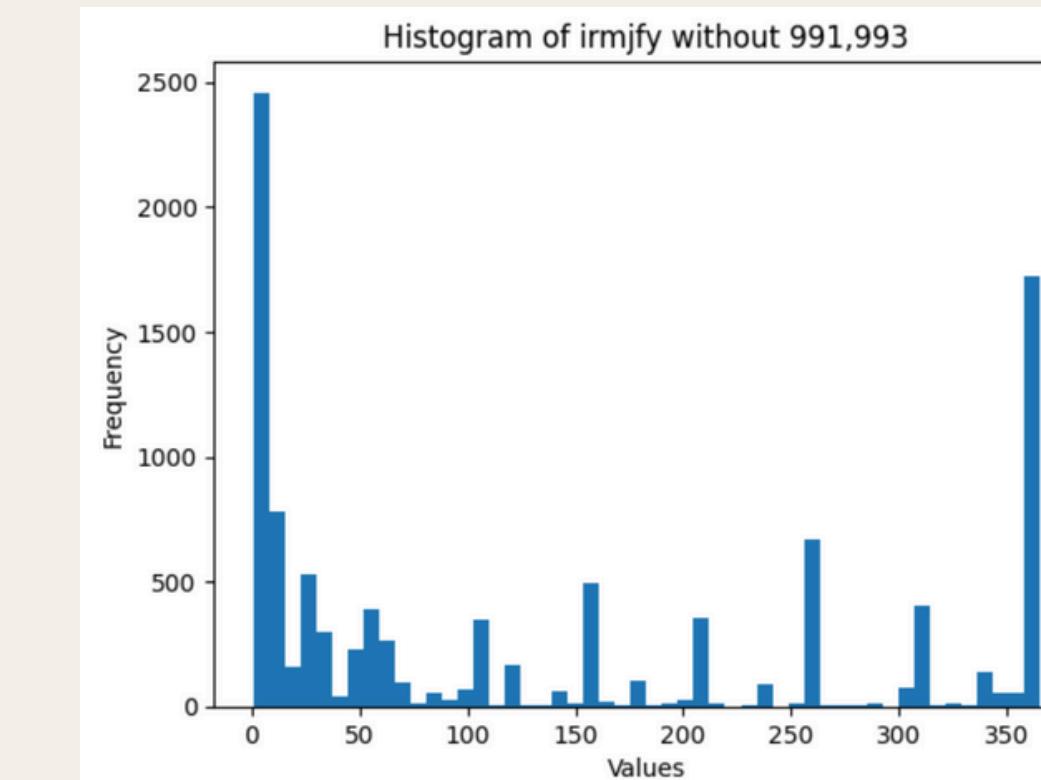
Have you ever, even once, used marijuana or any cannabis product?			
<i>(MJ01, MJREF)</i>			
<u>MJEVER</u>	Len : 2 EVER USED MARIJUANA/CANNABIS	Freq	Pct
1 = Yes.....	26078	44.15	
2 = No.....	32937	55.76	
94 = DON'T KNOW.....	38	0.06	
97 = REFUSED	16	0.03	
<i>(MJDAY30A, MR30EST)</i>			
<u>IRMJFM</u>	Len : 2 MARIJUANA FREQUENCY PAST MONTH - IMPUTATION REVISED	Freq	Pct
RANGE = 1 - 30	9730	16.47	
91 = NEVER USED MARIJUANA	32976	55.83	
93 = DID NOT USE MARIJUANA PAST MONTH	16363	27.70	
<i>(MJYRTOT)</i>			
<u>IRMJFY</u>	Len : 3 MARIJUANA FREQUENCY PAST YEAR - IMPUTATION REVISED	Freq	Pct
RANGE = 1 - 365	14271	24.16	
991 = NEVER USED MARIJUANA	32976	55.83	
993 = DID NOT USE MARIJUANA PAST YEAR	11822	20.01	

데이터 전처리 & EDA

최종 모델에는 **irmjfy**가 주로 사용되었다. irmjfy는 지난해 마리화나 사용 빈도를 나타낸 연속형 변수이다. 일년 중 마리화나 사용 일수를 1부터 365 사이로 나타내었고, 평생 마리화나를 사용하지 않은 사람은 991, 지난해 마리화나를 사용하지 않은 사람은 993으로 처리되었다. 아래는 irmjfy의 히스토그램이다.
(Regression-irmjfy 만 사용 / Classification-전체 약물종류 모두)



전체 데이터를 나타낸 히스토그램



991, 993 값을 제외한 히스토그램

모델 생성 및 평가 – Regression

사용한 X변수 (35개): 41개 X변수들 중 의미 중복 등을 제외하고 선택

```
X_cols=['PDEN10','COUTYP4','ATIND102','CATAG2','CATAG3','CATAG6','CATAG7','PREGAGE2','NEWRACE2','ENRLCOLLFT2','ENRLCOLLST2','IRWKSTAT18',
        'EDFAM18','NRCH17_2','IRHHSIZ2','IRKI17_2','IRHH65_2','catage','drvintage','drvindetage','sexage','eduhighcat','irwrkstat',
        'imother','ifather','govtprog','income','snysell','snystole','snyattak','snfamjев','snrlgsvc','snrlgimp','snrldcsn','snrlfrnd']
```

사용한 Y변수: irmjfy(지난해 마리화나 사용 일수)

[사용한 주요 Method]

- Random Forest Regressor
- Box Cox 또는 Yeo Johnson
- Pipeline
- Bayesian Optimization
- Cross Validation

모델 생성 및 평가 – Regression

Random Forest Regressor

- 예측 성능을 향상시키고 과적합을 줄이는 데 효과적인 모델

Box Cox 또는 Yeo Johnson

- 데이터의 정규성을 개선하는 도구
- 로그 변환보다 성능이 더 좋게 나와 채택

Pipeline

- 전처리, 모델 학습과 평가를 하나로 묶어 일관성 있게 처리하는 도구

Basyesian Optimization

- High level의 계산이 필요한 함수에서 최적의 매개변수를 빠르게 찾는 도구
- Grid Search에 비해 성능이 더 좋게 나와 채택

Cross Validation

- 모델의 과적합 유무를 평가하는 도구

모델 생성 및 평가 - Regression

Random Forest

X변수 41개를 모두 사용, Y변수 이상치(991, 993) 처리 전

R^2: -6.8 ~ -2.7

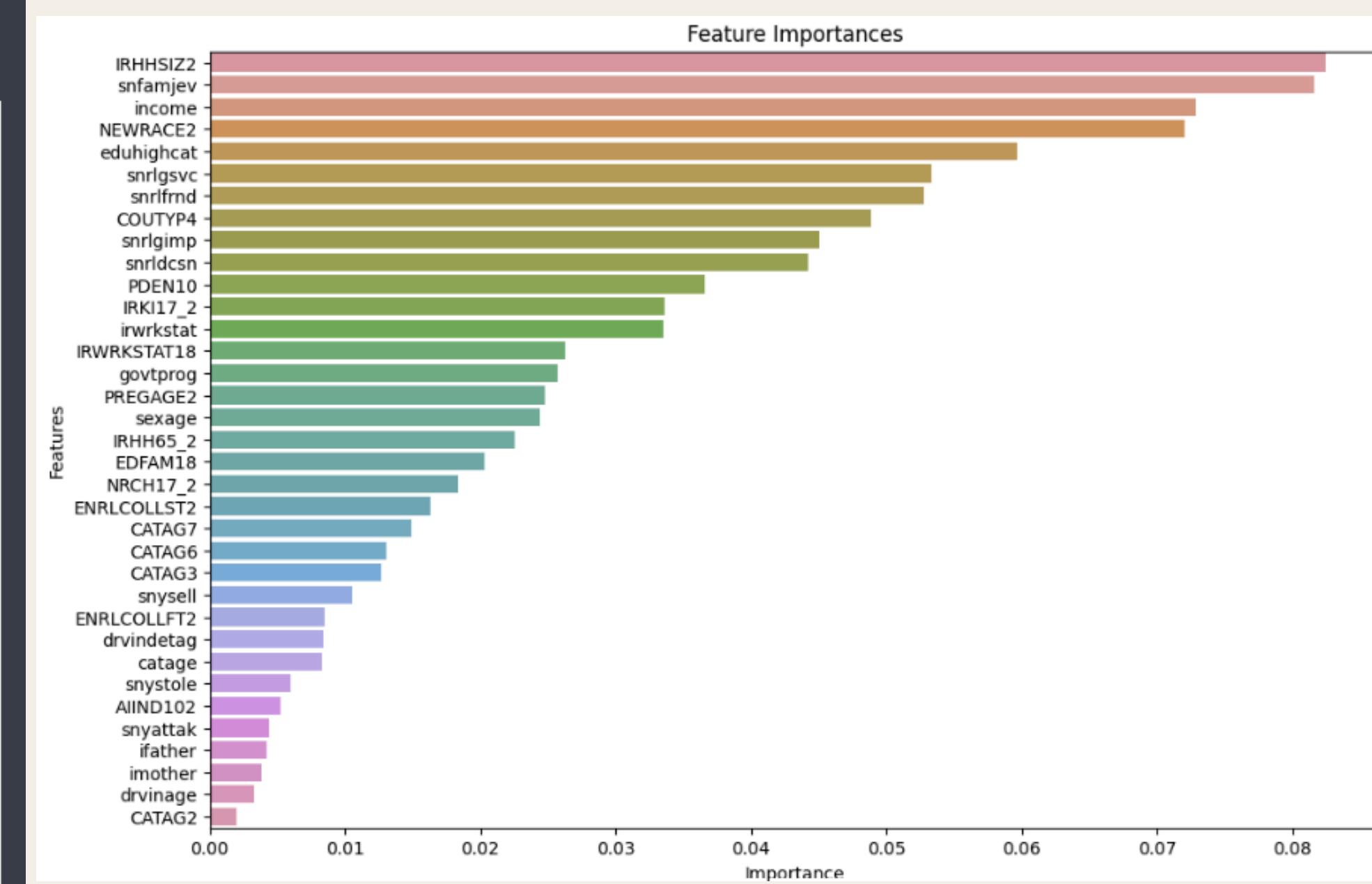
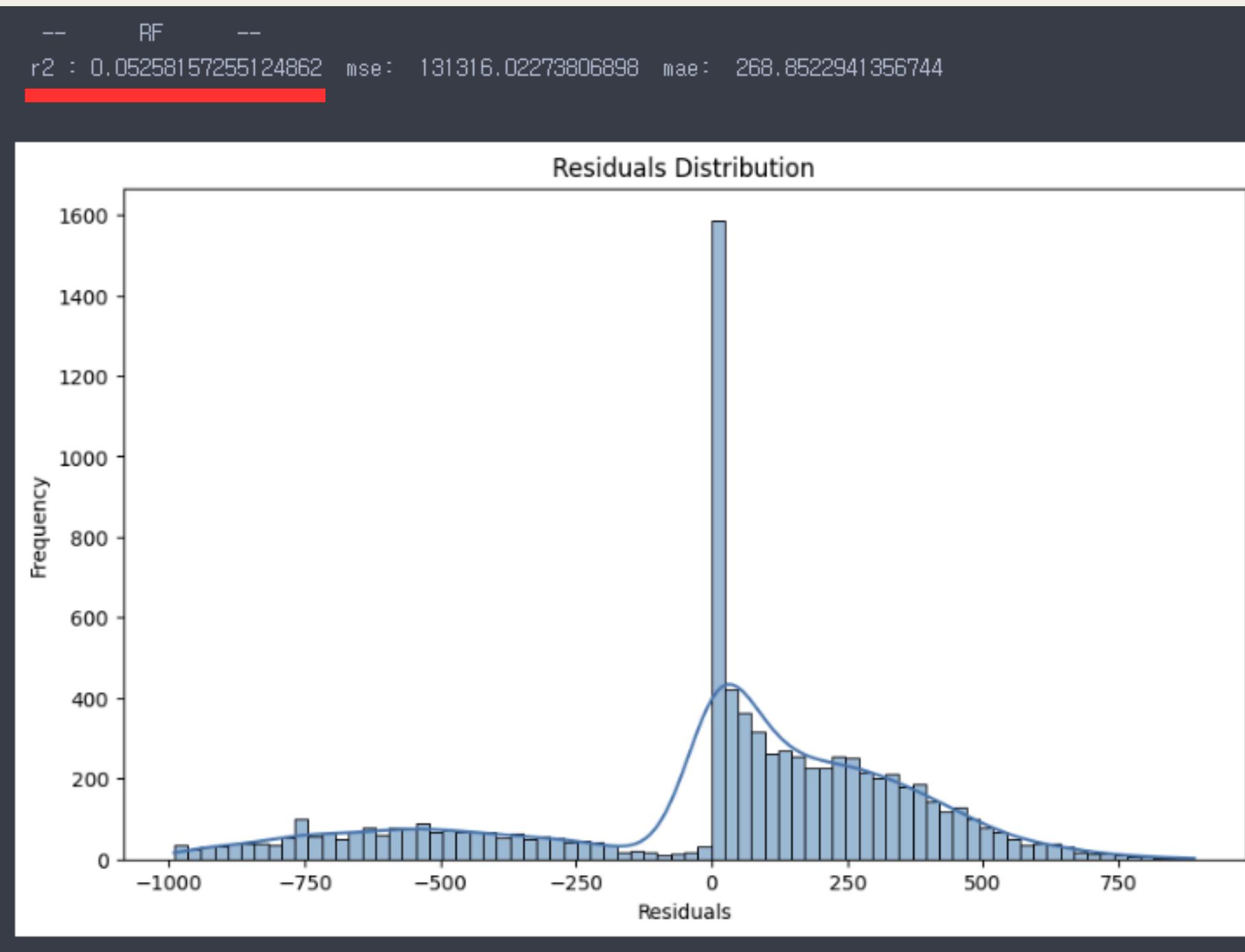
```
--      RF      --
r2 : -2.7811845883930375  mse: 129950.2186195157  mae: 270.78700655992225
--      RF      --
r2 : -3.9977561596187012  mse: 19359.514154773533  mae: 42.01976246426852
--      RF      --
r2 : -6.8748614325213575  mse: 3182.438509603308  mae: 6.984610833819736
--      RF      --
r2 : -3.7222460753816833  mse: 37492.369709966726  mae: 77.55957801146175
--      RF      --
r2 : -4.287476400712461  mse: 13130.464081404  mae: 29.4634919321835
```

