Prediction Of Flight delay

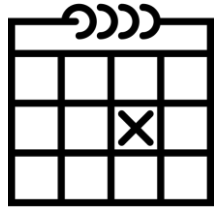
Excellent In Analysis

Prediction Of Flight delay

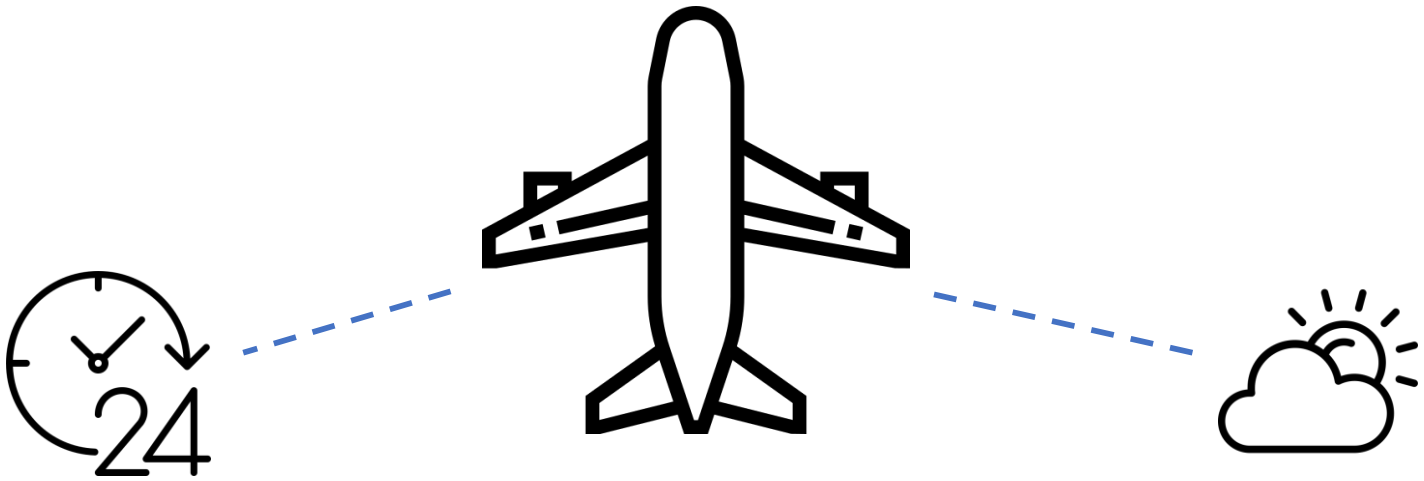
- 프로젝트 분석 배경
- 데이터 수집 & Feature Engineering
- 데이터 분석
- 예측 & 기대효과

1

프로젝트 분석 배경



항공운항데이터,항공기상데이터 등을 활용한 항공지연 예측 모형 개발

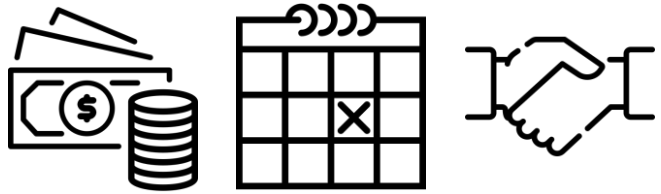


예측 모형의 필요성

순위	항공사명	평균 지연율
1	말레이시아항공	48.14%
2	선전항공	48.13%
3	피치항공	45.30%
4	중국동방항공	43.49%
5	중화항공	41.11%
6	대한한공	39.48%
7	제주항공	34.89%
8	에어차이나	33.90%
9	아시아나항공	33.87%
10	비엣제트항공	33.71%

<인천공항 항공사 평균 지연율 TOP10>
출처 : 국토교통부,항공정보포털시스템(2015.1.1~2017.7.31)

항공사의 피해



고객의 피해

다양한 변수에 따른
정확한 항공 지연 예측 모델을 통해 사전에 방지

기본 제공 데이터 소개 (1)

AFSNT → 2017. 1. 1 ~ 2019. 6. 30
항공편에 대한 Data

AFSNT_DLY → 2019. 9. 16 ~ 2019. 9. 30
예측해야 할 Data

SFSNT → 2017. 7 ~ 2019. 9
운항시즌에 대한 Data

기본 제공 데이터 소개 (1)

AFSNT → 2017. 1. 1 ~ 2019. 6. 30
항공편에 대한 Data

AFSNT_DLY → 2019. 9. 16 ~ 2019. 9. 30
예측해야 할 Data

SFSNT → 2017. 7 ~ 2019. 9
운항시즌에 대한 Data

기본 제공 데이터 소개 (2)

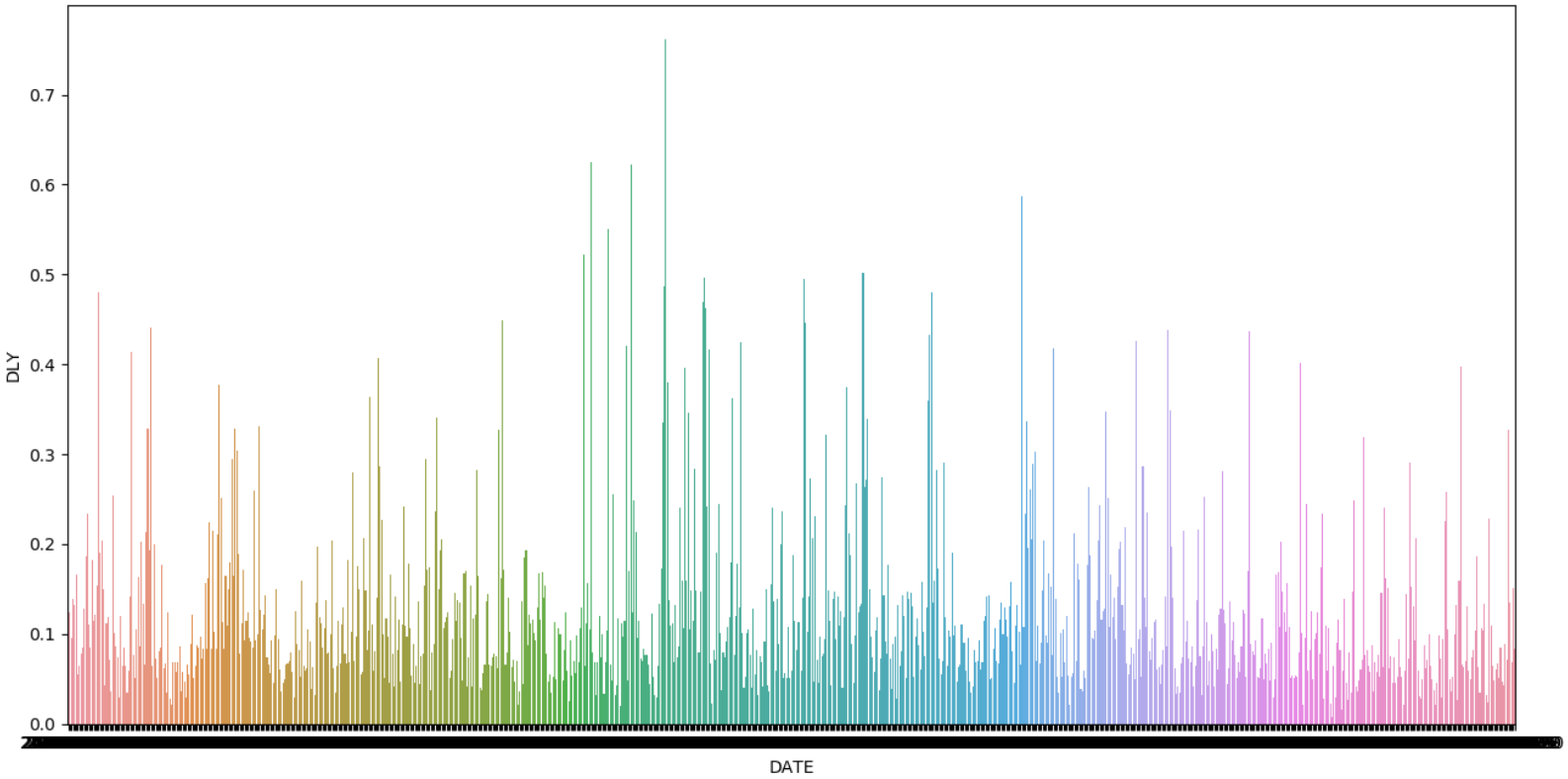
시간 데이터	
SDT_YY	연
SDT_MM	월
SDT_DD	일
SDT_DY	요일
STT	계획시각
ATT	실제시각

항공 데이터	
ARP	공항
ODP	상대공항
FLO	항공사
FLT	편명
REG	등록기호
AOD	출도착
IRR	부정기편
DLY	지연여부
DRR	지연사유
CNL	결항여부
CNR	결항사유

2

데이터 수집 & Feature Engineering

Raw data (1)



DLY : 지연여부

AFSNT에서 운항 지연율 12%

Class Imbalance Problem

Raw data (2)

IRR
- Y : 부정기편
- N : 정기편

→ 예측 대상은 스케줄이 미리 짜여진 '정기편'
즉, 부정기편 row 삭제

CNL/R
- CNL : 결항여부
- CLR : 결항사유

→ 결항된 항공편은 대회문제에서 제외

REG : 등록기호

→ 예측 데이터에 등록기호가 제시되지 않아
등록기호 col 삭제

Raw data (3)

DRR : 지연사유 → 예측 데이터에서 지연사유를 확인불가
지연사유 col 삭제

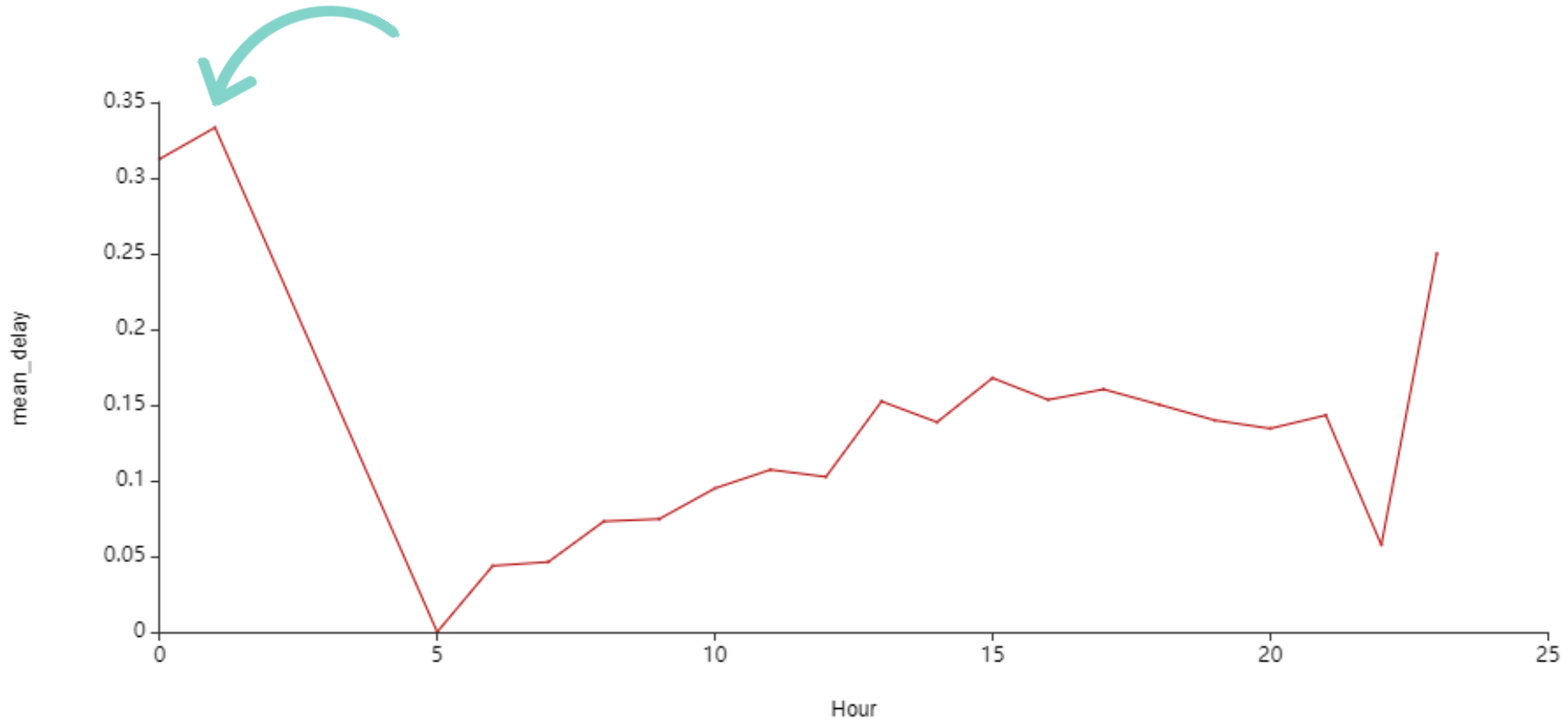
코드(설명)	지연 원인(%)
C02 (연결 지연)	90.58
A (기상 지연)	2.55
B (공항 지연)	0.46
D (항로 지연)	0.82
Z (기타 지연)	0.58

