



DELAY

# 항공 운항 데이터를 활용한 “항공 지연 예측”

P.O.P(Plan of Plane) 김태균, 김민우, 김승확, 이준호, 이정은



2019 빅콘테스트  
2019 BIG CONTEST



KOREA AIRPORTS  
CORPORATION



# CONTENTS

개요	활용데이터 정의	데이터 전처리 및 탐색적 데이터 분석	모델 선정	예측 결과	참고자료
01	02	03	04	05	06
1. 분석 목표	1. 활용데이터 정의	1. 데이터 전처리 2. 탐색적 데이터 분석	1. 분석기법별 비교 2. 모델선정기준 3. LightGBM	1. 예측 결과	1. 참고자료

“

당신의 소중한 여행, 비행기는 지연될까요? 안될까요?

”

“비행기 정시성 확보는 여러 사회분야에 도움이 됨”

어떤 요소가 비행기 지연여부와 지연확률에 영향을 미칠까?

항공요소별 영향력의 크기

기상요소별 영향력의 크기

항공요소와 기상요소의 결합

“공항이 비행기 지연에 미치는 영향은?”

“항공사가 비행기 지연에  
미치는 영향은?”


“평균풍속이 비행기 지연에  
미치는 영향은?”

“항공기상이 비행기 지연에  
미치는 영향의 크기는?”

“항공요소와 기상요소가 주어졌을 때  
비행기 지연 확률은?”

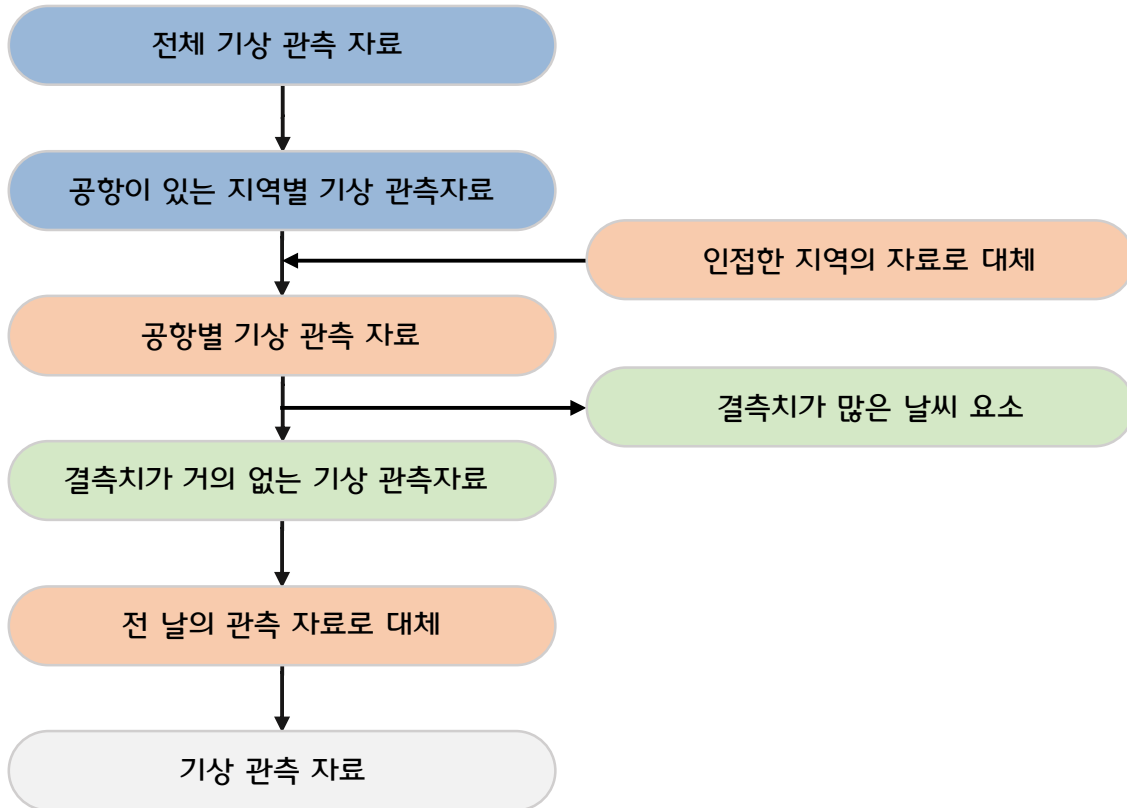
## 활용데이터 정의 | 2-1. 활용데이터 정의

분석에 활용한 데이터의 종류는 크게 1) 항공 데이터 2) 기상 데이터 3) 항공기상데이터이다.

활용 데이터 정보			
출처			
내용	항공 데이터 (AFSNT.csv)	기상 데이터	항공기상데이터
항목	연도, 월, 일, 요일, 공항, 상대공항, 출도착, 지연여부, 계획시각	관측지역, 관측날짜, 평균기온, 평균풍속, 상대습도	관측지역, 관측날짜, 평균해면기압, 평균기온, 최고기온, 최저기온, 평균 이슬점, 평균상대습도, 최소상대습 도, 평균풍속, 최대풍속, 최대풍향, 최대순간풍속, 최대순간풍향
기간	2017.01.01 ~ 2019.06.30		

기상 데이터 전처리는 1) 데이터 추출, 2) 결측치 대체, 3) 변인 제외 세 가지의 과정으로 진행하였다.

### 데이터 전처리 흐름도



### 데이터 전처리 흐름의 단계별 구분

#### ○ 데이터 추출

원래의 데이터에서 필요한 데이터 추출

#### ○ 결측치 대체

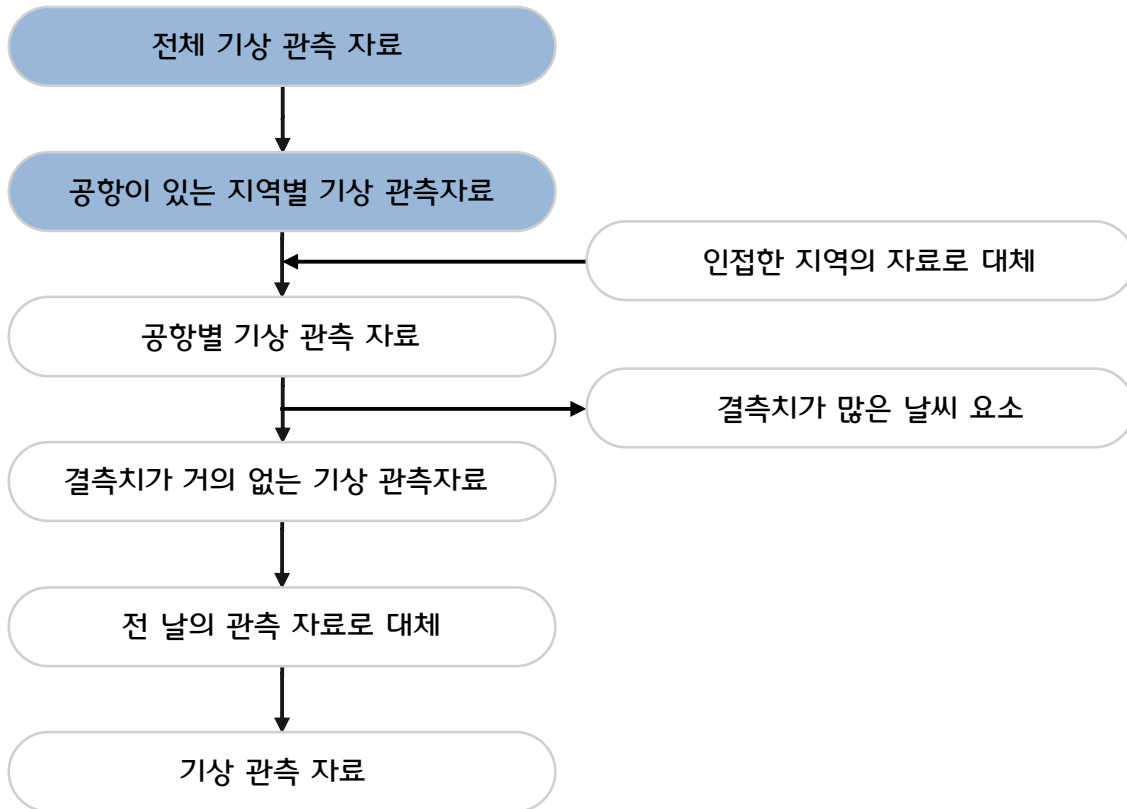
필요한 자료 및 데이터가 없을 경우 다른 것으로 보간

#### ○ 변인 제외

다수의 결측치가 있는 날씨요소인 경우 제외

기상 데이터에서 필요한 데이터를 이용하기 위해 항공 데이터(AFSNT.csv)에서 공항이 있는 지역의 날씨 데이터 추출하였다.

### 데이터 전처리 흐름 중 데이터 추출



### 데이터 추출 방안

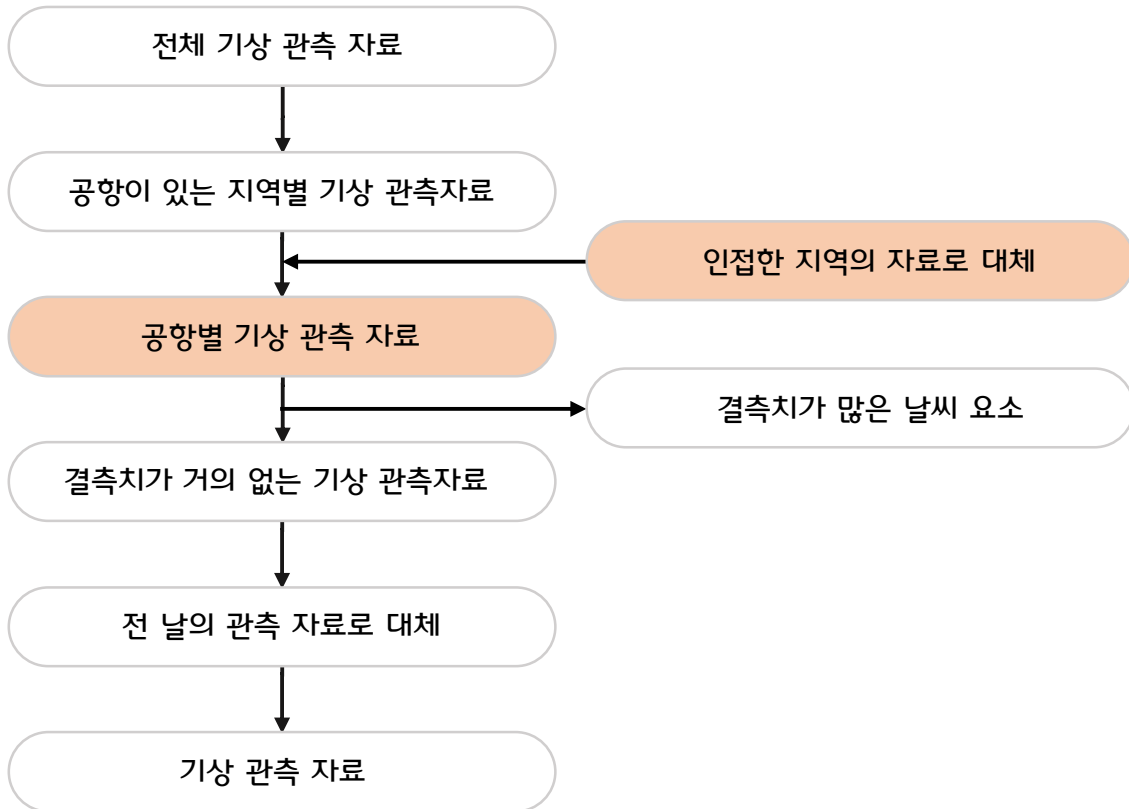
〈 공항이 있는 지역명 〉

공항	지역	공항	지역
ARP1	김포	ARP9	여수
ARP2	김해	ARP10	양양
ARP3	제주	ARP11	포항
ARP4	광주	ARP12	사천
ARP5	울산	ARP13	군산
ARP6	청주	ARP14	원주
ARP7	무안	ARP15	인천
ARP8	대구		

전국 기상 관측 자료에서 공항이 있는 지역의 기상 관측 자료를 추출

데이터를 대체하여도 편향이 생기지 않는 수준에서 데이터를 대체하였다.

