Drug Recommendation

21. 09. 28 (화)

김준기 배홍직 이상목 이준병 이현준　　　　발표자 : 배홍직

Contents

01

데이터 분석 및 처리

- 데이터 분석
- 데이터 가공

데이터 분석

	review	condition	drugName	rating
0	My son is halfway through his fourth week of ...	ADHD	Guanfacine	8
1	I used to take another oral contraceptive whi...	Birth Control	Lybrel	5
2	This is my first time using any form of birth...	Birth Control	Ortho Evra	8
3	He pulled out but he cummed a bit in me I too...	Emergency Contraception	Levonorgestrel	1
4	Abilify changed my life There is hope I was o...	Bipolar Disorde	Aripiprazole	10
...	...	...	...	...
94439	I was in a very bad place at the time I start...	Bipolar Disorde	Geodon	3
94440	This is the third med I ve tried for anxiety ...	Depression	Vortioxetine	2
94441	I was super against taking medication I ve st...	Anxiety	Ativan	9
94442	I have only been on Tekturna for NUM days Th...	High Blood Pressure	Tekturna	7
94443	This would be my second month on Junel I ve b...	Birth Control	Junel 1.5 / 30	6

94444 rows × 4 columns

```
top_df['drugName'].value_counts()

Levonorgestrel                                     3626
Etonogestrel                                       3314
Ethinyl estradiol / norethindrone                  2551
Nexplanon                                          2149
Ethinyl estradiol / norgestimate                   1943
                                                   ...
Capzasin                                              1
Hydralazine / hydrochlorothiazide / reserpine         1
Larin 24 Fe                                           1
Goody's Extra-Strength Headache Powders               1
Children's Motrin                                     1
Name: drugName, Length: 1129, dtype: int64
```

상위 20개 condition에 대한 데이터는 총 94,444개
이에 사용된 약의 종류는 1,129개

데이터 가공

	review	condition	drugName	rating
0	My son is halfway through his fourth week of ...	ADHD	Guanfacine	8
1	This is my first time using any form of birth...	Birth Control	Ortho Evra	8
2	Abilify changed my life There is hope I was o...	Bipolar Disorde	Aripiprazole	10
3	I had been on the pill for many years When my...	Birth Control	Ethinyl estradiol / levonorgestrel	8
4	I have taken anti depressants for years with ...	Depression	L-methylfolate	10
...	...	...	...	...
55969	I started taking this medication NUM years a...	Insomnia	Zolpidem	10
55970	I just got diagnosed with type NUM My doctor...	Diabetes, Type 2	Invokana	9
55971	Used with my throat cancer and it helped numb...	Chronic Pain	Roxicodone Intensol	10
55972	Had panic attacks and social anxiety starting...	Panic Disorde	Venlafaxine	9
55973	I was super against taking medication I ve st...	Anxiety	Ativan	9

55974 rows × 4 columns

```
top_df['drugName'].value_counts()

Levonorgestrel                          2251
Etonogestrel                            1367
Phentermine                             1285
Ethinyl estradiol / norethindrone       1012
Nexplanon                                846
                                         ...
Colesevelam                                1
NovoLog Mix 70 / 30                        1
Avita                                      1
Ethchlorvynol                              1
NuLYTELY                                   1
Name: drugName, Length: 1025, dtype: int64
```

평가가 좋은 약을 추천해주기 위해 rating이 8이상인 데이터만 사용
rating이 8이상인 데이터는 총 55,974개
사용된 약의 종류는 1,025개

