

2022 기업 연계 프로젝트

시계열 데이터 기반 모델링 개발

이상 탐지 & 미래 예측

TEAM HunHun Jieun ho~!



이현호(멘토)



정지훈



정지은



임 훈



시계열 프로젝트

전체 프로젝트는 1차와 2차 두번의 프로젝트로 진행 됨

I. 1차 프로젝트

- 1차 프로젝트 개요
- 1차 프로젝트 결과

II. 2차 프로젝트 개요

- 프로젝트 배경
- 데이터 설명

III. 데이터 엔지니어링

- 파생 변수

IV. 모델링

- 모델 비교
- 모델 develop

V. 결론



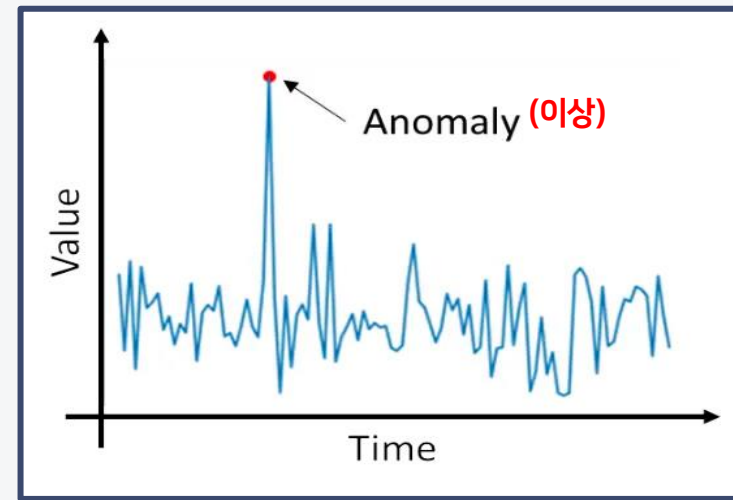
Index

I . 1차 프로젝트

- 1차 프로젝트 개요
- 1차 프로젝트 결과

1차 프로젝트 개요 전해 탈지 공정 중 품질 이상 탐지

- 주제 - 센서 데이터를 이용한 제품 이상 탐지
- 목적 - 생산제품의 이상을 생산 과정 중 조기에 탐지할 수 있는 모델 제작
- 기대효과 - 인건비 절약, 품질검사 비용과 시간 절감, 품질예측 가능 등
- 가설 설정 - 공정 변수들은 **기준치**를 벗어날 경우 품질 이상을 야기한다.



프로젝트 배경

전해탈지

: 세정 공정으로 오염물의 제거, 금속 표면 활성화

데이터

: 온도, 전류, pH, 시간 등 날짜, 공정 단위(Lot) 별 측정

공정 중 발생하는 문제

- 공정 중 발생한 이상으로 인해 불완전 세정 또는 열화 발생
- **전류, 시간, pH, 온도** 간의 복합적 상관관계를 고려한 공정 진행 필요

전처리

결측치 제거

: 시계열 데이터이기 때문에 선형 보간법 사용

이상치 제거

: 이상탐지 모델이므로 이상치 활용을 위해 처리 하지 않음

정규화

: Minmax, Standard, Robust

1차 프로젝트 모델링

파생변수

편차 - 각 변수 별 평균으로 부터의 편차

Lot 별 평균 - 공정 단위인 Lot 별 각 변수를 그룹화 한 평균

정상 여부

- 종속 변수 (Y)
- 정상 : 0
- 이상 : 1

이상 분포

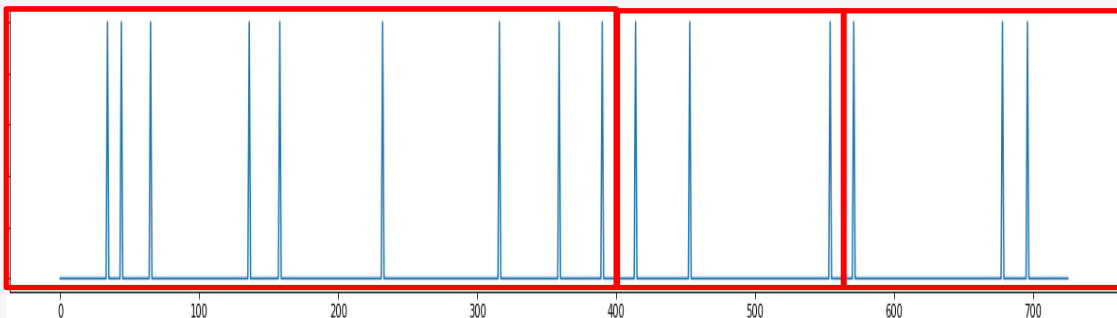
- 본래 값들이 이상 분포에 속할 경우 1,
- 정상 분포에 속할 경우 0