### 데이터 전처리 흐름 중 결측치 대체



### 결측치 대체 지역

#### 1. 지역 기상관측자료가 없는 경우

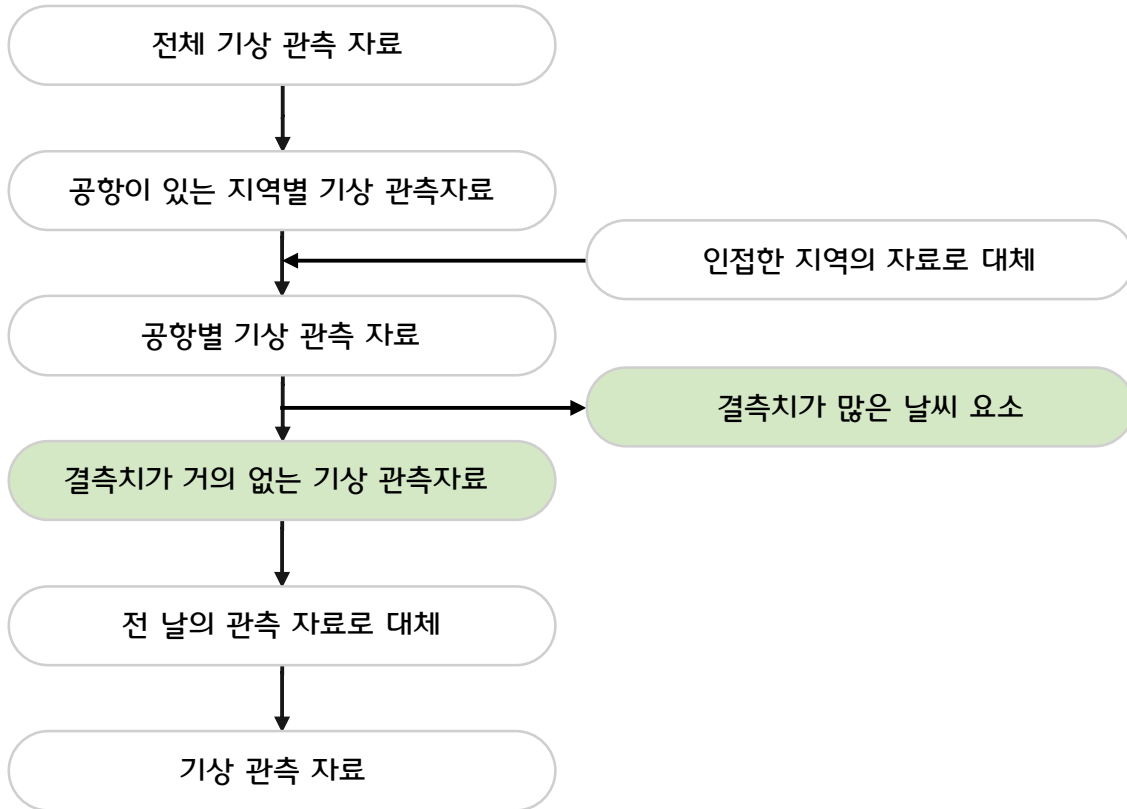
김포 → 서울  
무안 → 목포  
양양 → 속초  
사천 → 진주

#### 2. 김해공항은 김해보다 부산과 지리적으로 더 가까워 김해날씨 대신 부산날씨를 사용 김해 → 부산



결측치의 대체가 불가능할 정도로 많은 경우 해당 변수를 제외하였다.

### 데이터 전처리 흐름 중 변인 제외

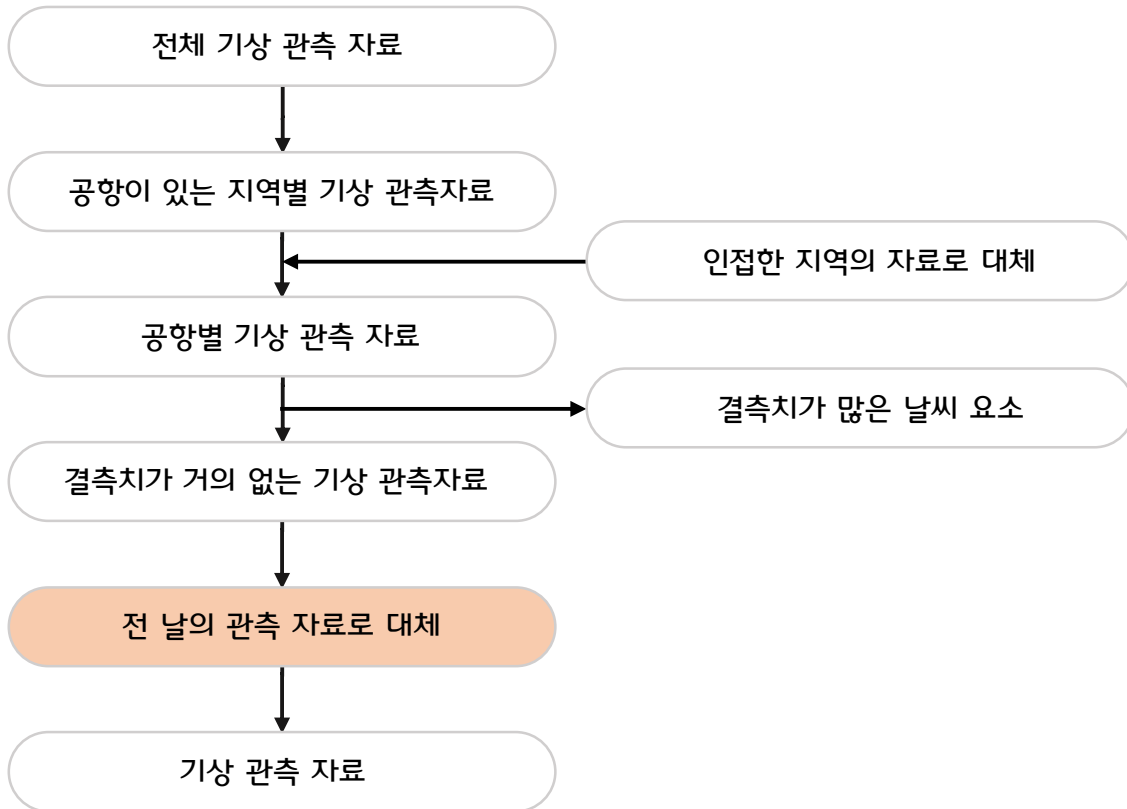


### 변인 제외 기준

다수의 결측값이 존재하는 날씨요소  
: 많은 결측치의 대체는 자료를 편향시킬 수 있으므로 제외

정확성 향상을 위해 결측치의 대체를 전 날의 관측값으로 대체하였다.

### 데이터 전처리 흐름 중 결측치 대체



### 결측치 대체 방법 예시

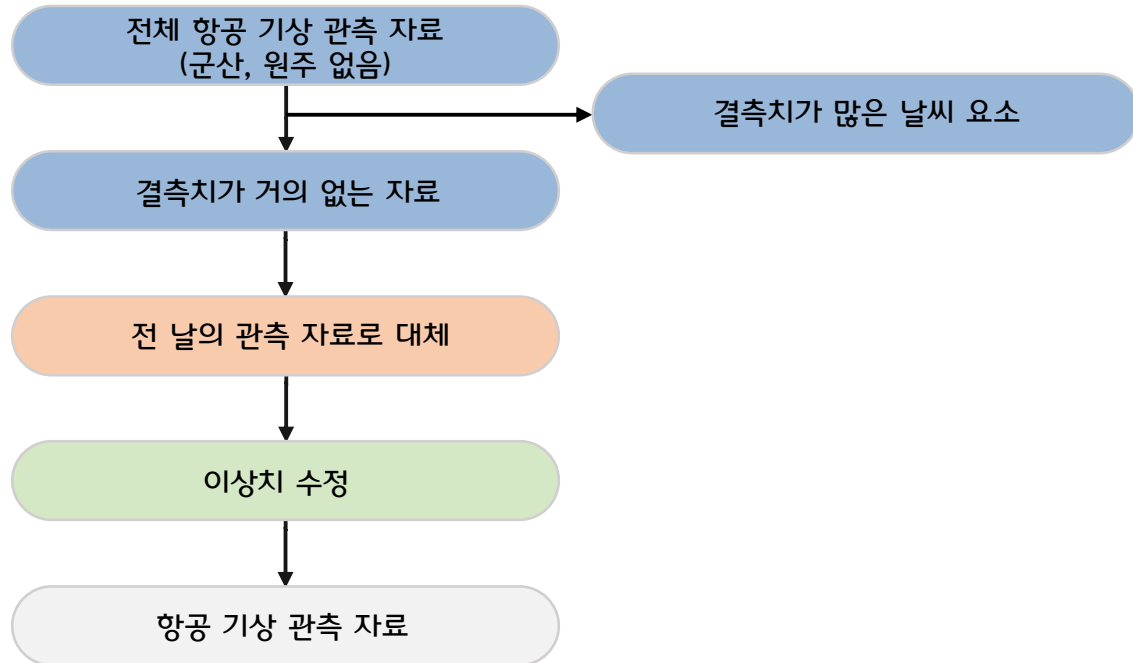
〈 2018년 목포 상대습도(%) 결측치 대체 예시 〉

	...	5월	...
1일	...	95.1	...
...	...	...	...
16일	...	92.0	...
17일	...	NA	...
18일	...	NA	...
...	...	...	...

연속적으로 결측치인 경우가 종종 존재하므로 결측치 전날 값과 다음 날과의 평균값을 사용하지 않고 전날의 값으로 대체

항공기상데이터 전처리는 1) 변인 제외, 2) 결측치 대체, 3) 이상치 수정 세 가지의 과정으로 진행하였다.

### 데이터 전처리 흐름도



### 데이터 전처리 흐름의 단계별 구분

#### ○ 변인 제외

다수의 결측치가 있는 날씨요소인 경우 제외

#### ○ 결측치 대체

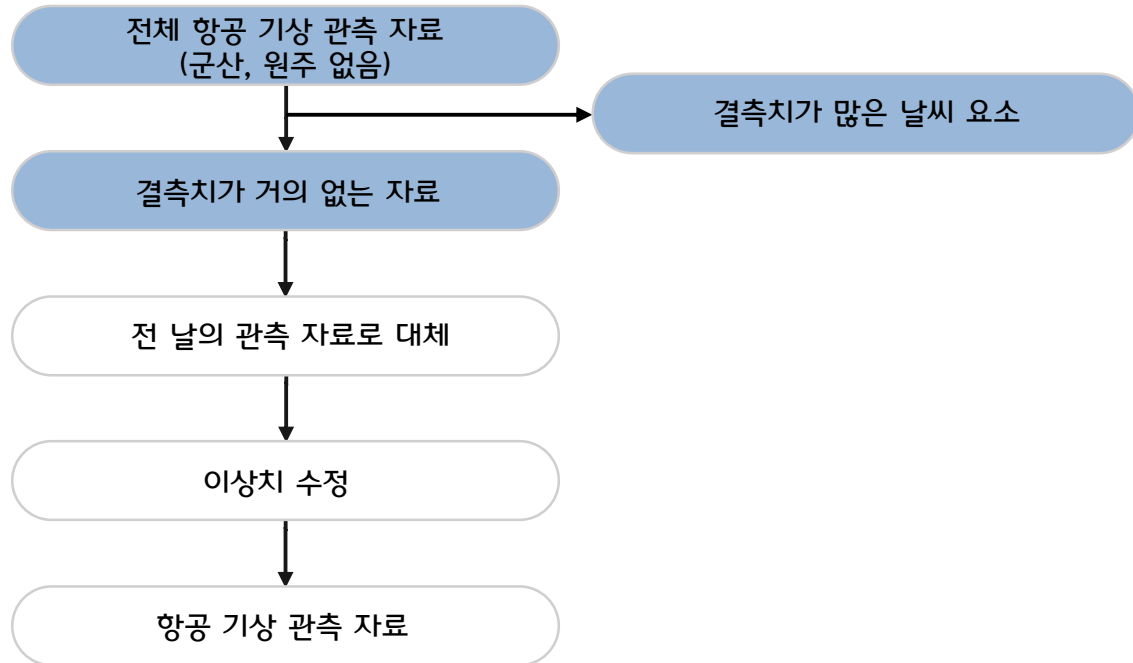
데이터가 없을 경우 다른 값으로 대체

#### ○ 이상치 수정

논리적 오류가 있는 자료인 경우 다른 값으로 수정

결측치의 대체가 불가능할 정도로 많은 경우 해당 변인을 제외하였다.

### 데이터 전처리 흐름 중 변인 제외

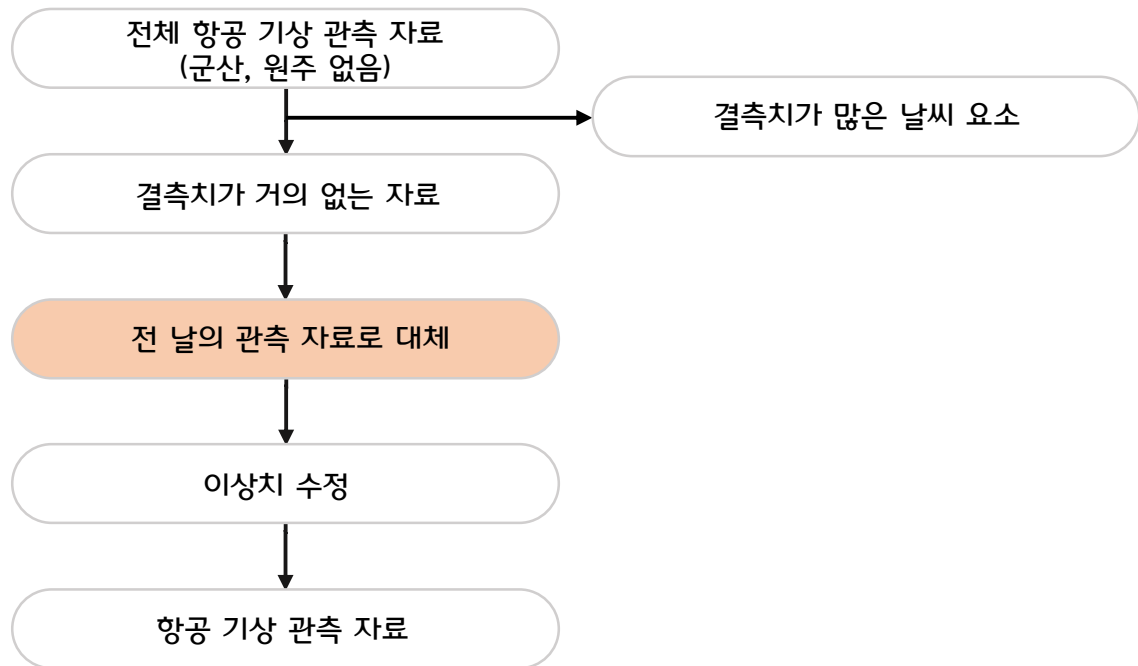


### 변인 제외 기준

다수의 결측값이 존재하는 날씨요소  
: 많은 결측치의 대체는 자료를 편향시킬 수 있으므로 제외

정확성 향상을 위해 결측치의 대체를 전 날의 관측값으로 대체하였다.

