By using aircraft operation data

Forecasting probability of flight delay

CONTENTS

1. Exploratory Data Analysis

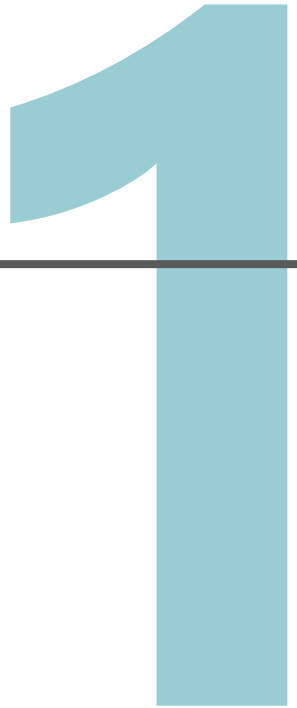
2. Create Features

- ✓ Foreign valuables
- ✓ Derived valuables

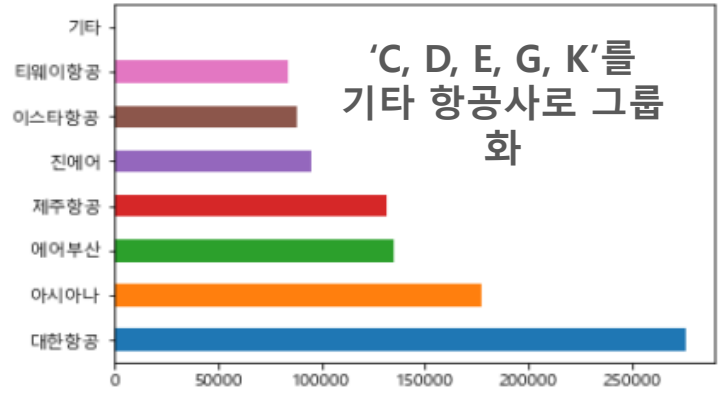
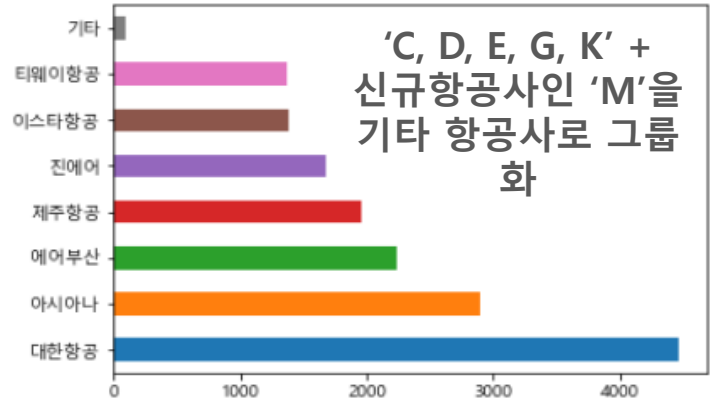
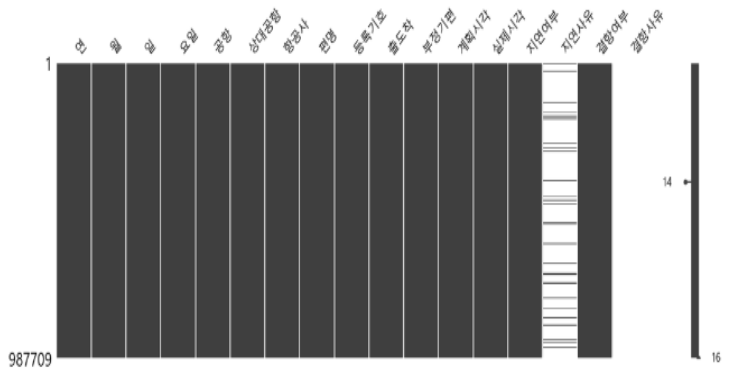
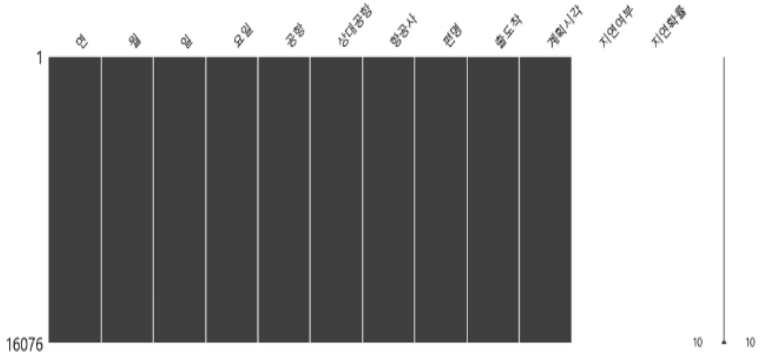
3. Modeling