모델 생성 및 평가 - Regression

Random Forest

선별한 X변수 35개를 사용, Y변수 이상치(991, 993) 처리 전

R^2: 0.05



모델 생성 및 평가 - Regression

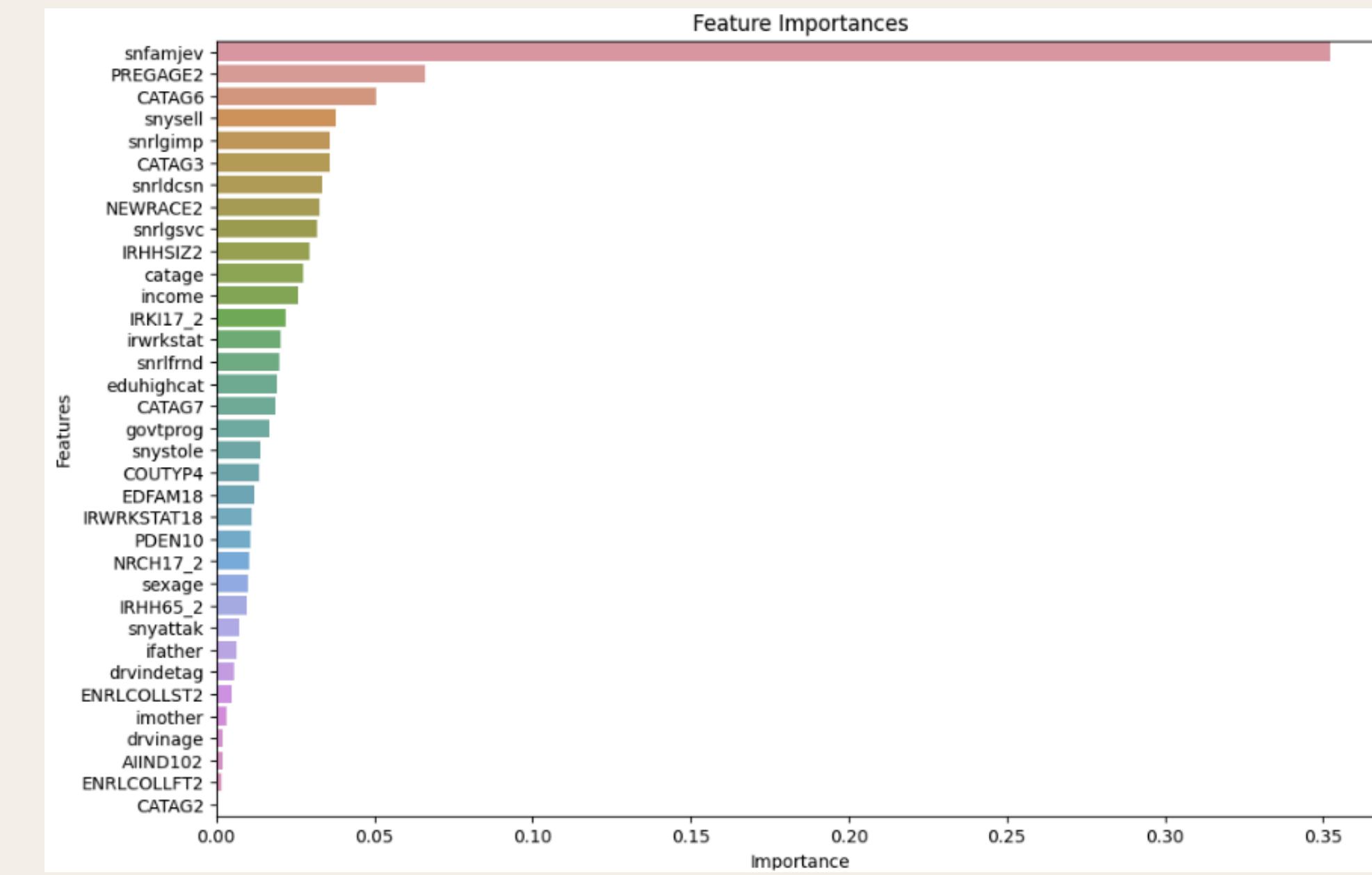
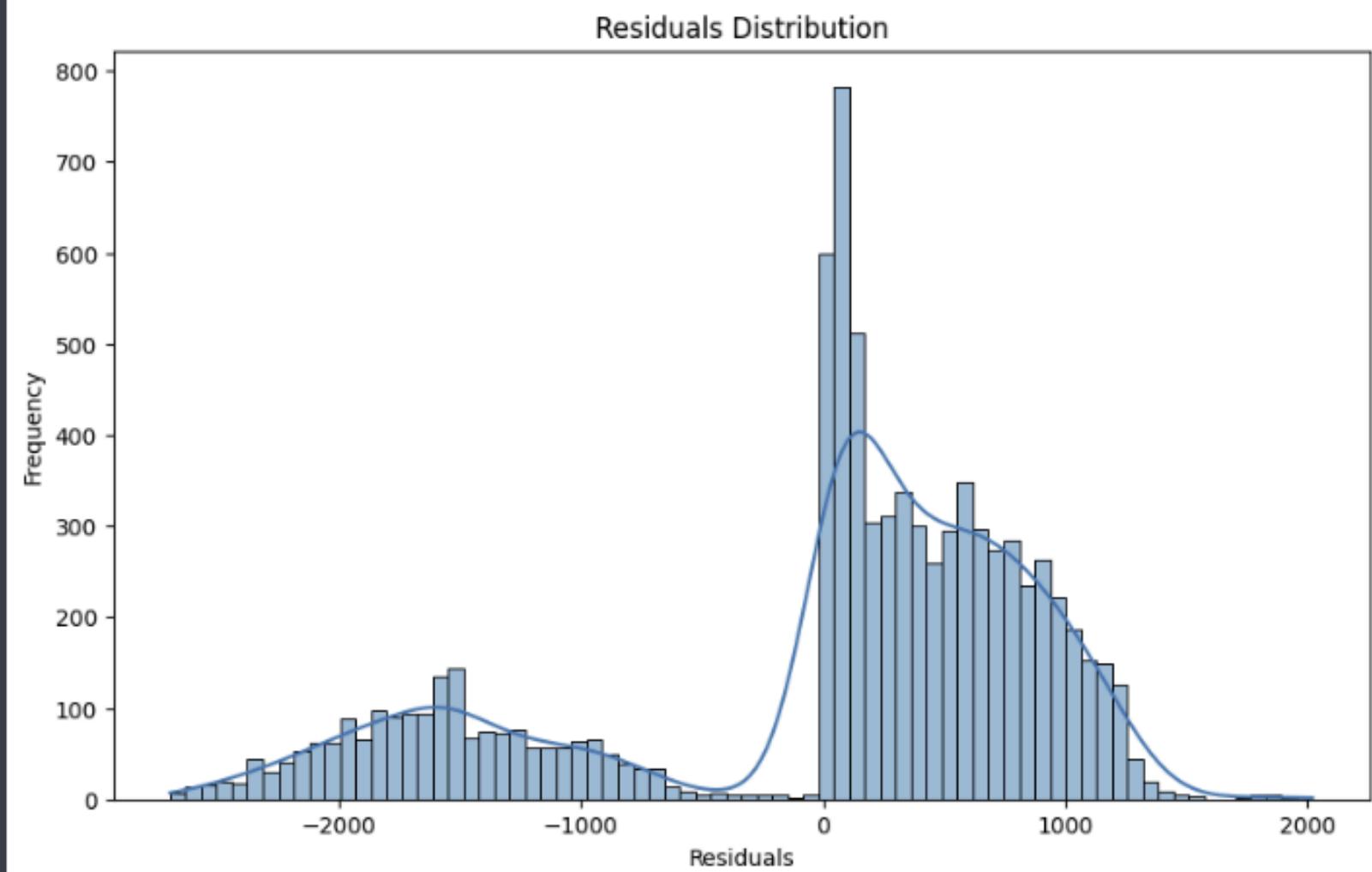
Box Cox, Pipeline, Bayesian Optimization으로 튜닝