데이터 가공

```
top_df_over10

Levonorgestrel                      2251
Etonogestrel                        1367
Phentermine                         1285
Ethinyl estradiol / norethindrone   1012
Nexplanon                            846
                                     ...
My Way                                11
Janumet                               11
Reglan                                11
Bontril Slow Release                  11
Flurazepam                            11
Name: drugName, Length: 453, dtype: int64
```

```
top_df_over50

Levonorgestrel                      2251
Etonogestrel                        1367
Phentermine                         1285
Ethinyl estradiol / norethindrone   1012
Nexplanon                            846
                                     ...
Geodon                                52
Suvorexant                            51
Amlodipine                            51
Risperidone                           51
Treximet                              51
Name: drugName, Length: 213, dtype: int64
```

```
top_df_over100

Levonorgestrel                      2251
Etonogestrel                        1367
Phentermine                         1285
Ethinyl estradiol / norethindrone   1012
Nexplanon                            846
                                     ...
Orlistat                             106
Lithium                              102
Milnacipran                          102
Lutera                               102
Temazepam                            101
Name: drugName, Length: 137, dtype: int64
```

drugName의 unique value가 1,025개
-> 리뷰 개수가 적은 drugName은 제거

데이터 가공

리뷰가 100개 이상인 drugName에 대한 데이터만 사용하여
총 42,745개의 데이터셋을 사용

데이터 가공

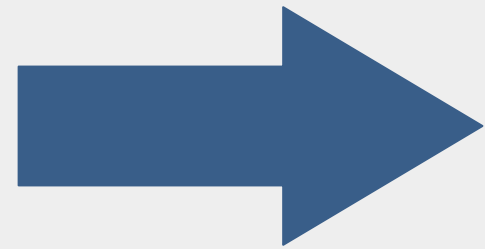|	condition	drugName
0	Birth Control	Etonogestrel, Ethinyl estradiol / norethindrone, Nexplanon, Levonorgestrel, Ethinyl estradiol / levonorgestrel, Ethinyl estradiol / norgestimate, Implanon, Mirena, Skyla, Lo Loestrin Fe, Ethinyl estradiol / etonogestrel, NuvaRing, Copper, Drospirenone / ethinyl estradiol, Medroxyprogesterone, Desogestrel / ethinyl estradiol, Depo-Provera, Sprintec, Ethinyl estradiol / norelgestromin, ParaGard, Norethindrone, Microgestin Fe 1 / 20, Loestrin 24 Fe, Tri-Sprintec, Aviane, Ortho Evra, Apri, Liletta, Lutera, TriNessa, Yaz, Ortho Tri-Cyclen Lo, Junel Fe 1 / 20, Xulane, Minastrin 24 Fe, Gildess Fe 1 / 20, Kyleena, Orsythia, Seasonique, Ethinyl estradiol / norgestrel, Drospirenone / ethinyl estradiol / levomefolate calcium, Reclipsen, Generess Fe, Chateal, Mononessa
1	Depression	Bupropion, Sertraline, Venlafaxine, Pristiq, Desvenlafaxine, Citalopram, Escitalopram, Vortioxetine, Duloxetine, Vilazodone, Lexapro, Cymbalta, Fluoxetine, Viibryd, Zoloft, Trintellix, Wellbutrin, Celexa, Mirtazapine, Prozac, Effexor, Aripiprazole, Wellbutrin XL, Abilify, Paroxetine, Effexor XR
2	Pain	Tramadol, Acetaminophen / hydrocodone, Oxycodone, Acetaminophen / oxycodone, Tapentadol, Diclofenac, Hydromorphone, Gabapentin, Nucynta, Dilaudid, Ketorolac, Oxymorphone, OxyContin, Percocet, Morphine, Methadone, Fentanyl, Amitriptyline, Norco, Toradol
3	Anxiety	Escitalopram, Alprazolam, Buspirone, Clonazepam, Lexapro, Hydroxyzine, Lorazepam, Xanax, BuSpar, Klonopin, Diazepam, Gabapentin, Ativan, Paroxetine, Mirtazapine, Venlafaxine, Paxil, Duloxetine, Cymbalta
4	Acne	Isotretinoin, Adapalene / benzoyl peroxide, Epiduo, Doxycycline, Accutane, Benzoyl peroxide / clindamycin, Minocycline, Tretinoin, Ethinyl estradiol / norgestimate, Drospirenone / ethinyl estradiol, Spironolactone, Adapalene, Ethinyl estradiol / norethindrone, Aczone, Dapsone, Duac
5	Bipolar Disorde	Lamotrigine, Lamictal, Quetiapine, Lurasidone, Latuda, Seroquel, Aripiprazole, Abilify, Lithium, Asenapine, Saphris, Divalproex sodium, Ziprasidone, Geodon, Olanzapine
6	Insomnia	Zolpidem, Trazodone, Ambien, Suvorexant, Quetiapine, Belsomra, Temazepam, Eszopiclone, Doxepin, Lunesta, Mirtazapine, Amitriptyline, Melatonin
7	Weight Loss	Phentermine, Lorcaserin, Belviq, Bupropion / naltrexone, Contrave, Adipex-P, Phentermine / topiramate, Qsymia, Liraglutide
8	Obesity	Bupropion / naltrexone, Contrave, Liraglutide, Phentermine, Saxenda, Lorcaserin, Belviq, Orlistat, Phentermine / topiramate, Qsymia
9	ADHD	Lisdexamfetamine, Vyvanse, Methylphenidate, Amphetamine / dextroamphetamine, Atomoxetine, Strattera, Concerta, Adderall, Guanfacine, Adderall XR
10	Diabetes, Type 2	Liraglutide, Victoza, Dulaglutide, Trulicity, Exenatide, Canagliflozin, Invokana, Bydureon
11	Emergency Contraception	Levonorgestrel, Plan B, Plan B One-Step
12	High Blood Pressure	Lisinopril, Losartan, Amlodipine, Nebivolol
13	Vaginal Yeast Infection	Miconazole, Tioconazole, Fluconazole
14	Abnormal Uterine Bleeding	Medroxyprogesterone, Depo-Provera, Levonorgestrel, Mirena, Ethinyl estradiol / norgestimate, Ethinyl estradiol / levonorgestrel
15	Bowel Preparation	Magnesium sulfate / potassium sulfate / sodium sulfate, Suprep Bowel Prep Kit, Polyethylene glycol 3350 with electrolytes, MoviPrep
16	ibromyalgia	Pregabalin, Savella, Milnacipran, Lyrica, Duloxetine, Cymbalta
17	Smoking Cessation	Varenicline, Chantix, Bupropion
18	Migraine	Sumatriptan, Rizatriptan
19	Anxiety and Stress	Citalopram, Celexa, Fluoxetine, Sertraline, Prozac