### 데이터 전처리 흐름 중 결측치 대체



### 결측치 대체 방법 예시

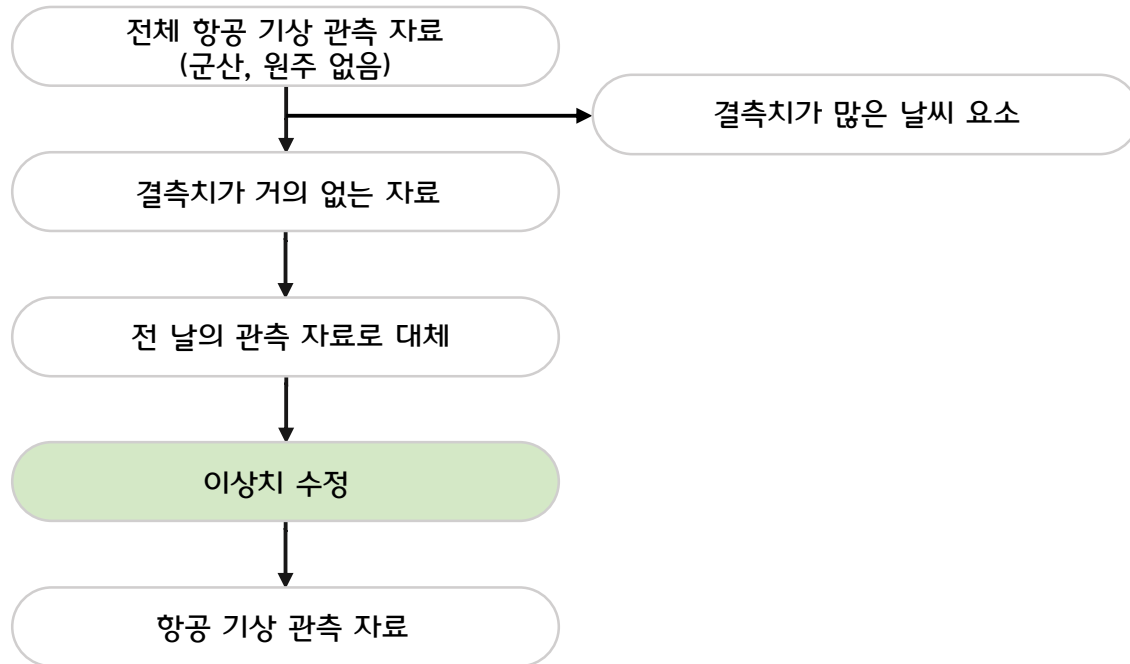
〈 2018년 대구공항 최고 기온(℃) 결측치 대체 예시 〉

	...	기온	...
20181201	...	12.0	...
...	...	...	...
20181222	...	14.8	...
20181223	...	NA	...
20181224	...	5.9	...
...	...	...	...

	...	기온	...
20181201	...	12.0	...
...	...	...	...
20181222	...	14.8	...
20181223	...	14.8	...
20181224	...	5.9	...
...	...	...	...

논리적 오류가 있는 값을 이상치로 판단하여 수정을 하였다.

### 데이터 전처리 흐름 중 이상치 수정



### 이상치 수정 방법 예시

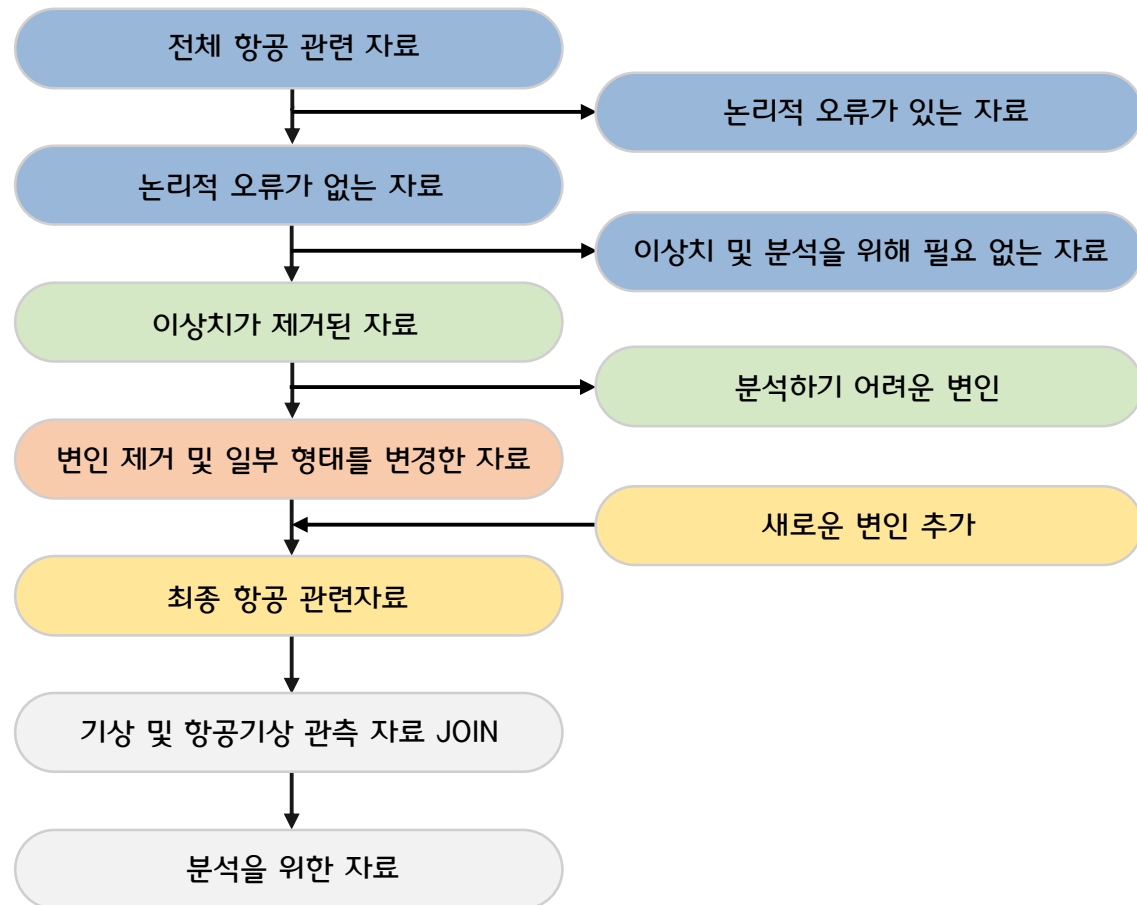
〈 청주공항 최대풍향(%) 이상치 수정 예시 〉

지점명	날짜	최대 풍속	지점명	날짜	최대 풍속
청주공항	2018 0701	31	청주공항	2018 0701	31
...	...	...	...	...	...
청주공항	2018 0706	90	청주공항	2018 0706	9
...	...	...	...	...	...

풍향은 0~36(0°~360°)의 값을 가짐  
: 주변지역의 당일 풍향을 참고하여 90을 9로 수정

항공 데이터 전처리는 1) 데이터 삭제, 2) 변인 제외, 3) 데이터 변형, 4) 새로운 변인 추가 네 가지의 과정으로 진행하였다.

데이터 전처리 흐름도



데이터 전처리 흐름의 단계별 구분

### ○ 데이터 삭제

이상치 및 논리적 오류가 있는 데이터 제거

### ○ 변인 제외

분석하기에 적합하지 않거나 분석하기 어려운 변인 제외

### ○ 데이터 변형

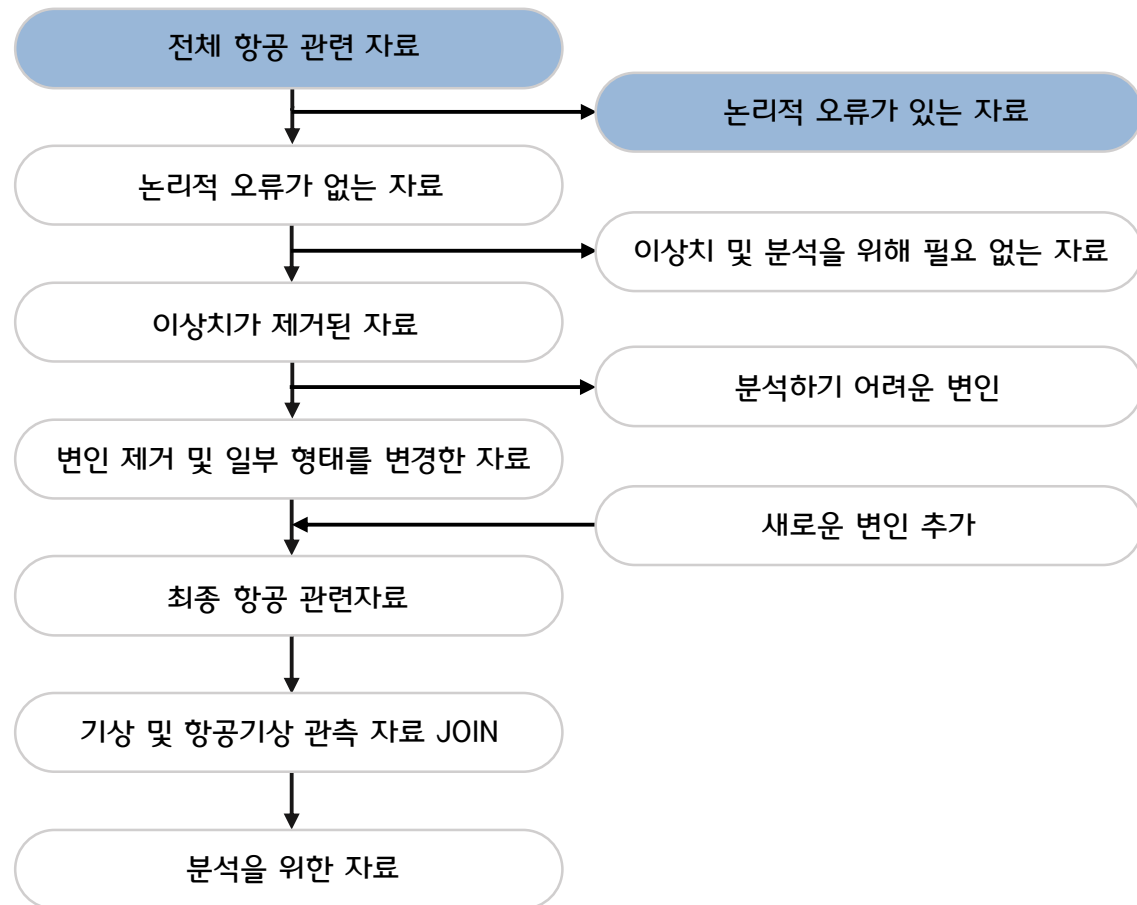
명목 척도의 데이터 수치화 및 간격 척도의 데이터 변형

### ○ 새로운 변인 추가

좀 더 나은 분석을 위해 새로운 변인 추가

지연시간을 구하여 논리적 오류를 찾아 그 데이터를 제거하였다.

### 데이터 전처리 흐름 중 데이터 삭제



### 논리적 오류를 찾는 과정

〈 지연시간을 계산하는 방법 〉

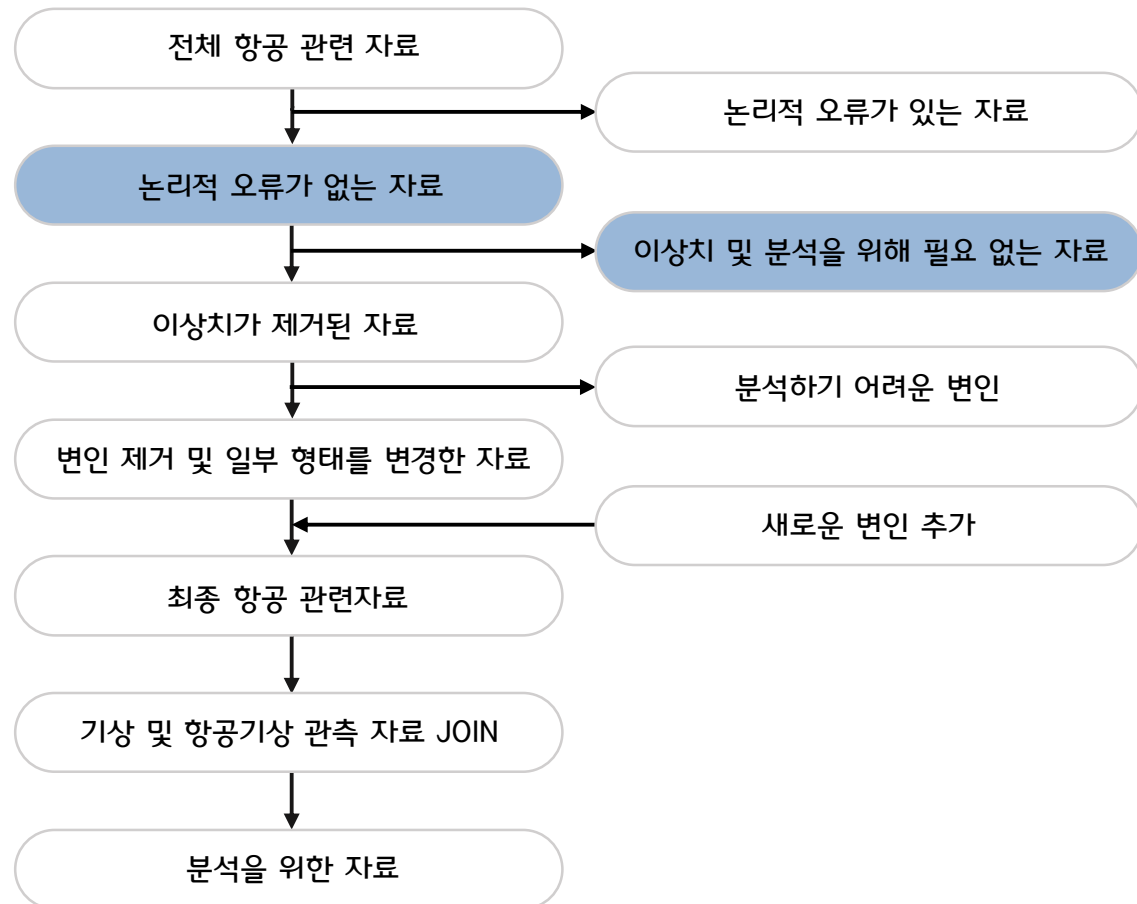
$$\text{DLYTIME(지연시간)} = \text{ATT(실제시각)} - \text{STT(계획시각)}$$

1. 지연시간이 30분 초과인데 지연여부가 'N' 인 경우
2. 지연시간이 30분 이하인데 지연여부가 'Y' 인 경우



이상치는 특정 범위에 너무 벗어나 있어 데이터 분석이나 모델링의 결과에 큰 영향을 미친다.  
따라서 이상치 판단 기준을 정하여 이상치 판별 후 제거하였다.

