빅콘테스트 퓨처스리그
항공지연 예측

공항노숙 이제 그만!

목차

1. 데이터 이해 및 EDA

데이터 통일

AFSNT(이후 train) 데이터와 AFSNT_DLY(이후 test) 데이터에 대해서 test 예측을 위해 **test 데이터 변수에 맞게 train 데이터의 변수를 통일** 하였고, A/C지연을 제외하면 **날씨 관련 지연**이 가장 많아 train 데이터에 대해 **매년 9/16~9/30 기간만 추출**

데이터셋 구분

test 데이터의 **M항공사**와 **test 데이터에서만 발견되는 FLT**의 경우 이후 모델링에서 각각 항공사와 FLT에 대한 변수를 사용할 수 없으므로 **전체 데이터셋과 구분**하여 진행

데이터 전처리 & EDA

train와 test 데이터에 대한 **전처리**를 진행하고 train 데이터에 대한 **EDA**를 바탕으로 **새로운 변수 생성**

1.1 데이터 통일(Train Data)

- 변수 통일
  REG, ATT, IRR, DRR, CNL, CNR 제거
- 데이터 범위 통일
  지연 요인 중에서 A/C 관련 지연을 제외하면 날씨에 관련된 지연이 가장 많은 것을 확인할 수 있었고, 월별 지연율이 상이하다는 것을 통해 사계절의 특징이 뚜렷한 우리나라에서 test 데이터에 대한 정확한 분석이 진행되려면 1년 중 같은 기간에 대해 분석이 진행되어야 한다고 판단

  -> 2017년 9/16 ~ 9/30, 2018년 9/16 ~ 30 사용

*** C10(제방빙작업)은 날씨에 의한 지연이라고 판단**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 987709 entries, 0 to 987708
Data columns (total 17 columns):
SDT_YY     987709 non-null int64
SDT_MM     987709 non-null int64
SDT_DD     987709 non-null int64
SDT_DY     987709 non-null object
ARP        987709 non-null object
ODP        987709 non-null object
FLO        987709 non-null object
FLT        987709 non-null object
REG        979446 non-null object
AOD        987709 non-null object
IRR        987709 non-null object
STT        987709 non-null object
ATT        987709 non-null object
DLY        987709 non-null object
DRR        118937 non-null object
CNL        987709 non-null object
CNR        8259 non-null object
```

Train Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16076 entries, 0 to 16075
Data columns (total 12 columns):
SDT_YY      16076 non-null int64
SDT_MM      16076 non-null int64
SDT_DD      16076 non-null int64
SDT_DY      16076 non-null object
ARP         16076 non-null object
ODP         16076 non-null object
FLO         16076 non-null object
FLT         16076 non-null object
AOD         16076 non-null object
STT         16076 non-null object
DLY         0 non-null float64
DLY_RATE    0 non-null float64
```

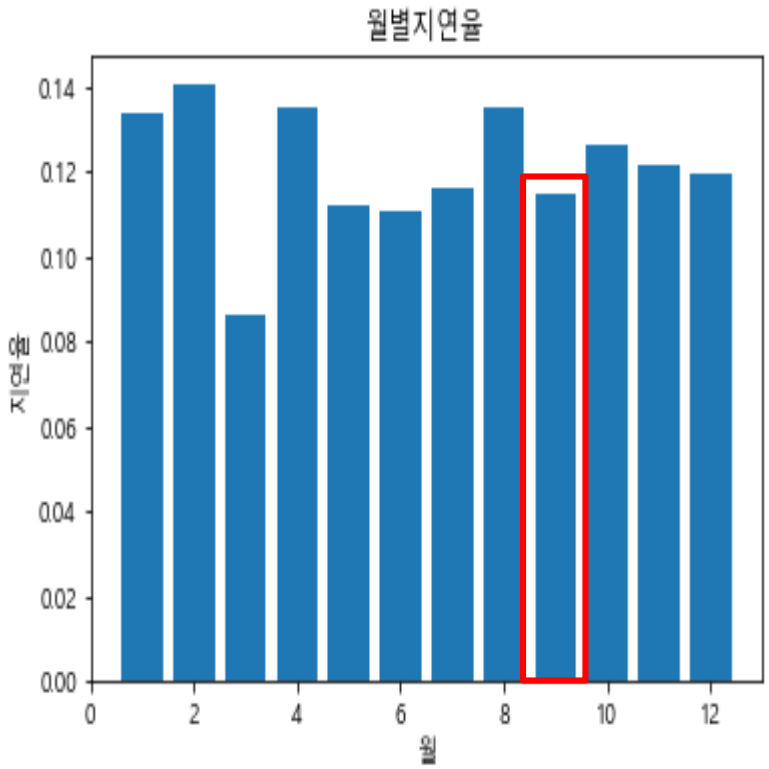
Test Data

1.1 데이터 통일(Train Data)

	DRR	DLY
18	C02	107738
17	C01	2031
0	A01	1524
25	C10	1227
32	D01	950
19	C03	907
29	C14	873
35	Z99	664
4	A05	596
10	B01	417

A/C지연 제외

	code	DLY
0	A	4267
2	C	2668
3	D	982
4	Z	698
1	B	553



A/C지연을 제외하면
날씨관련 지연이 가장 많음

월별로 지연율 상이

1.2 데이터셋 구분(Test Data)

- M 항공사
- Test 데이터에만 있는 FLT 데이터

Train과 Test의 FLO

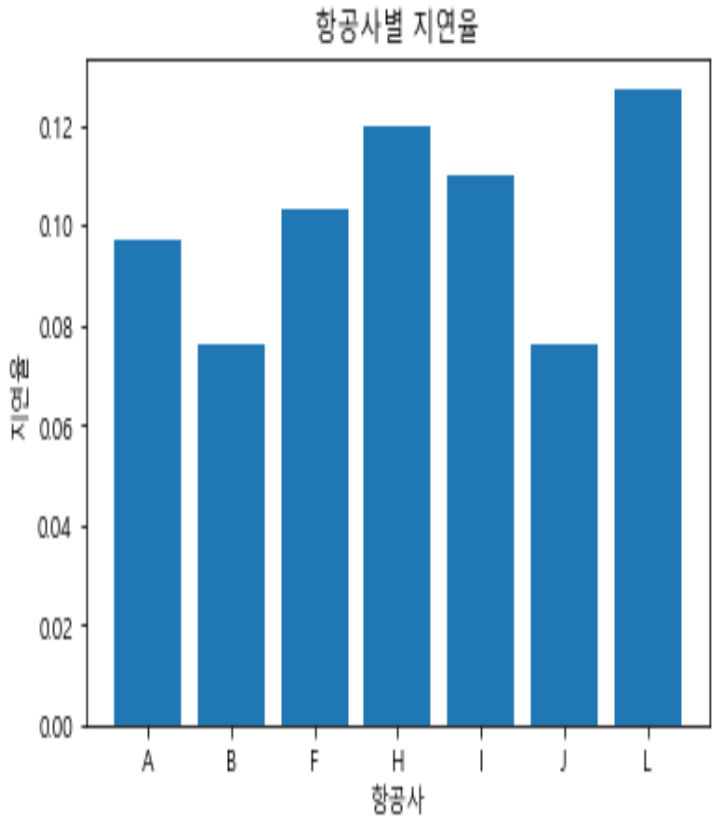
```
array(['J', 'B', 'F', 'A', 'H', 'I', 'L'], dtype=object)

array(['L', 'J', 'M', 'B', 'A', 'H', 'F', 'I'], dtype=object)
```

Test Unique FLT

```
array(['J1610', 'J1611', 'J1605', 'H1207', 'J1258', 'J1259', 'I1582',
       'I1583', 'J1011', 'J1807', 'A1125', 'J1812', 'J1413', 'J1856',
       'J1016', 'J1809', 'J1857'], dtype='<U5')
```

	FLT	DLY2
731	L1932	0.066667
730	L1931	0.050000
729	L1907	0.400000
728	L1906	0.283333
727	L1905	0.300000
726	L1904	0.266667
725	L1903	0.133333
724	L1902	0.166667
723	L1901	0.066667
722	L1816	0.066667