연결 지연을 비롯한

여러 지연사유를 보완할

추가 변수 필요

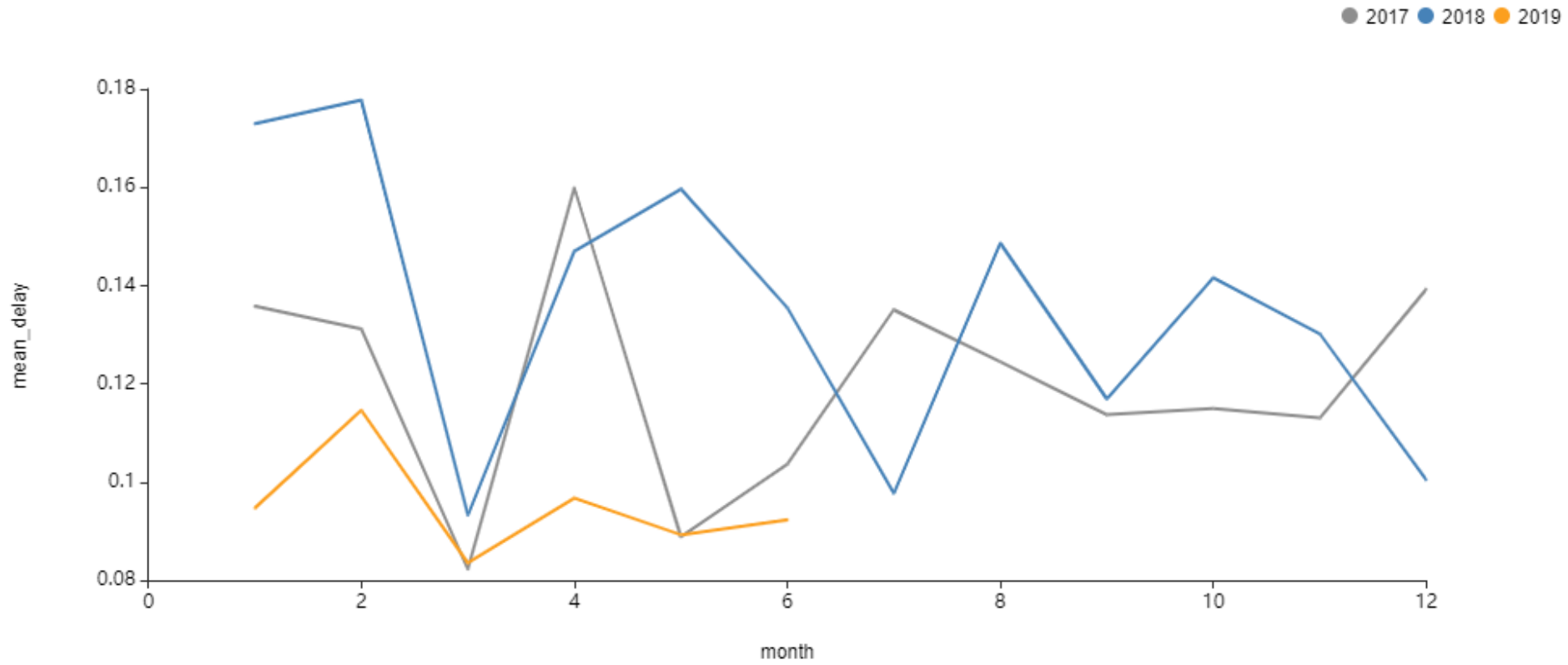
Raw data 시각화 (1)



시간별로 평균 지연율에 차이가 있음

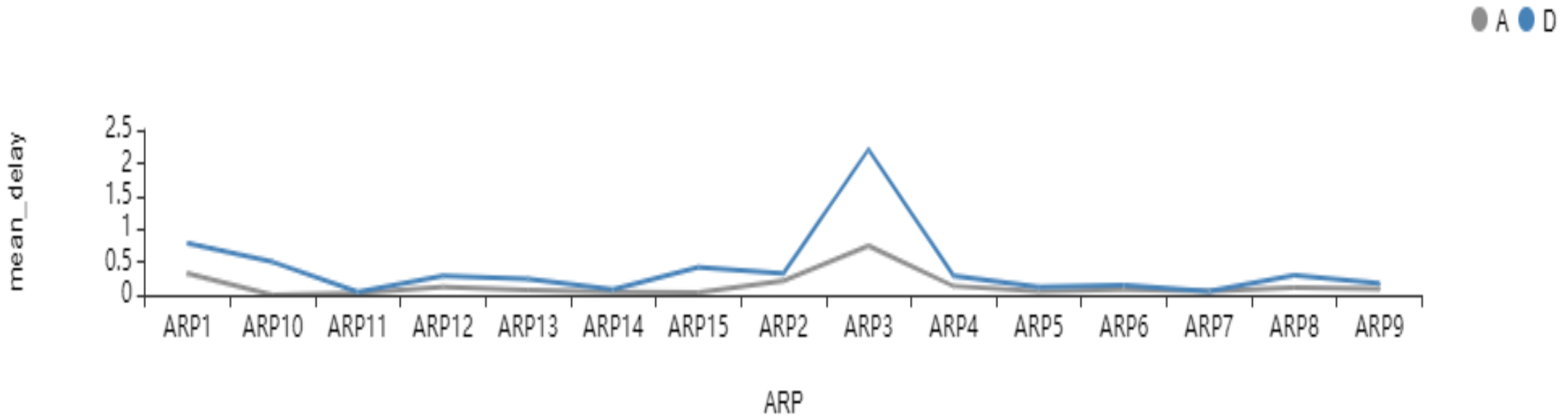
특히 **새벽시간**에 높은 지연율을 보임

Raw data 시각화 (2)

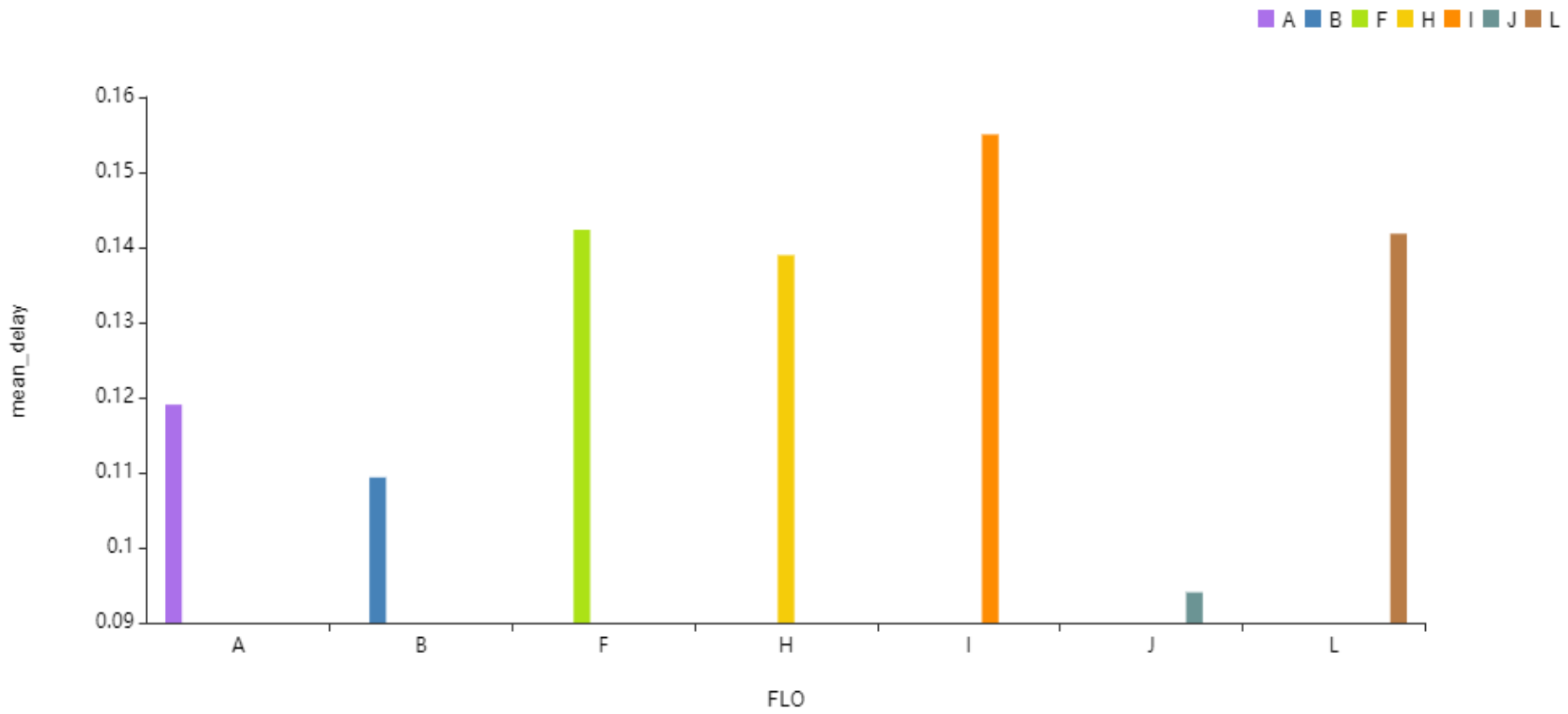


예측대상인 2019년의 평균 지연율이 전년대비 낮음

Raw data 시각화 (3)