상위 20개의 condition중에서 리뷰가 100개를 초과하는 drugName 분포 리스트
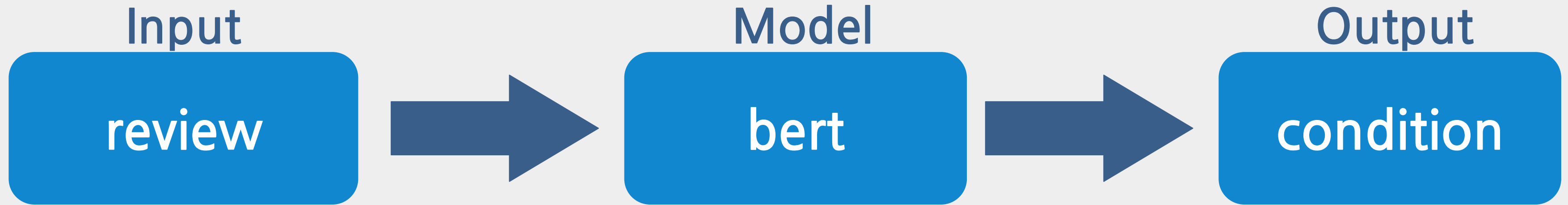
02

모델 적용

condition 분류

Input

Output

review

adhd: 0.6
birth control: 0.7
pain: 0.3
depression: 0.9
anxiety: 0.8
…
bipolar disorder: 0.5

review 문장을
적절한 condition으로
분류한다.