항공사와 편명별로 지연율 상이 → 데이터셋 구분

1.3 데이터 전처리 & EDA

시간관련 변수

- **변수생성**

  SDT_YY, SDT_MM, SDT_DD, STT 변수를 이용하여 YMD라는 datetime 형식의 변수 생성

  ex) 2017-09-16 09:30:00

  이후 YMD 변수를 이용해 hour, minute 변수를 추가적으로 생성

- **데이터 범위 통일**

  Train 데이터의 경우 Test 데이터에 없는 시간대가 존재해 제거

- **시간대별 지연율 계산**

  주로 오후 시간대에 지연율이 높고 반대로 오전 시간대에 지연율이 낮음.

```
array(['9', '10', '12', '16', '19', '20', '13', '15', '18', '7', '8',
       '11', '14', '22', '17', '21', '6', '23', '0'], dtype=object)

array(['9', '7', '14', '13', '20', '19', '16', '15', '8', '18', '21',
       '22', '10', '11', '12', '6', '17'], dtype=object)
```

	hour	DLY2
3	13	0.134905
11	21	0.133484
6	16	0.131725
7	17	0.131231
5	15	0.122708
8	18	0.116444
9	19	0.110119
4	14	0.099440
10	20	0.099439
0	10	0.081519
1	11	0.080675
15	8	0.071232
16	9	0.067449
2	12	0.064182
12	22	0.037037
14	7	0.029751
13	6	0.025564

1.3 데이터 전처리 & EDA

경로관련 변수

- **변수생성**
  ARP, ODP 변수를 이용하여 경로라는 변수 생성
  ex) 3_6
- **외부 데이터 이용**
  한국공항공사의 공항별 통계를 통해 ARP, ODP에 대해 각 공항을 대치
- **ARP, ODP별 지연율 계산 -> 경로별 지연율 계산**
  출발 및 도착공항별 지연율을 계산하였을 때, 공항에 따른 지연율 차이가 있었지만 지연율이 높은 공항의 특징을 확인하기 어려웠으나 경로별로 진행했을 때, 3공항(제주)이 포함된 경로가 높은 지연율 보임

* Train 데이터와 동일한 기간의 통계자료를 통해
  Train 데이터에서의 각 공항별 운행횟수와 비교하는 방식
* 앞의 통일 과정으로 ARP10(양양) 제거됨

공항명	운항(편수)		
	도착	출발	계
김포	6,182	6,206	12,388
김해	4,861	4,866	9,727
제주	7,841	7,837	15,678
대구	1,475	1,476	2,951
광주	558	558	1,116
무안	390	389	779
청주	874	872	1,746
양양	11	11	22
여수	216	216	432
울산	267	267	534
사천	83	83	166
포항	28	28	56
군산	93	93	186
원주	49	48	97
인천	17,790	17,776	35,566
합 계	**40,718**	**40,726**	**81,444**

1.3 데이터 전처리 & EDA

경로관련 변수

	ARP	공항명	DLY2
13	ARP9	여수	0.030952
9	ARP5	울산	0.032787
1	ARP11	포항	0.033333
11	ARP7	무안	0.033333
3	ARP13	군산	0.053333
10	ARP6	청주	0.061224
6	ARP2	김해	0.070476
0	ARP1	김포	0.076990
4	ARP14	원주	0.083333
8	ARP4	광주	0.083411
12	ARP8	대구	0.092117
2	ARP12	사천	0.106250
5	ARP15	인천	0.115385
7	ARP3	제주	0.127993

출발공항별 지연율

	ODP	공항명	DLY2
1	ARP11	포항	0.025000
5	ARP15	인천	0.036199
11	ARP7	무안	0.044444
9	ARP5	울산	0.047131
13	ARP9	여수	0.066667
7	ARP3	제주	0.084768
3	ARP13	군산	0.086667
8	ARP4	광주	0.099344
6	ARP2	김해	0.103700
2	ARP12	사천	0.106250
0	ARP1	김포	0.107205
12	ARP8	대구	0.124889
10	ARP6	청주	0.128015
4	ARP14	원주	0.133333

도착공항별 지연율

	경로	DLY2
29	4_1	0.190476
2	12_3	0.175000
22	3_2	0.152311
19	3_12	0.150000
11	1_4	0.142857
27	3_8	0.133796
21	3_14	0.133333
25	3_6	0.128015
18	3_1	0.127383
5	15_2	0.127329

3공항(제주)이 포함된 경로가 지연율 높음

2.1 시계열 분석을 통한 날씨예측

왜 날씨를 예측하려고 했는가 ?

	DRR	DLY
18	C02	107738
17	C01	2031
0	A01	1524
25	C10	1227
32	D01	950
19	C03	907
29	C14	873
35	Z99	664
4	A05	596
10	B01	417

A/C지연 제외

	code	DLY
0	A	4267
2	C	2668
3	D	982
4	Z	698
1	B	553

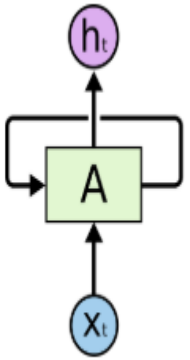
A/C지연을 제외한,
최초지연을 경우,
날씨관련 지연이 가장 많음

2.1 시계열 분석을 통한 날씨예측

1. 왜 시계열을 사용했는가?

날씨는 연속적으로 측정가능한 변수이며,
분석대상으로 설정한 데이터 또한, 1시간이라는 일정한 시간을
간격으로 구성되어 있었기 때문이다.

2. 왜 LSTM를 사용했는가?



LSTM은 순환 신경망으로 은닉층의 상태를 업데이트할 때,
시계열 데이터에서의 패턴을 감지할 수 있기 때문이다.

2.2 LSTM을 통한 날씨예측

분석대상 데이터

공항날씨 : 양양, 무안, 인천, 제주, 김포, 여수, 울산

제일 가까운 관측소 : 원주, 청주, 포항, 사천(진주), 김해, 군산, 광주, 대구

총 15개 지역(공항)

2.2 LSTM을 통한 날씨예측

분석 프로세스

각 공항의 시간별 기온, 강수량, 전운량, 풍속 살펴보기

각 공항의 일별 강수량, 전운량 살펴보기

각 공항의 일별 전운량을 통한 target 시점의 일별 전운량 예측

2.2 LSTM을 통한 날씨예측

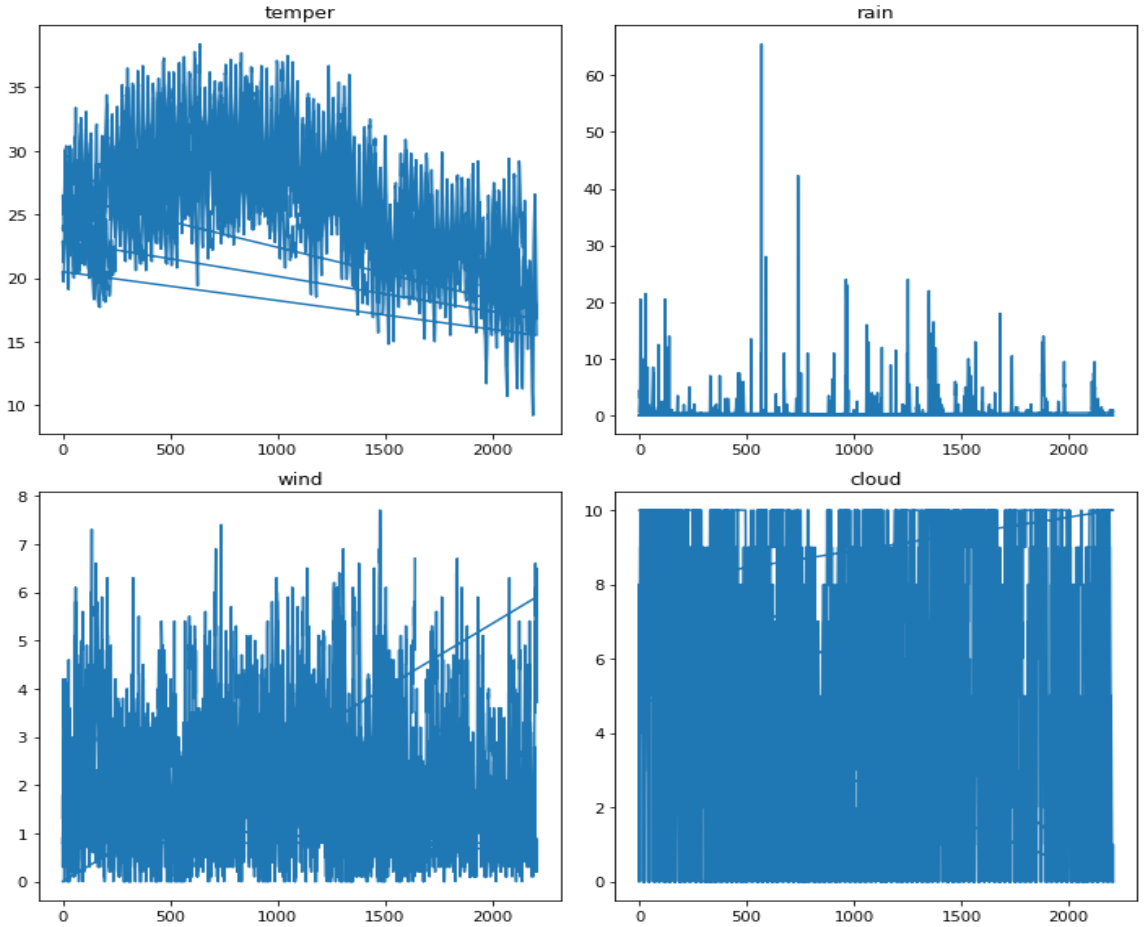
각 공항의 시간별 기온, 강수량, 전운량, 풍속 살펴보기

시간별 분석이 어려운 이유:

다음은 대구의 2016~2019년의 7월부터 9월까지의(~2019.09.08) 기온, 강수량, 풍속, 전운량을 나타내는 그림이다.