선별한 X변수 35개를 사용, Y변수 이상치(991, 993) 처리 전

R^2: 0.16 / Cross Validated Mean R^2: 0.15

```
Cross-validated R2 scores: [0.15249426 0.14443538 0.15271039 0.15205158 0.1549076]
Mean R2 score: 0.15131984298632184
```

```
-- RF --
r2 : 0.16768478928807862 mse: 910277.9879588167 mae: 736.6990444374444
```



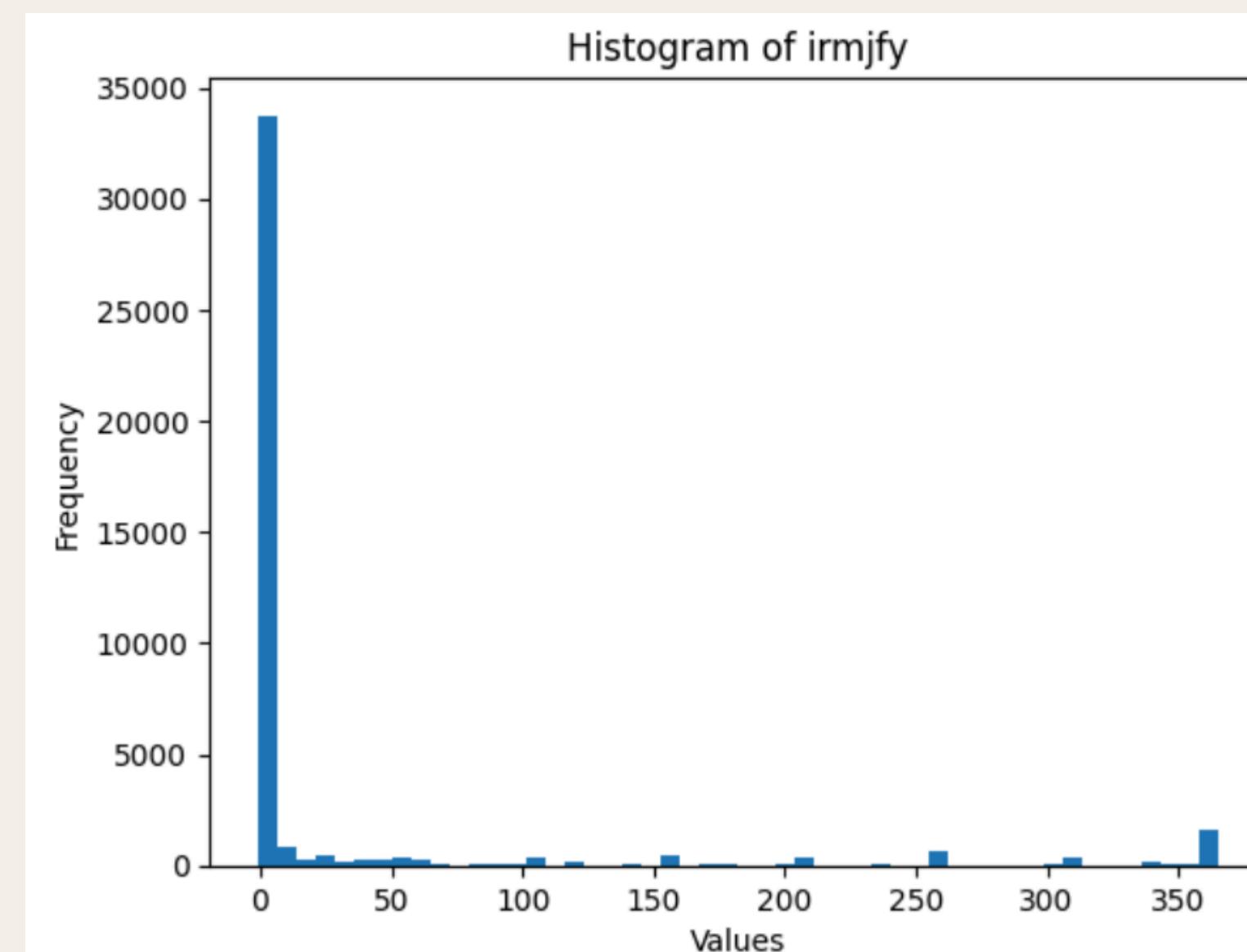
모델 생성 및 평가 – Regression

Y변수 이상치 처리:

마리화나 사용 전혀 안함을 991에서 -1로,

지난해 마리화나 사용 안함을 993에서 0으로 처리

```
df['irmjfy_2'] = df['irmjfy'].replace({991: -1, 993: 0})
```



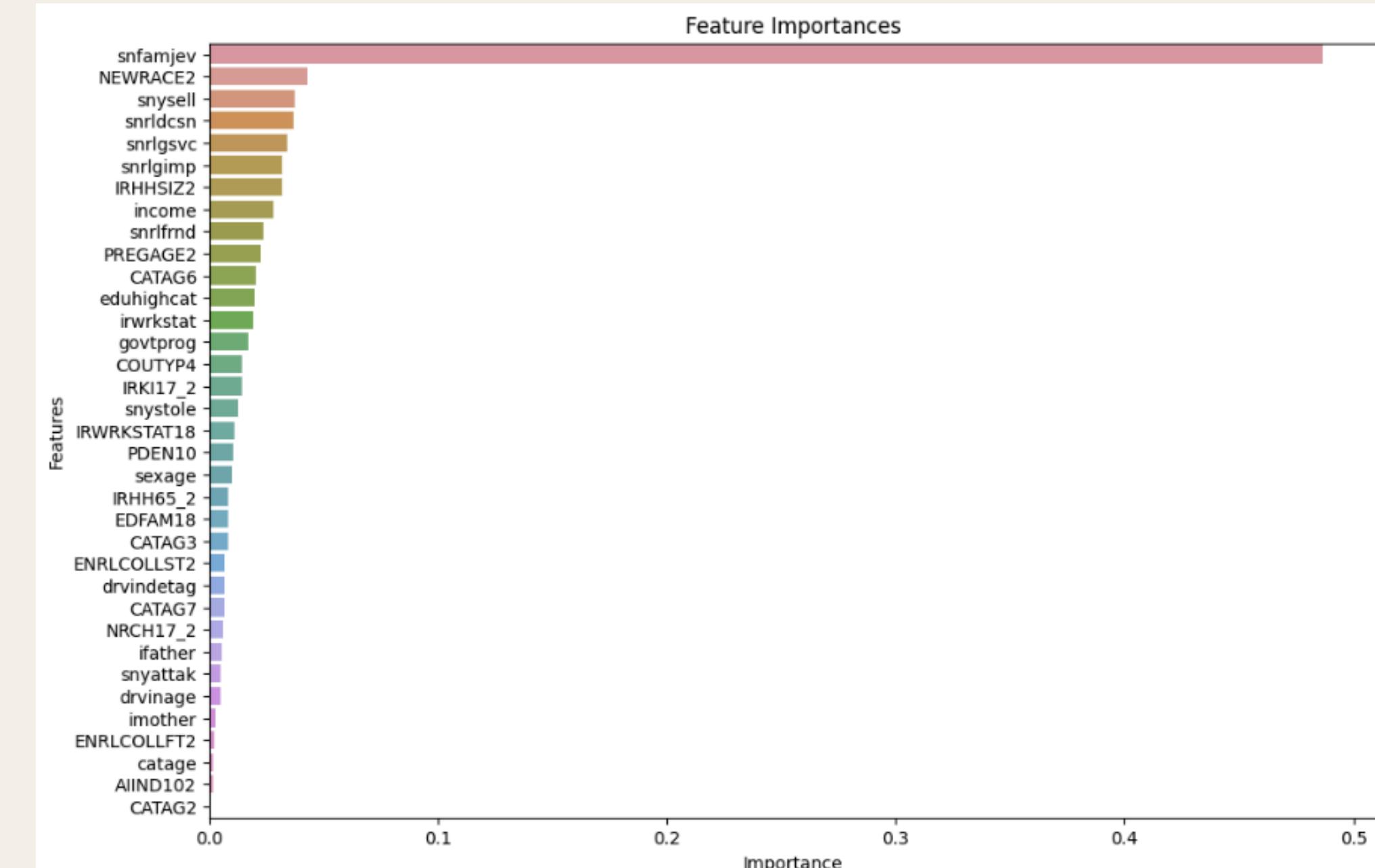
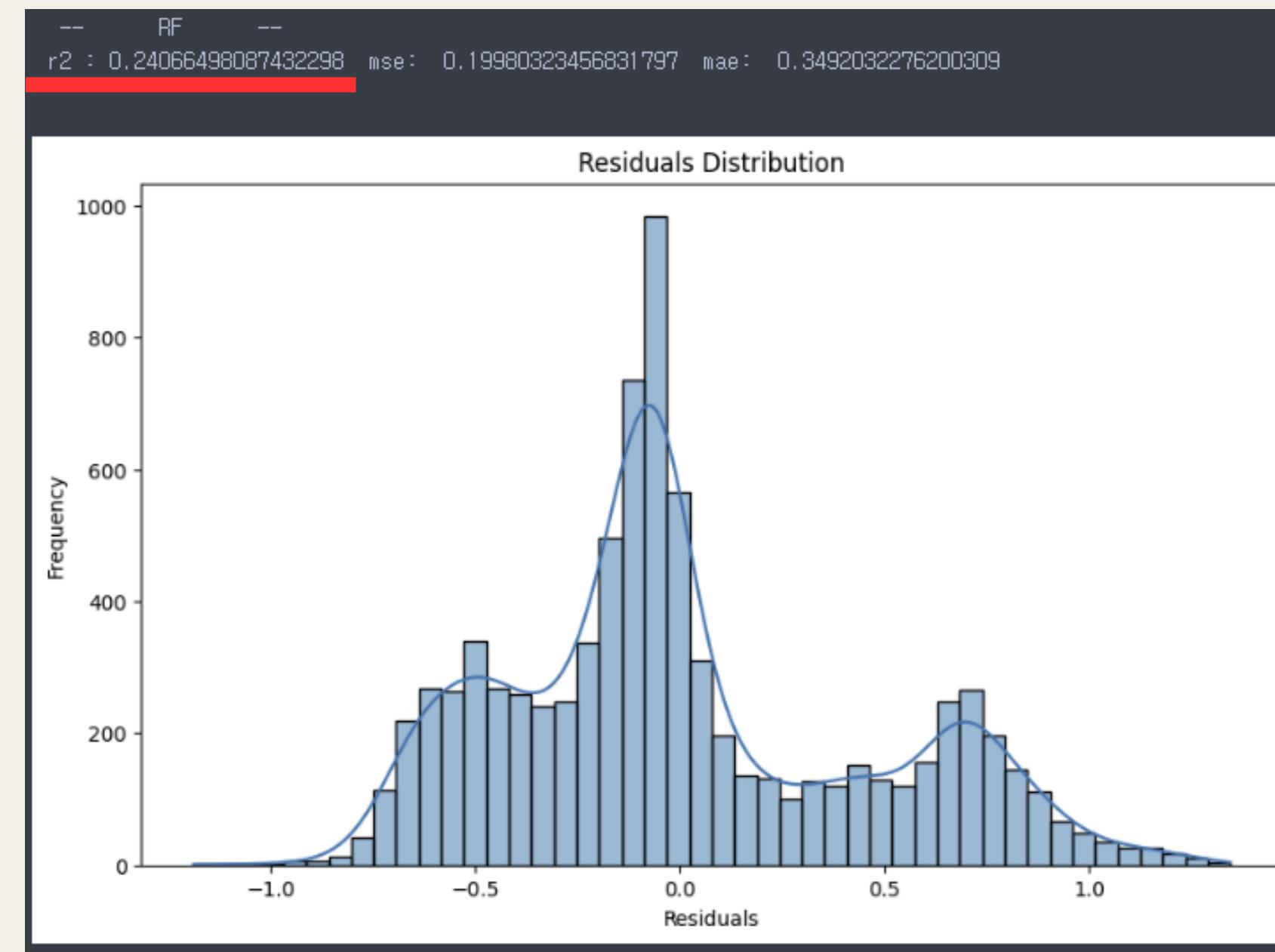
모델 생성 및 평가 – Regression

