Drug Recommendation

21. 10. 19 (화)

김준기 배홍직 이상목 이준병 이현준

발표자 : 김준기

Contents

데이터 분석 및 처리 01 데이터 정제
데이터 추가

모델 적용 02 LightGBM을 이용한 약 성분 예측

문제점 03 condition을 이용한 drugName 성분
예측에서 정확도가 낮은 문제

01

데이터 분석 및 처리

- 데이터 정제
- 데이터 추가

데이터 정제

	A	B	C
1		drugName	성분
2	Levonorgestrel	2251	Levonorgestrel
3	Etonogestrel	1367	Etonogestrel
4	Phentermine	1285	Phentermine
5	Ethinyl estradiol / norethindrone	1012	Ethinyl estradiol, Norethindrone
6	Nexplanon	846	Etonogestrel
7	Ethinyl estradiol / norgestimate	780	Ethinyl estradiol, Norgestimate
8	Ethinyl estradiol / levonorgestrel	690	Ethinyl estradiol, Levonorgestrel
9	Varenicline	678	Varenicline
10	Escitalopram	639	Escitalopram
11	Clonazepam	597	Clonazepam
12	Sertraline	589	Sertraline
13	Citalopram	585	Citalopram
14	Alprazolam	578	Alprazolam
15	Bupropion	576	Bupropion
16	Mirena	571	Levonorgestrel
17	Chantix	555	Varenicline
18	Bupropion / naltrexone	549	Bupropion, Naltrexone
19	Liraglutide	536	Liraglutide
20	Contrave	521	Bupropion, Naltrexone
21	Skyla	506	Levonorgestrel
22	Plan B	504	Levonorgestrel
23	Implanon	498	Etonogestrel
24	Lexapro	489	Escitalopram
25	Duloxetine	425	Duloxetine
26	Gabapentin	424	Gabapentin
27	Magnesium sulfate / potassium sulfate / sodium sulfate	423	Sodium sulfate
28	Lorcaserin	417	Lorcaserin
29	Isotretinoin	409	Isotretinoin
30	Belviq	404	Lorcaserin
31	Venlafaxine	401	Venlafaxine
32	Copper	399	Copper
33	Suprep Bowel Prep Kit	397	Sodium sulfate
34	Tramadol	392	Tramadol
35	Fluoxetine	388	Fluoxetine
36	Cymbalta	388	Duloxetine
37	Lamotrigine	383	Lamotrigine
38	Xanax	382	Alprazolam
39	Oxycodone	376	Oxycodone
40	Quetiapine	369	Quetiapine
41	Plan B One-Step	357	Levonorgestrel
42	Buprenorphine / naloxone	354	Buprenorphine, Naloxone
43	Drospirenone / ethinyl estradiol	352	Drospirenone, Ethinyl estradiol
44	Desvenlafaxine	347	Desvenlafaxine
45	Klonopin	341	Clonazepam
46	Zoloft	333	Sertraline
47	NuvaRing	326	Ethinyl estradiol, Etonogestrel