시간별로 변하는 변동폭이 너무 커, 시계열 분석에 적합하지 않은 것을 알 수 있다

일별 평균분석 실행



2.2 LSTM을 통한 날씨예측

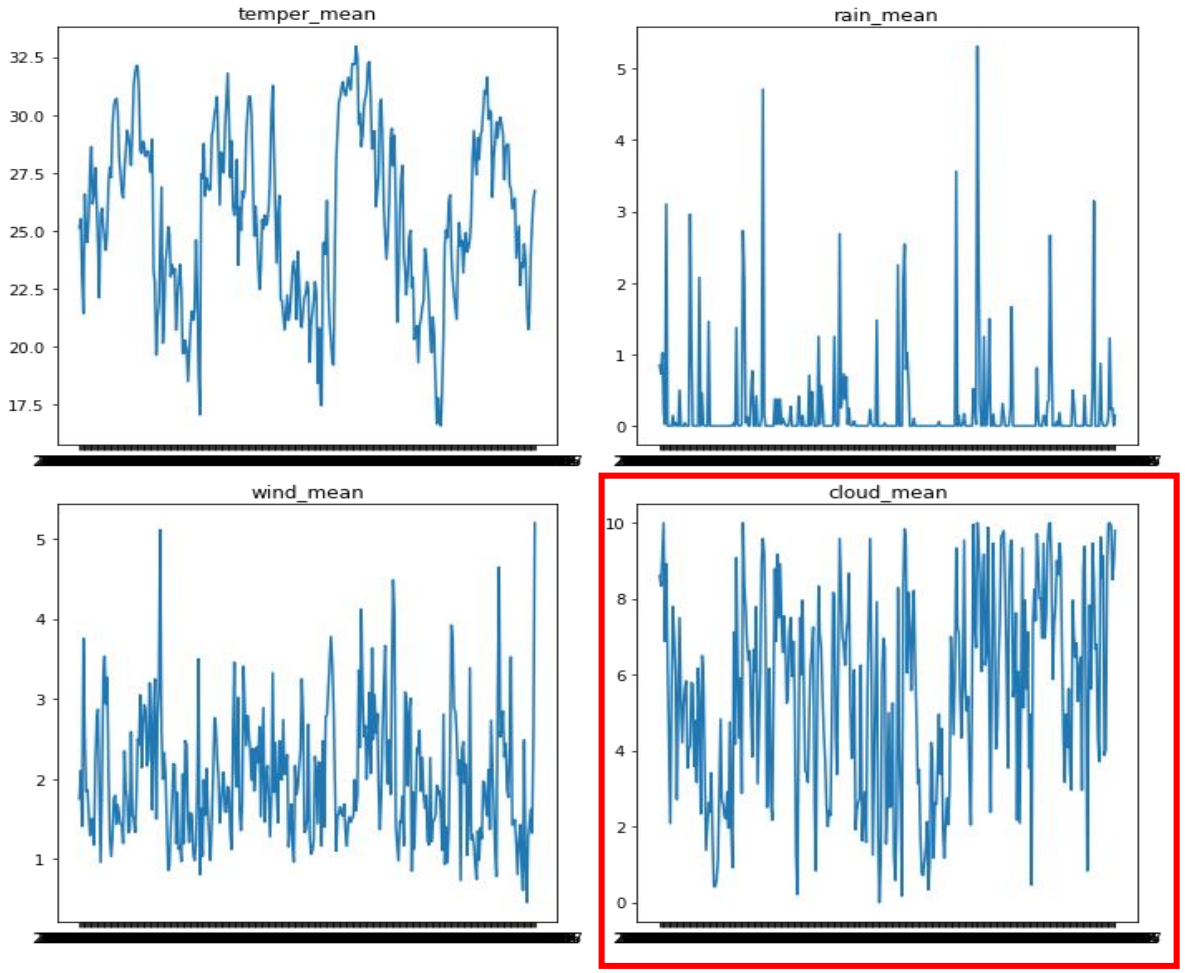
각 공항의 시간별 기온, 강수량, 전운량, 풍속 살펴보기

전운량을 선택한 이유:

전운량을 경우에 비행에 지장을 줄만한 높은 값을 가진 경우가 많았기 때문이다.

(온도의 경우에는 비행에 이상을 줄만한 수준이 없다고 판단하였다.)