Train-valid-test 분할

Train

Valid

Test



Model 별 Score

score	Logistic regression	Isolation forest	Deep Learning
Acc	0.99	0.99	0.99
Precision	1.00	1.00	0.49
Recall	0.83	0.83	0.5
F1-score	0.9	0.9	0.5

머신 러닝 기반

딥 러닝 기반

결론

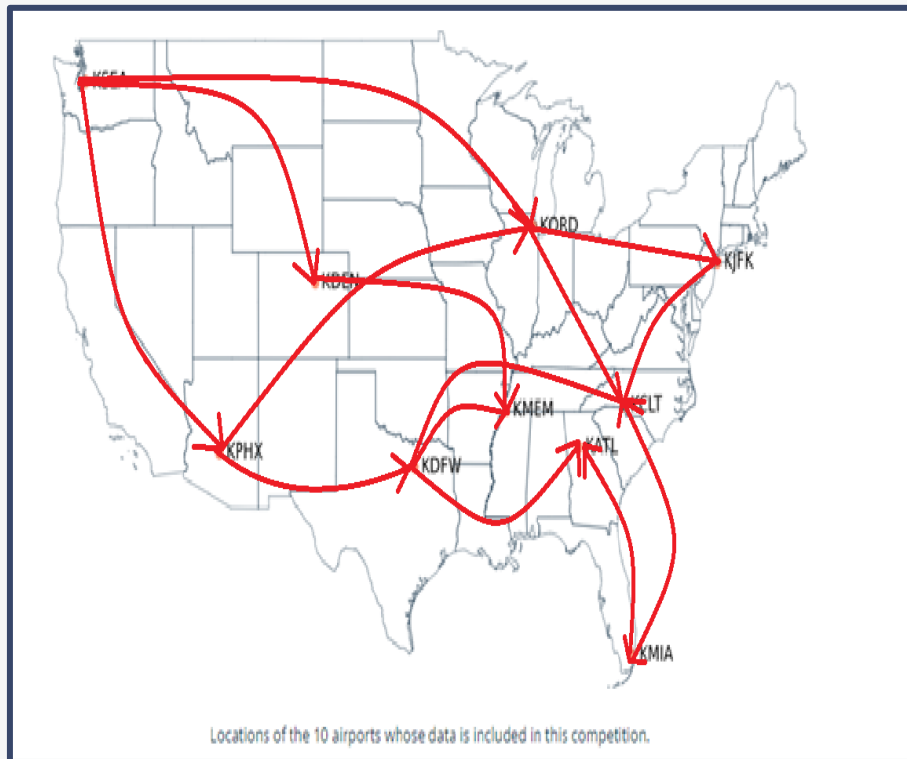
- 데이터가 많을 수록 유리한 딥러닝은 데이터가 적을 때 성능이 떨어지는 것으로 보임
- 머신러닝 기반 모델의 결과가 유의미하다고 판단
- 가설 기반 파생 변수를 사용한 모델이 유효하므로 가설이 유효하다고 판단

II . 2차 프로젝트 개요

- 프로젝터 배경
- 데이터 설명

프로젝트 배경

Run-way Functions: Predict Reconfigurations at US Airports (Open Arena) Funded by NASA



- 미국 'DRIVENDATA'에서 진행된 대회
- 주제 - 항공 시계열 데이터를 이용한 **Config** 예측
- 목적 - 항공 교통과 날씨의 실시간 데이터를 이용하여 활주로 구성 변경 사항 예측
- 기대효과 - 항공기 이착륙 지연이 최소화 되도록 활주로 구성 변경사항을 예측해
비용 절감, 에너지 절약, 영공 네트워크 지연 완화
- 예측 대상 - 10개의 공항 별 config
- 가설 - 시간, 요일, 기상 환경에 따라 활성화 되는 Config가 변화한다.