Yeo Johnson, Pipeline, Bayesian Optimization으로 튜닝

선별한 X변수 35개를 사용, Y변수 이상치(991, 993) 처리 후

R^2: 0.24 / Cross Validated Mean R^2: 0.23

```
Cross-validated R2 scores: [0.22881192 0.22290317 0.22993281 0.23379402 0.240757]
Mean R2 score: 0.2312397828050056
```



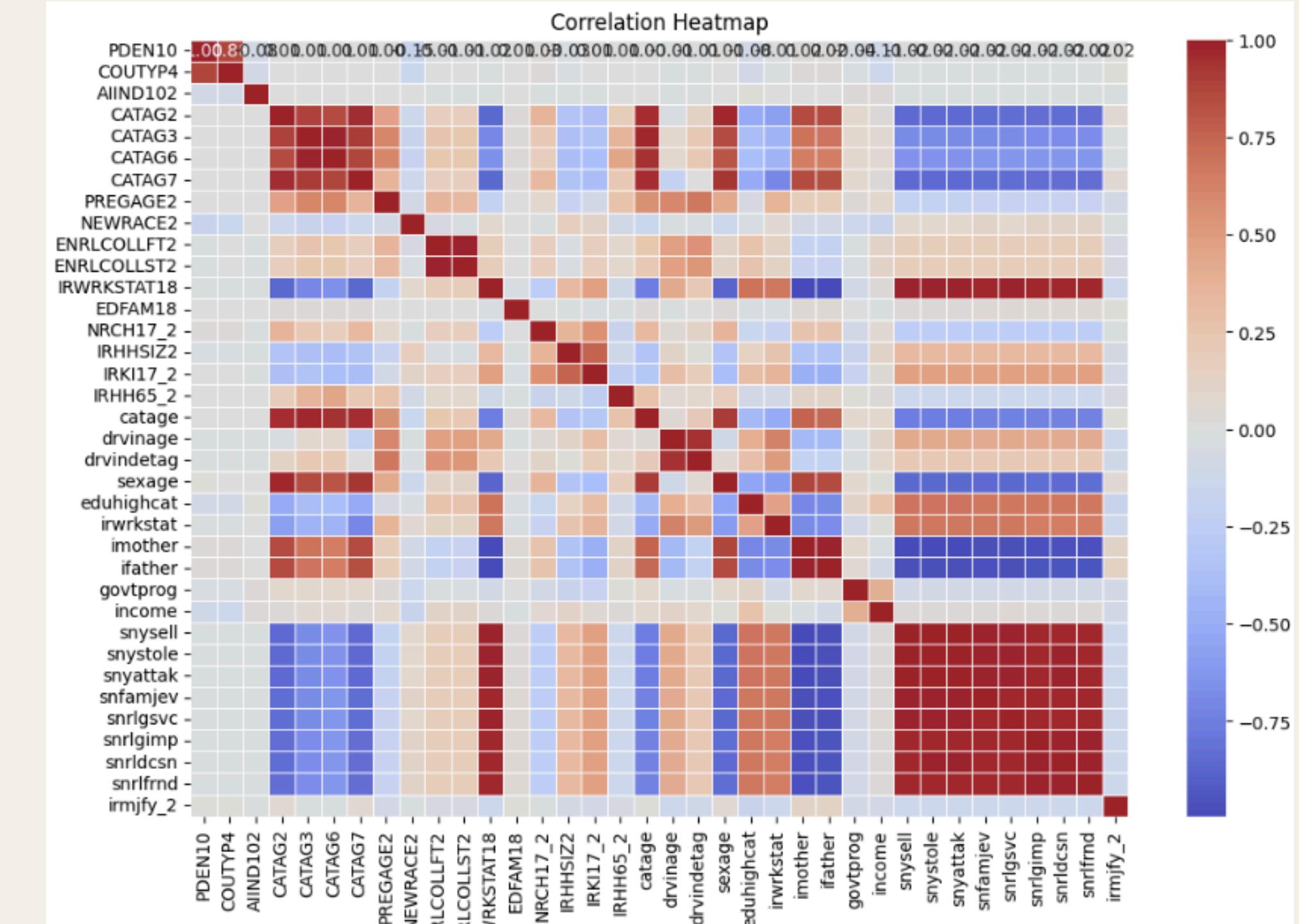
모델 생성 및 평가 - Regression

추가) 상관계수를 통해 X변수 걸러내기

eduhighcat	0.157704
drvintage	0.140152
snrlgsvc	0.130803
ifather	0.128249
IRWRKSTAT18	0.127463
imother	0.127279
snrldcsn	0.127003
snrlgimp	0.126259
snrlfrnd	0.126189
snyattak	0.123660
snystole	0.123364
snysell	0.122514
snfamjev	0.119912
irwrkstat	0.115523
drvindettag	0.107422
income	0.105860
IRKI17_2	0.094309
PREGAGE2	0.090847
govtprog	0.070562
sexage	0.067154
IRHHSIZ2	0.065406
CATAG7	0.064638
CATAG2	0.061007
ENRLCOLLST2	0.058015
IRHH65_2	0.057882
ENRLCOLLFT2	0.055986
CATAG6	0.023169
COUTYP4	0.017491
NEWRACE2	0.017090
PDEN10	0.017026
catage	0.015332
EDFAM18	0.012145
CATAG3	0.010257
AIIND102	0.009263
NRCH17_2	0.007759