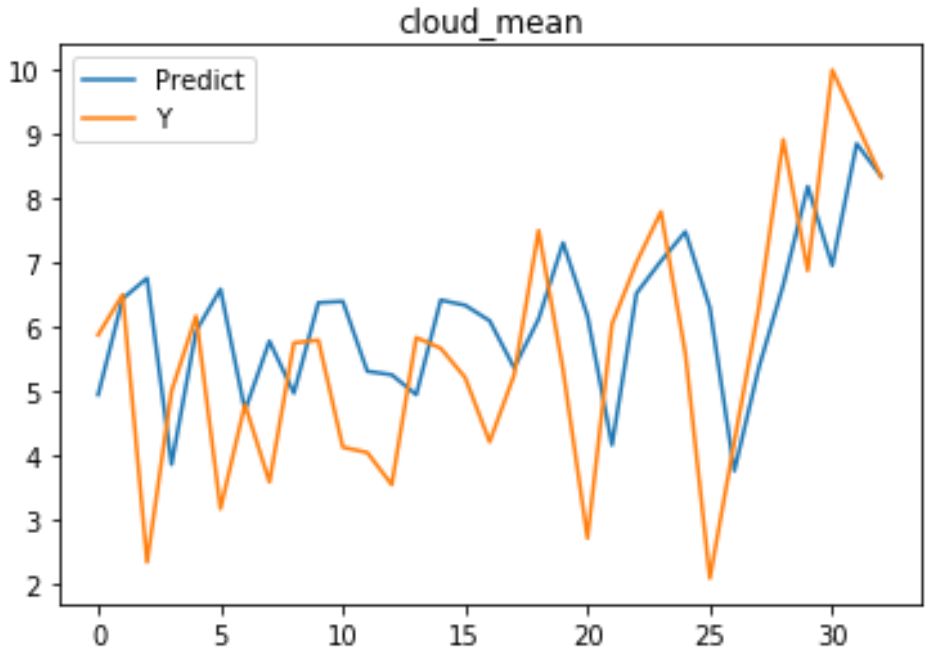
각 공항의 일별 전운량을 통한 target 시점의 일별 전운량 예측



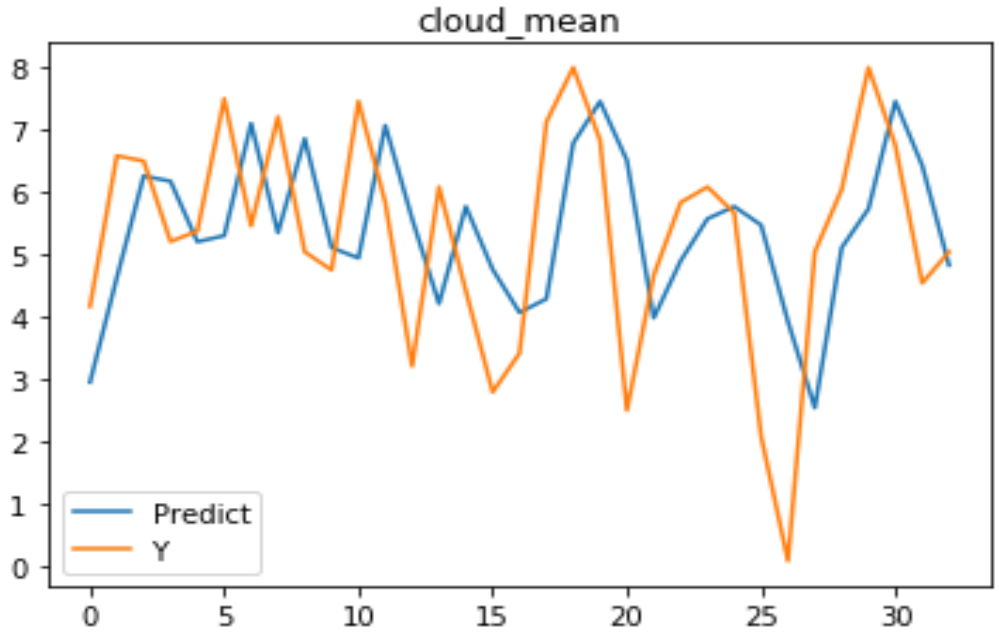
2.2 LSTM을 통한 날씨예측

LSTM 통한 전운량 예측

대구

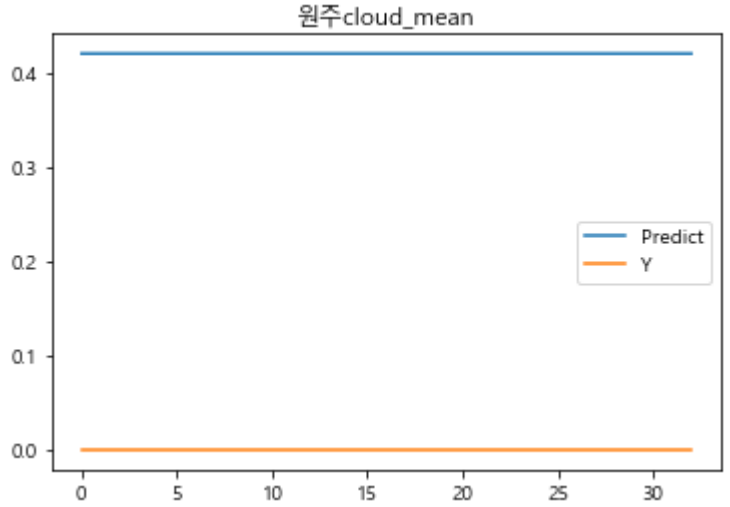


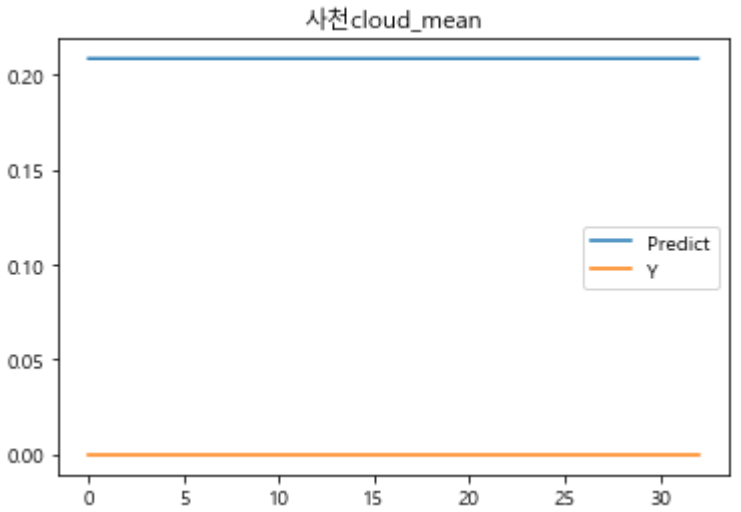
인천

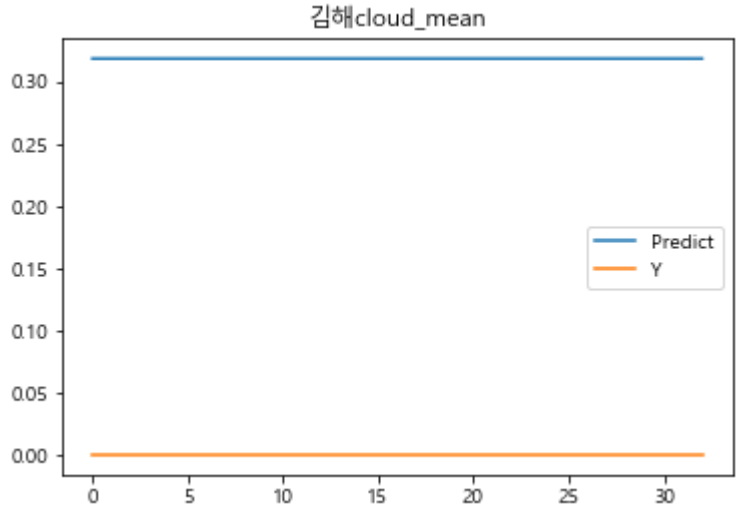


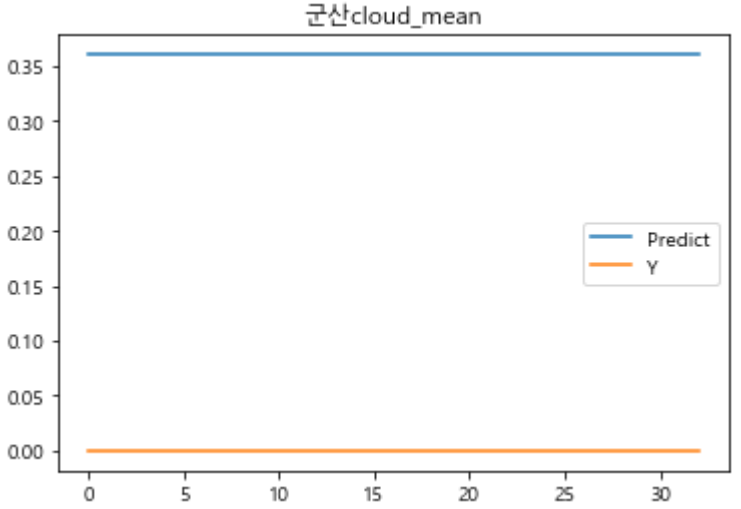
2.2 LSTM을 통한 날씨예측

원주, 사천, 김해, 군산 등의 4개의 공항에서의 예측 X

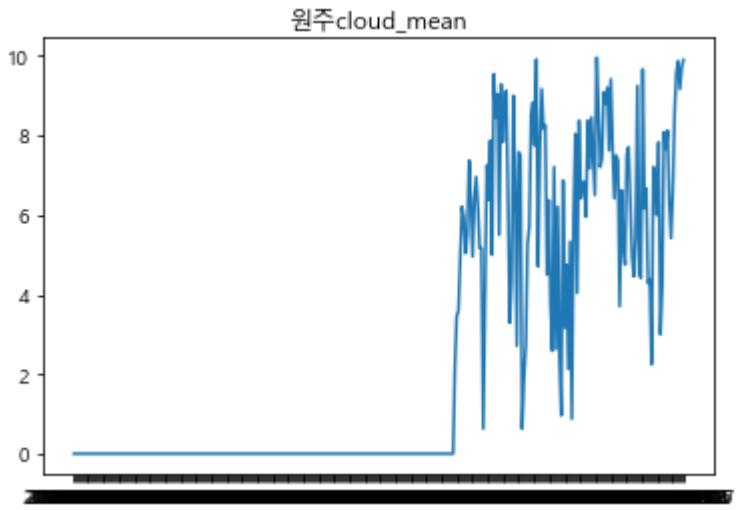


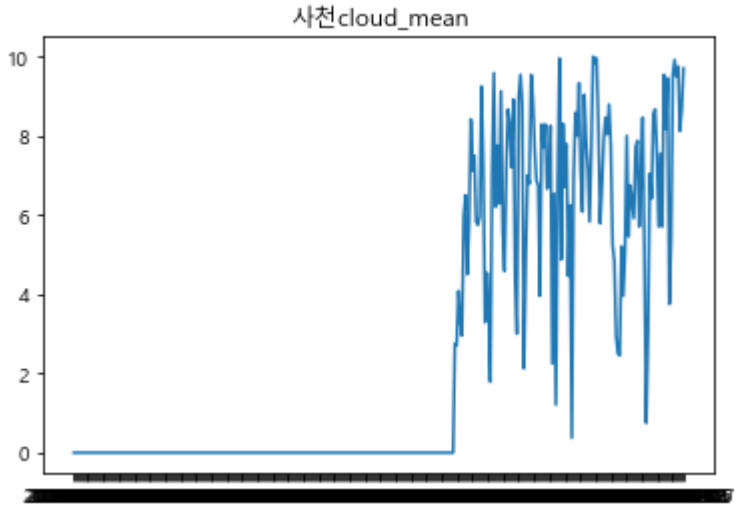


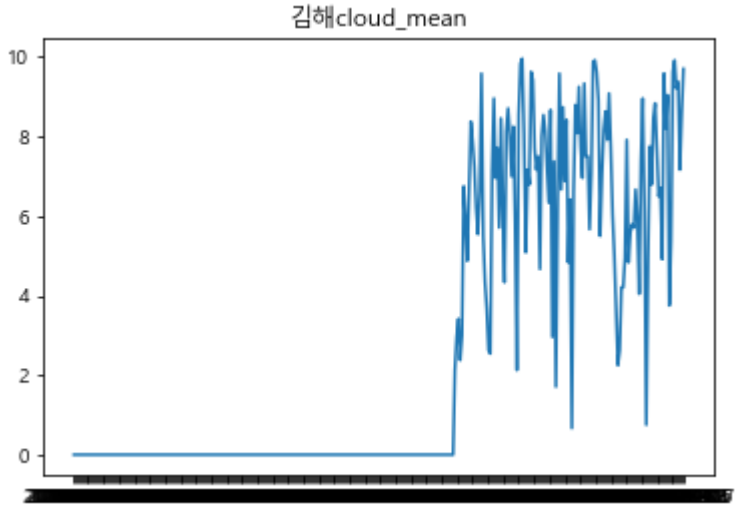


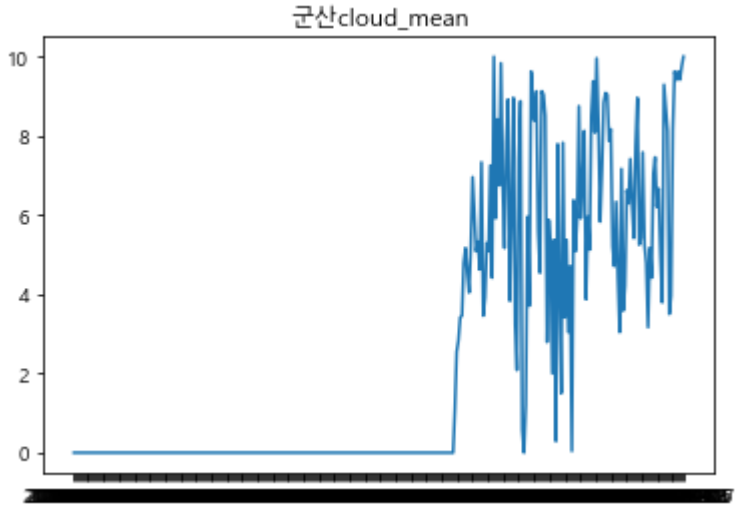


2.2 LSTM을 통한 날씨예측









WHY?

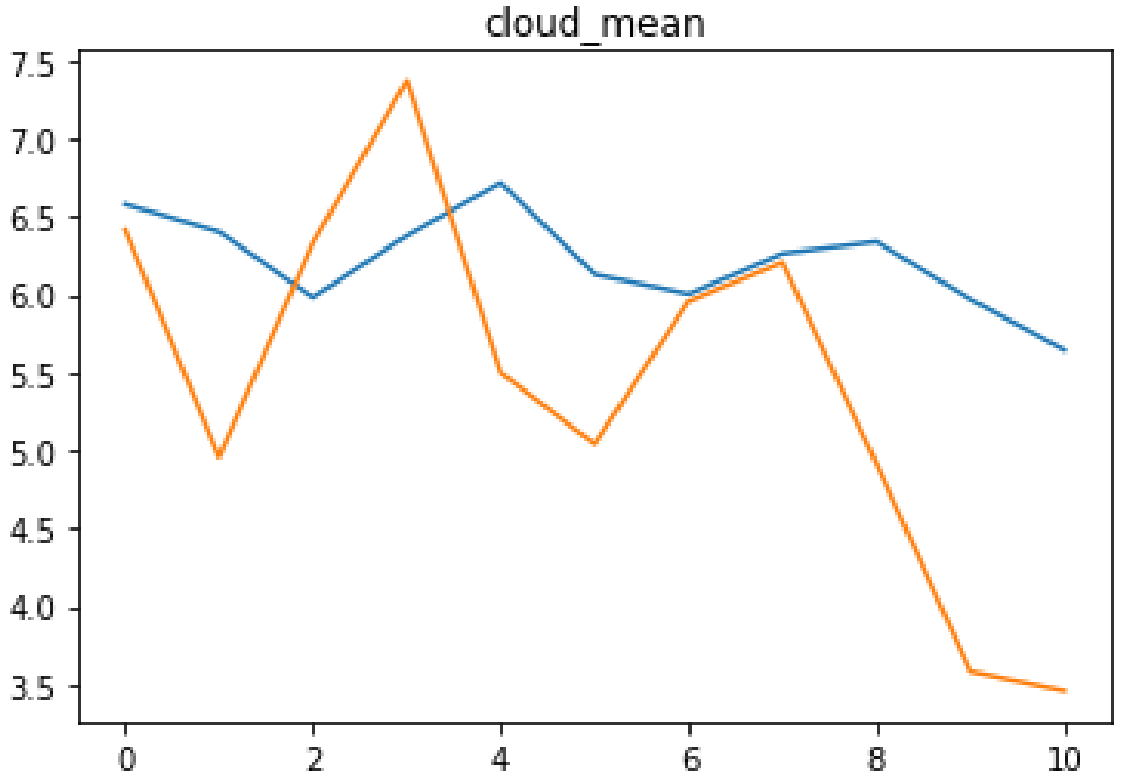
2018년 8월 이후로
전운량에 대한 측정시작

2018년 8월 이후의
데이터만 학습

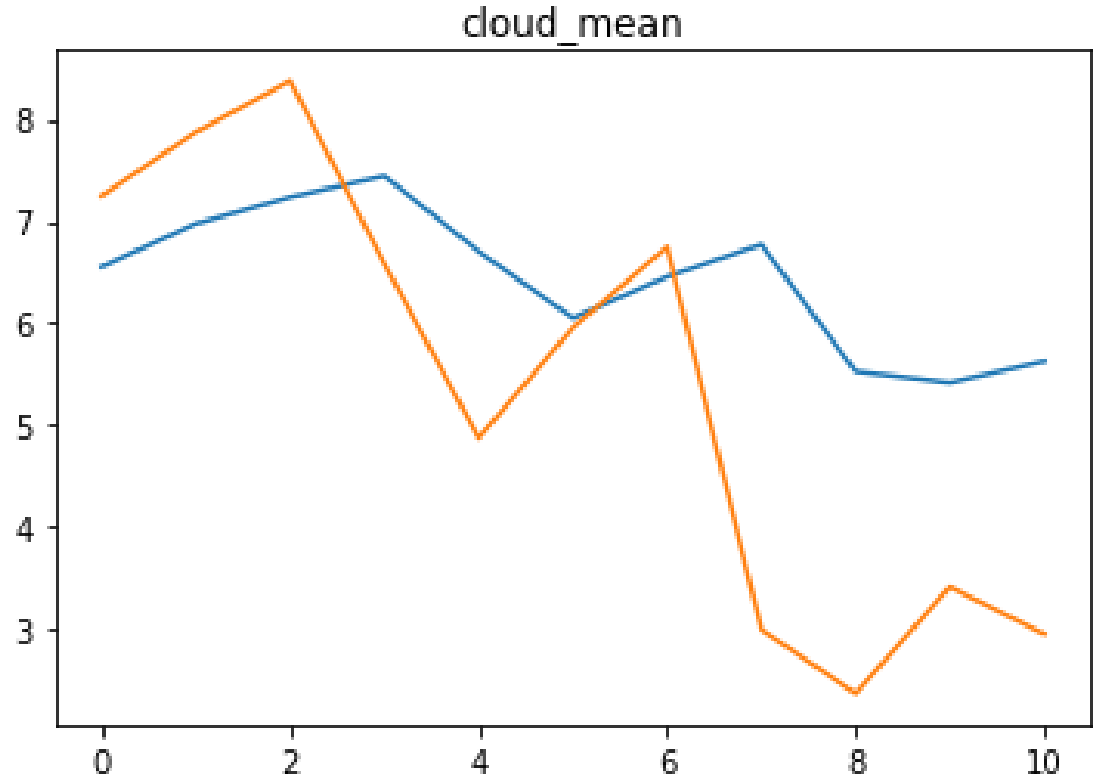
2.2 LSTM을 통한 날씨예측

원주, 진주, 김해, 군산 등의 4개의 공항에 대한 전운량 예측

원주



김해



3.1 변수 생성