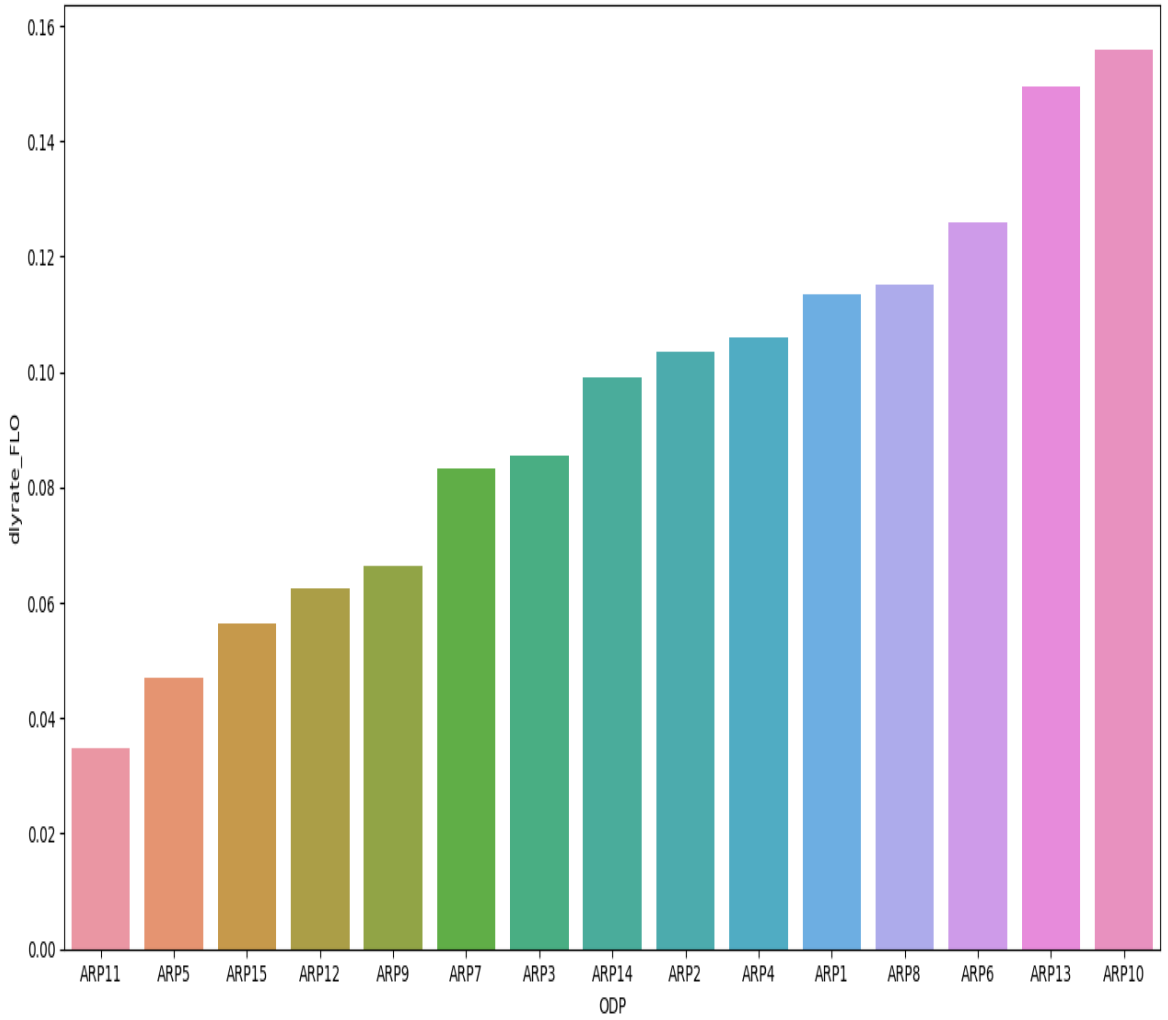


제주공항에서 출발하는 경우가 **다른 경우와 비교했을 때 상대적으로 평균지연율이 높음**

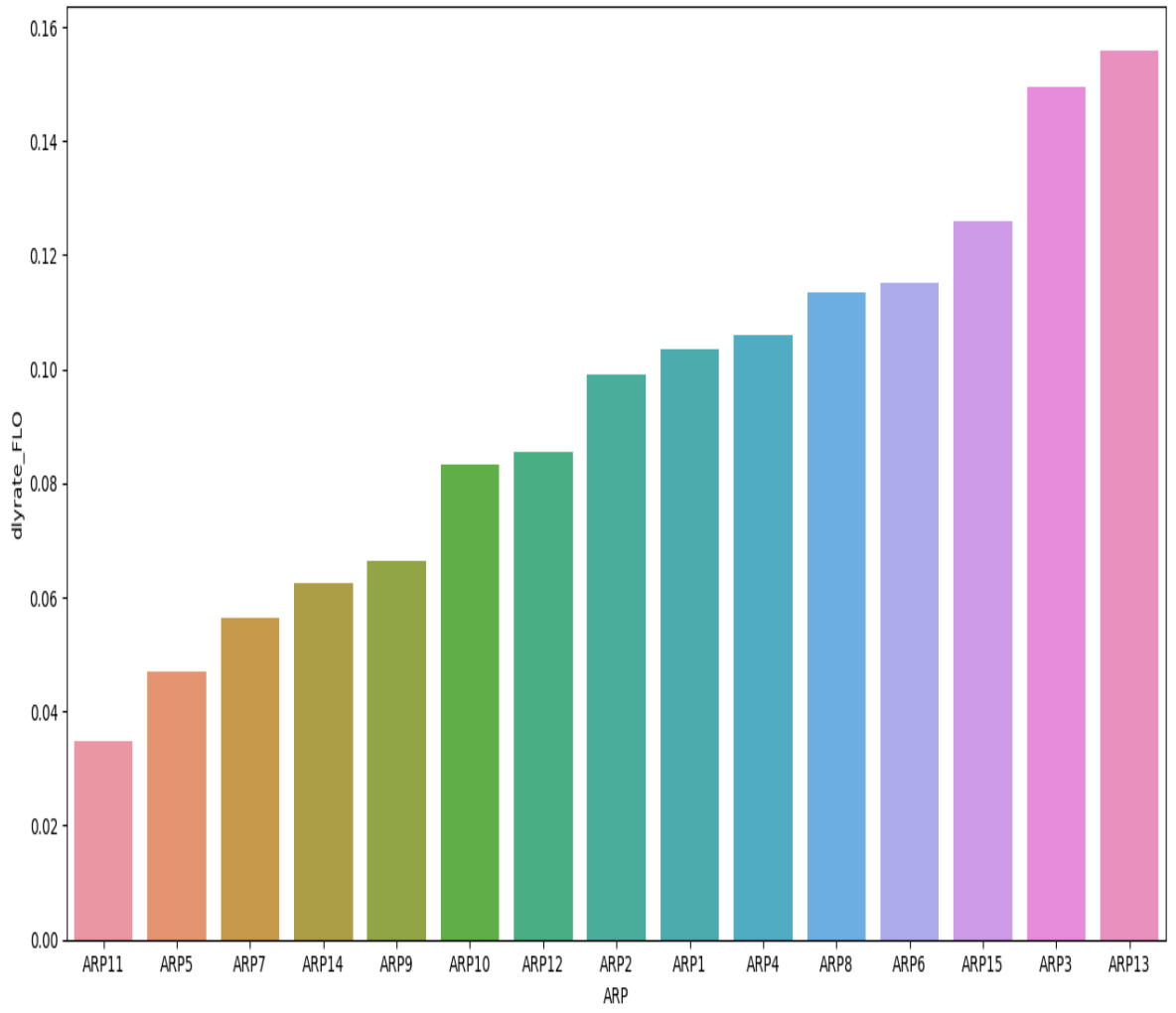


항공사 별 지연율

Raw data 시각화 (5)



도착공항 별 지연율



출발공항 별 지연율

Preprocessing (1)

한국 공항공사 운항스케줄을 참고하여
비식별화 처리된 공항 유추

<출처 : 한국공항항공사 공공데이터 개방>

ARP1 → 김포
ARP2 → 김해
ARP3 → 제주
⋮ ⋮
ARP15 → 인천

Preprocessing (2)

STT

분 절삭 후 이산형 Data로 간주

SDT_YY
SDT_MM
SDT_DY

One-Hot Encoding

AOD
ARP
ODP
FLO

범주/이산형 Data → 수치형 Data

데이터 수집 - 공항기후 (1)

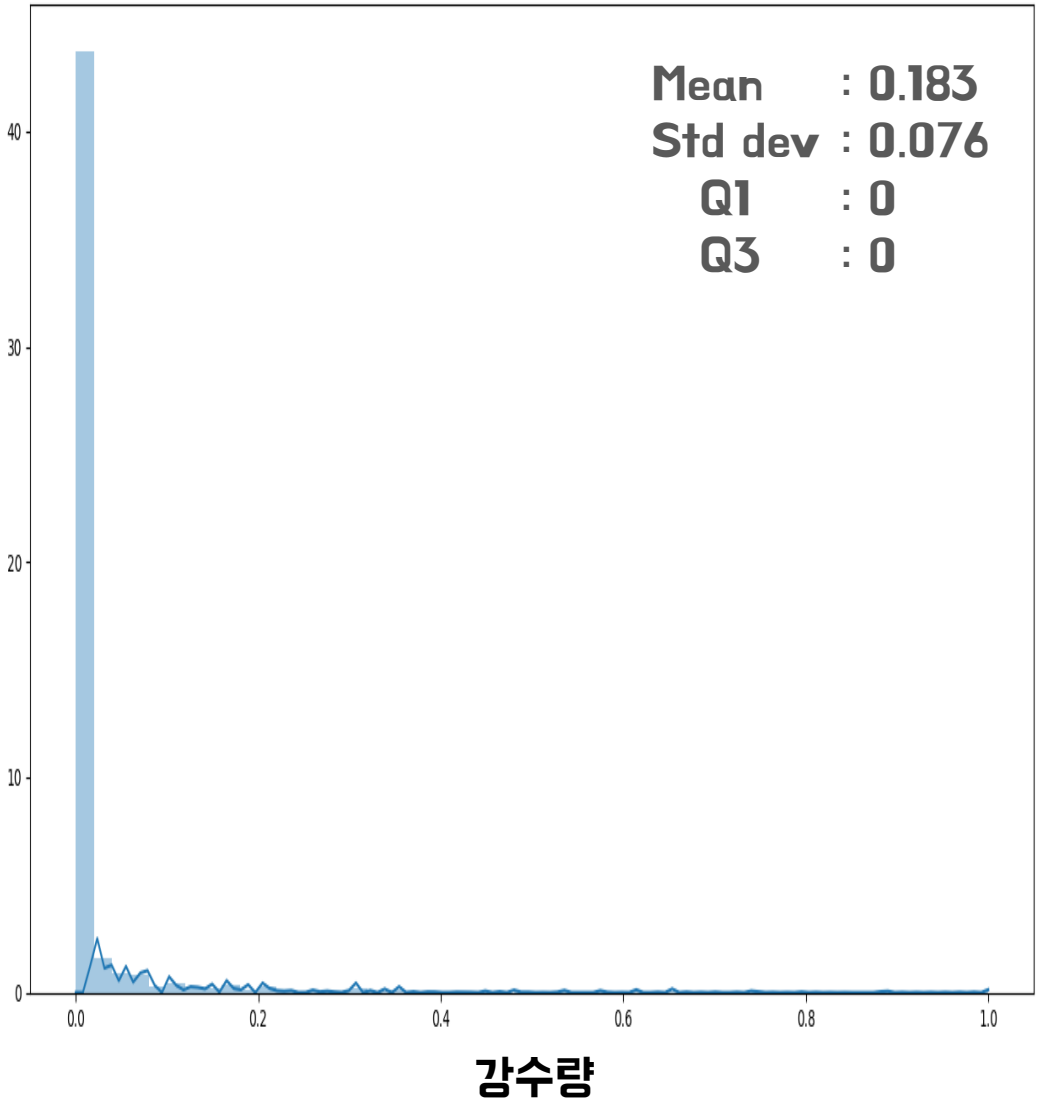
항공운항지원 기상서비스에서
공항 별 공항기후데이터 수집

연결지연의 원인이
기상지연일 확률이 가장 높음

특정 공항의 데이터가 없는 경우
가까운 공항을 기준으로 값 대입

2019. 9월 데이터는
2016~2018의
각 9월 데이터 평균으로 대체

데이터 수집 - 공항기후 (2)