### 데이터 전처리 흐름 중 데이터 삭제

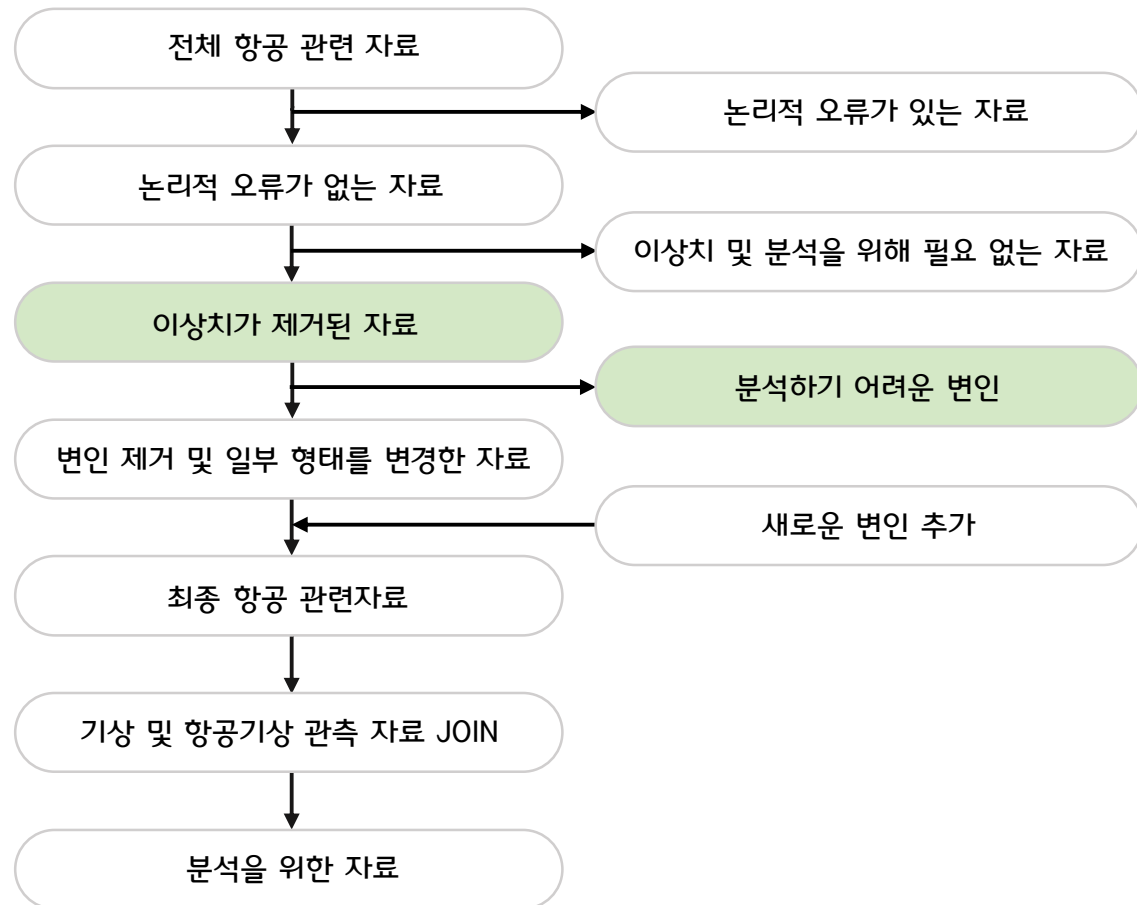


### 이상치 판단 기준

1. 이상치(outlier)  
: 지연시간이 -20 이하인 데이터를 조기출발 편 또는 부정기 페리편(공기비행, 승객이 탑승하지 않는 항공편)으로 간주하고 이를 이상치로 판단
2. 분석을 위해 필요 없는 자료  
: 결항여부(CNL)가 'Y' 인 경우에 지연여부(DLY)가 'N' 이지만 정상 운행한 경우 또한 아니기 때문에 제거

분석에 적합하지 않은 변인을 제외하였다.

### 데이터 전처리 흐름 중 변인 제외



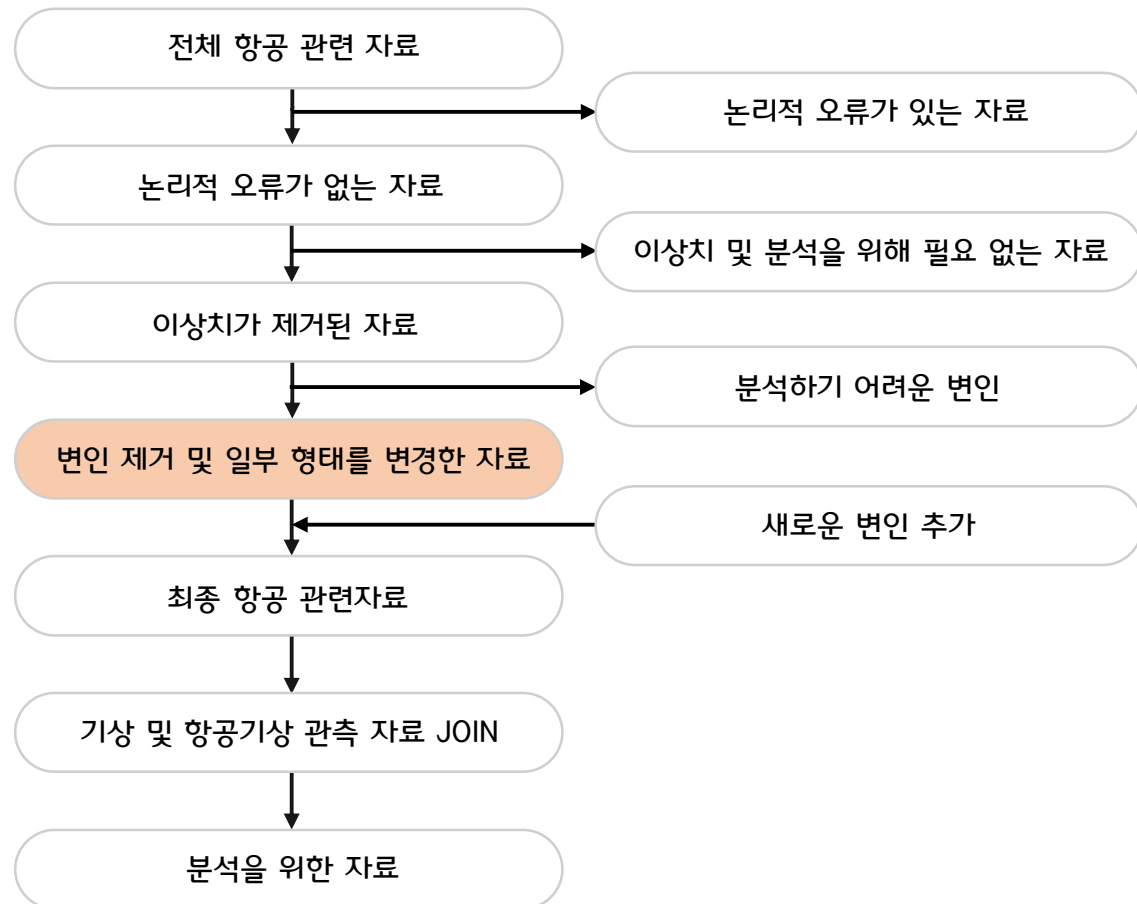
### 분석에 적합하지 않은 변인

: FLO(항공사), FLT(편명), REG(등록기호),  
IRR(부정기편), ATT(실제시각), DRR(지연사유),  
CNL(결항여부), CNR(결항사유)

1. 변인의 범주가 너무 많아 분석하기 어렵다고 판단한 경우
2. 예측할 데이터(AFSNT\_DLY.csv)에 없으면서 예측해서 데이터에 추가하여 사용하기 어려운 경우

분석을 위해 명목형자료를 수치화 하였고 순서형자료와 연속형자료의 형태를 분석의 용이함을 위해 변경하였다.

### 데이터 전처리 흐름 중 데이터 변형

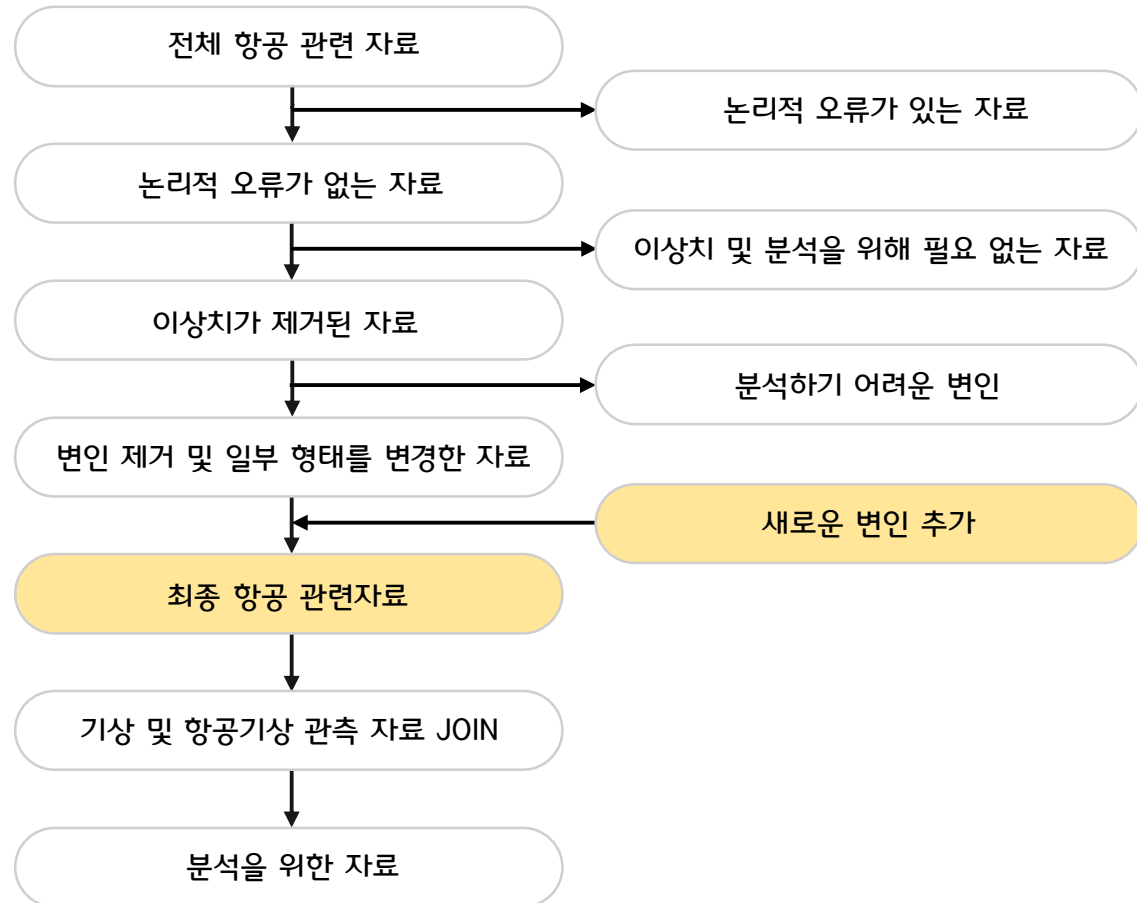


### 데이터 변형한 변인

1. 명목 척도의 질적 데이터 수치화  
: SDT\_DY(요일), ARP(공항), ODP(상대공항), AOD(출도착), DLY(지연여부)
2. 데이터의 형태 변환  
: STT(계획시각), SDT\_YY(연)

시간이 누적됨에 따라 지연율이 달라질 수 있는 가능성을 고려하였다.