- ✓ Feature selection
- ✓ Model selection

Exploratory Data Analysis

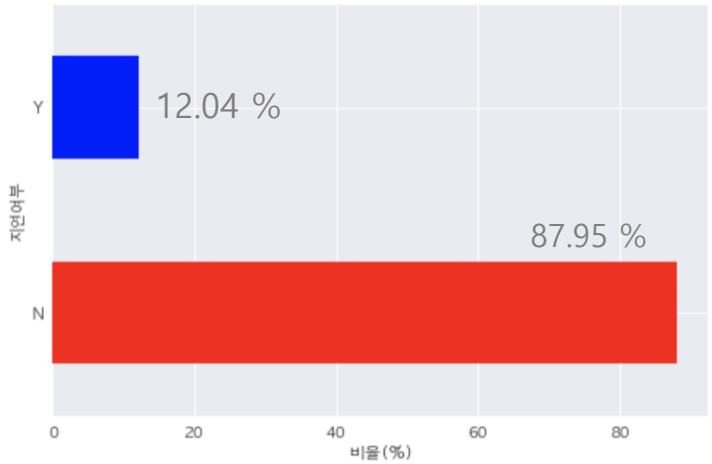


EDA

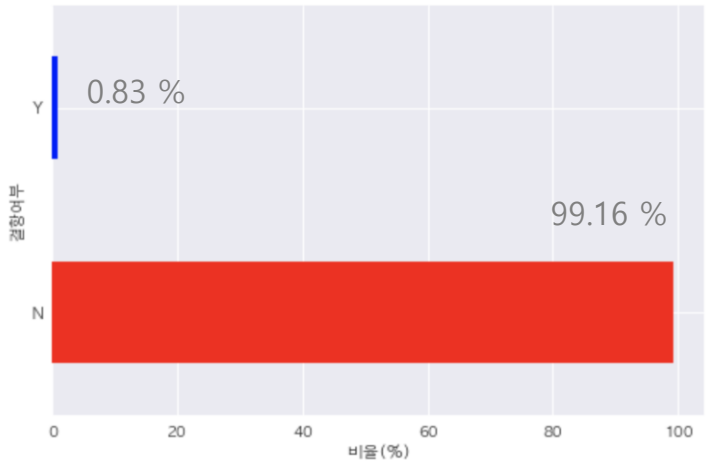
	AFSNT	AFSNT_DLY
데이터 크기	(987709, 17)	(16076, 12)
항공사	 ‘C, D, E, G, K’를 기타 항공사로 그룹화	 ‘C, D, E, G, K’ + 신규항공사인 ‘M’을 기타 항공사로 그룹화
결측치 데이터	 지연사유에만 결측치 존재	 지연여부, 지연확률(target)에만 결측치 존재