상위 12개만 사용

Visualization
using by
Heatmap

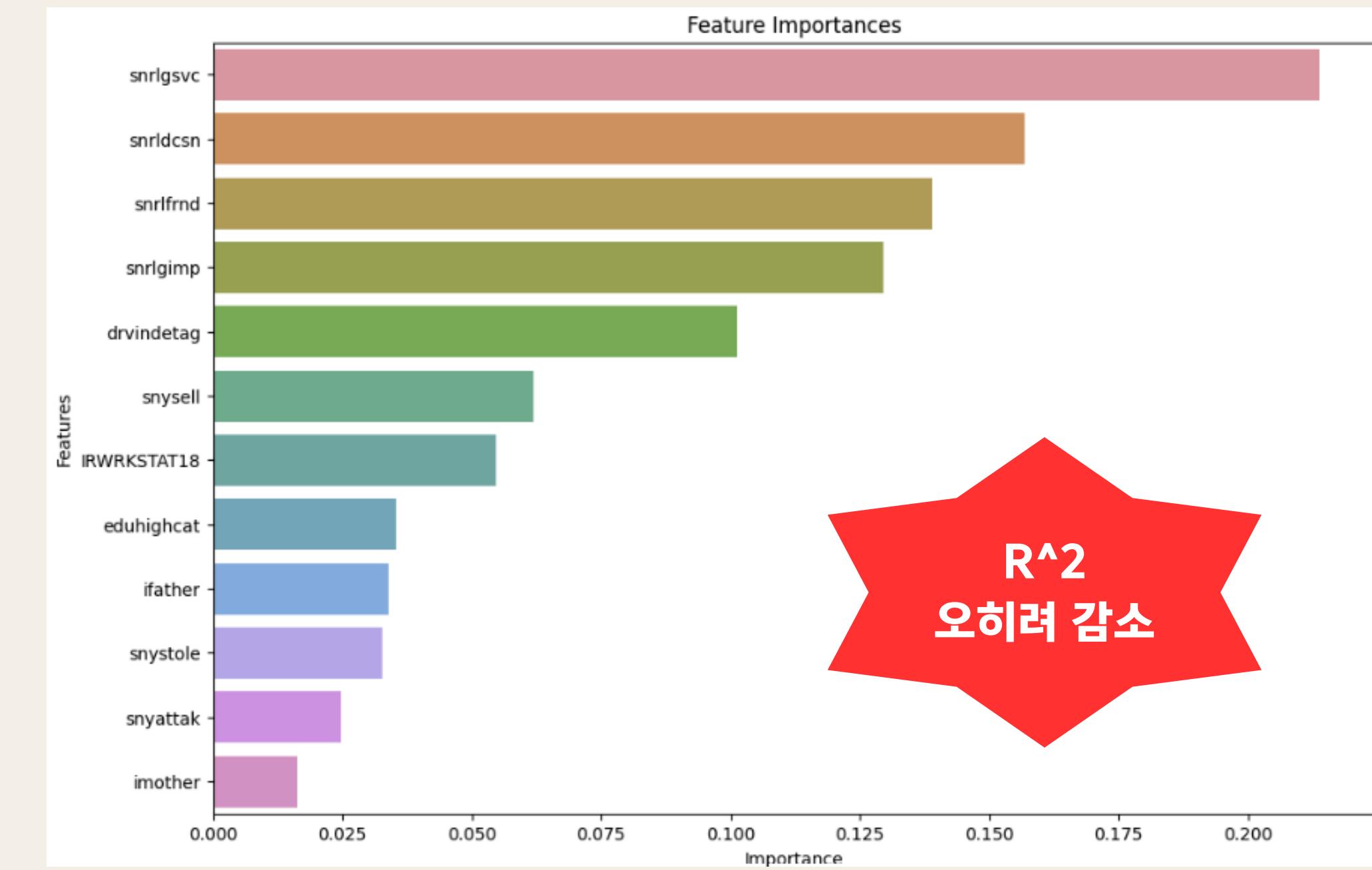
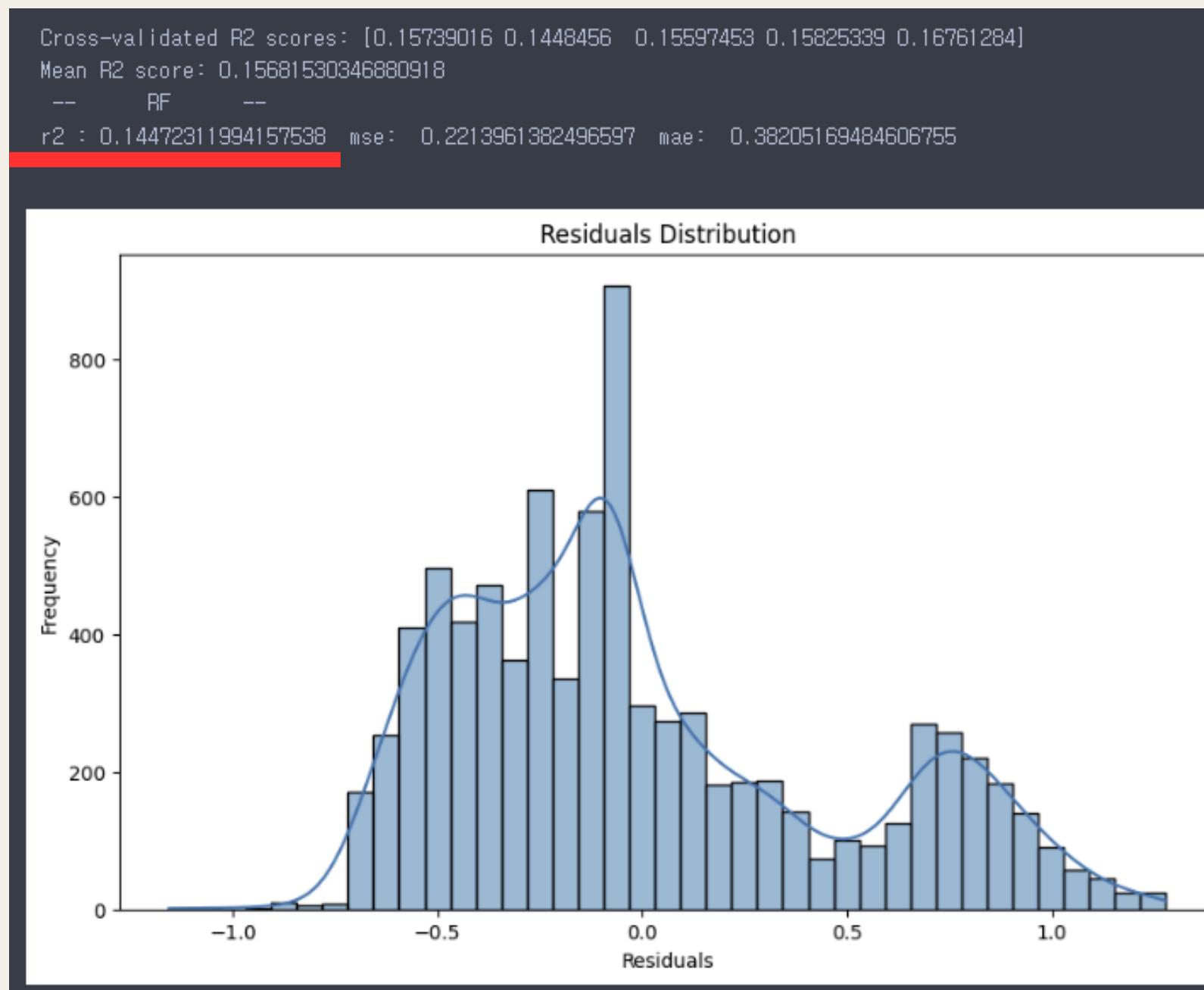


모델 생성 및 평가 – Regression

이상치 처리한 Y변수 사용

Yeo Johnson, Pipeline, Bayesian Optimization으로 튜닝

R²: 0.14 / Cross Validated Mean R²: 0.15



모델 생성 및 평가 – Regression

R² 정리

아무것도 안 했을 때: **-6.8 ~ -2.7**

변수 선별 후

아무것도 안 했을 때: **0.05**

Box Cox, Pipeline, Bayesian Optimization으로 튜닝했을 때: **0.16**

Y변수 이상치 처리 후

Yeo Johnson, Pipeline, Bayesian Optimization으로 튜닝했을 때: **0.24**

상관계수를 통해 X변수 걸러낸 후

Yeo Johnson, Pipeline, Bayesian Optimization으로 튜닝했을 때: **0.14**

모델 생성 및 평가 - Classification

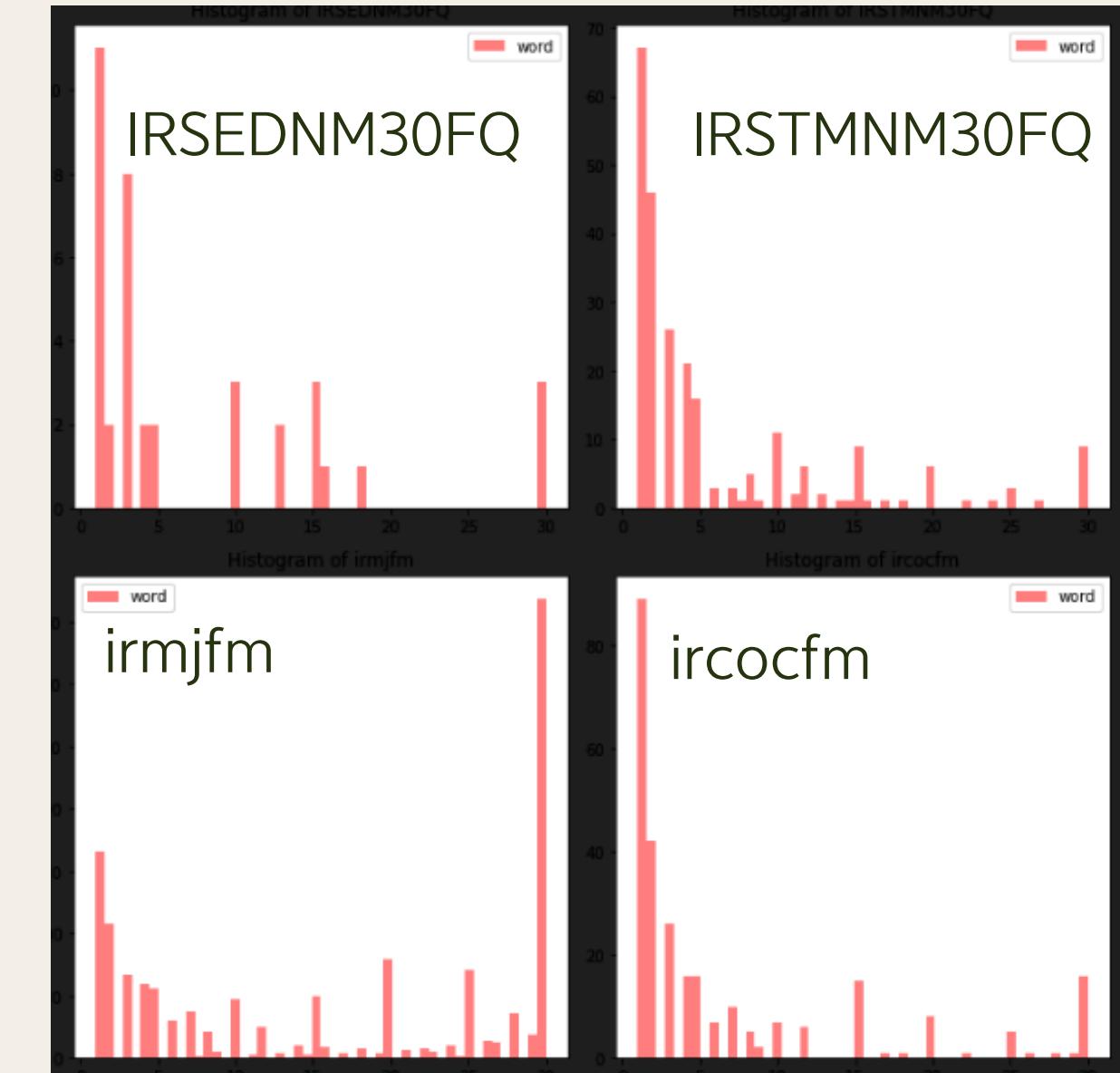
- Regression의 성능이 예상보다 저조해서 Classification을 이용해 구체적인 약물 이용빈도보다는, 그 빈도를 범주화한 새로운 중증도 지표를 예측하는 모델을 만들면 어떨까 싶어 분류 모델도 만들어보았다.
- Classification에서는 모든 약의 종류를 각각 종속변수로 두고 독립적으로 모델을 만들었다.
- 처음 모델을 만들었을 때 성능이 좋지 않아 성능을 올리기 위해 여러 방법들을 시도해보았다.

모델 생성 및 평가 – Classification

EDA



이상치 제거 후
분포



91,93,991,993(마약을 근 한달/일년간 하지 않았거나 평생 안한 사람들)의 비율이 극단적으로 높아 제대로 분포를 확인하기 어려웠다

대체적으로 마약 투여 횟수가 적은 사람들의 비중이 높은 것을 확인할 수 있었다

모델 생성 및 평가 – Classification

레이블 인코딩(1)

```
def encode_month(x):
    if x == 91:
        return -1
    elif x == 93:
        return 0
    elif 1 <= x <= 10:
        return 1
    elif 11 <= x <= 20:
        return 2
    elif 21 <= x <= 30:
        return 3
    else:
        return None # 조건
```

```
def encode_year(x):
    if x == 991:
        return -1
    elif x == 993:
        return 0
    elif 1 <= x <= 100:
        return 1
    elif 101 <= x <= 200:
        return 2
    elif 201 <= x <= 300:
        return 3
    elif 301 <= x <= 400:
        return 4
    else:
        return None # 조건
```

- 처음에는 마약을 평생동안 한 번도 안한 사람들(93,991)과 마약을 최근 한달 동안 안한 사람(93), 일년 동안 안한 사람(993)을 구분했다.
- 레이블 인코딩과 원핫인코딩 사이에서 고민했는데, 해당 y변수는 숫자간 순서가 의미 있는 **ordinal data**이기 때문에 레이블 인코딩을 사용했다. 또한 찾아보니 랜덤포레스트 기반 모델에서는 원핫인코딩을 권장하지 않는다고 하여 그냥 레이블 인코딩을 진행했다.