### 데이터 전처리 흐름 중 새로운 변인 추가



### 새로운 변인 생성 방법

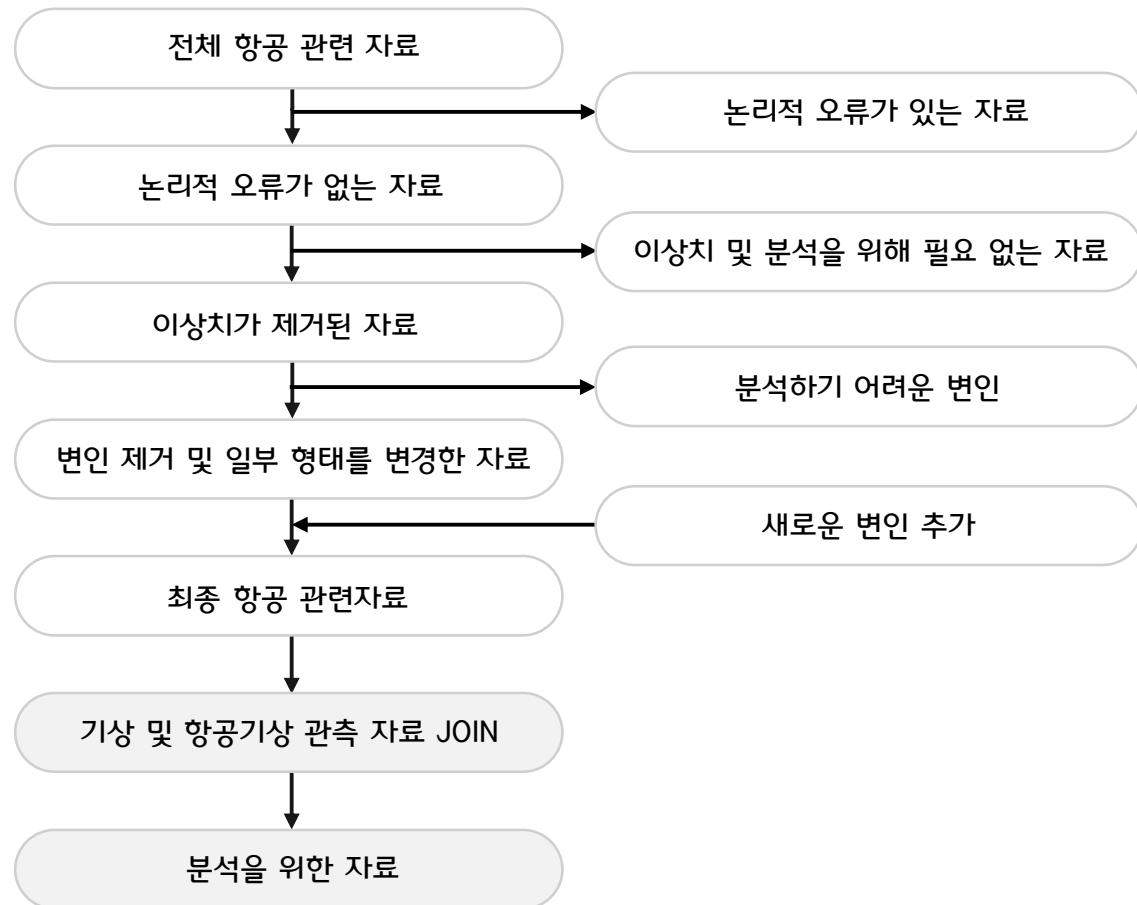
< all\_time 변인 생성 방법 >

$all\_time = 연 + 월 + 일 + 계획시각$

: 시간이 지남에 따라 지연율이 영향을 줄 수 있다고 가정하여 생성

분석을 위한 자료는 항공기상데이터 존재 유무의 기준으로 두 개의 자료로 나누었다.

### 데이터 전처리 흐름도



### 분석을 위한 자료

: 군산공항과 원주공항은 항공기상데이터가 존재하지 않으므로 기상데이터를 활용하여 예측

#### 1. data1

: 군산공항, 원주공항을 제외하고 항공기상데이터가 포함된 데이터

#### 2. data2

: 모든 공항이 존재하고 항공기상데이터가 제외된 데이터

data1은 변인 24개와 하나의 종속변수로 구성되어 있으며 data2는 변인 12개와 하나의 종속변수로 구성되어 있다. V~은 항공 데이터, T~은 시간 관련 데이터, W~은 기상 데이터이고 A~는 항공기상데이터이다.

data1 변수 설명(군산공항, 원주공항 제외)

	변수명	변수설명	변수명	변수설명
종속변수	V8	지연여부		
변인	V1	연	A1	평균해면기압(hPa)
	V2	월	A2	평균기온(°C)
	V3	일	A3	최고기온(°C)
	V4	요일	A4	최저기온(°C)
	V5	공항	A5	평균이슬점(°C)
	V6	상대공항	A6	평균상대습도(%)
	V7	출도착	A7	최소상대습도(%)
	T1	계획시각	A8	평균풍속(kts)
	T2	all_time	A9	최대풍속(kts)
	W1	평균기온(°C)	A10	최대풍향(10deg)
	W2	평균풍속(km/h)	A11	최대순간풍속(kts)
	W3	상대습도(%)	A12	최대순간풍향(10deg)

data2 변수 설명

	변수명	변수설명
종속변수	V8	지연여부
변인	V1	연
	V2	월
	V3	일
	V4	요일
	V5	공항
	V6	상대공항
	V7	출도착
	T1	계획시각
	T2	all_time
	W1	평균기온(°C)
	W2	평균풍속(km/h)
	W3	상대습도(%)

범주형 데이터(V1,V2,V3,V4,V5,V6,V7)의 상관관계를 알아보기 위해 cramer's V 계수를 살펴보았다.

Cramer's V 계수

Cramer's V 계수  
: 2개 이상의 범주로 나눈 집단 간의 상관계수를 구하는 경우  
: 변수들이 연속성 여부와 측정치들의 분포가 정상이든 편향된 분포이든 상관없이 적용가능  
: 계수는 0에서 1사이의 값을 가지며 1에 가까울수록 상관관계가 높고 0에 가까울수록 상관관계가 없음

V1,V2,V3,V4,V5,V6,V7의 Cramer's V 계수

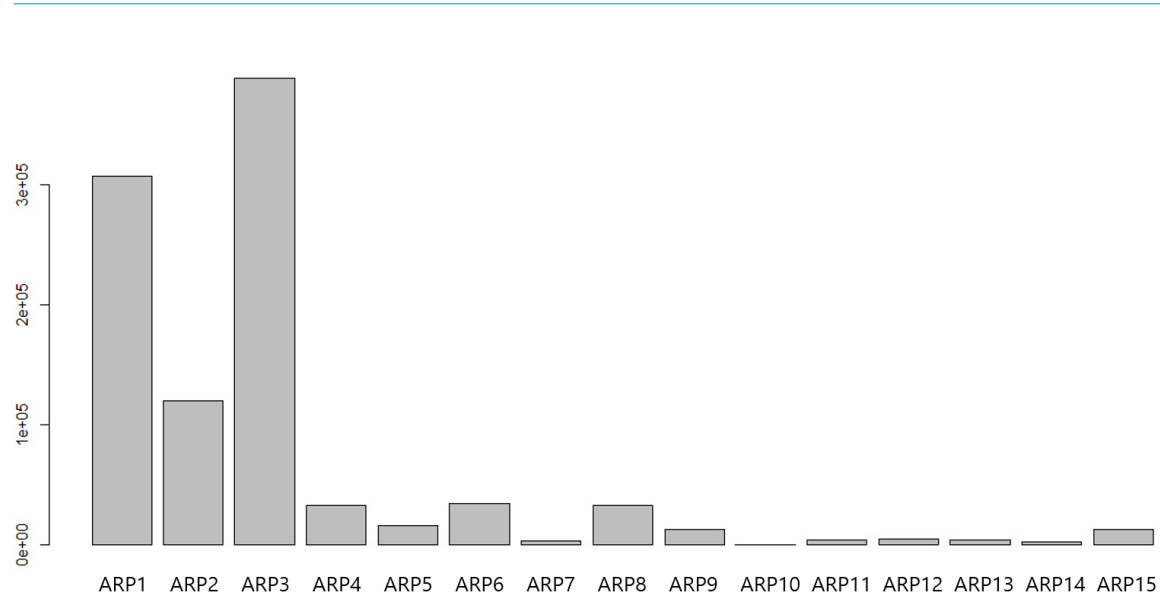
	V1	V2	V3	V4	V5	V6	V7
V1	1.000000000	0.289533098	0.011858008	0.007682510	0.020582418	0.020052013	0.001122119
V2	0.289533098	1.000000000	0.044103213	0.031607952	0.007274500	0.007019566	0.001029701
V3	0.011858008	0.044103214	1.000000000	0.068826074	0.002621573	0.002902226	0.000558127
V4	0.007682510	0.031607952	0.068826074	1.000000000	0.010507068	0.010467396	0.000651688
V5	0.020582418	0.007274500	0.002621573	0.010507068	1.000000000	0.253455508	0.005190829
V6	0.020052013	0.007019566	0.002902226	0.010467396	0.253455508	1.000000000	0.004743756
V7	0.001122119	0.001029701	0.000558127	0.000651688	0.005190830	0.004743757	1.000000000

→ Cramer's V 계수가 모두 0.3이하이므로 범주형 데이터간의 상관관계가 없다고 판단

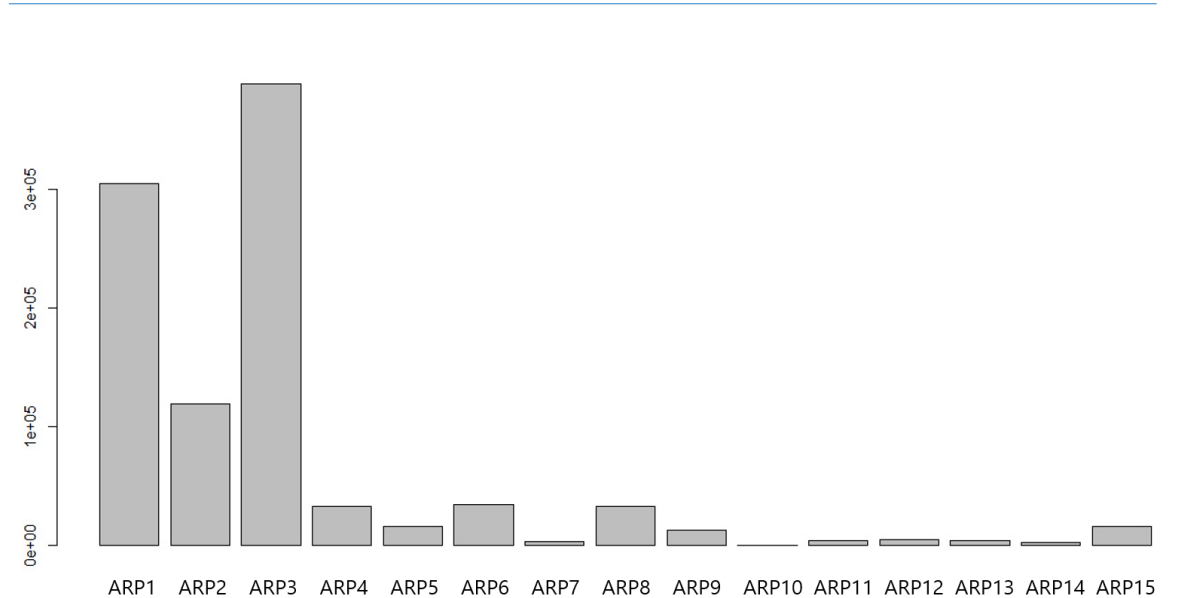
## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

범주형 데이터인 공항(V5) 데이터와 상대공항(V6) 데이터의 빈도표를 plot으로 각각 살펴보았다.

공항(V5)의 plot



상대공항(V6)의 plot



→ V5에서 가장 높게 나온 공항은 ARP3이며 가장 적게 나온 공항은 ARP10임

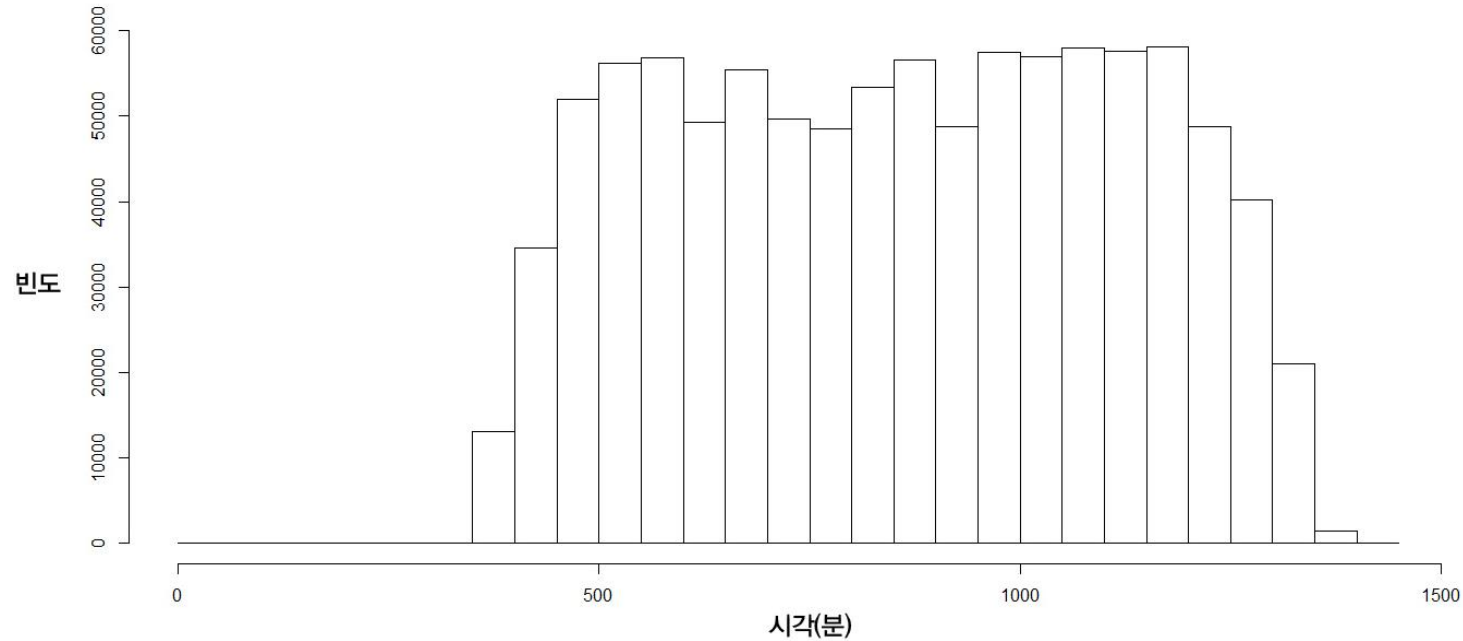
→ V6에서 가장 높게 나온 공항은 ARP3이며 가장 적게 나온 공항은 ARP10임



## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

수치형 데이터인 계획시각(T1) 데이터의 빈도를 히스토그램으로 살펴보았다.