EDA

<지연여부 비율>



<결항여부 비율>



Target의 비율이 불균형

↓

Oversampling이 필요함

AFSNT

<지연사유 >

1. C02 (90.58%) : A/C 연결 지연
2. C01 (1.70%) : A/C 정비
3. A01 (1.28%) : 안개
4. C10 (1.28%) : 제방빙작업

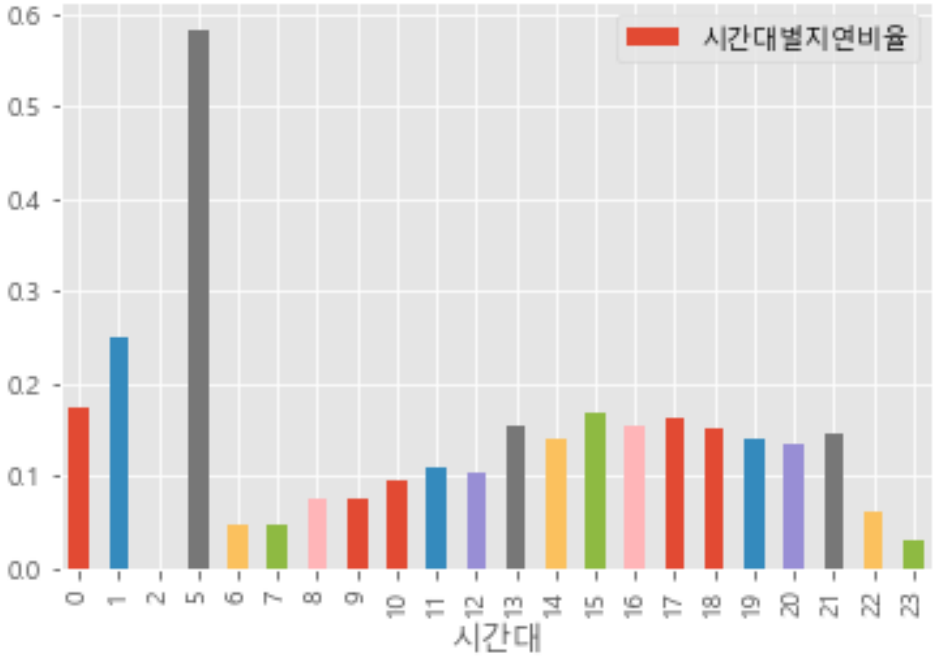
<결항사유 >

1. A04 (22.19%) : 태풍
2. C02 (18.28%) : A/C 접속
3. A02 (16.85%) : 강설
4. A05 (13.10%) : 강풍

✓ **지연사유로는 A/C 접속이, 결항사유로는 주로 기상이 중요**

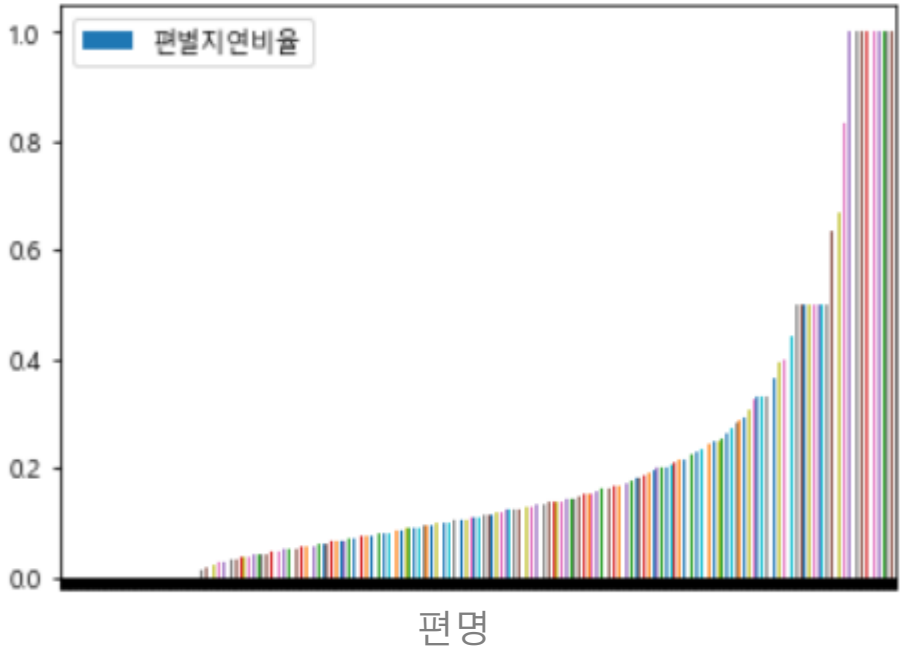
EDA

<시간대별 지연비율>



시간대별로 지연 비율 차이가 존재

<편명별 지연비율>



항공편별로 지연 비율 차이가 존재

Create Features

Foreign valuables
Derived valuables

2

Create Features

Foreign valuables (Whether)

Problem

Test와 Train 데이터 셋에 모두 기상 데이터를 넣기 위해선

2019년 9월 **미래의 기상데이터를 처리**해야 한다

지역	도시	11일 (수)	12일 (목)	13일 (금)	14일 (토)	15일 (일)	16일 (월)	17일 (화)	18일 (수)	기온범위
서울 · 인천 · 경기도	서울	22 / 29	21 / 27	21 / 28	21 / 29	21 / 28	20 / 28	19 / 27	17 / 26	그래프
	인천	23 / 28	21 / 26	21 / 27	22 / 28	22 / 28	21 / 27	20 / 26	18 / 25	그래프
	수원	20 / 29	20 / 27	21 / 28	21 / 28	21 / 28	18 / 27	19 / 27	16 / 26	그래프
	파주	20 / 28	18 / 26	18 / 28	18 / 28	18 / 28	17 / 27	17 / 27	15 / 26	그래프
	이천	19 / 28	18 / 28	19 / 28	19 / 29	19 / 29	18 / 28	17 / 27	14 / 25	그래프
	평택	21 / 29	19 / 27	20 / 28	20 / 28	20 / 28	18 / 27	18 / 27	16 / 26	그래프

기상청에서 제공하는 중기예보(10일 예보)에는
일별 최저, 최고 기온만 제공한다

Solution

기상 데이터는 예측 불가하므로 시간별 평균값을 기상 예보 데이터로 대체하는 것은 무의미하다고 판단

=> 4년동안의 기상 데이터의 최저, 최고 값을 구해 **feature에 범위**를 생성하기로 결정!

Create Features

Foreign valuables (Whether)