예측할 9월의 날씨데이터가

이상치에 가까운 값이 예보되면

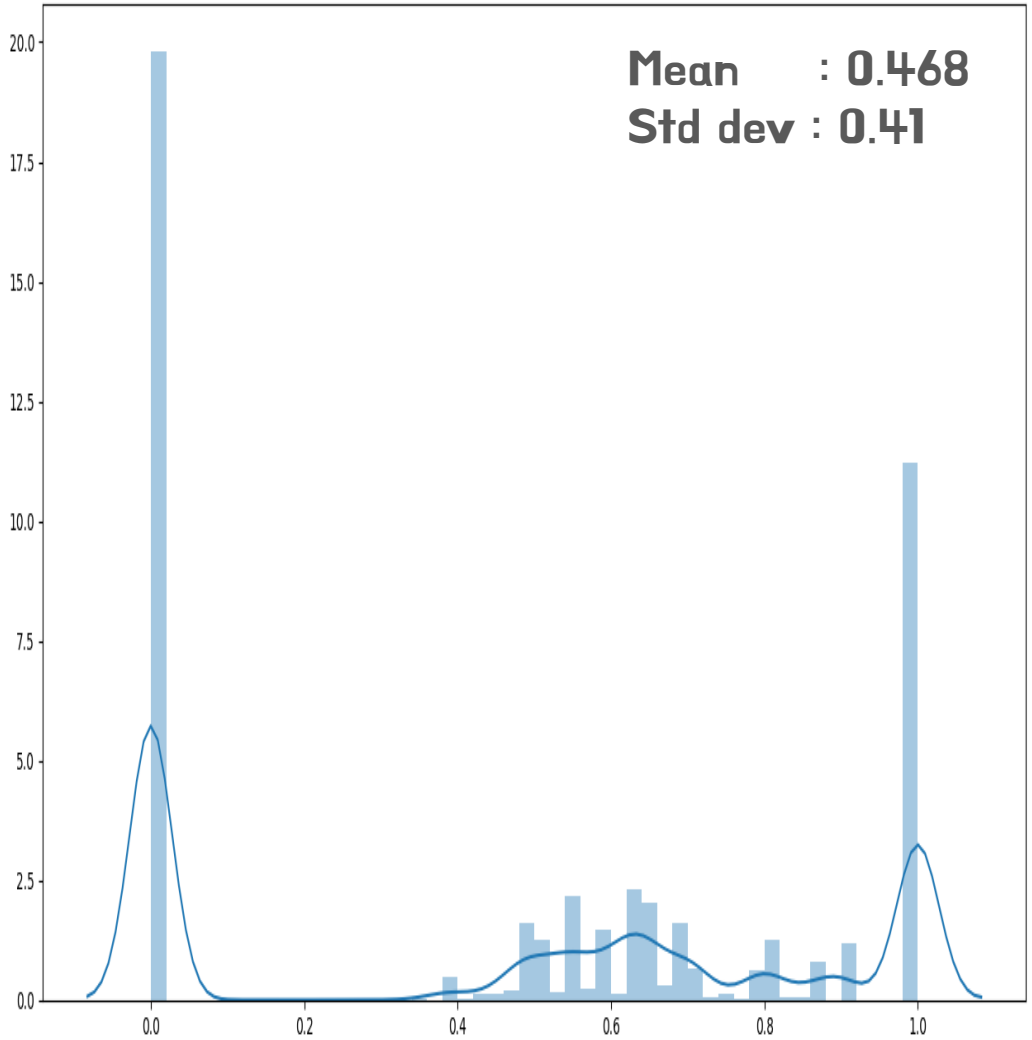
예측성능이 낮아지는 것을

방지하기 위해,

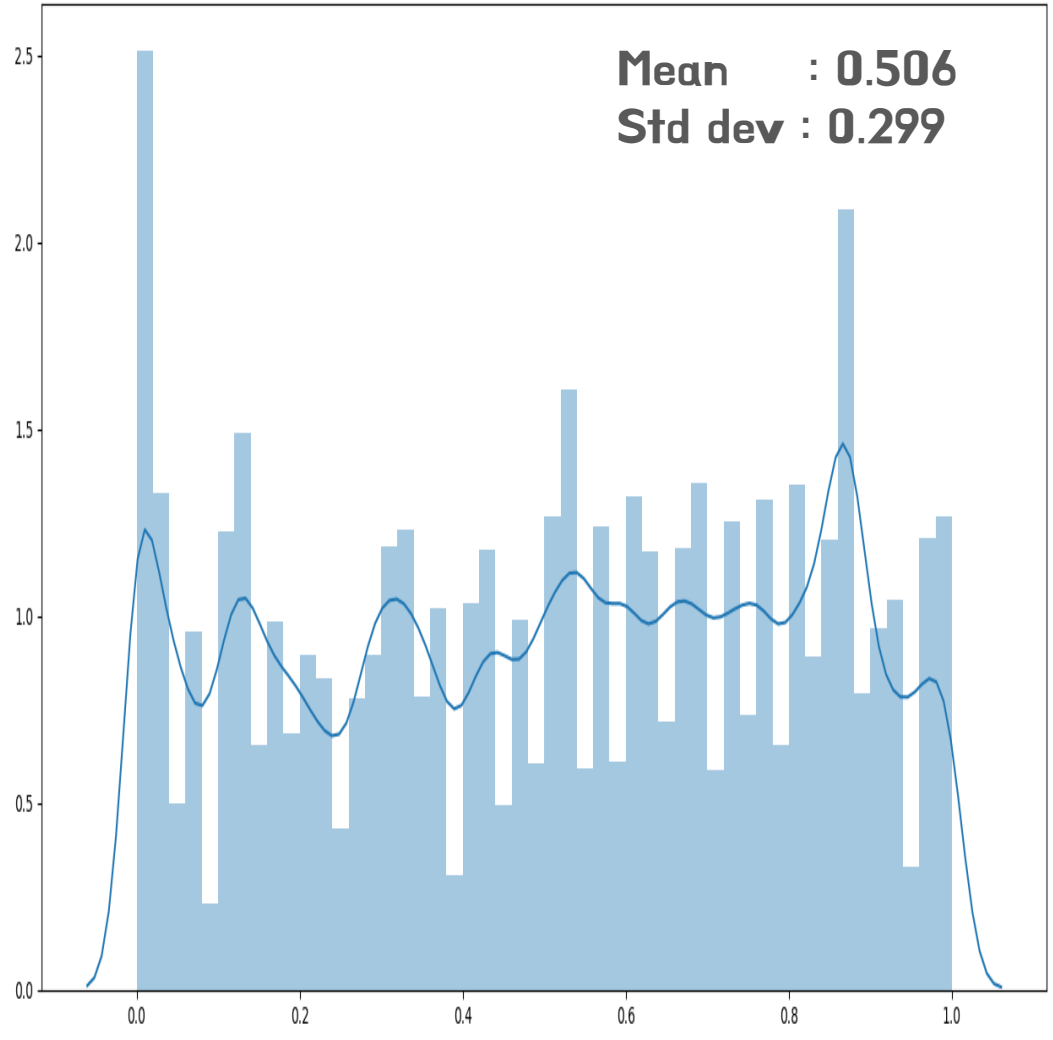
날씨변수 중에서

분산이 큰 변수 3개 제거

데이터 수집 - 공항기후 (3)