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	SDT_YY	SDT_MM	SDT_DD	SDT_DY	ARP	ODP	FLO	FLT	REG	AOD	IRR	STT	ATT	DLY	DRR	CNL
2	2017	1	1	일	ARP3	ARP6	J	J1955	SEw3NzE4	D	N	10:05	10:32	N		N
3	2017	1	1	일	ARP3	ARP6	J	J1954	SEw3NzE4	A	N	9:30	9:31	N		N
4	2017	1	1	일	ARP3	ARP6	J	J1956	SEw3NzE4	A	N	12:45	13:03	N		N
5	2017	1	1	일	ARP3	ARP6	J	J1957	SEw3NzE4	D	N	13:25	14:09	Y	C02	N
6	2017	1	1	일	ARP3	ARP6	J	J1958	SEw3NzE4	A	N	16:10	16:31	N		N
7	2017	1	1	일	ARP3	ARP6	J	J1959	SEw3NTk5	D	N	16:45	17:21	Y	C02	N
8	2017	1	1	일	ARP3	ARP6	J	J1960	SEw3NTk5	A	N	19:30	19:43	N		N
9	2017	1	1	일	ARP3	ARP6	J	J1961	SEw3NTk5	D	N	20:35	20:52	N		N
10	2017	1	1	일	ARP2	ARP3	J	J1015	SEw3NzA2	A	N	17:05	17:03	N		N
11	2017	1	1	일	ARP1	ARP3	J	J1242	SEw3NzA2	D	N	20:25	20:36	N		N
12	2017	1	1	일	ARP1	ARP3	J	J1257	SEw3NzA4	A	N	12:40	12:44	N		N
13	2017	1	1	일	ARP1	ARP3	J	J1220	SEw3NzA4	D	N	13:25	13:41	N		N