데이터 설명 – 대회에서 주어진 데이터 (약 100GB)

Tree 구조

[/content/kat1](#)

```

|   |-- kat1_airport_config.csv.bz2
|   |-- kat1_arrival_runway.csv.bz2
|   |-- kat1_departure_runway.csv.bz2
|   |-- kat1_etd.csv.bz2
|   |-- kat1_first_position.csv.bz2
|   |-- kat1_lamp.csv.bz2
|   |-- kat1_mfs_runway_arrival_time.csv.bz2
|   |-- kat1_mfs_runway_departure_time.csv.bz2
|   |-- kat1_mfs_stand_arrival_time.csv.bz2
|   |-- kat1_mfs_stand_departure_time.csv.bz2
|   |-- kat1_tbfm_scheduled_runway_arrival_time.csv.bz2
|   |-- kat1_tfm_estimated_runway_arrival_time.csv.bz2

```

파일명	데이터 내용
_Airport_config	활주로 구성
_Runway	시간 별 이착륙 시 사용되는 활주로
_First_position	예측시간 별 이착륙 시 사용되는 활주로
_Lamp	기상데이터 ex)온도, 풍향, 풍속, 강우, 낙뢰
_Runway_time	실제 이착륙시 활주로에 도달한 시간
_Stand_time	실제 이착륙시 게이트에 도달한 시간
_Scheduled_runway_time	예정된 이착륙 시간
_Estimated_runway_time	예측된 이착륙 시간

동일한 데이터 형태가 10개의 공항 별로 존재



세분화 된 활주로, 항공, 기상 데이터

데이터 설명 – configuration(예측 대상, 종속변수)

공항이름	Config 개수
katl : 26	
kclt : 13	
kden : 42	
kdfw : 31	
kjfk : 14	
kmem : 31	
kmia : 28	
kord : 38	
kphx : 18	
ksea : 12	

config	active
ksea:D_16C_A_16C_16R	0.0
ksea:D_16C_A_16L_16R	0.0
ksea:D_16L_A_16C	0.0
ksea:D_16L_A_16C_16L	0.0
ksea:D_16L_A_16C_16R	0.0
ksea:D_16L_A_16L_16R	0.0
ksea:D_16L_A_16R	0.0
ksea:D_34C_A_34C_34L	0.0
ksea:D_34R_A_34C	0.0
ksea:D_34R_A_34C_34R	0.0
ksea:D_34R_A_34L_34R	1.0
ksea:other	0.0

Ksea공항 예시 - 12개의 class

예시

10개의 공항 별로 예측해야 할 대상의 개수가 다름

- Katl공항은 26개의 config 중 1개 예측
- Kclt공항은 13개의 config 중 1개 예측
- Ksea공항은 12개의 config중 1개 예측