계획시각(T1)의 plot

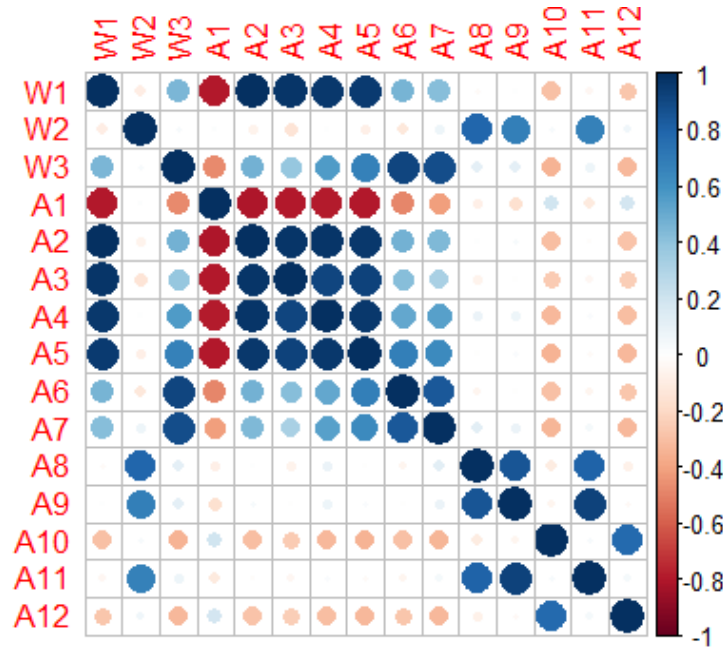


- 시간대별로 운행하는 항공편의 수가 어느정도 차이가 있음
- 하루 중 국내선 항공 운행이 가장 많은 시간대는 16시~20시

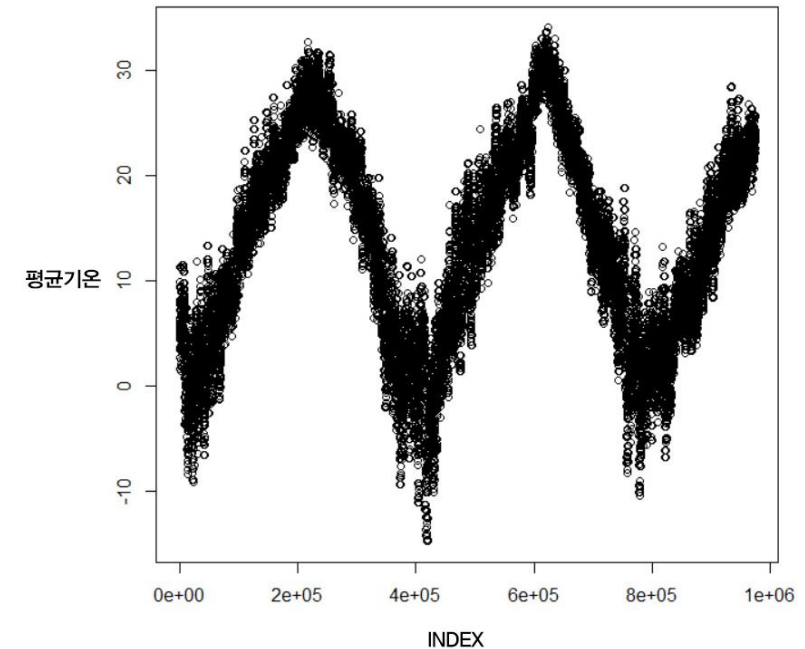
## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

수치형 데이터(W1,W2,W3,A1,A2,A3,A4,A5,A6,A7,A8,A9,A10,A11,A12)의 상관관계를 알아보기 위해 correlation plot을 그려 보았다. 평균기온(W1) 데이터의 plot 또한 살펴보았다.

Correlation plot



평균기온(W1)의 plot



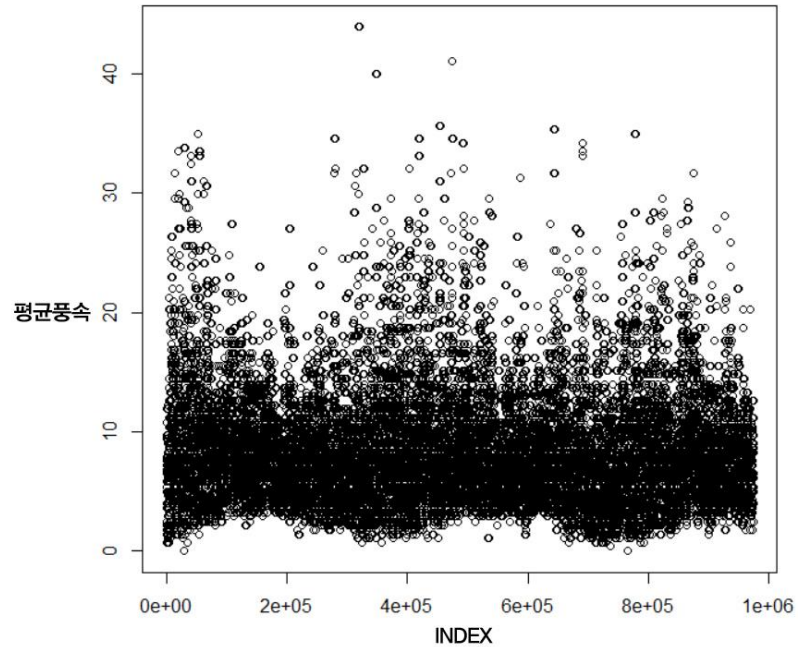
→ 계절성(seasonality)으로 인해 상관관계수가 높게 나온 경우가 있음

→ 평균기온은 계절성(seasonality)을 보임

## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

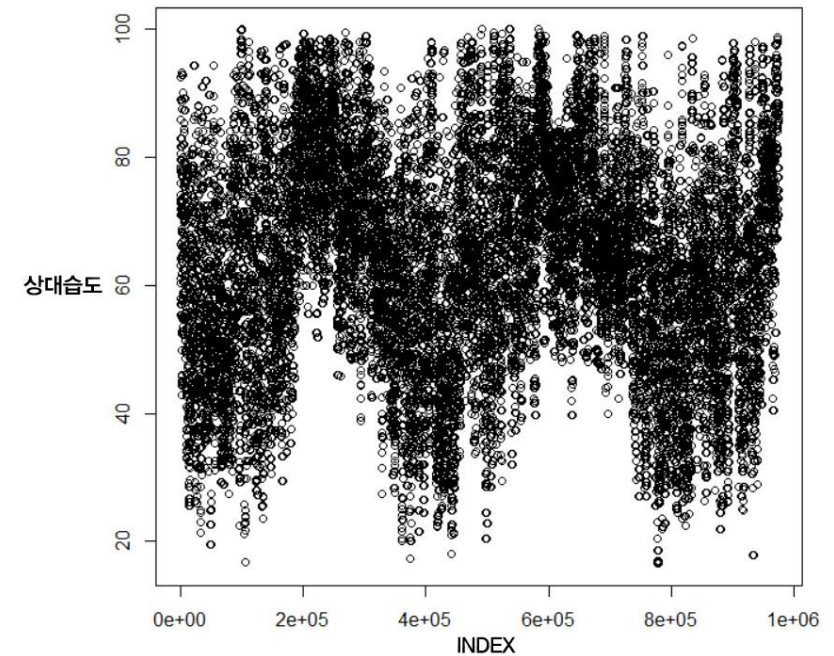
평균풍속(W2) 데이터와 상대습도(W3) 데이터를 알아보기 위해서 plot을 각각 살펴보았다.

평균풍속(W2)의 plot



→ 평균풍속은 대부분 0~20사이에 있으며 최대값은 43.9임

상대습도(W3)의 plot

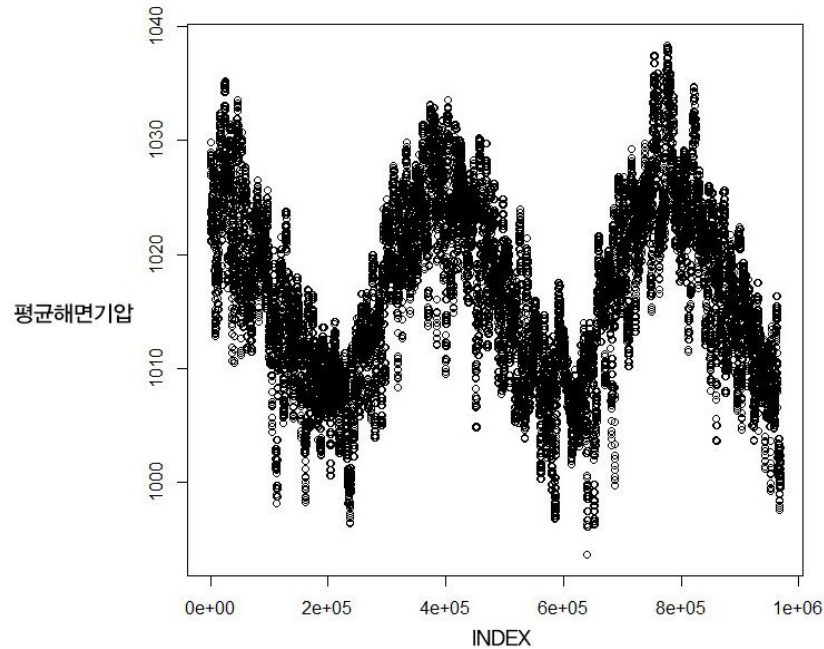


→ 상대습도는 약한 계절성(seasonality)을 보임

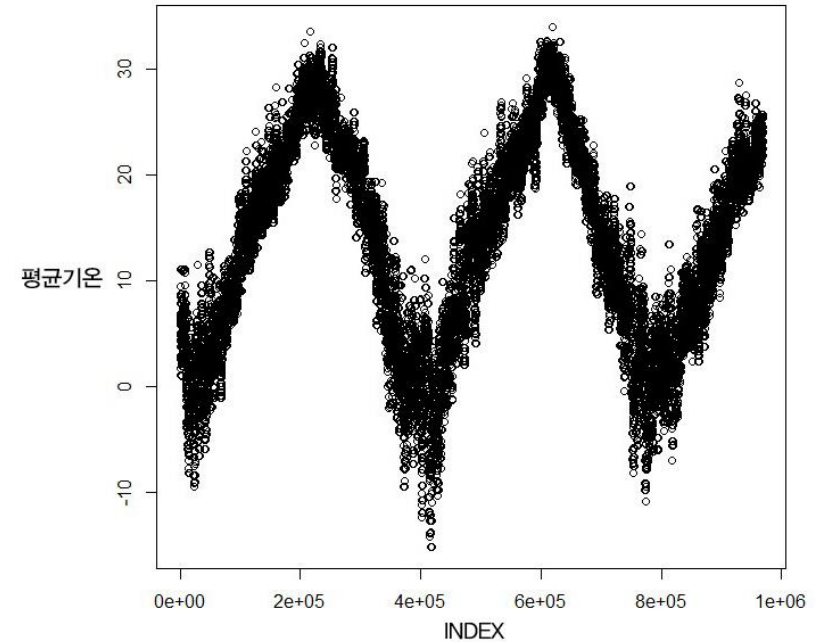
## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

평균해면기압(A1) 데이터와 평균기온(A2) 데이터를 알아보기 위해서 plot을 각각 살펴보았다.

평균해면기압(A1)의 plot



평균기온(A2)의 plot

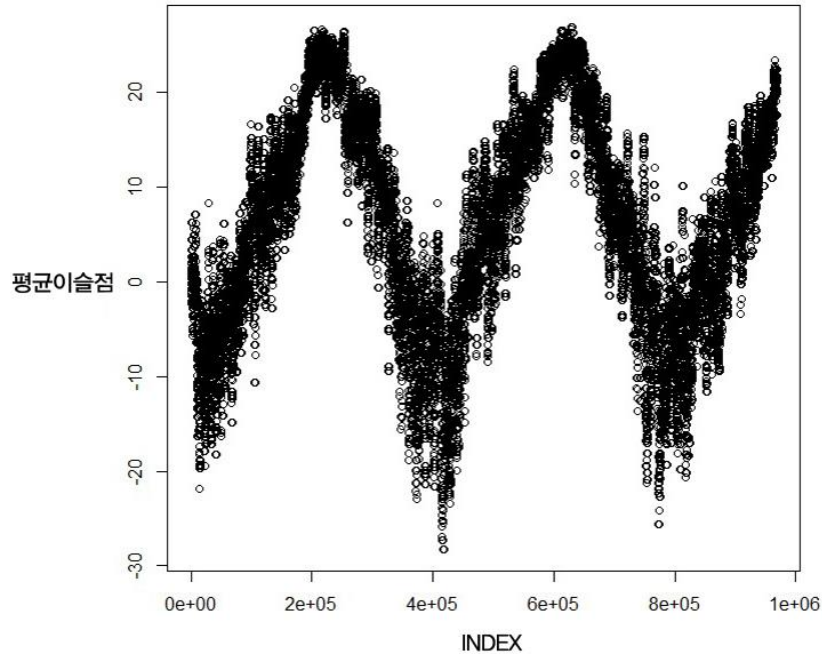


- 평균해면기압과 평균기온은 계절성(seasonality)을 보임
- 평균해면기압과 평균기온의 상관계수는 -0.8이며 강한 음의 상관관계를 보임

## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

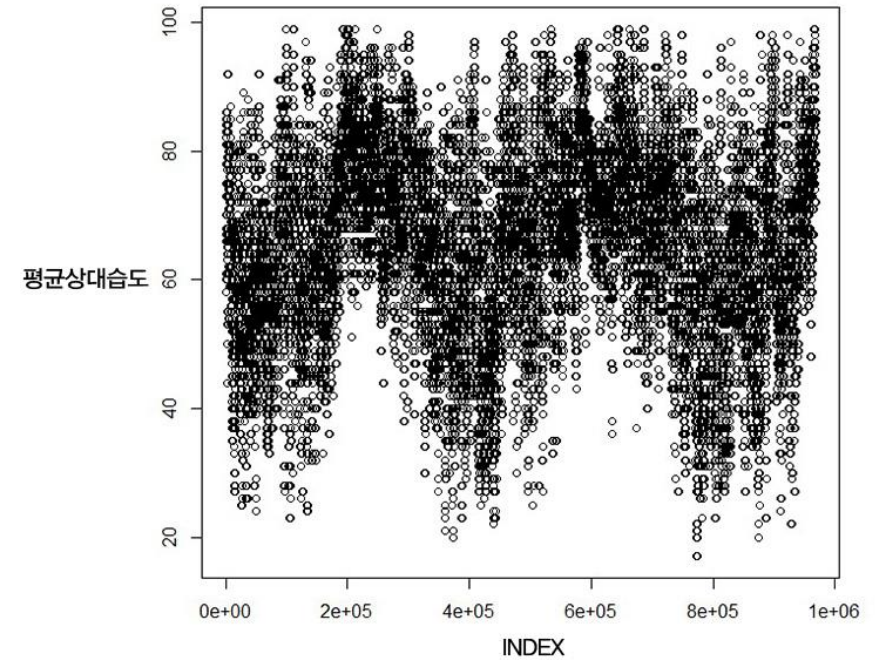
평균이슬점(A5) 데이터와 평균상대습도(A6) 데이터를 알아보기 위해서 plot을 각각 살펴보았다.