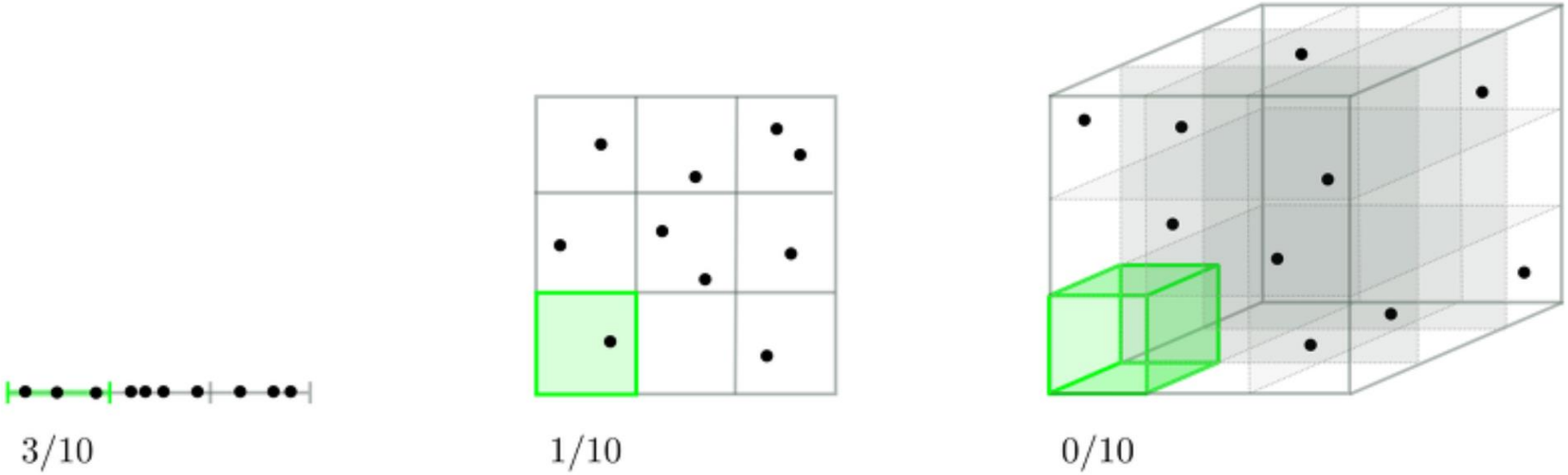
항공기 실적 데이터의 문제점
1. 범주형 변수가 많음.
2. 문자열 변수가 많음

→ 일반적인 방법
'원-핫 인코딩 '

→ 너무 많은 변수가 생성
'차원의 저주'

3.1 변수 생성

'차원의 저주 '



데이터의 차원이 증가할수록 데이터 포인트 간의 거리 또한 증가하게 되므로, 이러한 데이터를 이용해 머신러닝 알고리즘을 학습 하게되면 모델이 복잡해 지게 된다. 따라서, 오버피팅(overfitting) 위험이 커진다.

※ 출처 https://excelsior-cjh.tistory.com/167 ※

그럼 어떻게.....?

3.1 변수 생성

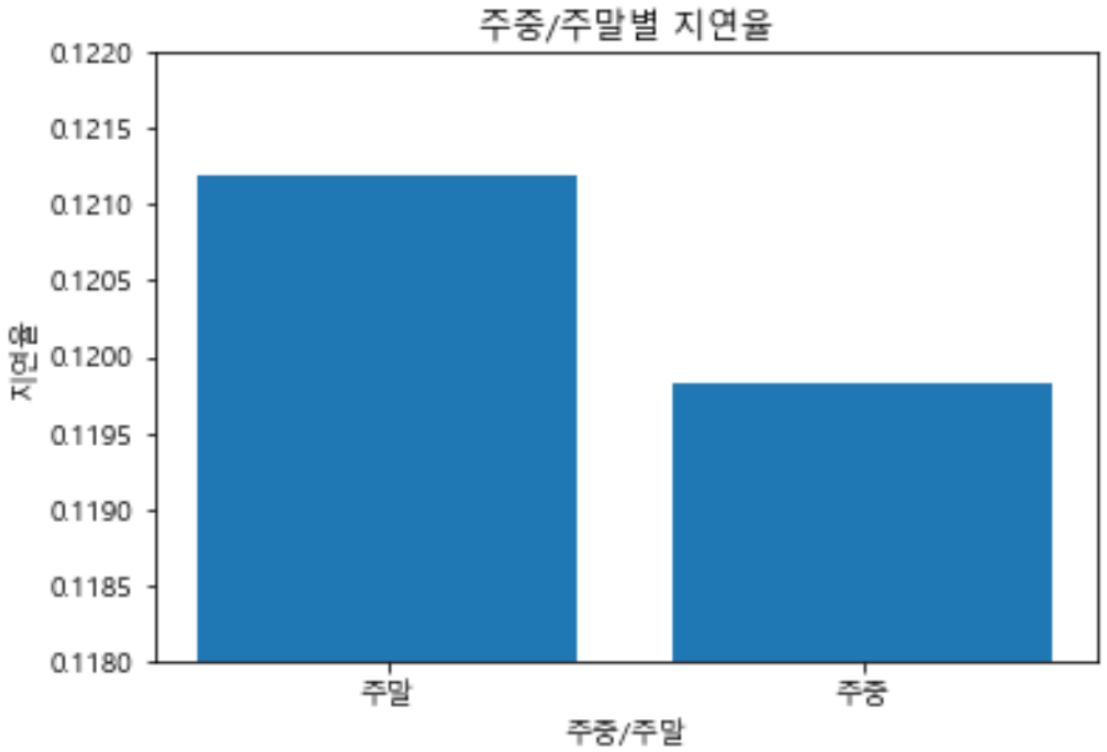
'수치적 특성 대치'

Groupby 변수 생성

요일별지연율
출발공항별지연율
도착공항별지연율
항공사별지연율
경로별지연율
편명별지연율
공항별 주중주말 지연비율
시간대별 지연건수
출발공항 초반후반 지연율
도착공항 초반후반 지연율
항공사크기별 지연율

3.1 변수 생성

'수치적 특성 대치'
(Groupby 변수 생성)



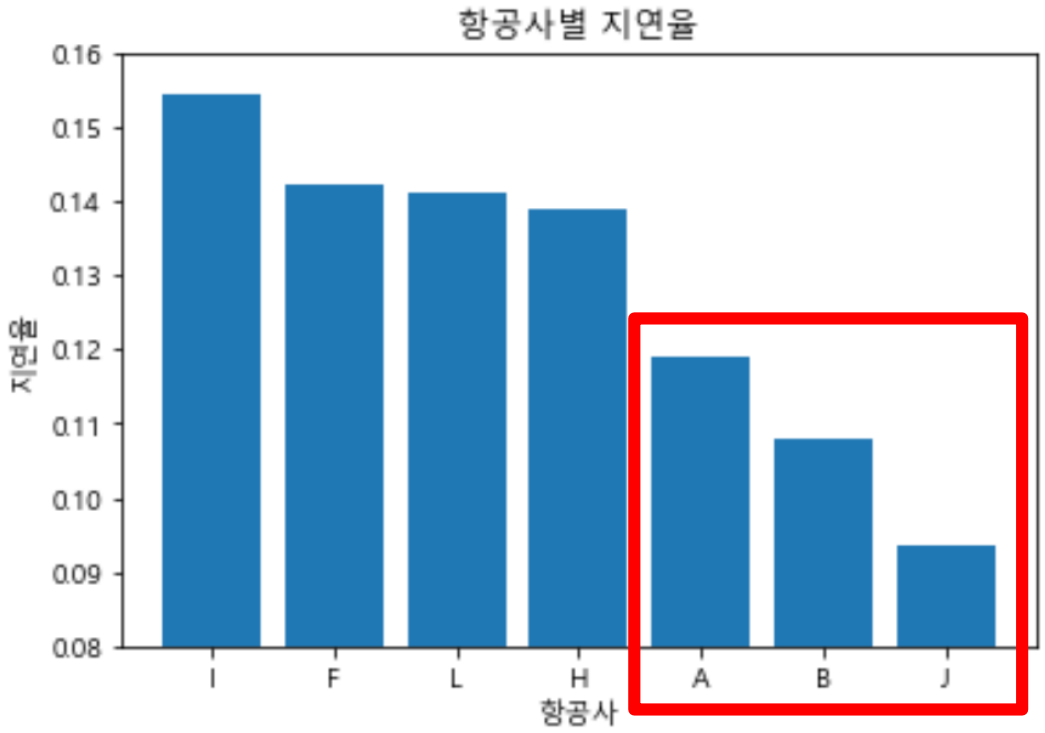
국내 여행은 금요일 출발 일요일 도착 계획이 많기에 따라서 주중이냐 주말이냐의 여부 또한 지연에 영향을 미칠것으로 판단.