데이터 수집

데이터 전처리

최종 데이터

기상자료개방포털 기상자료개방포털 정부혁신 보다나은 정부

- 김해, 광주, 청주, 대구, 포항, 사천, 군산, 원주, 울산, 여수, 양양

항공기상청 API 사용

- 김포, 제주, 인천

기상 데이터(기온, 습도, 시정, 풍속, 해면기압, 현지기압) **결측치 처리**

=>그날의 평균 또는 그날의 데이터가 통째로 없으면 그 달의 평균으로 대체

각 기상 데이터의 **최저,최고**를 feature로 만들어 최종 기상 파생 변수 **완성**

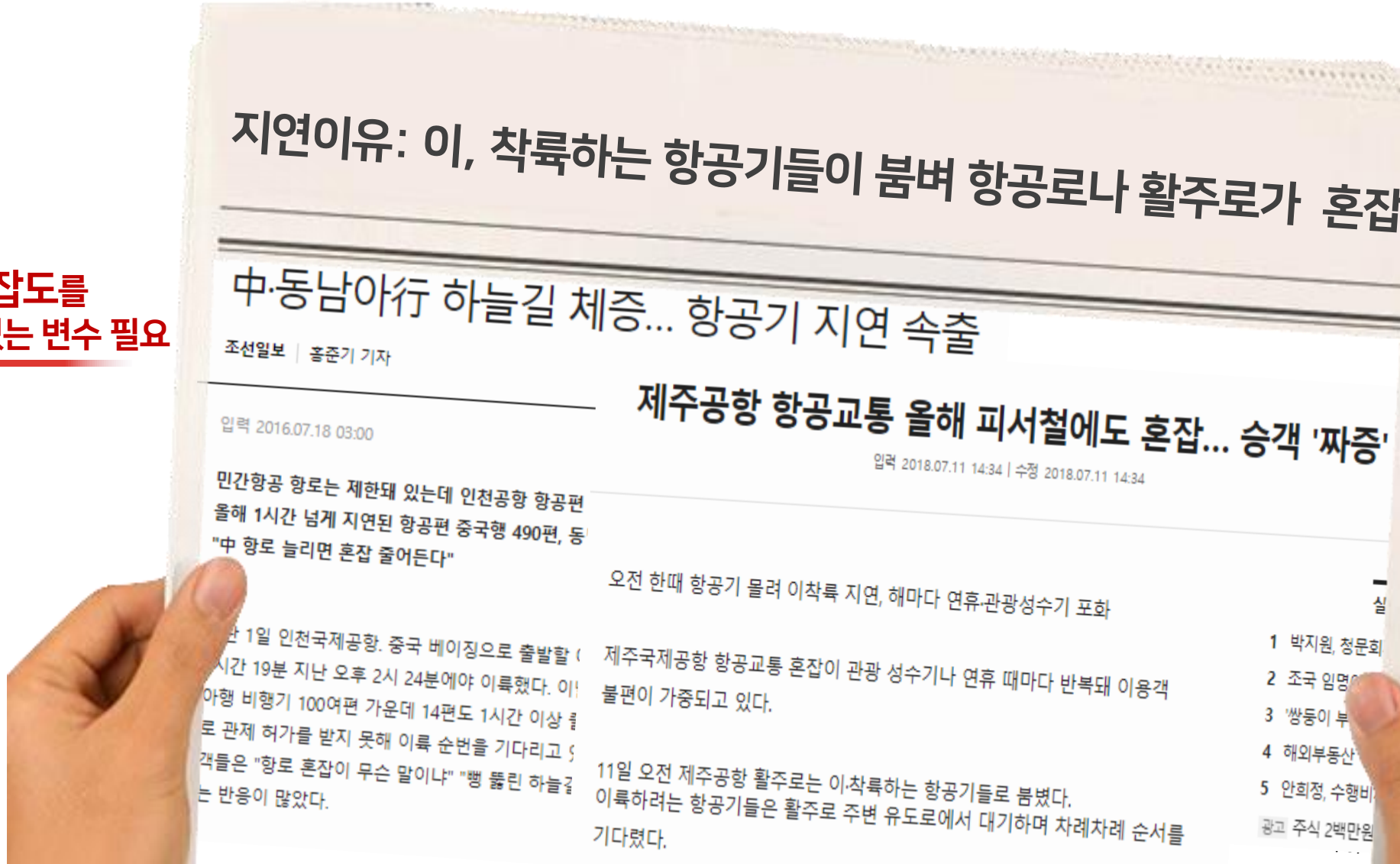
=> 최저_기온, 최저_습도 , 최저_풍속 , 최저_해면기압, 최저_현지기압, 최저_시정,
최고_기온, 최고_습도 , 최고_풍속 , 최고_해면기압, 최고_현지기압, 최고_시정

공항	월일	시간대	최저_기온	최저_습도	최저_시정	최고_습도		최고_시정	최고_풍속	최고_해면기압	최고_현지기압
ARP1	01월 01일	1	-7	25	280	87	• • •	1000	13	10334	10317
ARP1	01월 01일	2	-8.4	31	260	87		1000	7	10335	10318
ARP1	01월 01일	3	-8.1	30	180	87		1000	13	10336	10319
ARP1	01월 01일	4	-8.1	24	160	87		1000	17	10332	10315

Create Features

Derived valuables (Degree of congestion)

공항별 혼잡도를
고려할 수 있는 변수 필요



Create Features

Derived valuables (Degree of congestion)

<공항 혼잡도 피쳐 생성 과정 개요>

Q: 계획 비행기 댓수 만으로는 공항별 수용 가능 대수를 고려할 수 없는데 어떻게 해야 할까?