평균이슬점(A5)의 plot



→ 평균이슬점은 계절성(seasonality)을 보임

평균상대습도(A6)의 plot



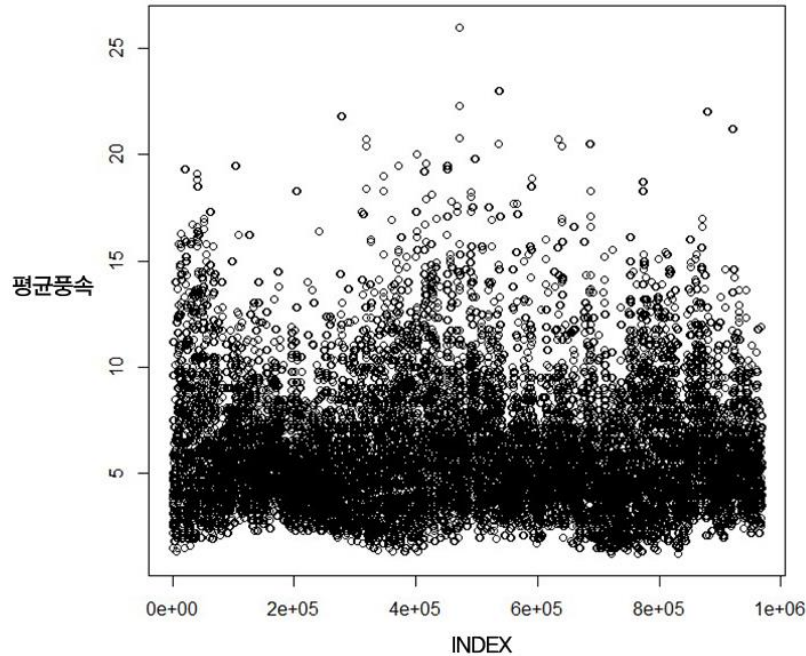
→ 상대습도는 약한 계절성(seasonality)을 보임



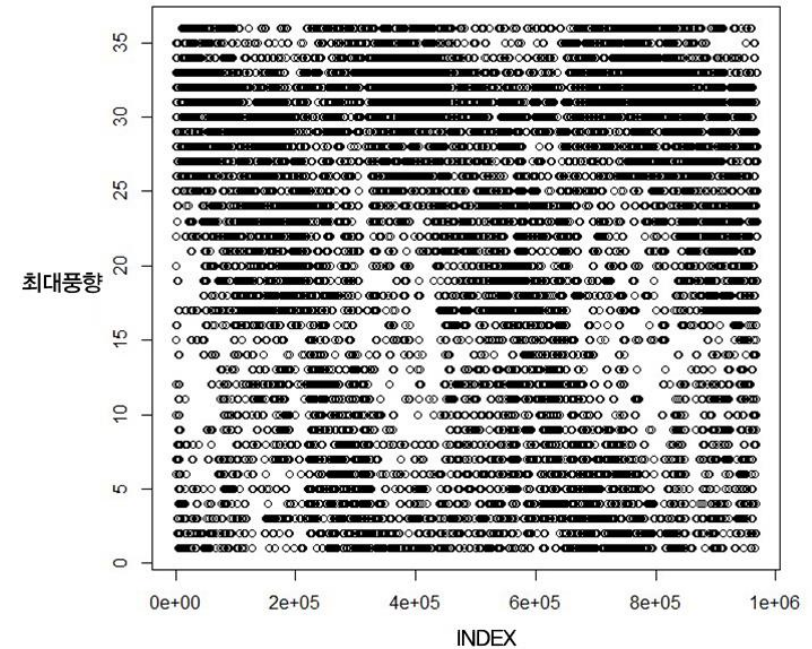
## 탐색적 데이터 분석 | 3-2. 탐색적 데이터 분석(EDA)

평균풍속(A8) 데이터와 최대풍향(A10) 데이터를 알아보기 위해서 plot을 각각 살펴보았다.

평균풍속(A8)의 plot



최대풍향(A10)의 plot



→ 평균풍속은 대부분 0~15사이에 있으며 최대값은 26

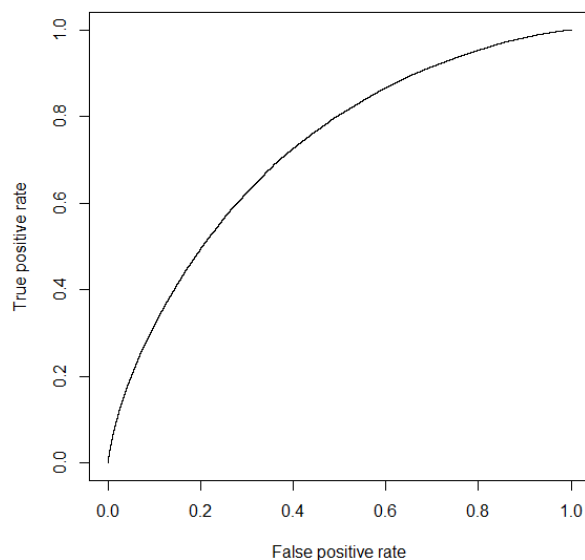
→ 평균풍속은 약한 계절성(seasonality)을 보임

→ 최대풍향은 약한 계절성(seasonality)을 보임

Logistic regression, Xgboost와 LightGBM의 ROC Curve, AUROC, 학습시간을 각각 살펴보았다.

### Logistic regression

ROC Curve



AUROC

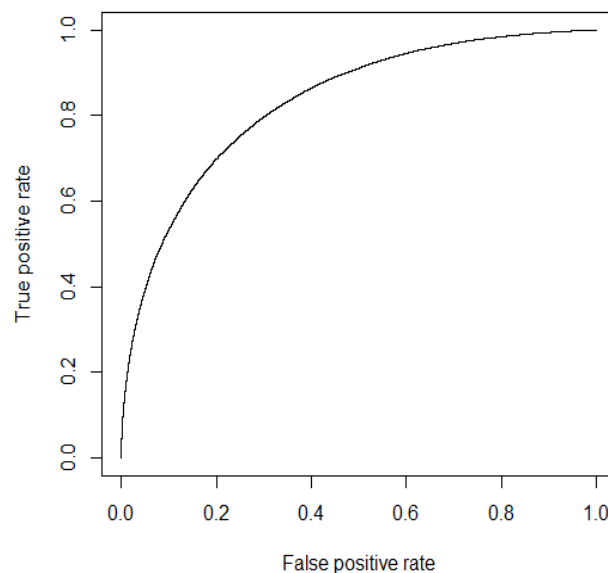
0.722

Training time

00:35:56

### Xgboost

ROC Curve



AUROC

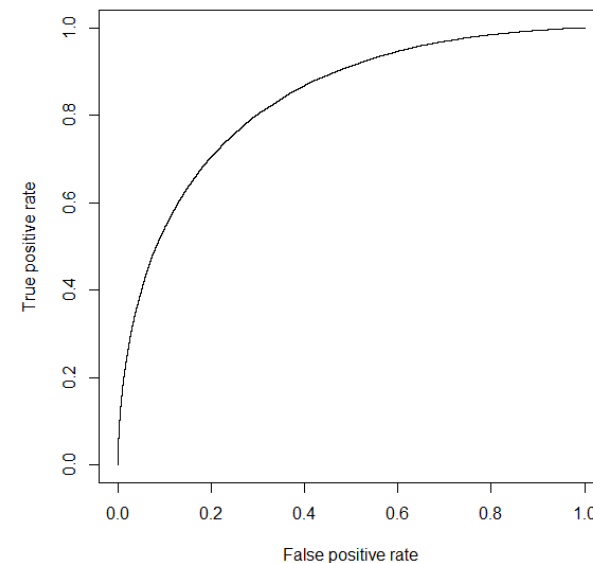
0.830

Training time

00:39:58

### LightGBM

ROC Curve



AUROC

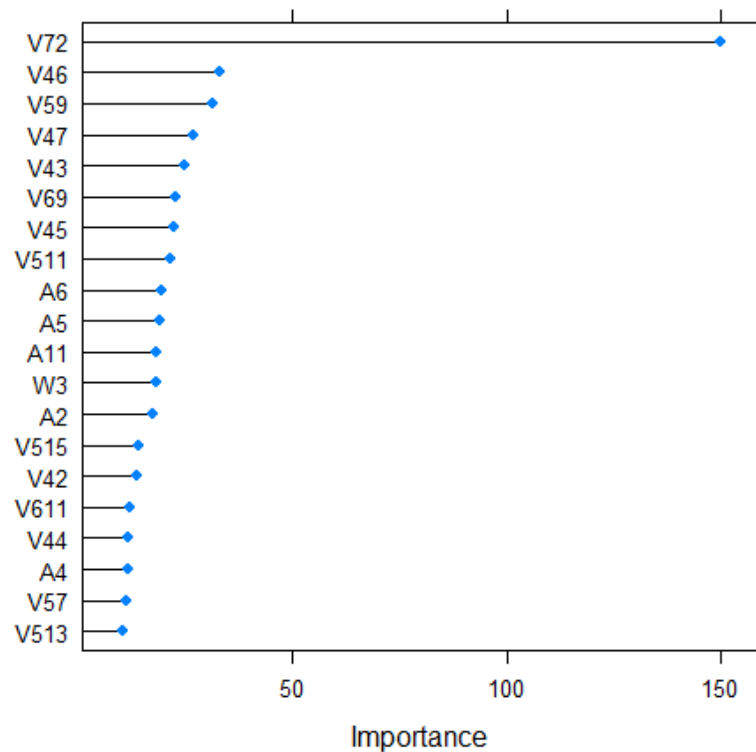
0.831

Training time

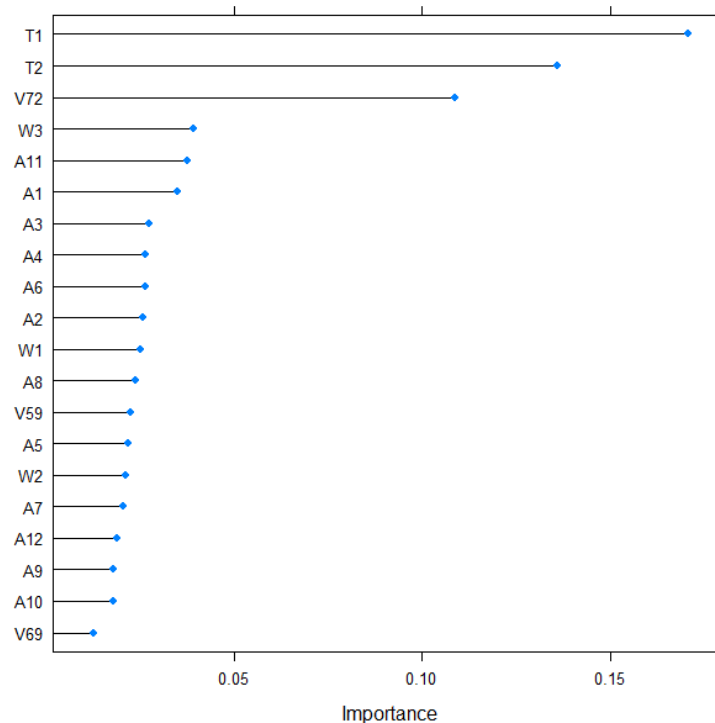
00:00:40

Logistic regression, Xgboost와 LightGBM의 data1의 변수 중요도를 각각 살펴보았다.  
가장 중요도가 높은 변수 20개를 나타내었다.