사용자에게 condition을
표시해주는 경우
가장 높은 확률을 가지는
condition을 표시한다.

모델 적용

Input
Model
Output

```
review   →   bert   →   condition
```

모델 검증 결과

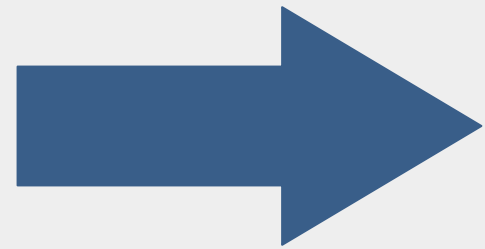| Step	Training Loss	Validation Loss	F1
500	1.184700	0.663190	0.715316
1000	0.604600	0.520742	0.779052
1500	0.492800	0.475694	0.807719
2000	0.432600	0.450248	0.815626

컨디션 예측 confusion matrix

	Birth Control	Depression	Anxiety	Weight Loss	Acne	Pain	Obesity	Emergency Contraception	Bipolar Disorde	ADHD	Smoking Cessation	Insomnia	Panic Disorde	Anxiety and Stress	Bowel Preparation	ibromyalgia	Chronic Pain	Diabetes, Type 2	Migraine	High Blood Pressure
Birth Control	1004	0	1	13	0	13	6	2	44	5	0	2	15	4	5	7	1	1	0	3
Depression	0	252	5	0	0	0	383	0	2	0	3	1	0	0	0	1	0	0	0	1
Anxiety	4	0	1701	1	0	0	108	0	8	1	1	6	3	0	1	10	0	3	0	0
Weight Loss	15	0	1	1432	20	27	5	0	290	3	2	16	40	2	2	47	2	2	0	2
Acne	9	0	0	131	177	11	2	0	214	1	0	3	9	1	4	8	1	0	0	2
Pain	20	0	0	53	1	1062	10	0	179	3	1	10	25	0	1	7	0	0		7
Obesity	1	86	95	1	1	5	9387	0	21	3	4	8	3	0	10	15	0	3	3	2
Emergency Contraception	0	0	0	0	0	0	1	623	0	4	1	1	3	0	2	4	0	0	0	0
Bipolar Disorde	26	0	3	199	52	150	19	1	2523	10	2	11	40	3	9	32	2	0	0	13
ADHD	0	0	0	2	0	1	7	1	13	709	0	17	3	1	21	10	1	7	13	1
Smoking Cessation	0	0	1	5	0	0	10	0	1	2	799	0	0	0	0	5	0	2	2	0
Insomnia	4	0	3	12	0	7	5	0	17	11	0	691	5	2	2	23	0	0	1	0
Panic Disorde	5	0	0	36	2	29	2	3	36	2	2	1	1074	3	0	32	1	0	1	
Anxiety and Stress	4	0	0	2	0	9	1	3	8	3	0	4	10	490	0	46	0	1	0	
Bowel Preparation	5	0	0	1	0	6	1	2	16	20	0	1	6	3	689	10	3	0	424	2
ibromyalgia	5	0	4	20	0	10	7	2	18	10	0	8	15	40	4	1902	0	8	1	45
Chronic Pain	2	0	0	3	1	0	0	1	17	1	0	2	4	0	9	3	617	0	0	
Diabetes, Type 2	0	0	1	3	0	3	2	2	5	4	0	4	4	0	2	8	0	773	0	
Migraine	9	1	0	4	0	4	8	1	16	20	0	5	2	0	326	2	0	1	848	
High Blood Pressure	3	0	3	7	91	6	1	0	41	1	0	8	10	0	1	106	0	0	1	391

drugname 예측

Input

adhd: 0.6
birth control: 0.7
pain: 0.3
depression: 0.9
anxiety: 0.8
...
bipolar disorder: 0.5

Output

drug

137 종류

이전 단계의 Output인 condition 분류 벡터를

Input으로 사용해 **drugname을 예측**한다.