	날짜	계획댓수	실제댓수	차이	시간대	공항	출도착
0	2017-01-01	10.0	8.0	2.0	6	2	1
1	2017-01-01	1.0	0.0	1.0	6	5	1
2	2017-01-01	2.0	1.0	1.0	6	12	1
3	2017-01-01	1.0	1.0	0.0	7	0	0
4	2017-01-01	1.0	1.0	0.0	7	0	1
5	2017-01-01	1.0	1.0	0.0	7	2	0
6	2017-01-01	11.0	12.0	-1.0	7	2	1
7	2017-01-01	9.0	9.0	0.0	7	3	1
8	2017-01-01	1.0	0.0	1.0	7	6	0
9	2017-01-01	1.0	0.0	1.0	7	9	1
10	2017-01-01	12.0	13.0	-1.0	7	12	0

A: **운항 예정 비행기 – 실제 운항 비행기** = 차이

계획 댓수와 실제 댓수의 차이를 통해
공항별 수용 가능 대수 고려 가능

한번 비행기 스케줄이 밀리면
다음 시간대의 비행기 스케줄에 영향을 미침

=>차이평균을 누적한다

	공항	출도착	월	시간대	차이	차이누적
0	0	0	1	6	-0.500000	-0.500000
1	0	0	1	7	-0.012821	-0.512821
2	0	0	1	8	0.215054	-0.297767
3	0	0	1	9	-0.217391	-0.515158
4	0	-0.5+-0.012821 = -0.512821				
5	0	0	1	11	-0.206522	-0.232549
6	0	0	1	13	-0.142857	-0.375407
7	0	0	1	15	0.290323	-0.085084
8	0	0	1	16	-0.172043	-0.257127
9	0	0	1	17	0.290323	0.033196
10	0	0	1	18	-0.204301	-0.171105
11	0	0	1	19	-0.294118	-0.465223

✓ **차이누적** 파생변수 완성

Create Features

Derived valuables (Probability of delay)

모든 항공기의
도착 시 평균 지연 시간

5.06 분

VS

출발 시 지연한 항공기의
도착 시 평균 지연 시간

33.47 분

출발 시 지연된 시간이 도착 시 지연에 영향을 미침

출발 시 지연된 시간(시간차)를 고려할 수 있는 변수 생성이 필요

Create Features

Derived valuables (Probability of delay)

"지연 확률이 높은 경우 **도착에서도 지연할 가능성**이 높다"

1. 출/도착 연결

"날짜, 항공사, 등록기호, 편명을 **기호라는 컬럼**에 한꺼번에 넣어 출, 도착을 묶어준다"

<기호 생성>

항공사	편명	등록기호	날짜	기호
1.0	1408.0	90.0	2017-01-01	1.0-1408.0-90.0-2017-01-01
1.0	1222.0	49.0	2017-01-01	1.0-1222.0-49.0-2017-01-01
1.0	1174.0	96.0	2017-01-01	1.0-1174.0-96.0-2017-01-01
1.0	1178.0	126.0	2017-01-01	1.0-1178.0-126.0-2017-01-01

출도착	기호	기호값
0	64631	2
1	64631	2
0	455755	2
1	455755	2

2. 출발 항공편 지연확률 예측

```
col = ['공항', '편별지연비율', '시간대',
       '월', '편별시간차이평균', '기호',
       '정시출발율', '차이누적', '기온',
       '습도', '풍속', '풍향', '해면기압',
       '현지기압', '시정']
```

```
model_departure.predict_proba(X_train_d_over)[:,1]
```

```
array([0.03, 0.95, 0.23, ..., 0.99, 0.86, 0.94])
```

3. 도착 항공편에 지연확률 feature 추가

"출발에서 구한 출발 지연확률을 기호를 기준으로 **도착 항공편 column**에 추가"

기호	지연확률
31114.000000	0.17
387554.000000	0.14
62306.000000	0.14
109534.000000	0.37
350326.000000	0.15

```
col_a = ['공항', '편별지연비율',
         '시간대', '월',
         '편별시간차이평균', '지연확률',
         '차이누적', '기온', '습도',
         '풍속', '풍향', '해면기압',
         '현지기압', '시정']
```

✓ **지연확률** 파생변수 완성

Create Features

Derived valuables (On-time departure rate)

생성 이유

최종 데이터

	공항	요일	15분이내	전체	정시출발율
0	0	Fri	969	2390	0.405
1	0	Mon	956	2486	0.385
2	0	Sat	1122	2332	0.481
3	0	Sun	1143	2385	0.479
4	0	Thu	1005	2472	0.407
5	0	Tue	1115	2427	0.459
6	0	Wed	974	2322	0.419

$$\text{정시 출발률} = \frac{\text{15분 이내에 출발한 댓수}}{\text{전체 댓수}}$$

- ✓ 항공사의 항공기 운항능력을 검증하는 대표적인 국제 지표로 작용
- ✓ 국제 지표인 만큼 출발 모델 feature로 사용하기로 결정

✓ **정시 출발률** 파생변수 완성

Create Features

Derived valuables (Select Features)

변수 선택

"날씨에 영향을 미치는 계절성 또는
휴가철을 고려하기 위해 "

✓ **월** 변수 선택

"지연사유 중 공항 자체의 결함을 고려하기 위해"

✓ **공항** 변수 선택

변수 생성

"항공편별 지연 여부에 가중치를 부여하기 위해"

✓ **편별지연비율** 파생변수 생성

"시간대별 지연 여부에 가중치를 부여하기 위해"