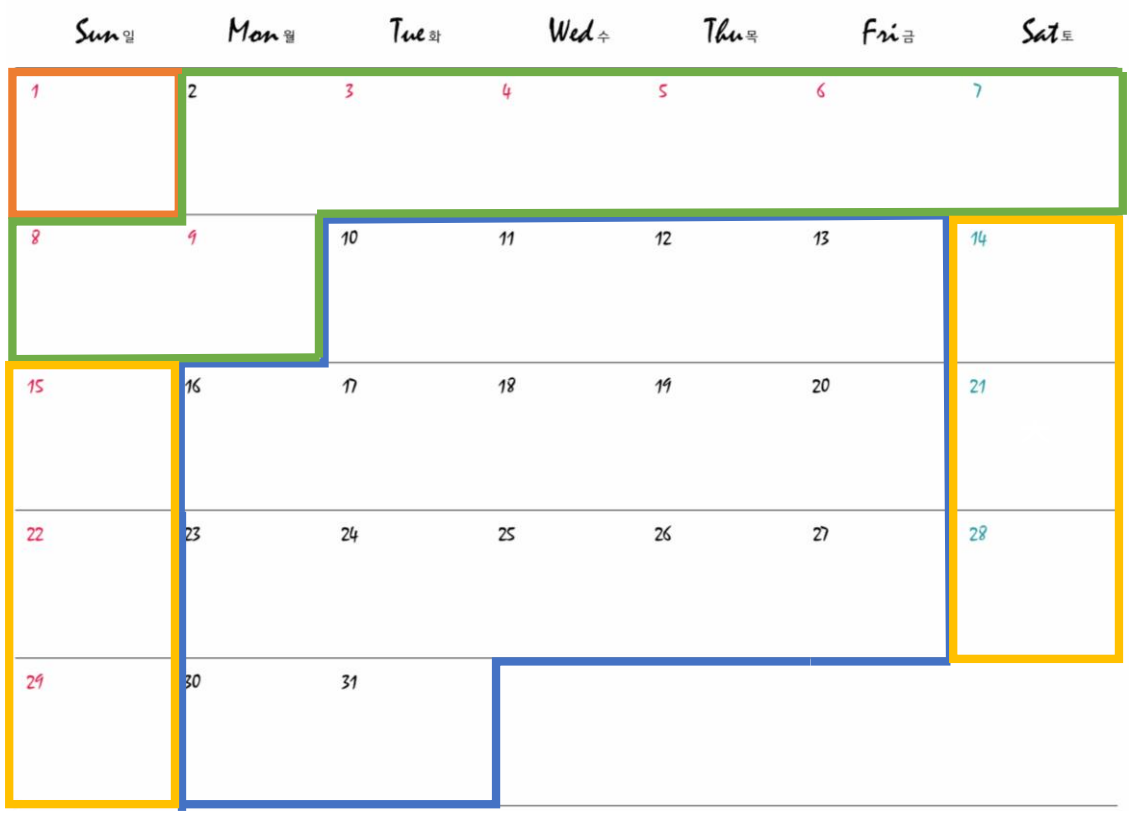


RVR



평균 운량

데이터 수집 - 휴일



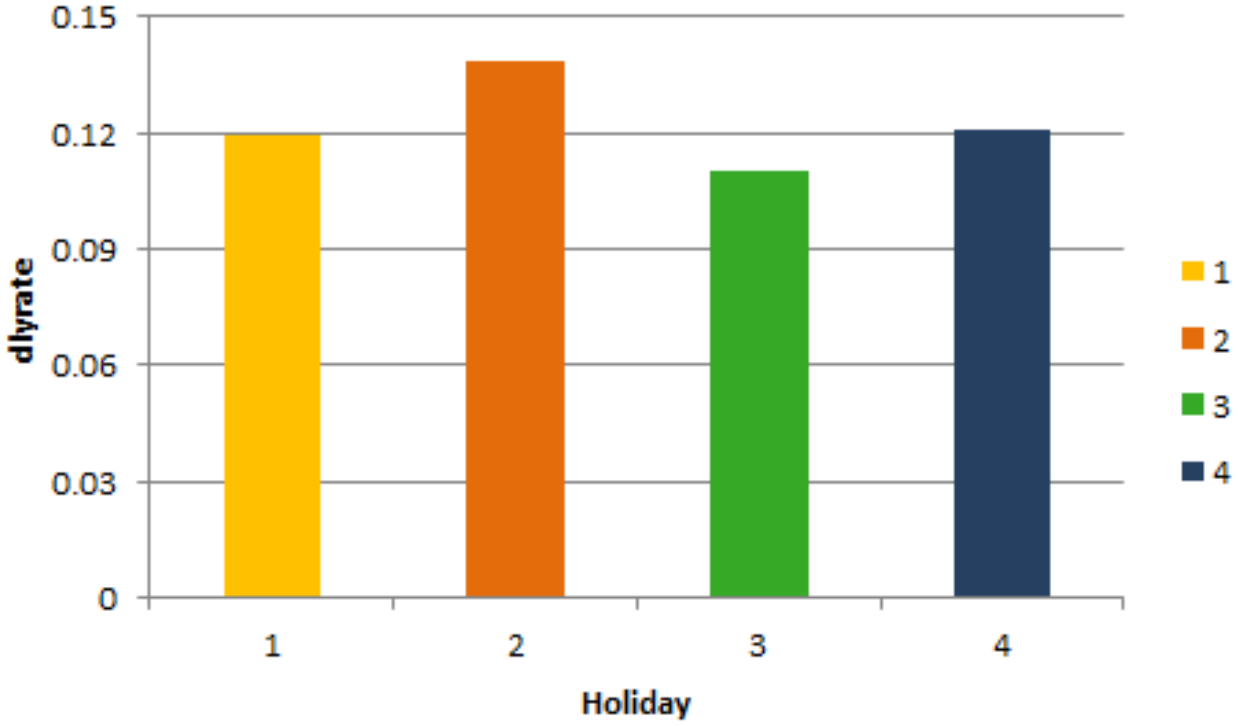
<2017. 10>

주말 / 공휴일 : 1

공휴일 전날 : 2

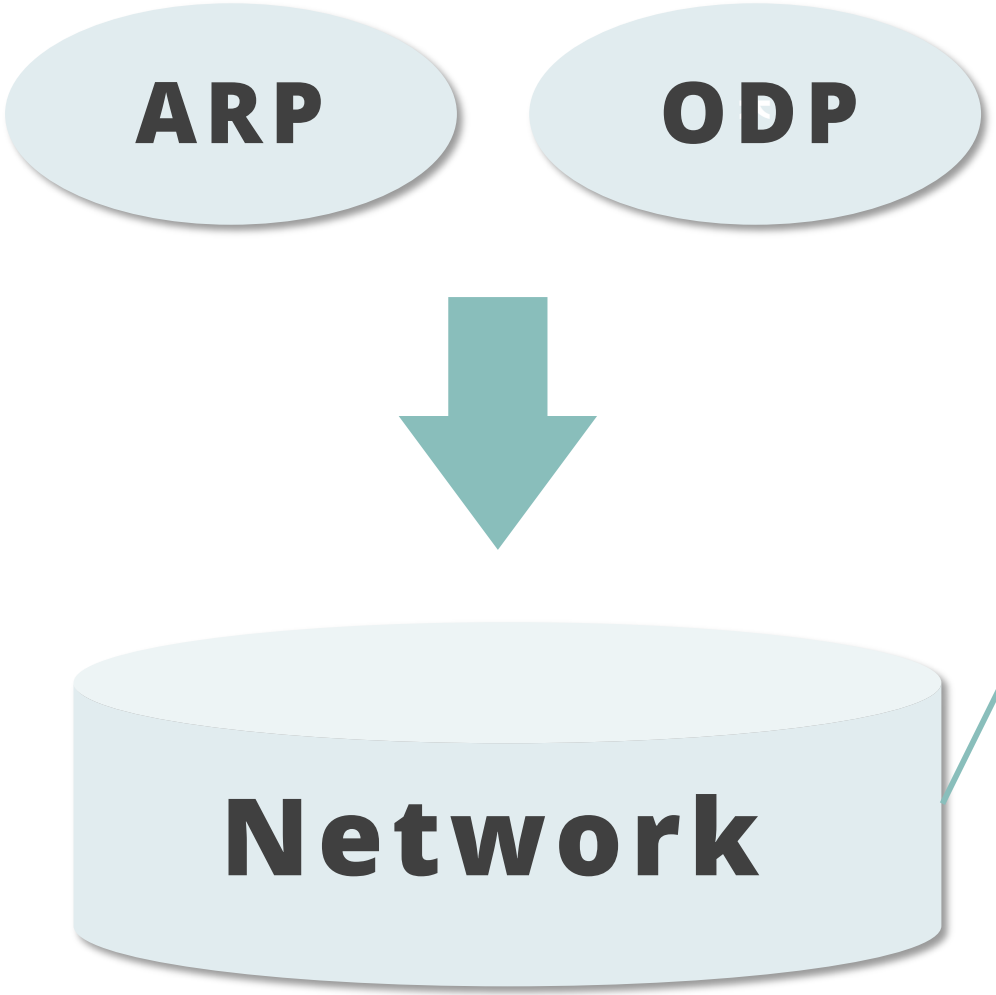
연휴 : 3

평일 : 4



예상과 달리 휴일 범주에 따라 지연율의 차이가 없으므로 변수로 채택하지 않음

파생변수 - Network Analysis



Degree Centrality

한 공항에 연결된 다른 공항의 개수

Edge

두 공항 사이의 edge 개수

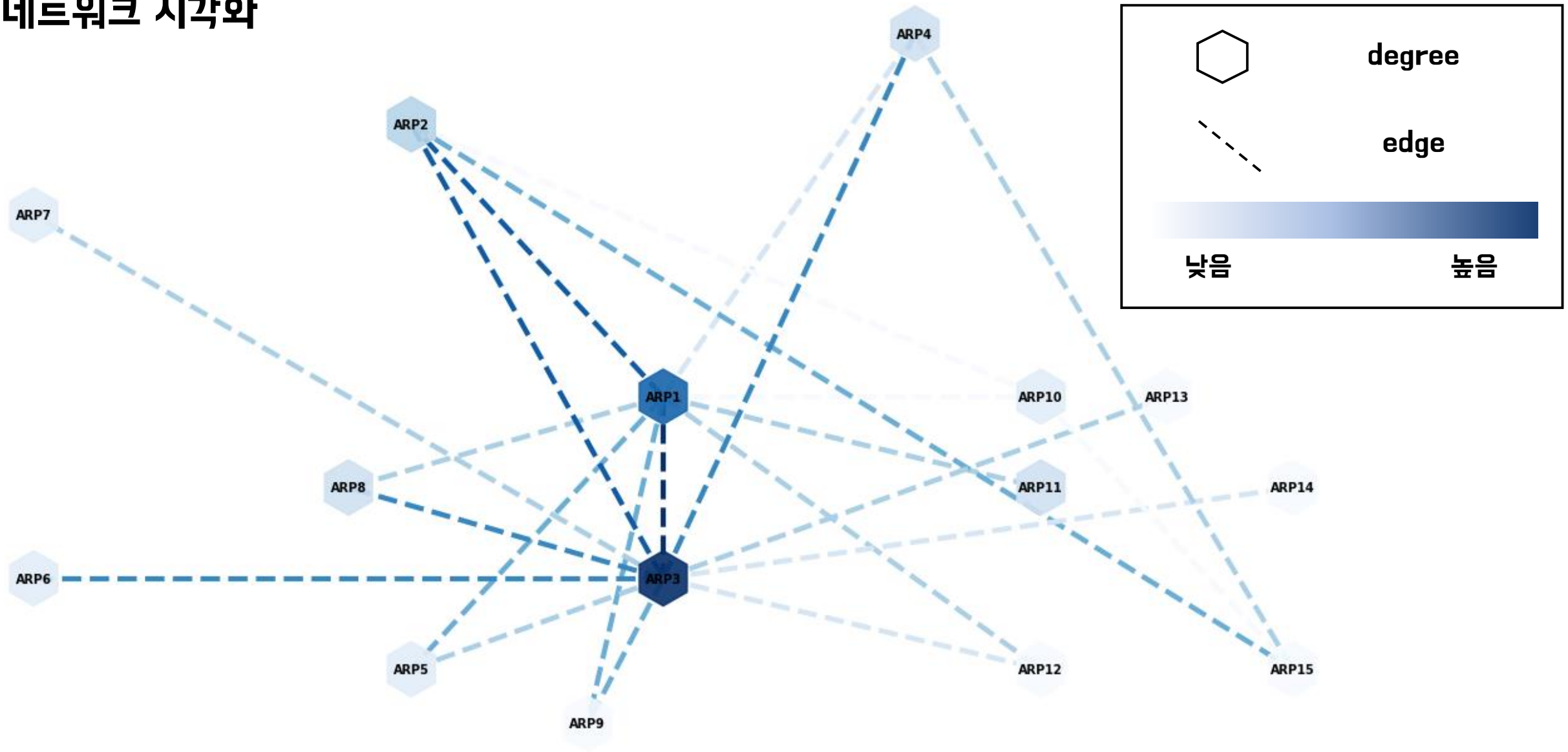
Betweenness Centrality

A 공항을 제외한 공항에서 다른 공항으로 최단거리로 이동할 때 A공항을 거쳐가는 비율

ex) 제주공항의 경우 대부분의 공항과 연결되어 있어 이 변수가 크다

높은 공항혼잡도

네트워크 시각화


ARP4
ARP2
ARP7
ARP1
ARP10
ARP13
ARP8
ARP11
ARP14
ARP6
ARP3
ARP5
ARP12
ARP15
ARP9
degree
edge
낮음
높음

파생변수 Accumulate (1)

SDT_YY	SDT_MM	SDT_DD	SDT_DY	ARP	ODP	FLO	FLT	REG	AOD	IRR	STT	ATT	DLY	DRR
2017	1	1	일	ARP1	ARP3	L	L1712	SEw4MjY4	A	N	12:50	12:47	N	
2017	1	1	일	ARP1	ARP3	L	L1711	SEw4MjM	D	N	12:55	13:16	N	
2017	1	1	일	ARP1	ARP3	L	L1717	SEw4MjY4	D	N	13:20	13:43	N	
2017	1	1	일	ARP1	ARP3	L	L1764	SEw4MDg	A	N	13:40	15:07	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1765	SEw4MDg	D	N	14:20	15:52	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1710	SEw4MDY	A	N	14:40	15:42	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1715	SEw4MDY	D	N	15:15	16:40	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1718	SEw4MjY4	A	N	16:30	16:40	N	
2017	1	1	일	ARP1	ARP3	L	L1766	SEw4MDg	A	N	17:25	18:20	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1723	SEw4MjY4	D	N	17:25	17:38	N	
2017	1	1	일	ARP1	ARP3	L	L1767	SEw4MDg	D	N	18:05	19:06	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1716	SEw4MjM	A	N	18:10	19:01	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1721	SEw4MjM	D	N	18:45	19:49	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1747	SEw4MjM	D	N	19:05	19:25	N	
2017	1	1	일	ARP1	ARP3	L	L1768	SEw4MDg	A	N	21:05	21:45	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1722	SEw4MjM	A	N	21:40	22:14	Y	C02
2017	1	1	일	ARP1	ARP3	L	L1748	SEw4MjM	A	N	22:15	22:43	N	
2017	1	1	일	ARP1	ARP3	L	L1724	SEw4MDY	A	N	22:20	22:49	N	
2017	1	1	일	ARP1	ARP5	A	A1603	SEw3Nzkw	D	N	19:15	19:27	N	
2017	1	1	일	ARP1	ARP5	A	A1604	SEw3Nzkw	A	N	21:35	21:36	N	