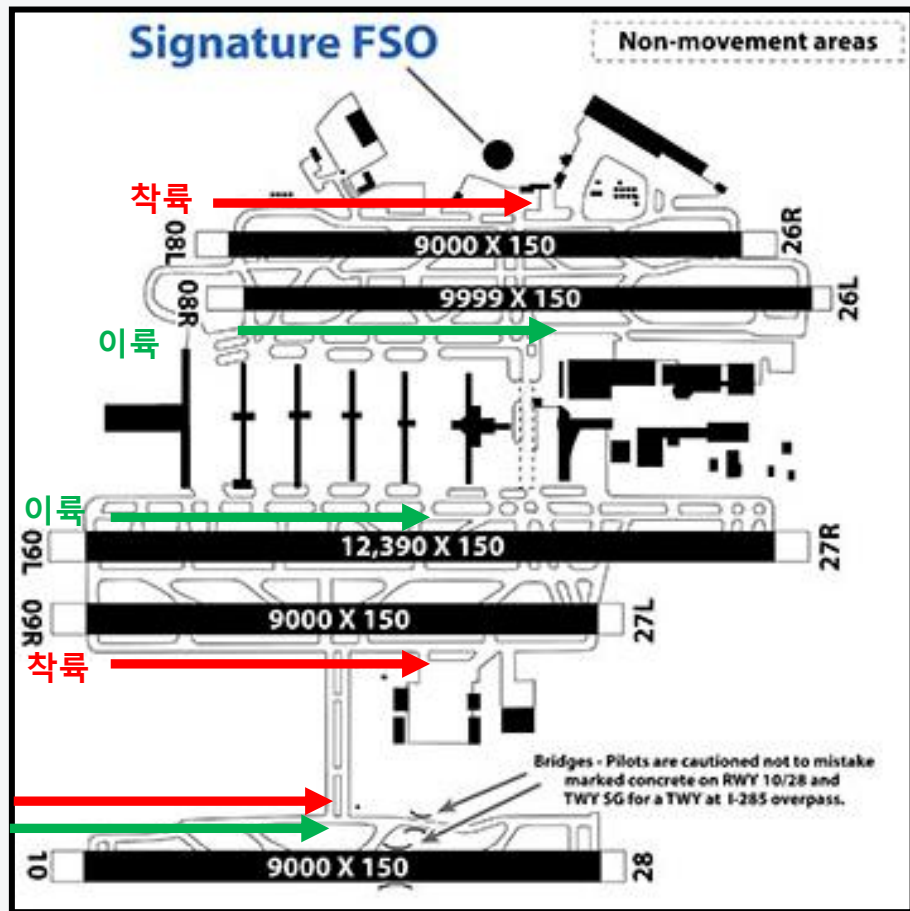
예측 대상

* 각 공항의 시간대 별 어떤 config가 활성화 될지 예측하는
다중 분류 문제

* 각 공항 별 config가 다르고, 다중 분류 개수도 다름

➡ **공항 별로 모델을 각각 만들어야 할 것으로 판단**

데이터 설명 – config 세부설명



Katl Airport Runway map

	airport	timestamp	lookahead	config	active
0	katl	2021-05-21T05:00:00	360	katl:D_10_8L_A_10_8L	0.0
1	katl	2021-05-21T05:00:00	360	katl:D_10_8R_9L_A_10_8L_9R	1.0
2	katl	2021-05-21T05:00:00	360	katl:D_10_8R_A_10_8R	0.0
3	katl	2021-05-21T05:00:00	360	katl:D_26L_27L_A_26R_27L_28	0.0
4	katl	2021-05-21T05:00:00	360	katl:D_26L_27R_28_A_26R_27L_28	0.0
5	katl	2021-05-21T05:00:00	360	katl:D_26L_27R_A_26L_27L_28	0.0

예시

katl:D_10_8R_9L_A_10_8L_9R

Departure 활주로

config 코드

Arrival 활주로

config 코드의 숫자 : 방위를 0~35까지로 나누어 표현
 config 코드의 알파벳 : L, R은 각각 왼쪽, 오른쪽
 D : Departure (출발)
 A : Arrival (도착)

III. 데이터 엔지니어링

- 전처리
- 기상 변수
- 파생 변수

데이터 전처리

시간간격 통일 (30분 단위)

통일 전

	timestamp
0	2020-11-01T01:11:00
1	2020-11-01T01:57:00
2	2020-11-01T02:53:00
3	2020-11-01T03:54:00
4	2020-11-01T04:52:00
...	...
12426	2021-10-31T20:56:00
12427	2021-10-31T21:53:00
12428	2021-10-31T22:04:00

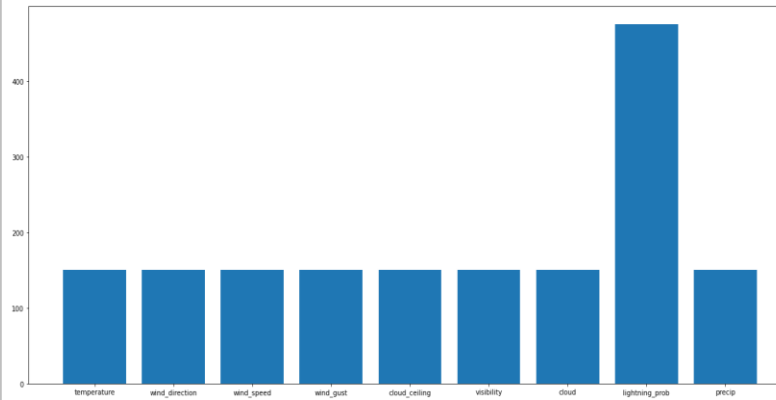
통일 후

	timestamp
0	2020-11-01 01:11:00
1	2020-11-01 01:57:00
2	2020-11-01 02:30:00
3	2020-11-01 03:00:00
4	2020-11-01 03:30:00
...	...
17512	2021-10-31 21:30:00
17513	2021-10-31 22:00:00
17514	2021-10-31 22:30:00