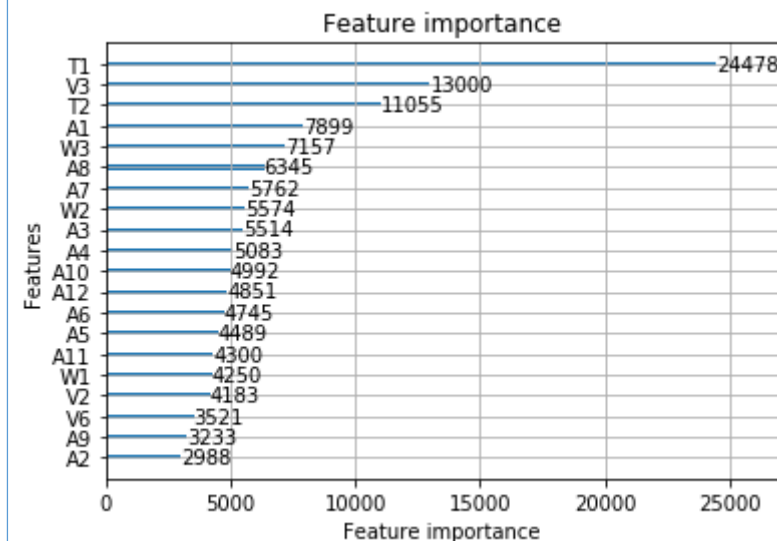
Logistic regression



Xgboost



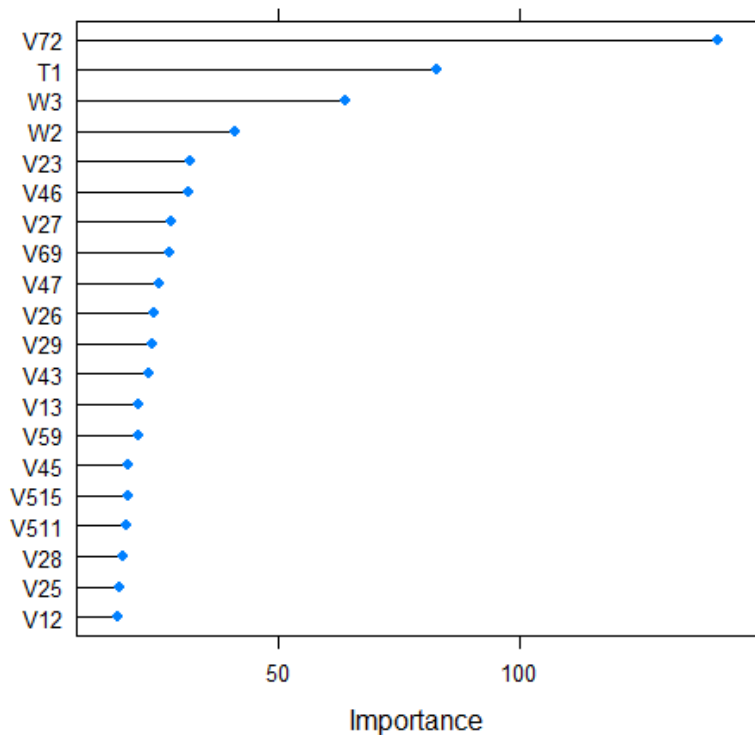
LightGBM



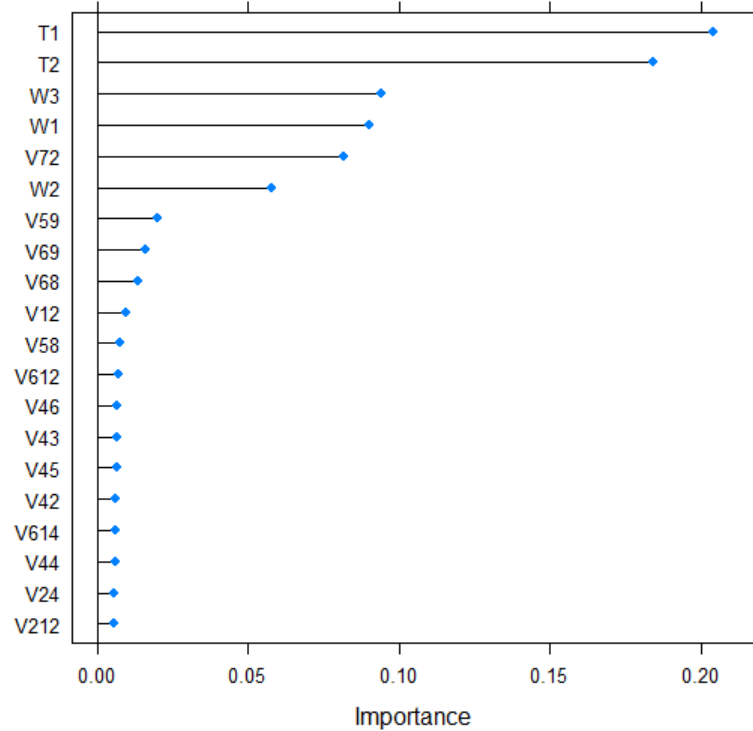


Logistic regression, Xgboost와 LightGBM의 data2의 변수 중요도를 각각 살펴보았다.  
가장 중요도가 높은 변수 20개를 나타내었다.(LightGBM은 12개)

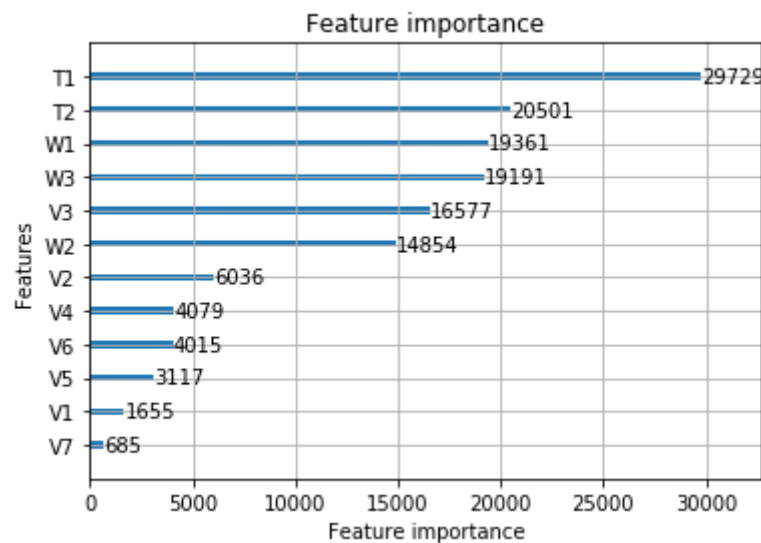
Logistic regression



Xgboost



LightGBM



Logistic regression, Xgboost와 LightGBM 중에서 세 가지의 기준으로 best model을 선정하였다.

### 모델선정기준

#### 1. 분석할 데이터와 모델의 적합도

: 분석할 데이터는 크기가 크고 Tabular format 데이터(X-Y Grid로 되어있는 데이터)

#### 2. 예측 정확도

: AUROC값

#### 3. 효율성

: train에 걸리는 시간

### LightGBM

- 크기가 큰 데이터 분석에 적합하다.
- 높은 예측 정확도를 가진다.
- 높은 효율성을 가진다.
- 적은 메모리 용량을 가진다.
- 기능상의 다양성이 있다.

→ Xgboost와 LightGBM의 AUROC값 즉, 예측 정확도는 비슷하나 효율성이 Xgboost에 비해 매우 높은 LightGBM을 best model로 선정하였다.

“

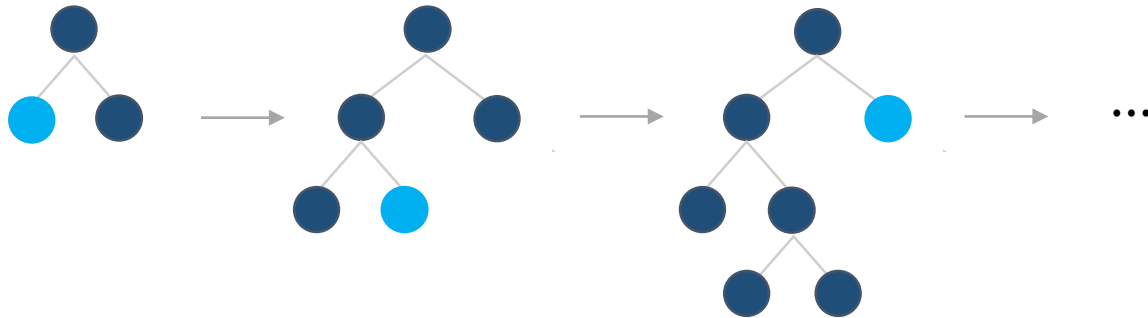
**LightGBM** is a fast, high-performance gradient boosting framework based on decision tree algorithm.

– LightGBM은 gradient boosting에 기반한 빠르고 효율적인 기법이다.

”

LightGBM은 다른 tree-based algorithm이 level-wise 형식으로 tree를 만드는 반면에 leaf-wise 형식으로 만든다.

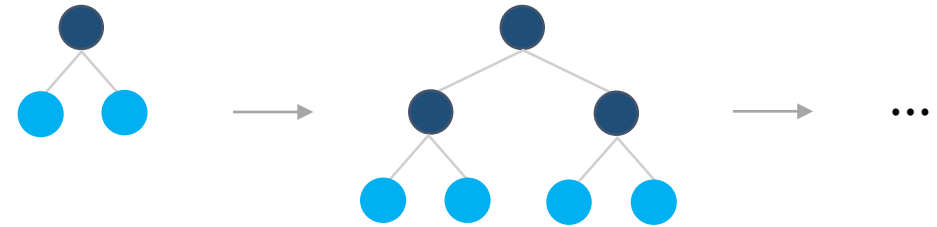
Leaf-wise tree growth



grow vertically

LightGBM

Level-wise tree growth



grow horizontally

other algorithm (Xgboost, ...)

→ leaf-wise 형식은 leaf를 max delta loss를 이용하여 선택한다.

→ 따라서 leaf-wise은 level-wise형식보다 loss를 줄일 수 있다.

LightGBM으로 예측할 때 사용한 주요 parameter를 정리하였다.

### LightGBM에서 사용한 parameter 종류 및 설명

〈 parameter(매개변수)와 argument(전달인자) 〉

parameter	argument	설명
max_depth	200	트리 모델의 최대 깊이
num_leaves	700	트리 모델의 복잡성을 제어하는 주요 매개 변수
num_boost_round	200	하나의 적합한 모델을 만들기 위한 최대 부스팅 횟수
early_stopping_rounds	10	하나의 적합한 모델을 만들기 위한 최소 부스팅 횟수
nfold	100	data를 fold할 횟수
learning_rate	0.05	각 트리가 최종 결과에 미치는 영향을 결정하는 값
objective	binary	종속변수의 형태를 지정

→ 그 외 대부분의 parameter는 기본값으로 사용하였다.

Anaconda Python 3-7 version으로 LightGBM을 사용하였다.

### 사용한 주요 packages

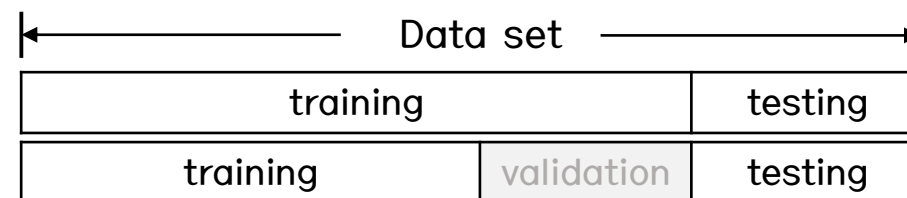
< packages >

packages	import한 이유
lightgbm	LightGBM을 활용하기 위해 사용
roc_auc_score	AUROC값을 계산할 때 사용
roc_curve	ROC Curve을 도출할 때 이용
train_test_split	train 데이터와 test 데이터를 분리할 때 사용
matplotlib	plot을 그리기 위해 사용
numpy	seed값을 고정할 때 사용
pandas	csv 파일 저장 및 불러오기를 할 때 사용

### Training & Predicting

#### 1. Training

: train data와 test data를 7:3으로 나누었다.



#### 2. Predicting

: type 1 error와 type 2 error를 고려하여 cutoff값을 설정한 뒤 지연여부를 예측하였다.

LightGBM으로 예측한 결과의 crosstab과 문제 데이터를 예측한 결과의 일부를 나타내었다.

crosstab

< 교차분석 >

	지연이 되지 않았다고 예측	지연이 되었다 고 예측
실제로 지연되지 않은 경우	192531	63728
실제로 지연된 경우	8942	27006

1. 실제가 0인데 0로 예측한 비율  
: 0.751

2. 실제가 1인데 1으로 예측한 비율  
: 0.751

예측결과(일부)

< AFSNT\_DLY.csv >

SDT_YY	SDT_MM	SDT_DD	...	DLY	DLY_RATE
2019	9	16	...	0	0.01
2019	9	16	...	0	0.04
2019	9	16	...	0	0.05
2019	9	16	...	1	0.22
2019	9	16	...	0	0.06
2019	9	16	...	0	0.05
2019	9	16	...	0	0.03
2019	9	16	...	0	0.04
2019	9	16	...	0	0.01
...	...	...	...	...	...

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu의 LightGBM: A Highly Efficient Gradient Boosting Decision Tree

<https://lightgbm.readthedocs.io/en/latest/Parameters.html>에서 Parameters 참고

<https://www.kaggle.com/questions-and-answers/59067>의 What is Light GBM vs XGBOOST?

<http://www.airport.co.kr/www/extra/stats/airportStats/layOut.do?cid=2015102917501542253&menuId=397>에서 공항명 참고

<https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial>의 Predicting flight delays

<https://www.kaggle.com/pranav84/lightgbm-fixing-unbalanced-data-lb-0-9680>의 LightGBM

Samprit Chatterjee and Ali S. Hadi(2006), Regression Analysis by Example. 4<sup>th</sup> Ed., Wiley



Network Meta-Analysis for Decision-Making

(공)저: Sofia Dias, A. E. Ades, Nicky J. Welton, Jeroen P. Jansen, Alexander J. Sutton