✓ **(계획)시간대** 파생변수 생성

"지연이 되지 않더라도 평균적으로
늦게 출발하는 경우를 고려하기 위해 항공편별
지연시간의 평균값으로 가중치 부여하기 위해"

✓ **편별 지연시간 평균** 파생변수 생성

Modeling

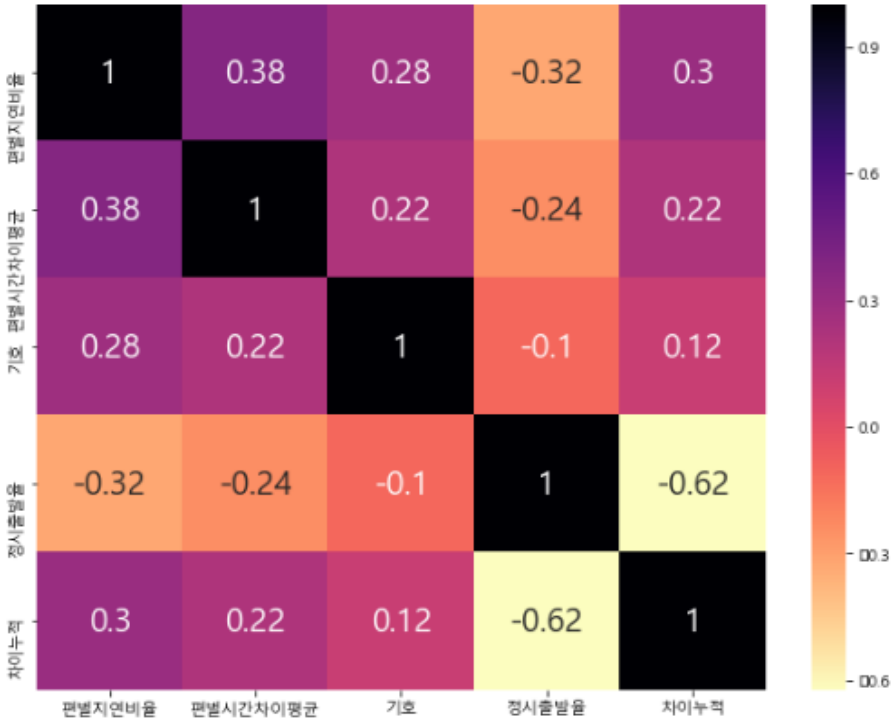
3

Feature selection
Model selection

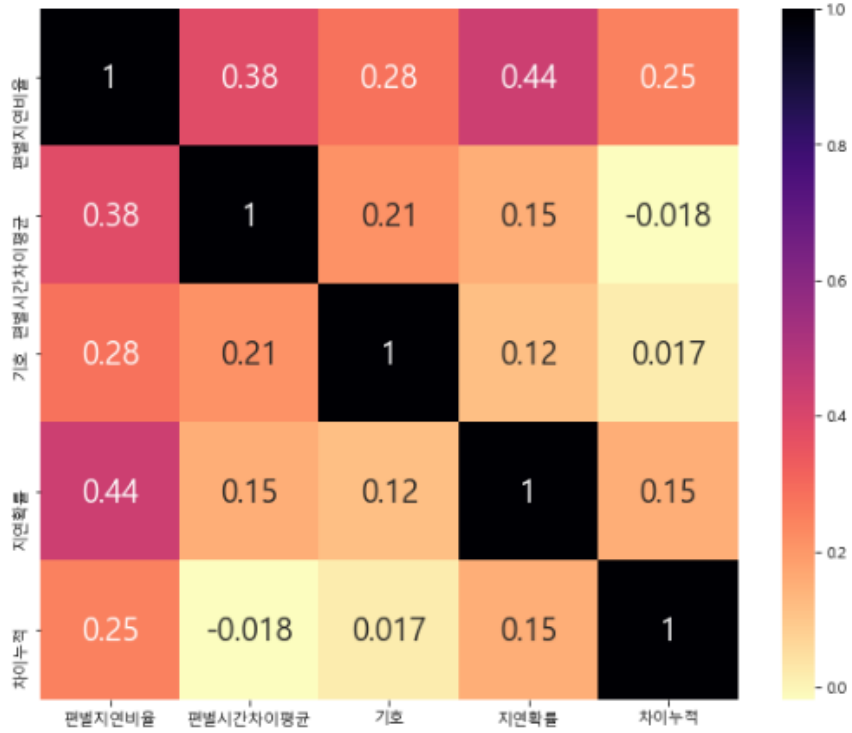
Modeling

Analysis of correlation

Departure



Arrival



기후를 제외한 feature 중 상관관계가 높은 것이 없으므로
다중공선성 의심을 하지 않아도 된다

Modeling

Feature selection

출발 모델

종속변수:
지연여부(범주형 변수)

Feature:
공항, 시간대, 월, 편별 지연비율, 편별 시간 차이평균, 기호, 정시 출발율, 차이누적,
최저_기온, 최저_습도 , 최저_풍속 , 최저_해면기압, 최저_현지기압, 최저_시정,
최고_기온, 최고_습도 , 최고_풍속 , 최고_해면기압, 최고_현지기압, 최고_시정