모델 생성 및 평가 – Classification

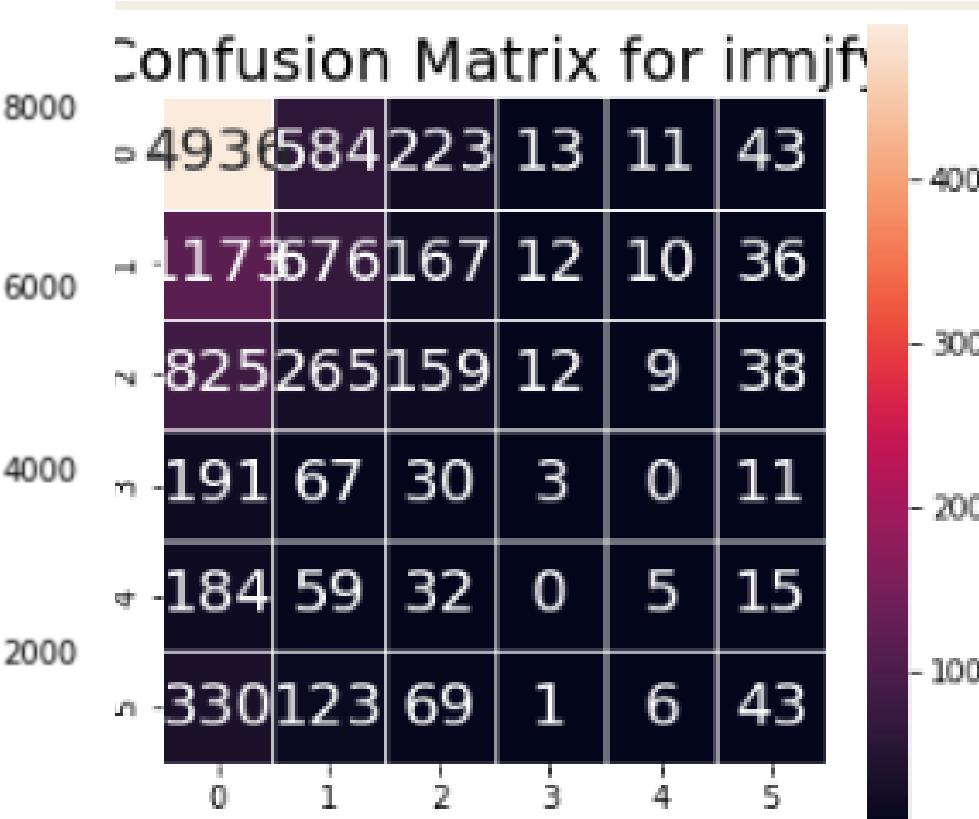
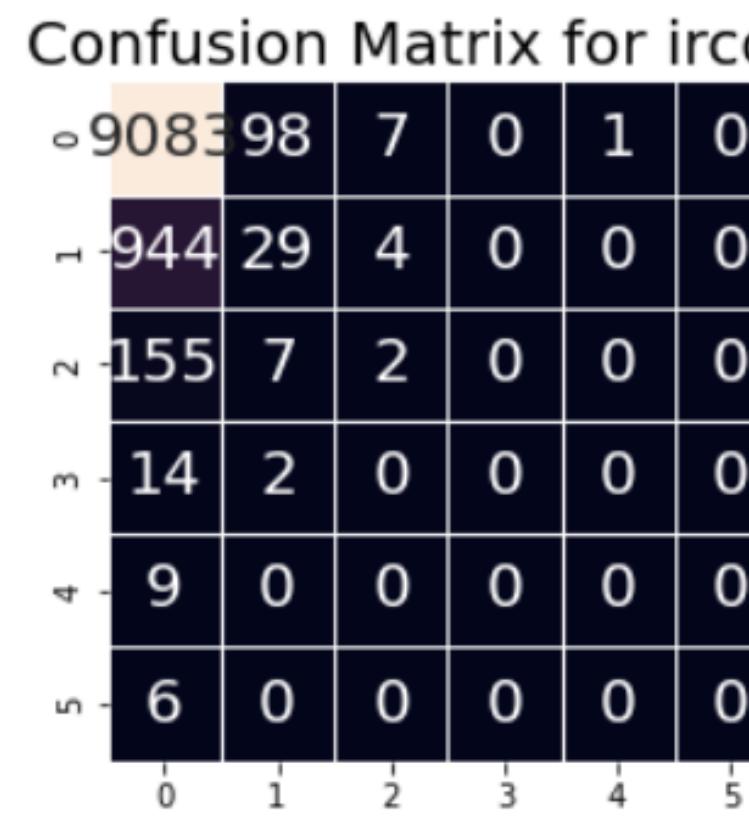
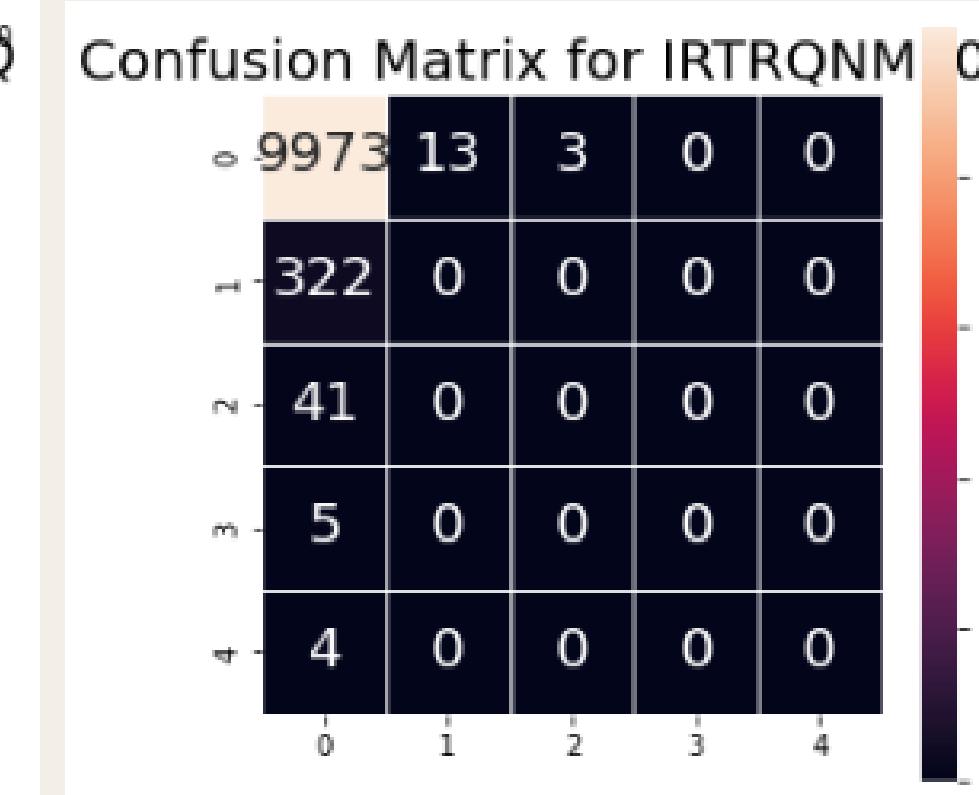
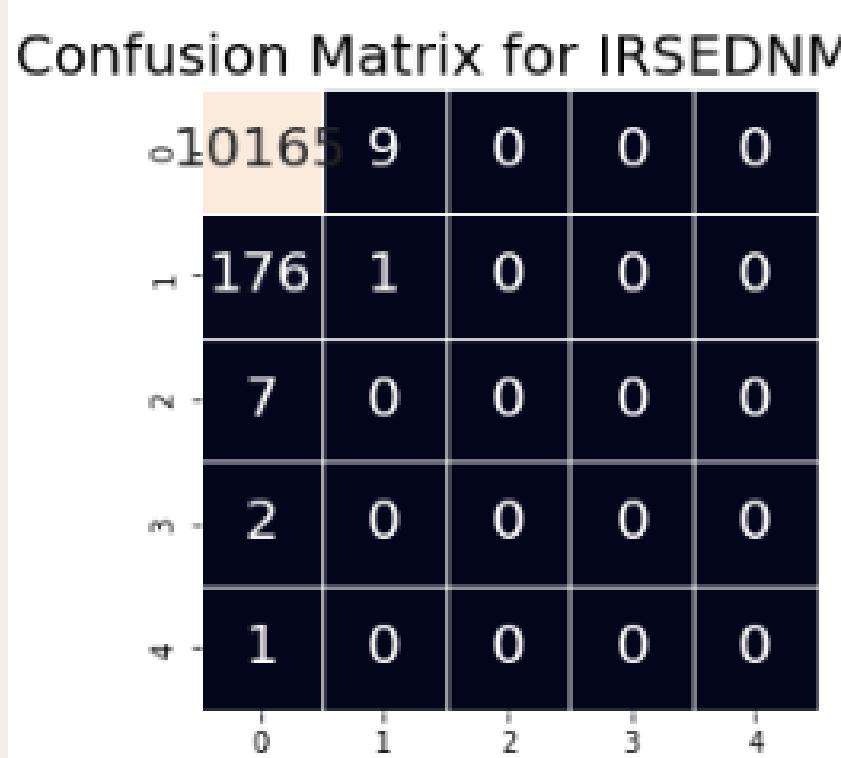
레이블 인코딩(1)–결과

- multiclass classification 문제이기에 어떤 평가지표를 사용해야하는지 잘 몰랐다. 따라서 여러 지표들을 종합적으로 고려하고, confusion matrix도 확인해보았다. 아래는 그 중 일부를 가져온 것이다. 다른 약물들도 아래 결과와 비슷한 결과가 나왔다.

	IRSEDNM30FQ	IRSTMNM30FQ	ircocfy	irmjfy
accuracy	0.981	0.949	0.897	0.561
balanced accuracy	0.200	0.200	0.171	0.233
f1	[0.990 0.010 0. 0. 0.]	[0.974 0.013 0. 0. 0.]	[0.936 0.052 0.022 0. 0.]	[0.734 0.351 0.159 0.017 0.029 0.113]

모델 생성 및 평가 – Classification

레이블 인코딩(1)–결과



- 심각한 class imbalance 문제가 있음을 확인할 수 있었다.
- accuracy는 높지만, 그 이유가 모델이 전부 -1 class로 분류했기 때문이었다.
- 이 문제를 해결하기 위해 1. 랜덤포레스트 하이퍼파라미터 변경, 2. SMOTE를 이용한 오버샘플링 두 가지 방법을 시도해보았다.