주말 (금,토,일) / 주중 (그 외)

-> 공항별(도착/출발)
주중/주말 지연율 구함

3.1 변수 생성

'수치적 특성 대치'
(Groupby 변수 생성)



항공사별 지연율을 보니 대체적으로 대형 항공사인 '아시아나 항공', '대한항공 ' 이 지연율이 낮다.

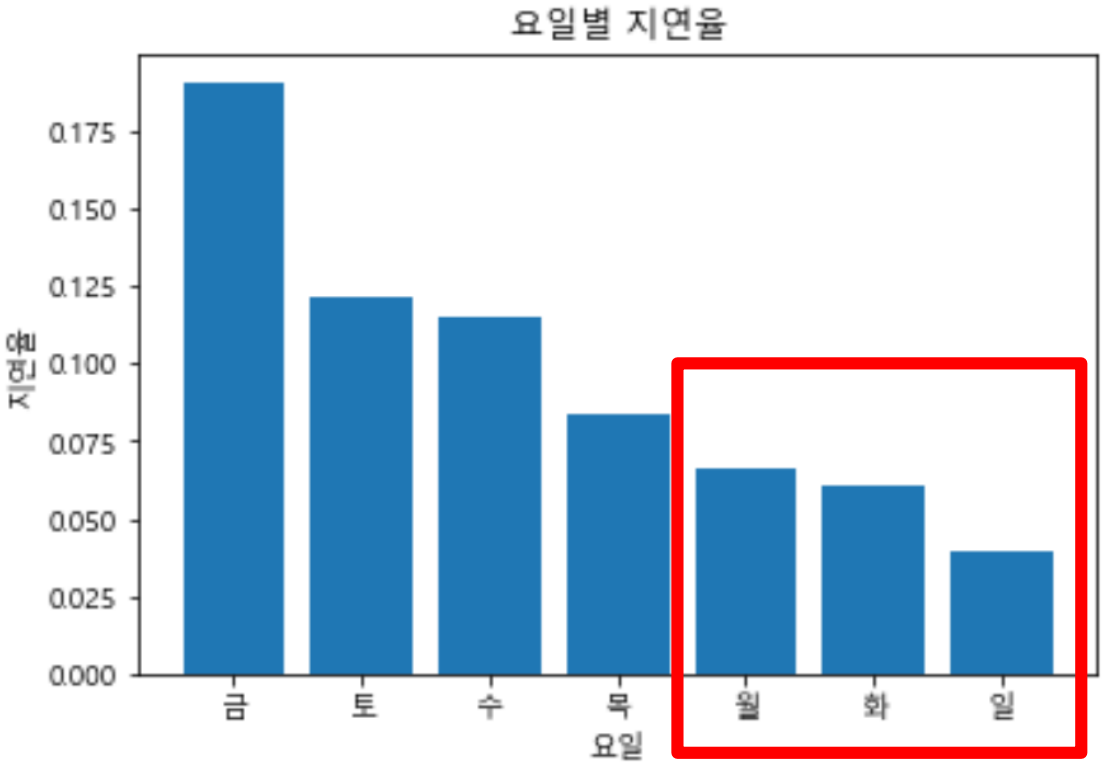
대형 항공사 (A, J) / 소형 (그 외)

-> 대형/소형 항공사별 지연율 구함

I : 진에어
F : 이스타항공
L : 티웨이항공
H : 제주항공
A : 아시아나항공
B : 에어부산
J : 대한항공

3.1 변수 생성

'수치적 특성 대치'
(Groupby 변수 생성)



국내 여행은 금요일 출발 일요일 도착 계획이 많기에 따라서 주중이냐 주말이냐의 여부 또한 지연에 영향을 미칠것으로 판단.

초반 (금,토,수,목) / 주중 (그 외)

-> 일주일(초반/후반)별 지연율 구함

3.2 데이터 모델링 및 앙상블

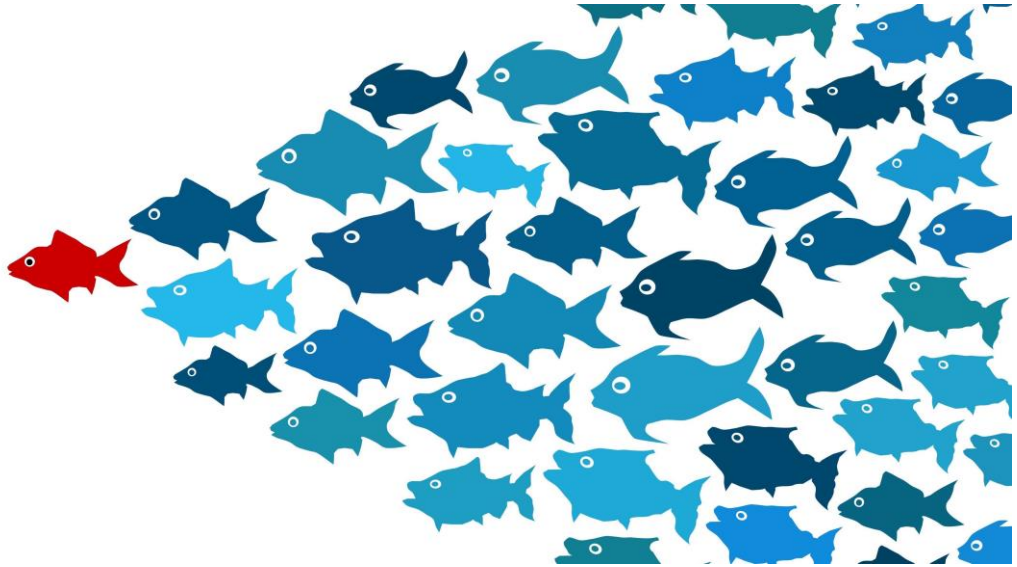
데이터 목적변수 확인

```
train.DLY.value_counts()
```

```
0    29450
1     3124
Name: DLY, dtype: int64
```

전체 데이터의 약 90%가 0.
즉, 전체 운행의 약 90%는 지연되지 않았다.
-> 데이터가 편향되어 있음

편향된 데이터의 해결방법
-> **'Imbalanced Learning'**



*** Imbalanced Learning이란?**

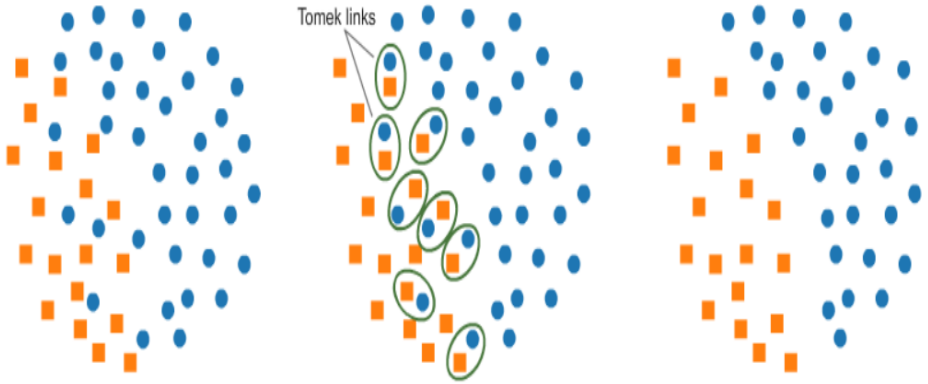
편향된 데이터에서 더 많은 값인 0을
줄이거나 더 적은 값인 1을 늘리는 기법