모델 적용

Input

review

Model
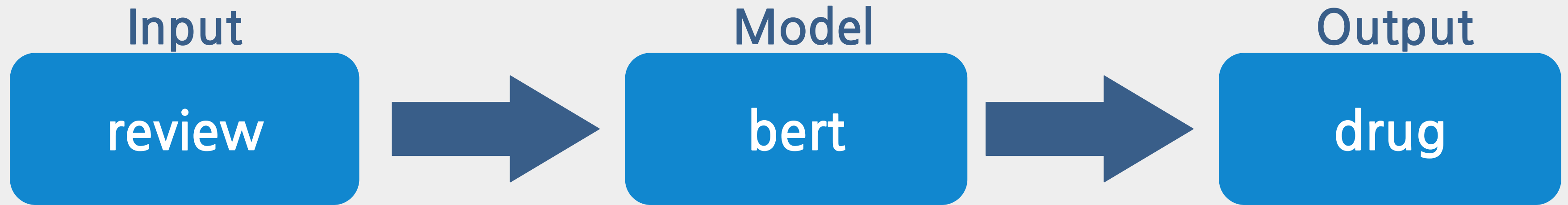
LGBM

Output

drug

모델 검증 결과

```
오차 행렬
[[ 2  0  0 ...  0  0  0]
 [ 0  1  0 ...  0  0  0]
 [ 0  0 10 ...  0  0  0]
 ...
 [ 0  0  0 ...  0  0  0]
 [ 0  0  0 ...  0  3  0]
 [ 0  0  1 ...  0  0 23]]
정확도: 0.2117, 정밀도: 0.1722, 재현율: 0.1634,    F1: 0.1615
```

drugName 분류

Input

review

Model

bert

Output

drug

모델 검증 결과

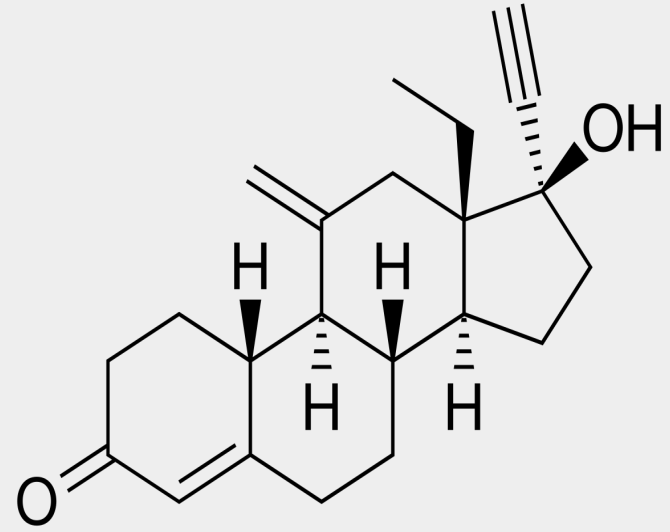| Step	Training Loss	Validation Loss	F1
500	3.667300	2.701502	0.219613
1000	2.499900	2.287778	0.273221

03

문제점

drugName 분류

Etonogestrel(성분 이름)

Etonogestrel을 사용하는 약

사실상 같은 약이지만, 여러가지 이름으로 리뷰 데이터에 존재함을 확인할 수 있었음

문제점

1. 상위 20개의 컨디션으로 데이터를 묶었을때, 1위와 2위 및 그 이하의 데이터들의 분포 개수 차이가 심해 불균형 했음. oversampling 또는 undersampling을 고려 중.

2. 사용자가 말하지 않은 증상을 물어보는 시나리오 문제
   - 특정 증상(symptom)은 데이터 셋에 존재하지 않음. 따라서 별도의 symptom table data를 만드는게 좋은 방법인가?
   - symptom table을 만들어도, 입력 증상과 매칭시키기 어려움

3. 20개의 컨디션으로 130개의 drugName 분류 문제