→(행: actual, 열: predicted)

모델 생성 및 평가 – Classification

레이블 인코딩(1) + 랜덤포레스트 하이퍼파라미터 변경

```
# Random Forest Classification
rt_clf = RandomForestClassifier(class_weight='balanced')
rt_clf.fit(X_train, y_train)
pred_y = rt_clf.predict(X_test)
```

class_weight : {"balanced", "balanced_subsample"}, dict or list of dicts, default=None

The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as $n_{samples} / (n_{classes} * np.bincount(y))$

class_weight 파라미터를 balanced로 설정해, 랜덤포레스트 모델이 샘플 개수가 적은 class에 더 큰 가중치를 두도록 했다.

결과

	IRSEDNM30FQ	IRSTMNM30FQ	ircocfy	irmjfy
f1	[0.989 0.009 0. 0. 0.]	[0.970 0.044 0. 0. 0.]	[0.934 0.054 0.019 0. 0. 0.]	[0.733 0.328 0.160 0.028 0.016 0.112]

이전과 큰 차이가 없었다.

모델 생성 및 평가 – Classification

레이블 인코딩(1)+SMOTE를 이용한 오버샘플링

```
min_class_size = min(class_counts)

if min_class_size > 1:
    smote_neighbors = min(min_class_size - 1, 5)
    if smote_neighbors < 1:
        smote_neighbors = 1 # Ensure smote_neighbors is at least 1
    smote = SMOTE(k_neighbors=smote_neighbors)
    try:
        X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
        print(f'Class counts after SMOTE:\n{pd.Series(y_train_resampled).value_counts()')
    except ValueError as e:
        print(f'SMOTE failed for feature {i}: {e}')
        X_train_resampled, y_train_resampled = X_train, y_train
else:
    print(f'SMOTE not applied for feature {i} due to insufficient samples in the minority class')
    X_train_resampled, y_train_resampled = X_train, y_train
```

Class	counts
-1	27564
0	27564
1	27564
3	27564
2	27564
4	27564

- SMOTE 방법을 이용해 class별 샘플 개수를 모두 majority 샘플 수와 동일하게 맞추었다.
- k_neighbors를 default인 5로 두니 그보다 더 작은 샘플사이즈를 가진 class에서 오류가 발생해 예외처리해줬다.

결과

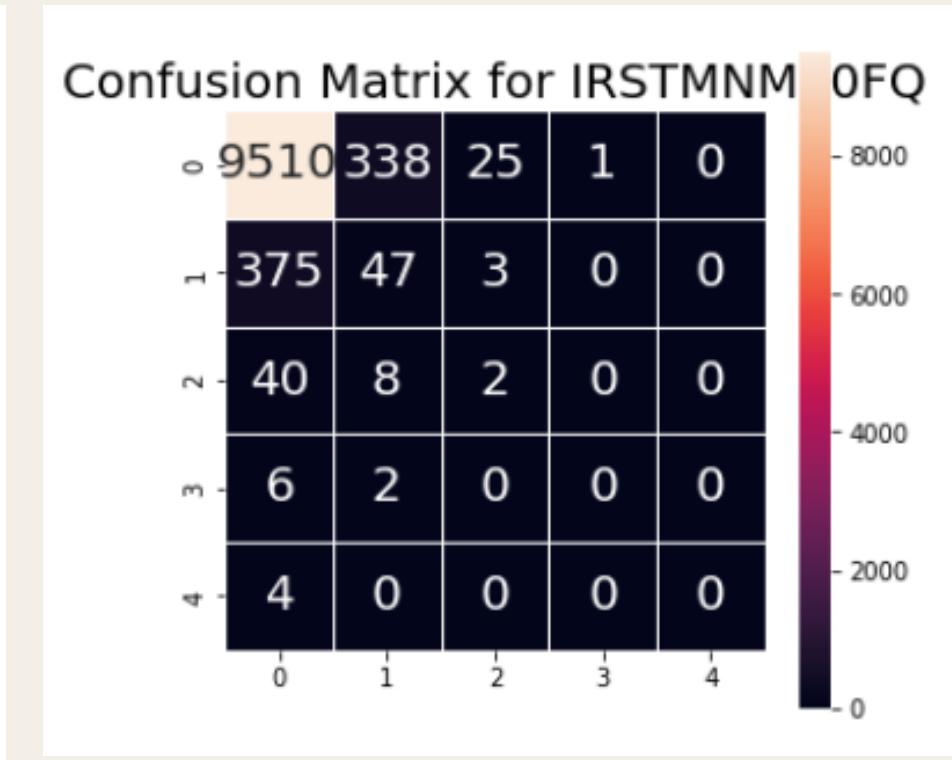
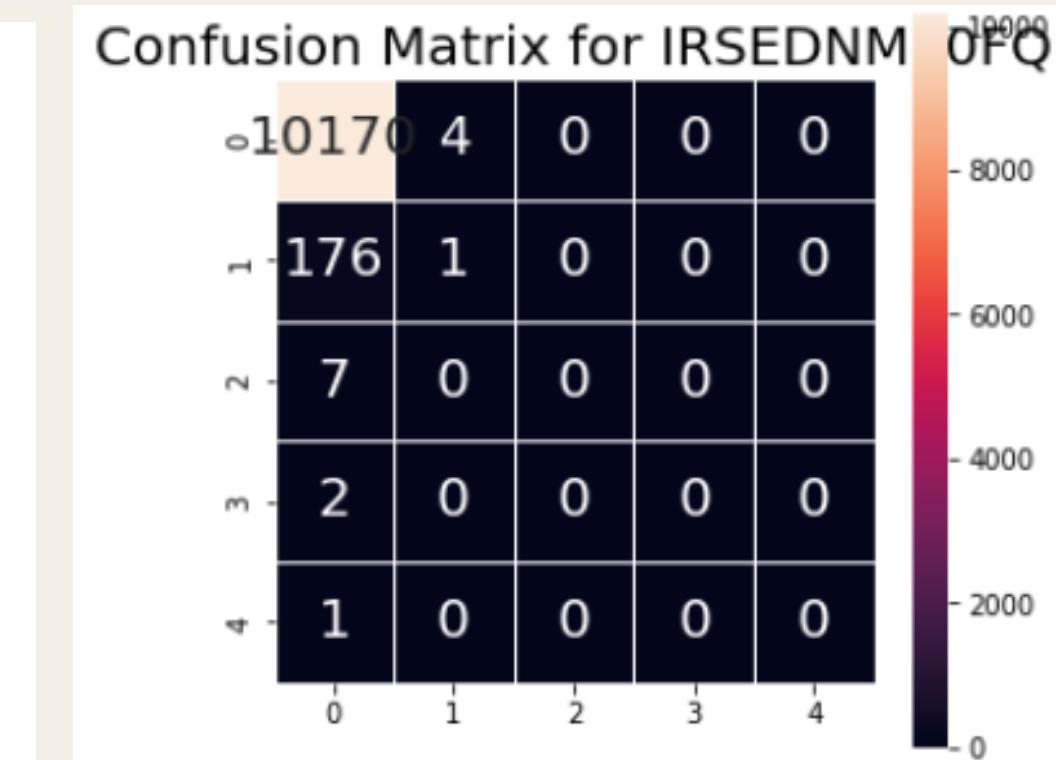
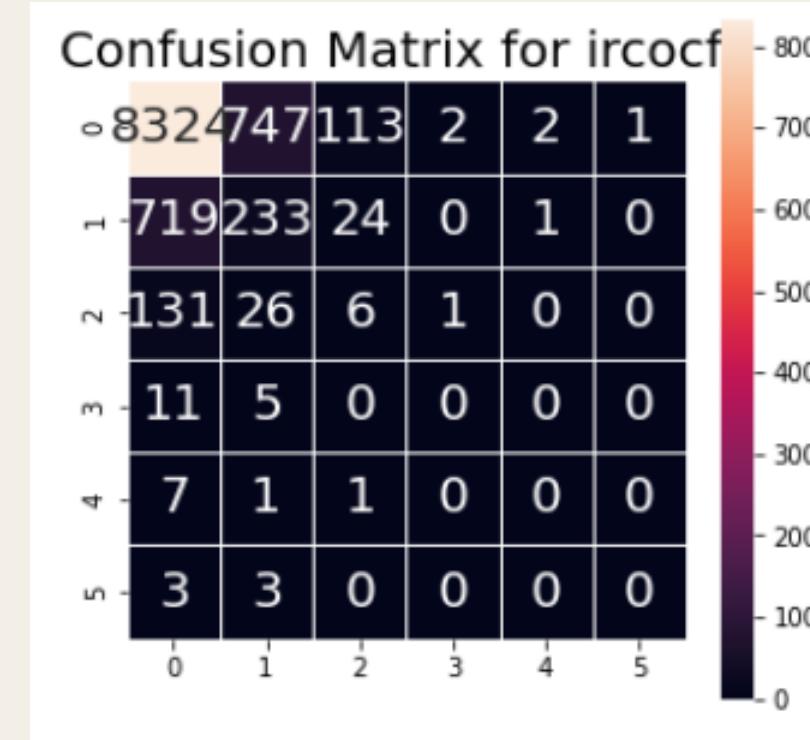
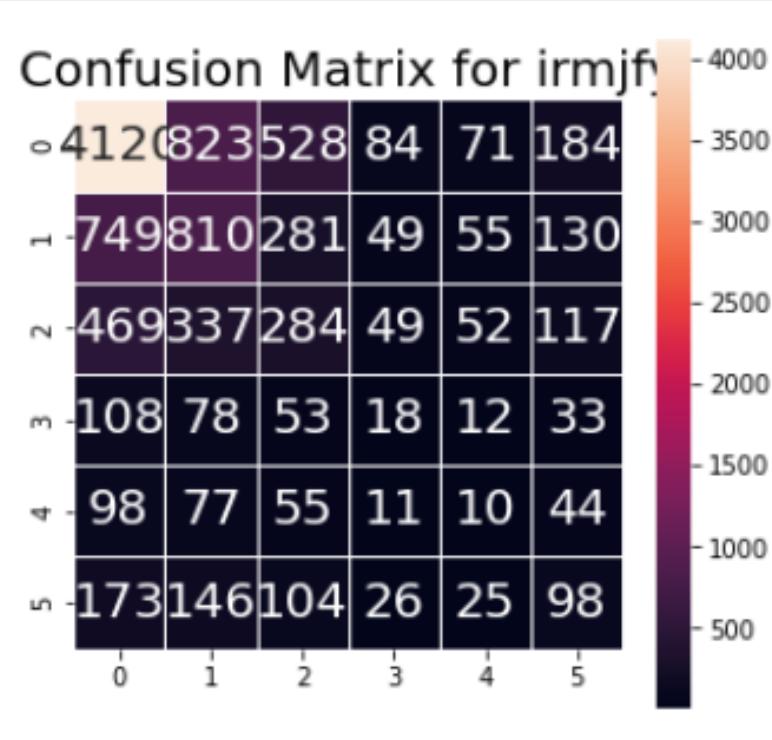
	IRSEDM30FQ	IRSTMNM30FQ	ircocfy	irmjfy
f1	[0.990 0.011 0. 0. 0.]	[0.960 0.114 0.05 0. 0.]	[0.905 0.233 0.038 0. 0.]	[0.714 0.372 0.217 0.066 0.038 0.166]

몇몇 약물들에서 class imbalance 문제가 조금 개선되었다.