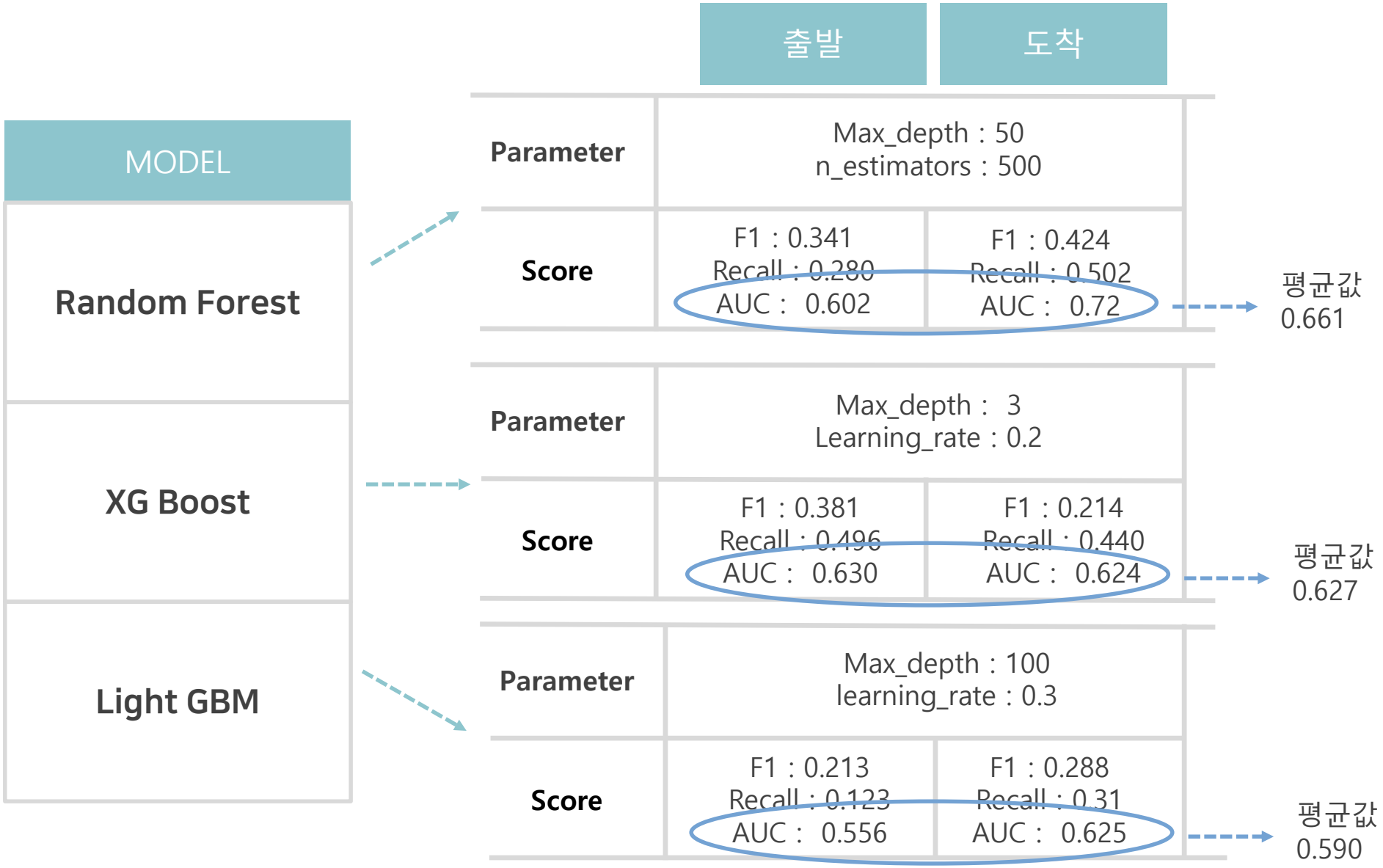
도착 모델

종속변수:
지연여부(범주형 변수)

Feature:
공항,시간대, 월, 편별 지연비율, 편별 시간 차이평균, 기호, 지연확률, 차이누적,
최저_기온, 최저_습도 , 최저_풍속 , 최저_해면기압, 최저_현지기압, 최저_시정,
최고_기온, 최고_습도 , 최고_풍속 , 최고_해면기압, 최고_현지기압, 최고_시정