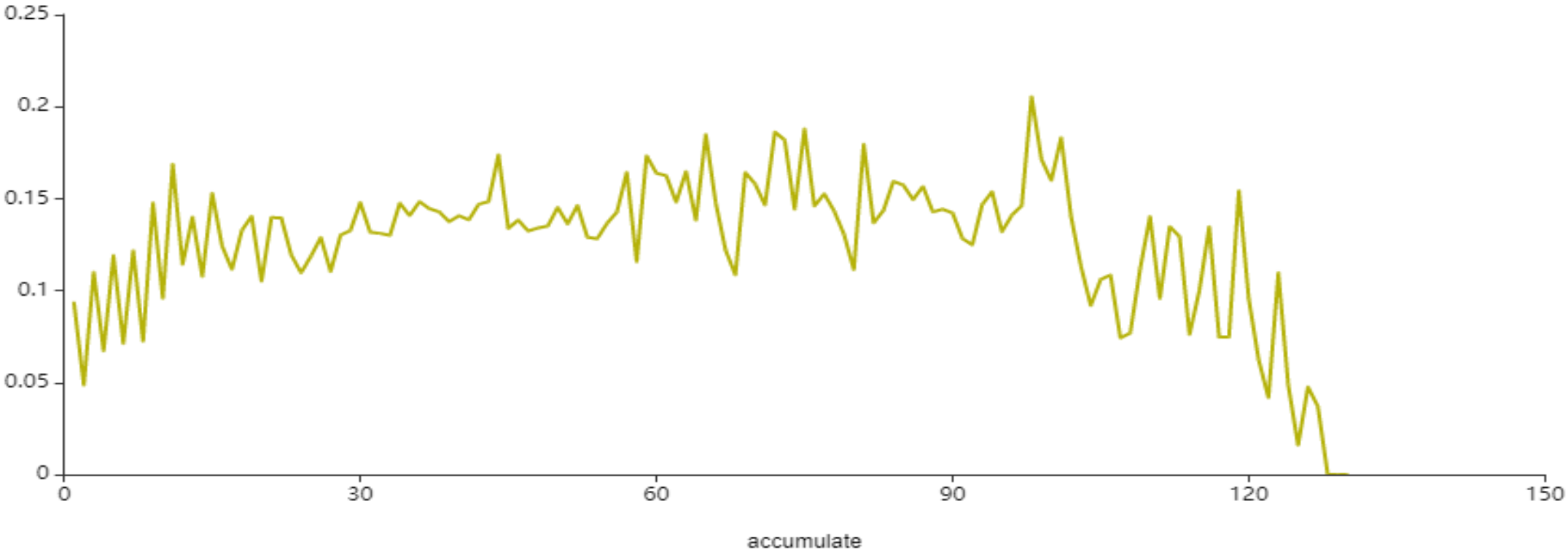
정렬 기준 순서

1. YY
2. MM
3. DD
4. ARP+ODP (ODP+ARP)
5. FLO
6. STT

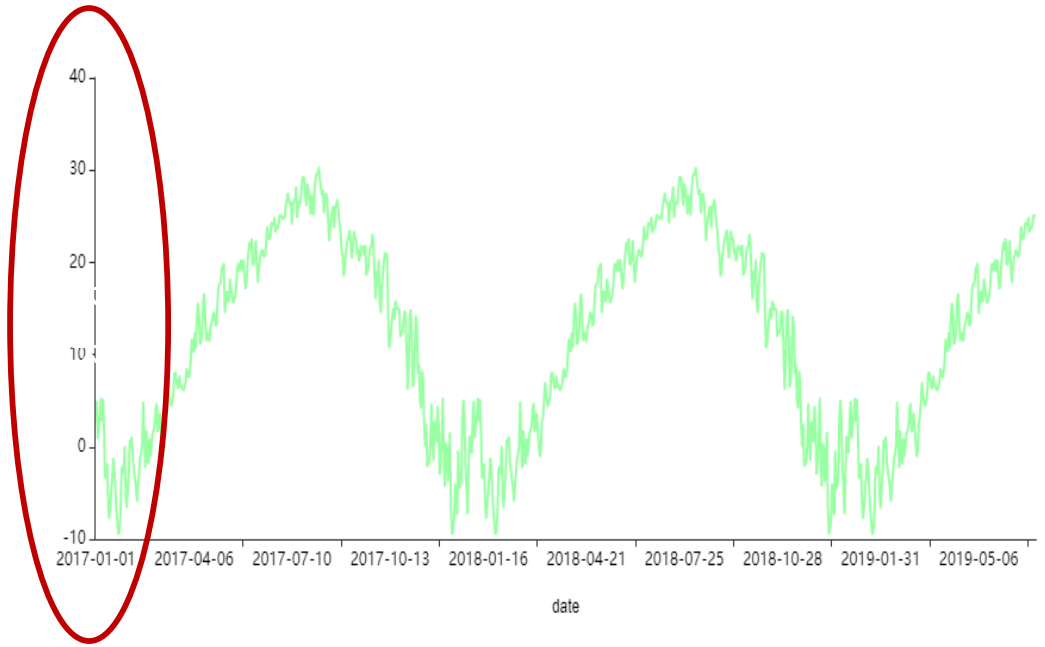
정렬 순서에 따라 데이터를 정렬하여 같은 날, 같은 항공사, 같은 편도가 연속되는 경우

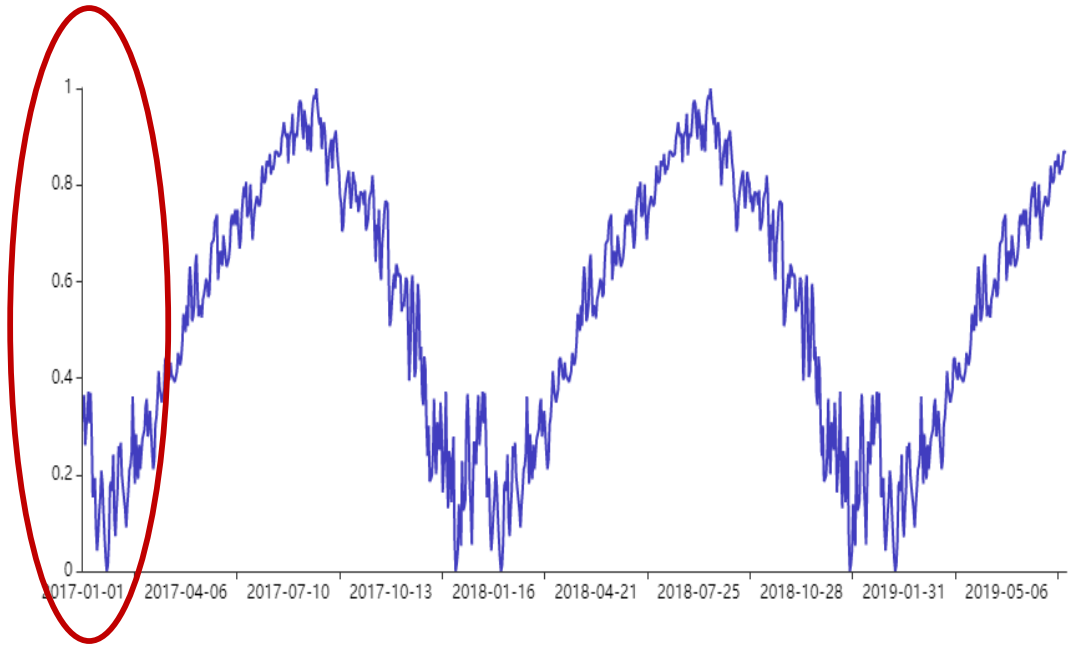
해당 편도의 시작 비행부터 카운트하여 변수를 생성

파생변수 Accumulate (2)