	A	B	C
94	Pregabalin	168	Pregabalin
95	Tretinoin	167	Tretinoin
96	Seroquel	167	Quetiapine
97	Viibryd	165	Vilazodone
98	Ativan	165	Lorazepam
99	Lurasidone	162	Lurasidone
100	Wellbutrin	161	Bupropion
101	Dulcolax	161	Bisacodyl
102	Aripiprazole	161	Aripiprazole
103	Vortioxetine	158	Vortioxetine
104	Dilaudid	156	Hydromorphone
105	Latuda	156	Lurasidone
106	Fentanyl	155	Fentanyl
107	Sumatriptan	155	Sumatriptan
108	Morphine	154	Morphine
109	Abilify	151	Aripiprazole
110	OxyContin	149	Oxycodone
111	Tri-Sprintec	148	Ethinyl estradiol, Norgestimate
112	Ortho Evra	147	Ethinyl estradiol, Norelgestromin
113	Hydroxyzine	144	Hydroxyzine
114	Lyrica	143	Pregabalin
115	Magnesium citrate	143	Magnesium citrate
116	Loestrin 24 Fe	142	Ethinyl estradiol, Norethindrone
117	BuSpar	140	Buspirone
118	Rizatriptan	137	Rizatriptan
119	Microgestin Fe 1 / 20	135	Ethinyl estradiol, Norethindrone, Ferrous fuma
120	Percocet	132	Acetaminophen, Oxycodone
121	Spironolactone	130	Spironolactone
122	Topiramate	130	Topiramate
123	Adderall	127	Amphetamine, Dextroamphetamine
124	Norethindrone	125	Norethindrone
125	TriNessa	124	Ethinyl estradiol, Norgestimate
126	Yaz	122	Drospirenone, Ethinyl estradiol
127	Concerta	119	Methylphenidate
128	Trintellix	118	Vortioxetine
129	Adapalene	118	Adapalene
130	Effexor	116	Venlafaxine
131	Wellbutrin XL	116	Bupropion
132	Exenatide	115	Exenatide
133	Minocycline	115	Minocycline
134	Aviane	115	Ethinyl estradiol, Levonorgestrel
135	Polyethylene glycol 3350 with electrolytes	115	Polyethylene glycol 3350 with electrolytes
136	Savella	114	Milnacipran
137	Tapentadol	114	Tapentadol
138	Dulaglutide	107	Dulaglutide
139	Orlistat	106	Orlistat
140	Lithium	102	Lithium
141	Lutera	102	Ethinyl estradiol, Levonorgestrel
142	Milnacipran	102	Milnacipran
143	Temazepam	101	Temazepam

상위 147개의 drugName들에 대해 generic name을 추가한 table 생성

데이터 정제

```
Levonorgestrel                          2251
Etonogestrel                            1367
Phentermine                             1285
Ethinyl estradiol / norethindrone       1012
Nexplanon                                846
                                         ...
Orlistat                                 106
Milnacipran                              102
Lithium                                  102
Lutera                                   102
Temazepam                                101
Name: drugName, Length: 143, dtype: int64
```

```
Levonorgestrel                          4189
Etonogestrel                            2711
Phentermine                             1570
Ethinyl estradiol, Norgestimate         1250
Varenicline                             1233
                                         ...
Tapentadol                               114
Dulaglutide                              107
Orlistat                                 106
Lithium                                  102
Temazepam                                101
Name: 성분, Length: 85, dtype: int64
```

기존 143개의 약 이름(drugName)을 약의 성분인 generic name으로 통일 시켜
유니크 값이 85개로 줄어듦

데이터 추가

	drugName	성분	review	condition	rating
0	Levonorgestrel	Levonorgestrel	My time with skyla has been an absolute breez...	Birth Control	10
1	Levonorgestrel	Levonorgestrel	I am NUM yrs old no kids and have never been...	Birth Control	10
2	Levonorgestrel	Levonorgestrel	Let me just tell you to those reading these r...	Emergency Contraception	9
3	Levonorgestrel	Levonorgestrel	Ok girls if your reading this it s because yo...	Emergency Contraception	10
4	Levonorgestrel	Levonorgestrel	I have had my Mirena for three months now Ins...	Birth Control	10
...	...	...	...	...	...
43981	Temazepam	Temazepam	I have taken temazepam now for NUM years Pro...	Insomnia	9
43982	Temazepam	Temazepam	Restoril helps me get to sleep and helps prev...	Insomnia	10
43983	Temazepam	Temazepam	This is the only thing that helps me fall asl...	Insomnia	10
43984	Temazepam	Temazepam	It works well for me No side effects	Insomnia	10
43985	Temazepam	Temazepam	Started taking Temazepam NUM mg two months a...	Insomnia	9

43986 rows × 5 columns

training data set에 성분 column을 추가하고, generic name table을 이용하여 데이터 추가

데이터 추가

	condition	성분	len
0	[Birth Control]	Levonorgestrel	13
1	[Birth Control]	Levonorgestrel	13
2	[Emergency Contraception]	Levonorgestrel	23
3	[Emergency Contraception]	Levonorgestrel	23
4	[Birth Control]	Levonorgestrel	13
...	...	...	...
44577	[Pain, Opiate Dependence]	Buprenorphine	2
44578	[Chronic Pain, Opiate Dependence]	Buprenorphine	2
44579	[Obesity, ibromyalgia]	Topiramate	2
44580	[Obesity, Bipolar Disorde]	Topiramate	2
44581	[ibromyalgia, Bipolar Disorde]	Topiramate	2

44582 rows × 3 columns

성분	len	condition_code	ondhot
Levonorgestrel	13	[6]	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...
Levonorgestrel	13	[6]	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...
Levonorgestrel	23	[11]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...
Levonorgestrel	23	[11]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...
Levonorgestrel	13	[6]	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...
...	...	...	...
Buprenorphine	2	[10, 4]	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...
Buprenorphine	2	[19, 4]	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
Topiramate	2	[16, 7]	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...
Topiramate	2	[16, 2]	[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
Topiramate	2	[7, 2]	[0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...