다른 데이터와의 연결을 위해 시간 간격을 통일 함

결측치 처리

기상 데이터의 컬럼 별 결측치 수



- 기상 데이터 외에 다른 데이터에는 결측치가 없음
- 선형 보간 법을 통해 보간

원핫 인코딩

ksea-D_16C_A_16C_16R	ksea-D_16C_A_16L_16R	ksea-D_34R_A_34L_34R	ksea-other	active
1	0	0	0	0.0
0	1	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

모든 Config에 원핫 인코딩을 진행 함

기상 데이터 (airport_Lamp)

국토교통부 예규 제 165 호

비행장시설(활주로) 설계 매뉴얼

Manual on Aerodrome Design (Runways)

2017. 4.

국토교통부

MINISTRY OF LAND, INFRASTRUCTURE AND TRANSPORT OF
KOREA

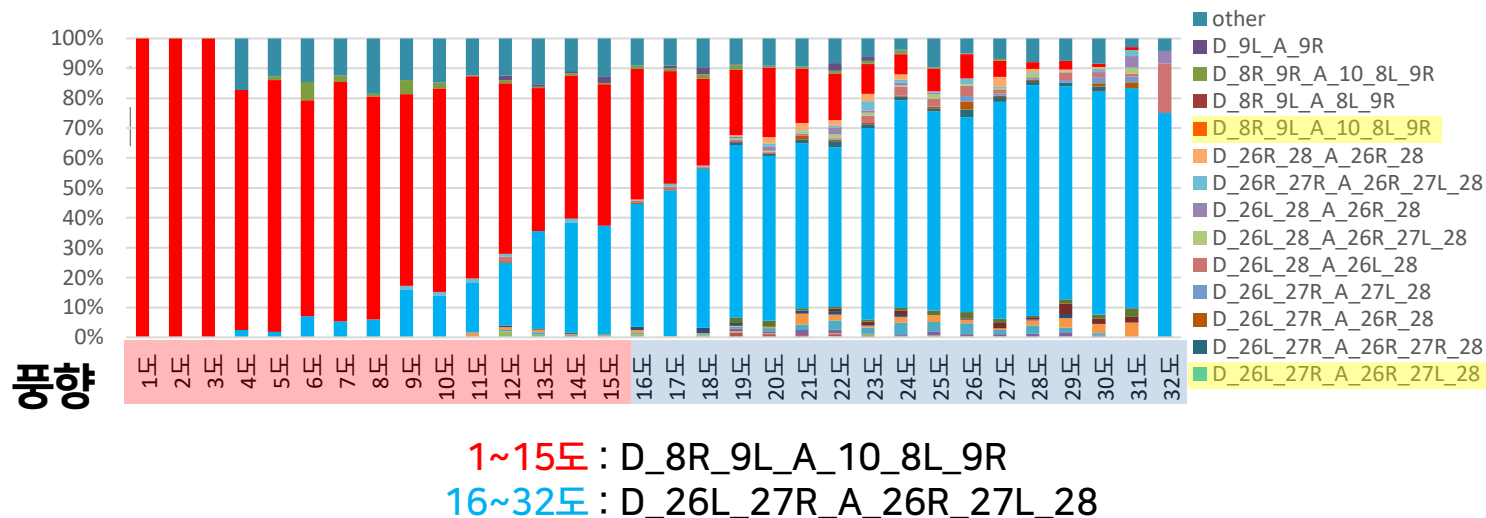
활주로 설계 매뉴얼에 따르면
온도, 풍향 -> 활주로 길이에 영향
풍향, 풍속 -> 활주로 방향에 영향