accumulate가 25까지 지연율이 계속해서 증가함
즉, 같은 편도에서 하루 누적 비행 횟수가 많아질수록
연결지연이 증가함을 유추할 수 있음





Min_Max Scaler를 이용하여

연속형 변수를 동일하게 0~1 의 범위로 스케일링

최종 feature 선택

기후 변수

Mean_wind
Min/Max/Mean_temp
Relative_humidity
Mean_sea
Visibility
Cloud_height

공항 변수

ARP/ODP
FLO
AOD
Accumulate

DLY

SNA 변수

Degree
Betweeness Centrality
edge

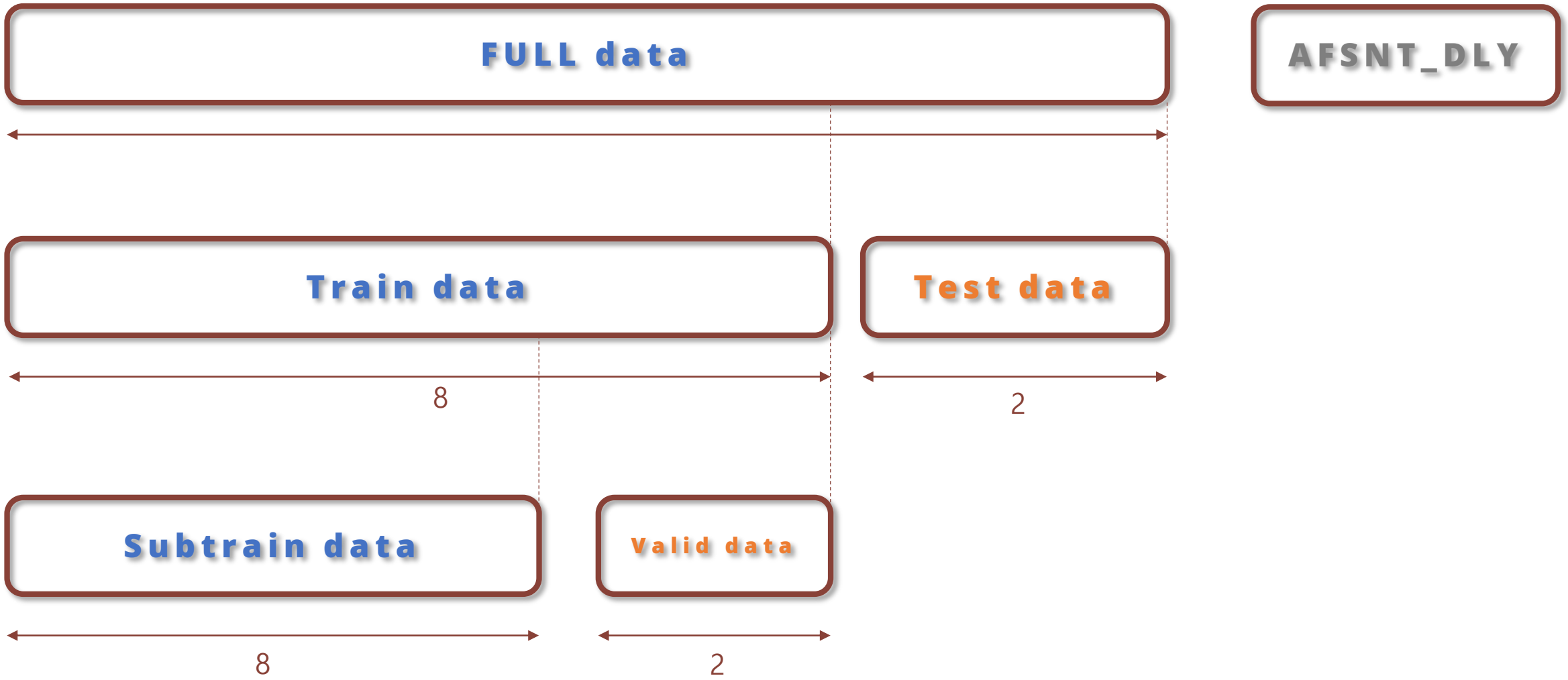
시간 변수

SDT_YY
SDT_MM
SDT_DY
STT

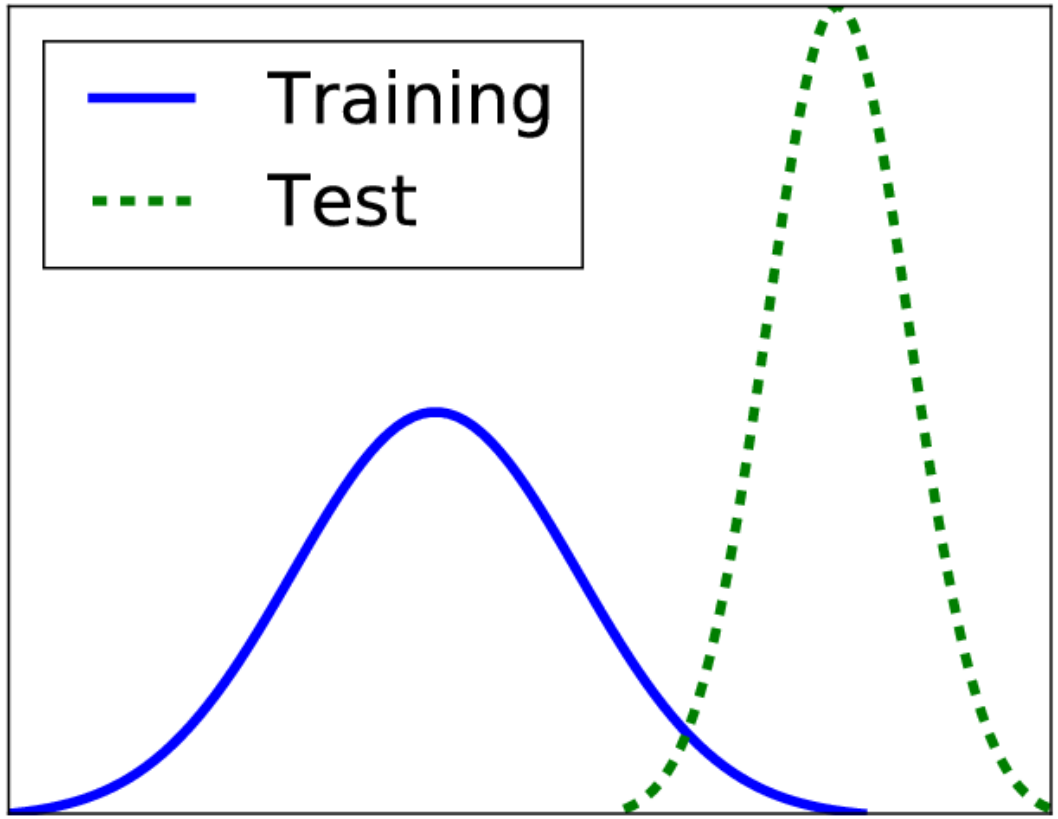
3

데이터 분석

3-way Hold out



Covariate Shift

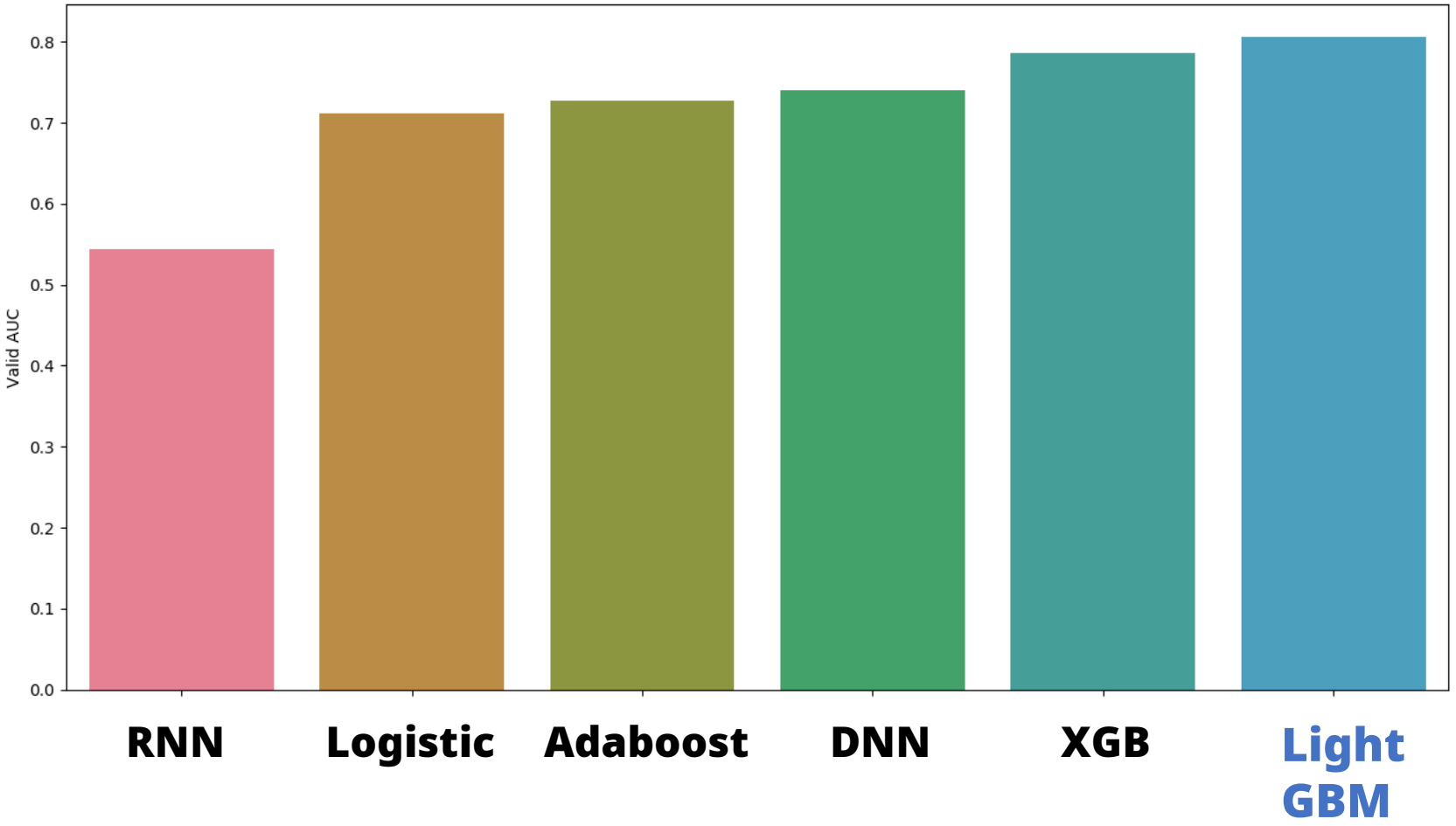


<Covariate Shift 문제 예시>

Train, Test 를 분류하는 알고리즘의

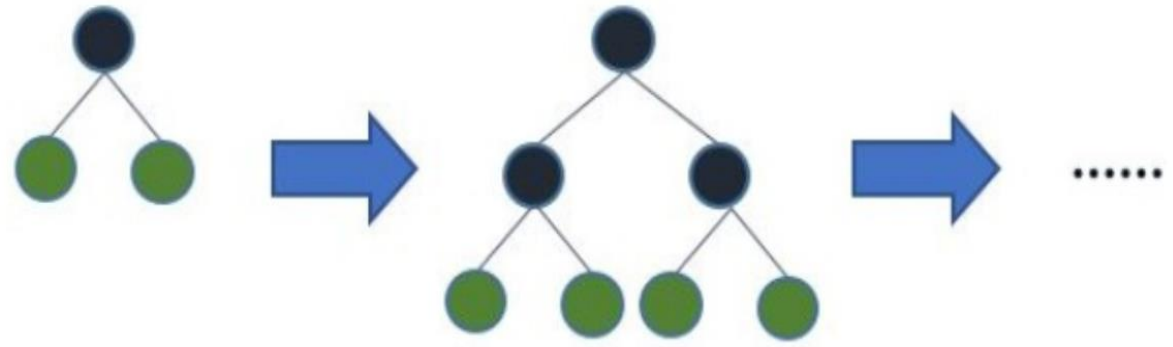
ROAUC 값이 0.5이므로 문제가 없다고 판단

최종 모델 선정


Valid AUC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
RNN
Logistic
Adaboost
DNN
XGB
Light GBM

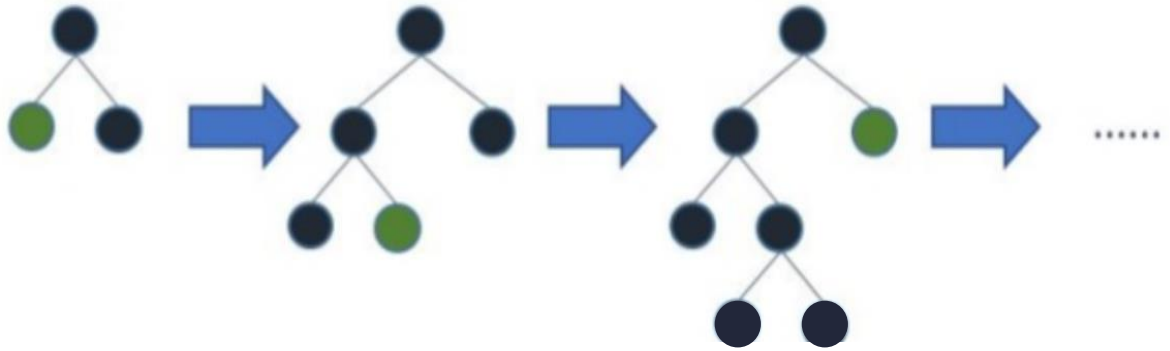
최종 모델 설명

XGBoost:

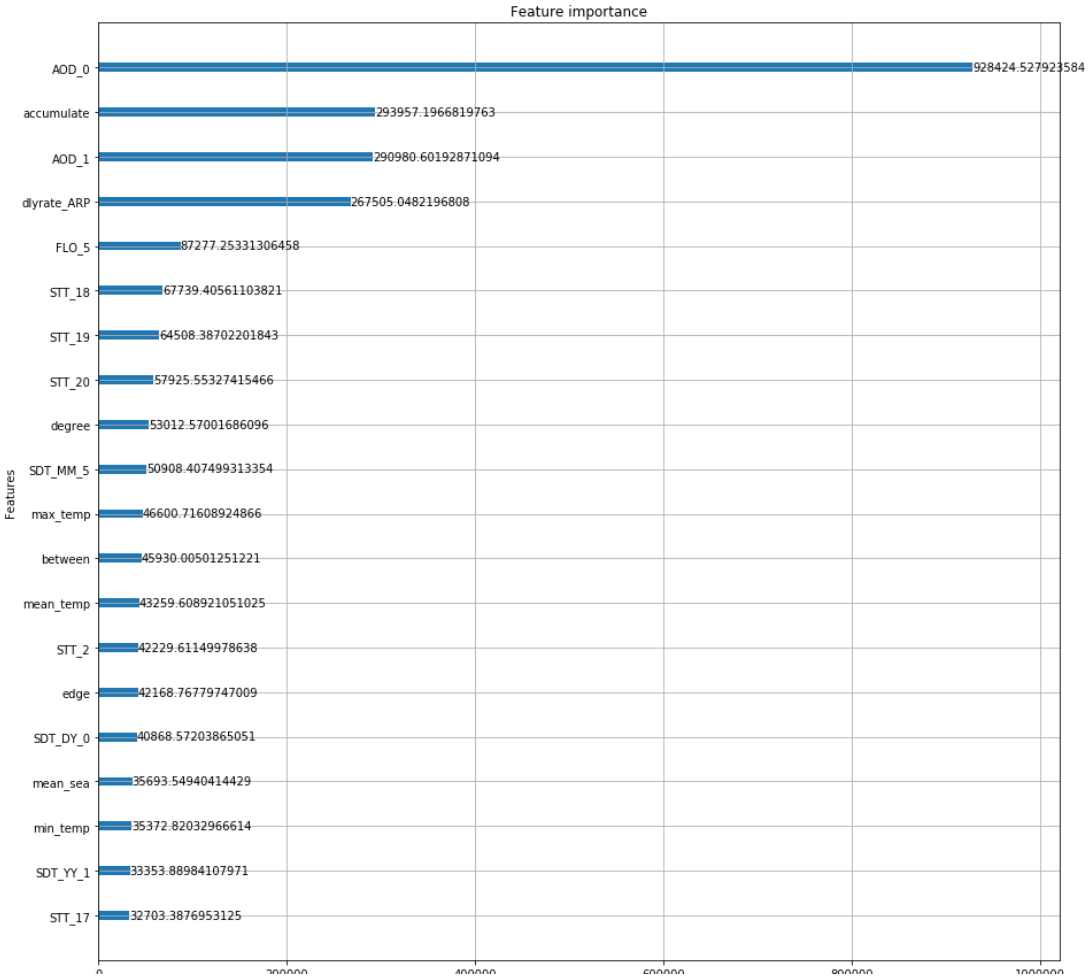


Level-wise tree growth

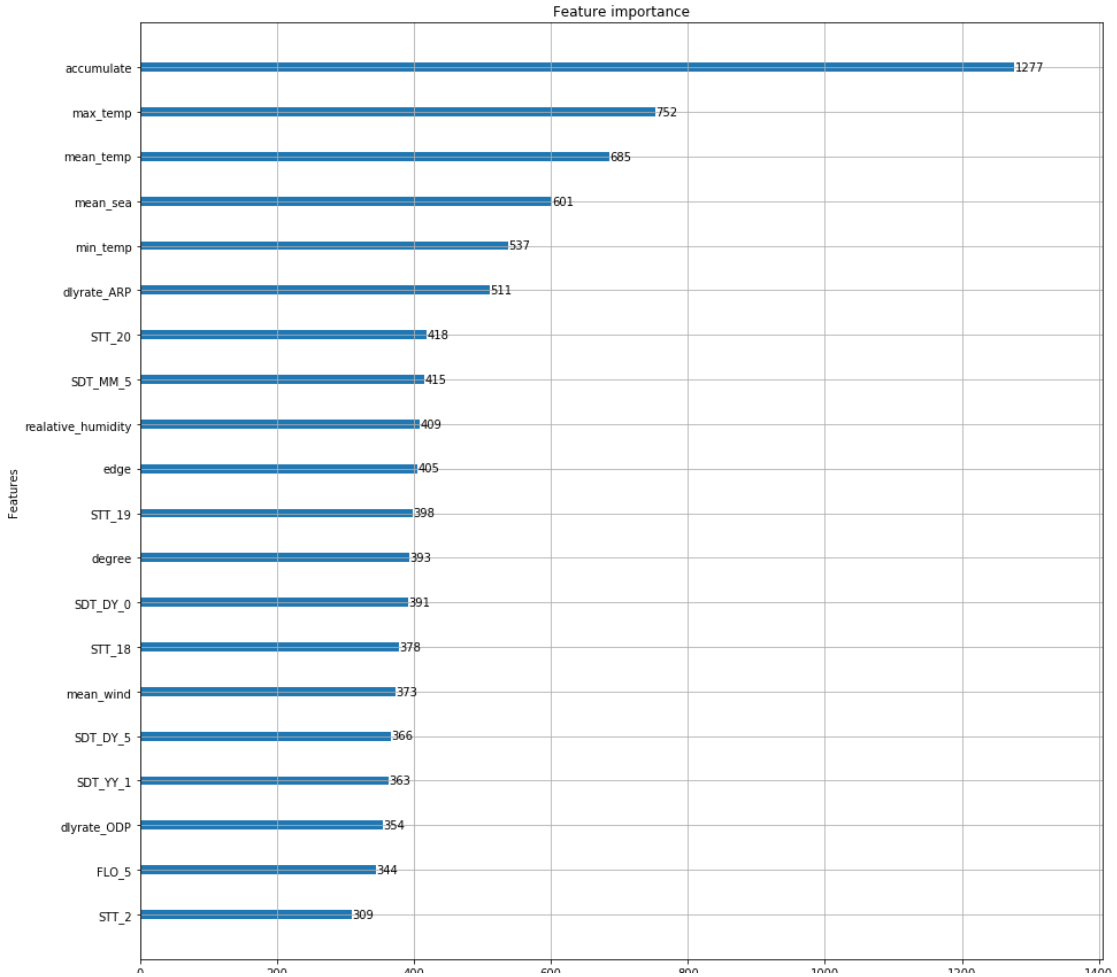
LightGBM:



Feature_Importance

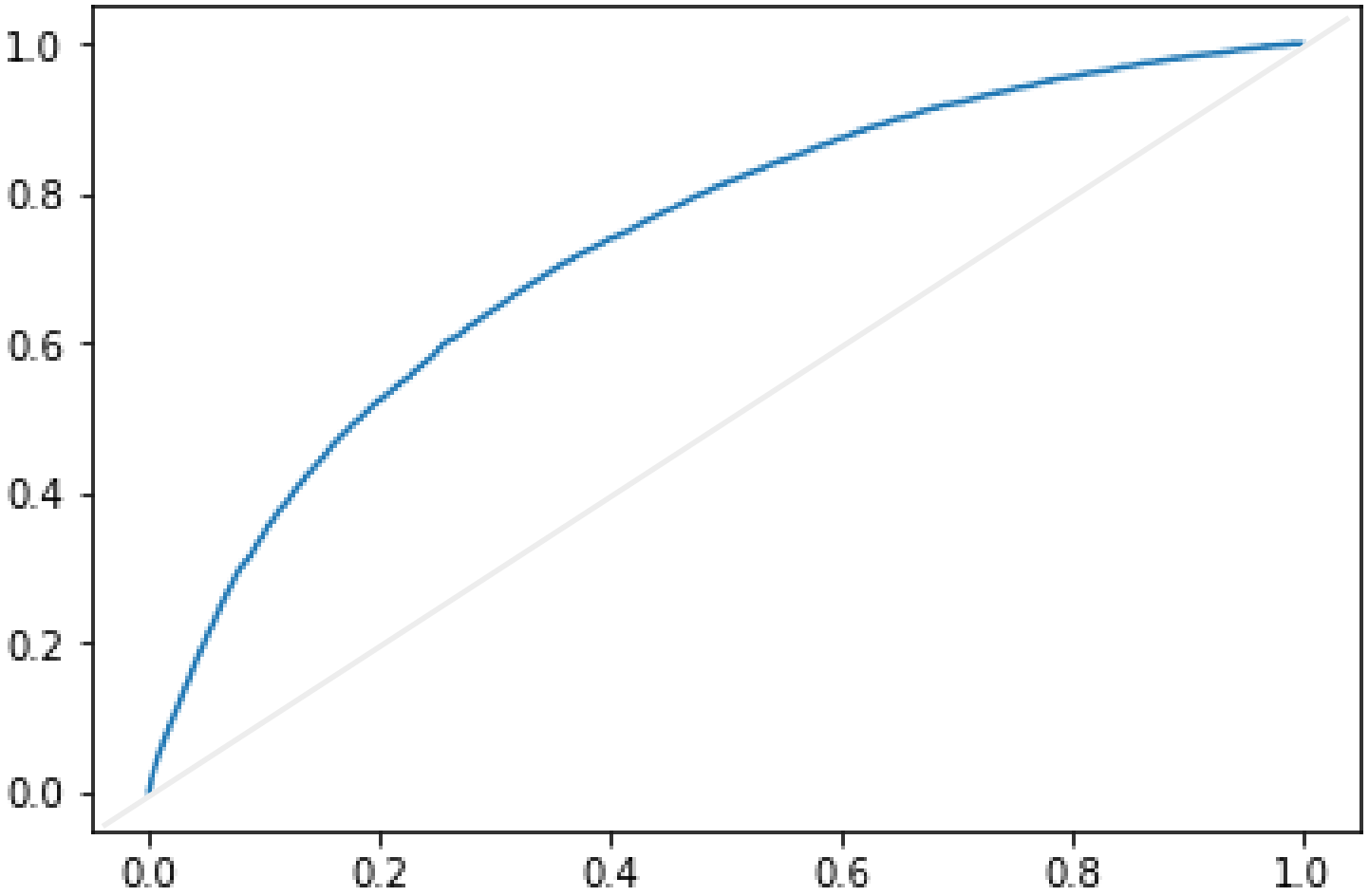


Gain



Weight

ROC Curve



예측확률로 계산한 Test set의 AUC : 0.734

Threshold에 따라 정확도 평가

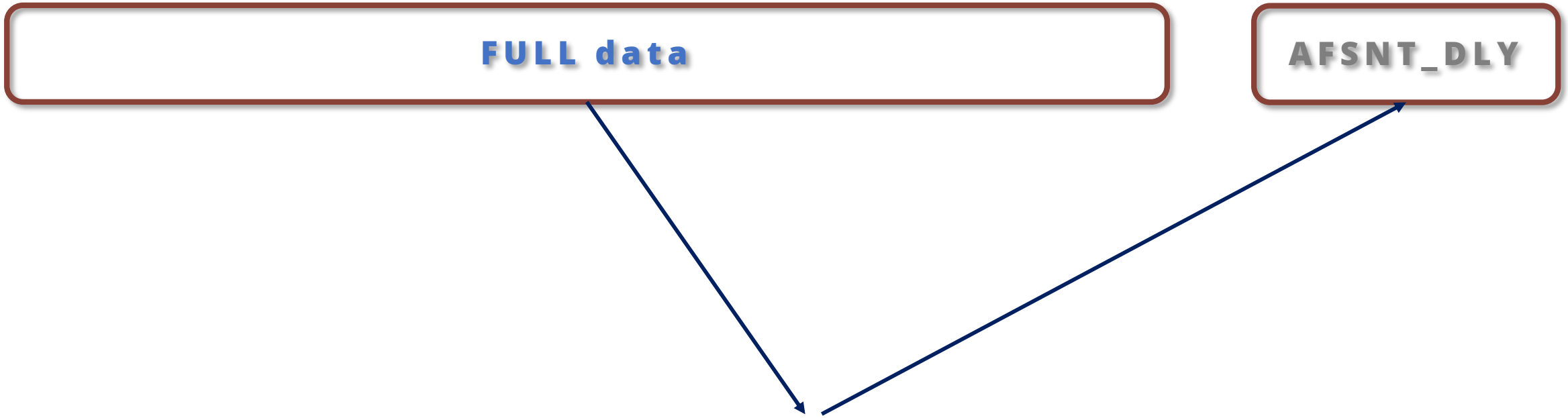
지표인 AUC를 가장 높일 수 있는

Best Threshold 를 찾기

4

예측
&
기대효과

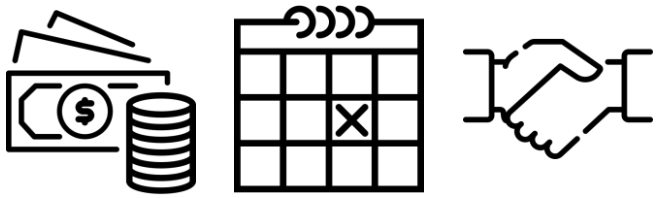
예측 모형의 기대효과



Best Threshold : 0.094125

예측 모형의 기대효과

항공사의 피해



예측모형의 활용으로
항공사는 **이미지상승** 및 지연으로 인한 손해 **예방**

고객의 피해

고객은 금전적/시간적 비용 **SAVE!**

사용 라이브러리

1. 전처리 및 시각화

SAMSUNG SDS
Brightics AI

matplotlib

Pandas

Seaborn

2. 지연율 예측 모델

Keras

LightGBM

dmlc
XGBoost

scikit
learn

TensorFlow

Q & A