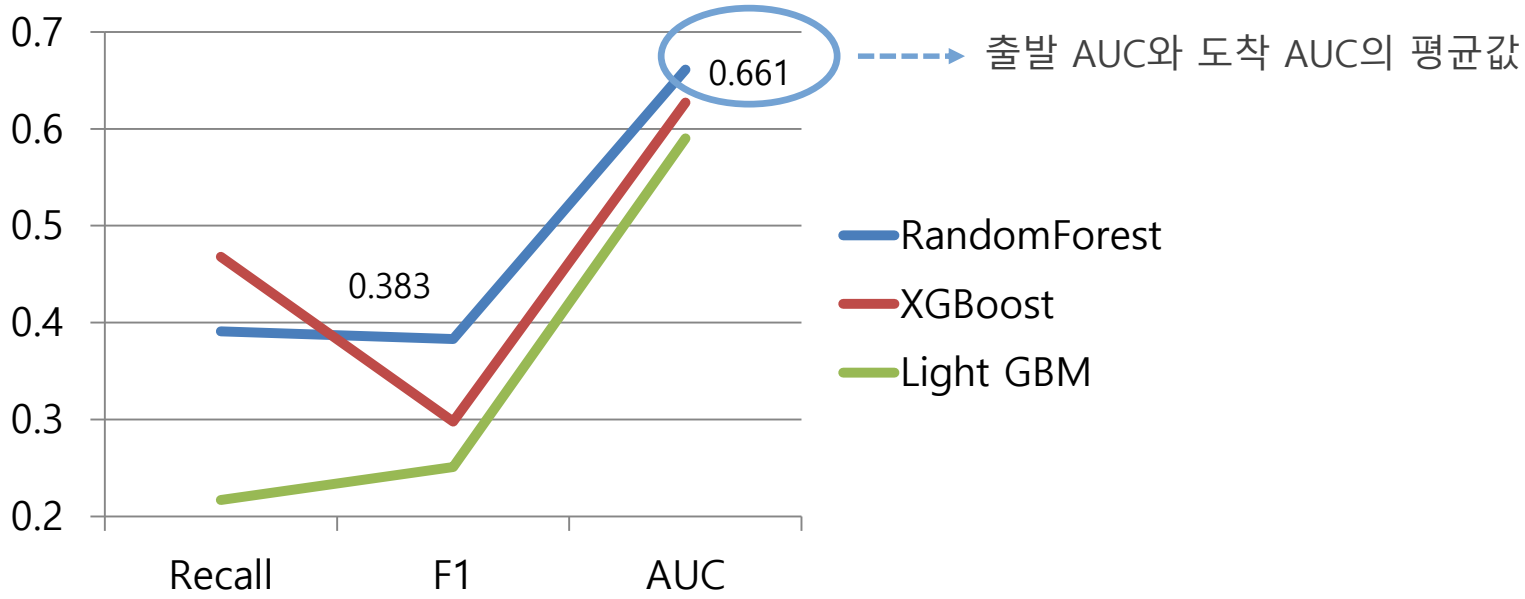
Modeling

Model selection



Modeling

Final Model



지연인데 지연이 아니라고 예측하는(FN)이 중요하므로 모델 평가 지표 중
AUC, F1, Recall값이 중요하다고 판단

세 개의 평가 지표 중에 RandomForest가 두개(auc, f1)의 평가지표가 높으므로
✔ **RandomForest(Max depth = 50, n_estimators = 500)을**
최종모델로 선택!

About more

- ✓ **각 공항별 기상데이터가 존재하지 않아 근처 지역의 기상 데이터로 대체하였기 때문에 약간의 오차가 발생**

  → TAF (항공기상예보)를 활용하여 해당 공항의 실시간 기상 예보 업데이트하여 해당 공항의 정확한 기상 데이터 확보가능

- ✓ **기상 데이터 결측치를 회귀분석 또는 시계열 모델로 처리하려 했으나 모델 성능이 떨어져 사용 불가**

  → 결측치 처리를 위한 회귀분석 또는 시계열 모델의 성능을 올린다면 최종 모델의 성능도 올라갈 가능성이 높음

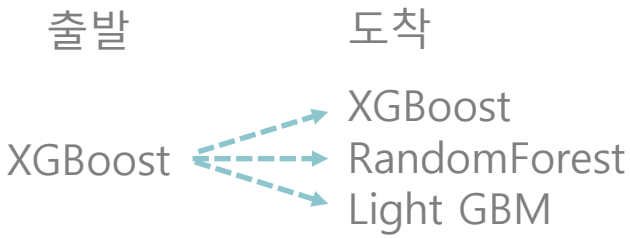
- ✓ **컴퓨터 성능의 문제로 각 모델 별로 GridSearch를 돌려보지 못함**

  → GridSearch로 최적의 parameter을 찾아 모델 성능을 향상시킬 수 있음

- ✓ **출/도착 모델 선정**

  → 출/도착 모델을 하나의 모델로 결정하는 것이 아닌 각각의 모델로 선택하면 구체적이고 더 높은 성능을 기대할 수 있음

EX)



Reference

데이터

- 항공데이터: 빅콘테스트 – 퓨처스리그
- 날씨: 항공기상청 공공데이터

http://amo.kma.go.kr/new/html/news/api.jsp
: 기상자료개방포털
https://data.kma.go.kr/cmmn/main.do;jsessionid

관련자료

- 지연, 결항 승객 불만

https://www.msn.com/ko-kr/money/topstories

- 항공 스케쥴 어플리케이션

https://play.google.com/store/apps/details?id=com.airportal&hl=ko

Thank you!