※ 이미지출처 https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2※

3.2 데이터 모델링 및 앙상블

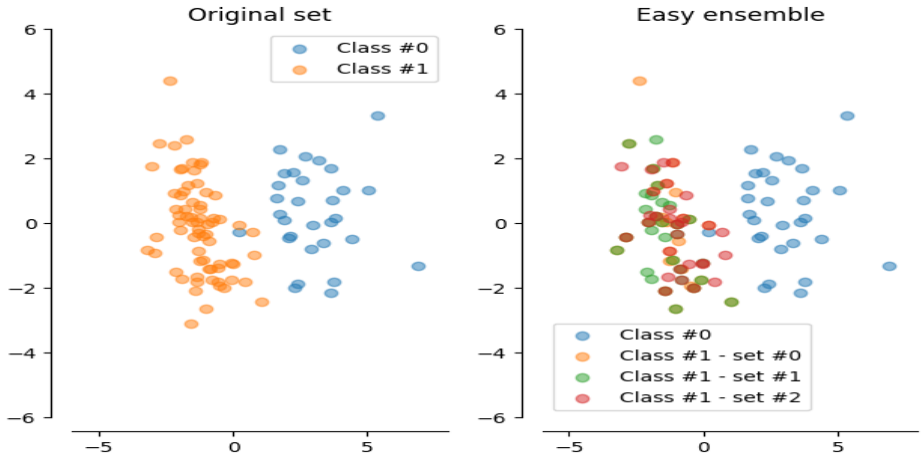
사용한 Imbalanced Learning 기법

1. TomekLink



마이너값(1)과 가장 가까이에
있는 메이저값(0)을 제거

2. EasyEnsemble



Random Undersampling을 반복적으로 적용하여
앙상블세트를 만들어 줌
이 방법은 랜덤한 부분집합을 반복적으로 선택하고
다른 세트의 앙상블을 만들어 준다

3.2 데이터 모델링 및 앙상블

Imbalanced Learning 결과

```
display(X_train2_train.shape)
display(X_res_tl.shape)
display(X_res_ee.shape)
```

(22801, 24) → 기존의 train_test_split의 결과 train set

(21995, 24) → TomekLink 적용 데이터

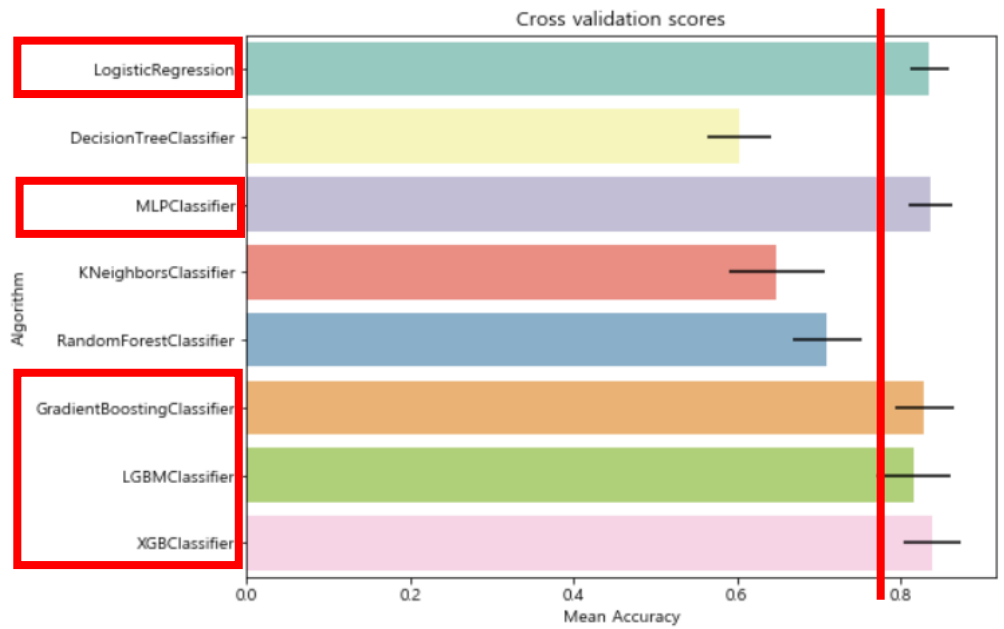
(44400, 24) → EasyEnsemble 적용 데이터

➡ 만들어진 3개의 데이터셋을 대상으로 모델링을 진행 후 앙상블

➡ 서로 다른 특성을 지닌 데이터를 앙상블 함으로써 과적합을 방지

3.2 데이터 모델링 및 앙상블

1. 모델선택



	CrossValMeans	CrossValerrors	Algorithm
7	0.836689	0.033855	XGBClassifier
2	0.836083	0.028197	MLPClassifier
0	0.835930	0.023380	LogisticRegression
5	0.828803	0.040388	GradientBoostingClassifier
6	0.816192	0.045193	LGBMClassifier

기본 데이터에 fitting 후 Cross validation score가 가장 높은 5개의 모델을 선택, 파라미터 튜닝 후 앙상블을 진행한다.

3.2 데이터 모델링 및 앙상블

2. 파라미터 튜닝

모델명	튜닝 전	튜닝 기본	튜닝 tl	튜닝 ee
GBC	0.828803	0.84126	0.88569	0.98947
XGB	0.836689	0.83321	0.88606	0.83321
LGBM	0.816192	0.84392	0.88366	0.93053
LR	0.835930	0.83702	0.85002	0.84558
MLP	0.828976	0.83684	0.87849	0.89987

몇몇 값들은 제외하고 대체적으로 성능(roc_auc score)이 향상되는 결과를 보여준다.
하지만 EasyEnsemble을 적용한 성능을 보면 과적합으로 의심이 가는 모델들이 존재한다.

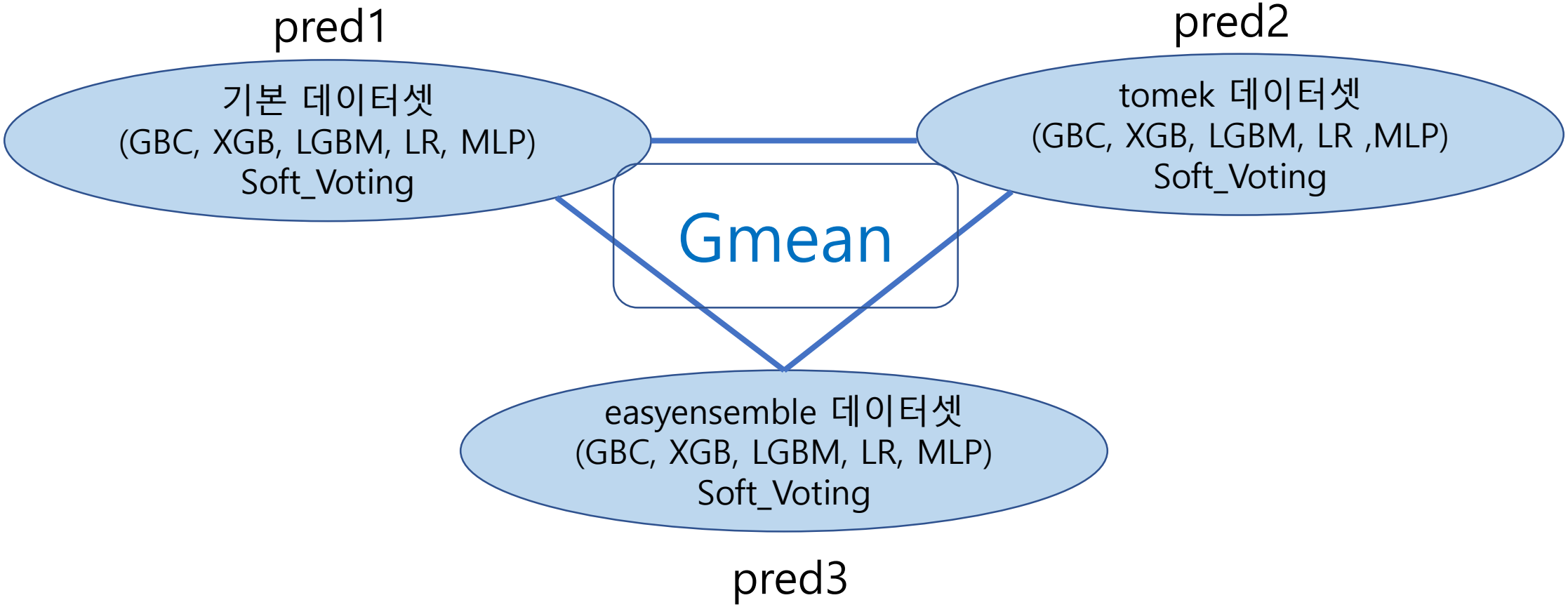
3.3 데이터셋별 성능

모델 성능

모델명	M항공사	Test에만 있는 편명	이외 데이터
GBC	0.74429	0.75467	0.84126
XGB	0.73108	0.74310	0.83321
LGBM	0.74312	0.75575	0.84392
LR	0.73929	0.76092	0.83702
MLP	0.74414	0.738933	0.836083

3.3 데이터셋별 성능

3. 앙상블



외부 참조 데이터

외부참조 데이터

데이터	출처	기준년도
기상 데이터	기상자료개방포털	2016~2019
공항별 통계	한국공항공사	-
항공사별 통계	한국공항공사	-

※ 한국공항공사 데이터를 통해 공항, 항공사를 유추함

분석도구