모델 생성 및 평가 – Classification

레이블 인코딩(1)+SMOTE를 이용한 오버샘플링

- 결과–Confusion Matrix 확인



- 대각행렬 주변 숫자들이 높은 걸로 보아, 모델이 비슷한 class들을 잘 구별하지 못해 성능이 낮게 나온 것일 수도 있다고 생각했다. 특히, -1과 0일 때를 잘 구별해내지 못한다는 걸 알게 되었다. 즉, 마약을 평생동안 한번도 안한 사람들과 최근 한달or일년동안 안한 사람들을 잘 구별하지 못하는 것이다.
- class를 개수를 줄이면 모델 성능이 나아질까 싶어 class를 재정의해보았다.

모델 생성 및 평가 – Classification

레이블 인코딩(2)+SMOTE를 이용한 오버샘플링

```
def encode_month(x):
    if x >= 90 :
        return 0
    elif 1 <= x <= 15:
        return 1
    elif 16 <= x <= 30:
        return 2
    else:
        return None # 조건에 맞지 않는

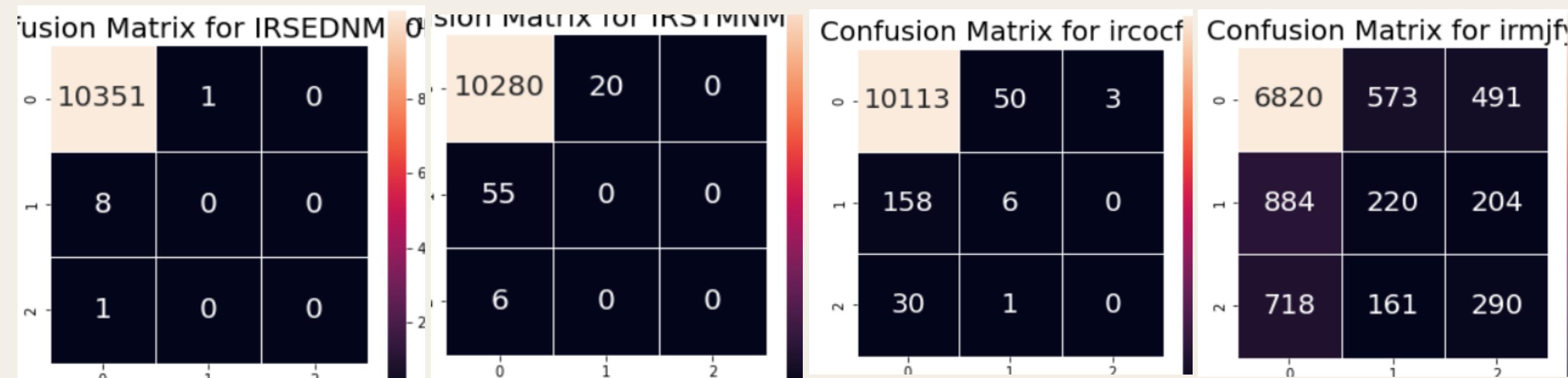
def encode_year(x):
    if x >= 900:
        return 0
    elif 1 <= x <= 100:
        return 1
    elif 101 <= x:
        return 2
    else:
        return None # 조건에 맞지 않는

for i in y_month:
    df[i]=df[i].apply(encode_month)

for i in y_year:
    df[i]=df[i].apply(encode_year)
```

결과

	IRSEDNM30FQ	IRSTMNM30FQ	ircocfy	irmjfy
f1	[0.999 0. 0.]	[0.996 0. 0.]	[0.988 0.05 0.]	[0.836 0.194 0.269]



- 오히려 class imbalance 문제가 더 심해진 것 같았다.
- Classification에서 여러 시도를 해본 결과 이 데이터는 구체적인 마약 투여 횟수가 이미 있으므로 회귀가 더 적합한 문제인 것 같다고 결론지었다.

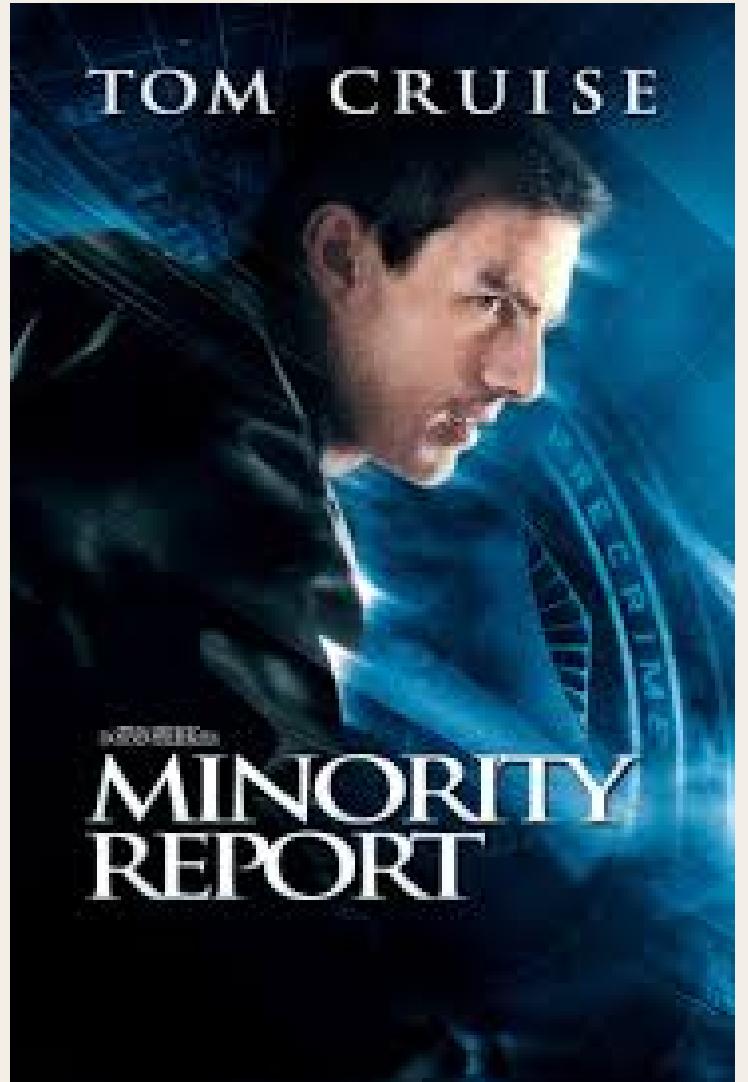
결과 분석

- 모델 생성에 사용된 X변수는 거주 지역, 나이, 학업 수준, 고용 상태, 가족 구성원 형태, 사회적 지원 프로그램 이용 여부, 가족 수입, 범죄 가담 여부, 마약과 종교에 대한 개인적 가치관에 대한 정보가 담겨있었다.
- 이러한 변수들을 이용해 지난해 마리화나 사용일을 예측하는 Regression 모델을 생성한 결과, R^2 최댓값이 **0.24** 정도로 설명력이 낮았다. 그래도 주요 피처들을 추출해보았는데, 마리화나 사용 빈도 예측에 중요했던 변수들은 아래와 같았다.
 - 대마 관련 어떤 제품이라도 1-2회 사용해본 성인에게 느끼는 감정
 - 인종
 - 과거 12달 동안, 불법 약물 판매 횟수
- 마약 사용 빈도를 범주화하여 classification 모델을 생성한 결과, 다소 오버피팅된 결과가 나와 여러 방법을 이용해 개선해보았다. **SMOTE**를 사용했을 때 약간 균일하게 개선된 결과가 나옴을 관찰할 수 있었다.

고찰 및 느낀점

- 지금껏 다뤄보지 않은 대용량 데이터를 직접 구하고 분석해보는 과정에서 주제 선정 후 실제 데이터 분석을 들어가려 했을 때 접근을 어떻게 해야 할지 조금 막막하기도 했고 무엇보다 모델 성능을 높이기가 정말 어려웠다. 로그 변환과 Box Cox, 그리고 서치과 베이지안 최적화 등등 이 데이터에 적합한 튜닝을 직접 실행하고 비교하며, 어떤 상황에 어떤 모델을 써야 좋은지 배울 수 있어서 유익한 경험이었다.
- R^2 가 0.24 정도밖에 안 나온 건 아쉬운 부분이지만, 현장에서 데이터를 다룬다면 이렇겠구나를 몸소 느낄 수 있었다.
- multiclass classification을 다루면서 binary class와 다른 점이 많다고 생각했고, 여러 라이브러리를 사용하고 파라미터를 계속 조정하면서 사용할 수 있는 방법은 정말 많고 정답은 없다고 느꼈다.

고찰 및 느낀점



- 마약 사용 빈도를 예측하는데 인종, 경제 상황, 교육 수준 등 의 데이터를 사용하는 것에 있어서의 윤리적 문제에 대해서 도 생각해 보았다. 개인적 특성을 알면 마약 중독자인지 판 단하는 모델을 사용하는 것이 영화 '마이너리티 리포트'에서 처럼 사람들간의 편견과 불신을 키울 수 있을 것 같다는 우려가 되었다.
- 데이터 분석 시 문제 정의 및 모델 생성에 있어서의 윤리적 책임 또한 항상 염두에 두어야 겠다는 깨달음을 얻었다.

참고문헌

[주제 선정]

- <https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates>
- <https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2023.html>
- https://www.umt.edu/rural-disability-research/focus-areas/disability_maps/state-maps/west-virginia.php

참고문헌

[변수 선택 근거]

- <https://www.bbc.com/korean/features-63275934>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6494986/>
- <https://typeset.io/questions/what-is-the-relationship-between-education-level-and-drug-54g49azopt>
- <https://pubmed.ncbi.nlm.nih.gov/12507796/>
- <https://www.helpguide.org/articles/addictions/substance-abuse-and-mental-health.htm>
- https://www.cdc.gov/nchs/pressroom/sosmap/drug_poisoning_mortality/drug_poisoning.htm

참고문헌

[모델 제작]

- <https://velog.io/@nomaday/Multiclass-Multilabel-Classification>
- <https://medium.com/@Jerrylzj/methods-for-multi-class-imbalanced-data-classification-574ab4b73d09>
- <https://sungkee-book.tistory.com/34>
- <https://medium.com/apprentice-journal/evaluating-multi-class-classifiers-12b2946e755b>

감사합니다