{'ADHD': 18,
 'Acne': 2,
 'Anxiety': 7,
 'Bipolar Disorde': 10,
 'Birth Control': 9,
 'Bowel Preparation': 19,
 'Chronic Pain': 12,
 'Constipation': 14,
 'Depression': 8,
 'Diabetes, Type 2': 13,
 'Emergency Contraception': 16,
 'High Blood Pressure': 15,
 'Insomnia': 5,
 'Migraine': 3,
 'Obesity': 1,
 'Opiate Dependence': 0,
 'Pain': 11,
 'Panic Disorde': 17,
 'Smoking Cessation': 6,
 'ibromyalgia': 4}

특정(약)이 사용되는 condition들에 대해 가능한 한 모든 조합을 구하고, training data set 생성

데이터 셋

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20	label
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	44
1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	44
2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	44
3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	44
4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	44
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
44577	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	10
44578	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	10
44579	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	76
44580	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	76
44581	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	76

44582 rows × 21 columns

상위 20개의 condition들이 attribute이고, target label이 약(성분)인 데이터

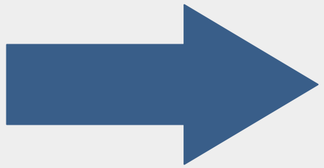
02

모델 적용

drugname 예측

Input

[adhd]
[birth control]
[pain]
…
[adhd, depression]
[adhd, anxiety]
…
[anxiety, pain, sore]
…



Output

generic drug name

85 종류

특정 약에 사용될 수 있는 condition의 조합들을

Input으로 사용해 drugname의 성분(generic)을 예측한다.

모델 적용

모델 검증 결과

```
오차 행렬
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
정확도: 0.4113, 정밀도: 0.0494, 재현율: 0.0745,    F1: 0.0530
```

대부분 0으로 예측함

모델 적용

condition 조합을 추가하지 않은
기존 training data set으로 학습한 경우

```
[[  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0   0]
 ...
 [  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0 146]]
정확도: 0.3162, 정밀도: 0.0927, 재현율: 0.1798,    F1: 0.1094
```

(drugName, 데이터 개수)

```
[(44, 4189),
 (35, 2711),
 (64, 1570),
 (34, 1250),
 (80, 1233),
 (32, 1154),
 (27, 1128),
 (13, 1070),
 (45, 1045),
 (4, 960),
 (16, 938),
 (70, 922),
 (30, 907),
 (15, 889),
 (12, 853),
 (49, 821),
 (52, 820),
 (26, 813),
 (81, 713),
 (42, 662),
 (38, 646),
 (43, 622),
 (17, 610),
 (46, 554),
 (68, 536),
 (3, 531),
 (62, 525),
 (65, 504),
 (84, 487),
 (63, 477),
 (48, 476),
 (24, 474),
```

03

문제점

drugName(성분) 예측

1. 평가지표가 없고 맞추지를 못한다

2. attribute가 condition이고, target이 generic drug name 예측 모델

   -> 예측을 잘 못하는게 데이터 셋이 불균형해서 문제인가?

   -> 이외에 가능성있는 문제

(drugName, 데이터 개수)

```
[(12, 126),
 (20, 126),
 (26, 62),
 (19, 62),
 (39, 62),
 (81, 62),
 (77, 30),
 (5, 30),
 (38, 14),
 (27, 14),
 (16, 14),
 (15, 14),
 (70, 14),
 (78, 14),
 (43, 6),
 (63, 6),
 (10, 6),
 (48, 6),
 (68, 6),
 (58, 6),
 (4, 6),
 (76, 6),
 (46, 2),
 (47, 2),
 (45, 2),
 (33, 2),
 (44, 2),
 (24, 2),
```