기상 데이터 필요하다고 판단

사용한 변수 : 온도, 풍향, 풍속, 번개, 강수

풍향에 따른 Config변화

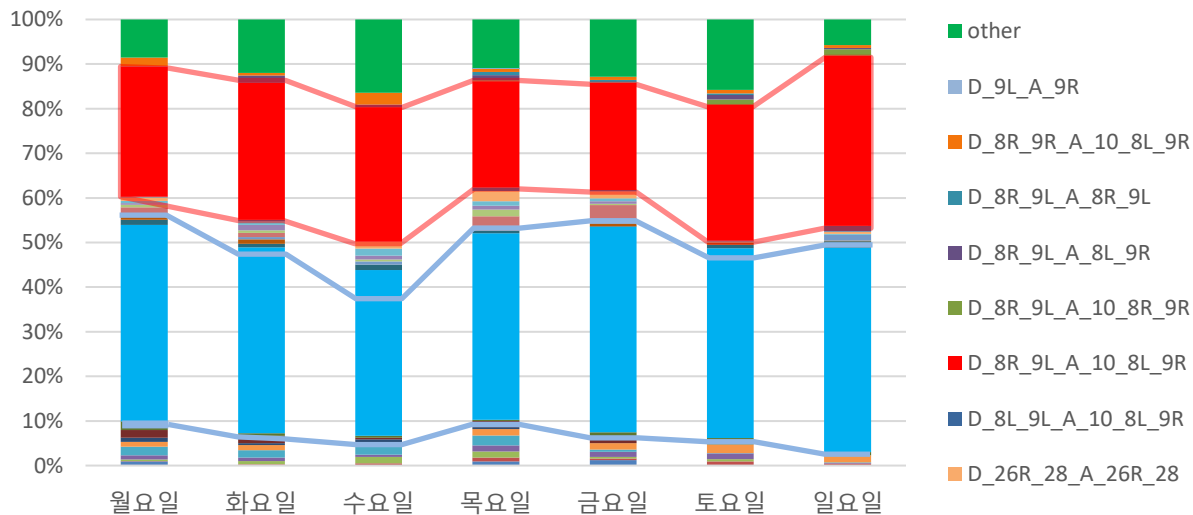


풍향에 따라 주로 활성화 되는 config가 다름

파생 변수 - 시간, 요일 활용

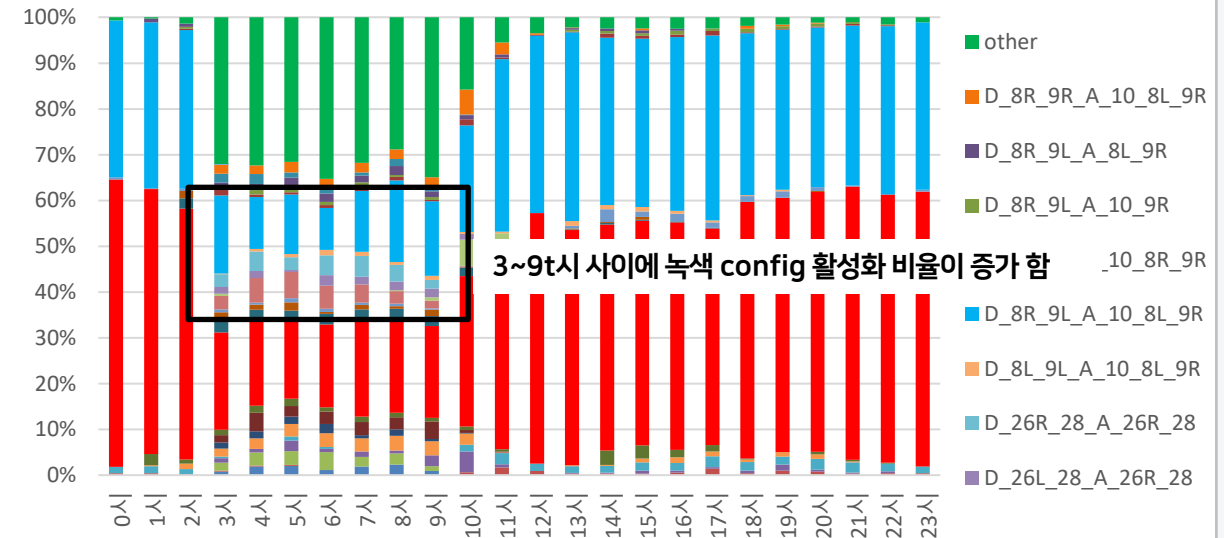
- 기존의 변수를 가공하여 새로운 변수 생성
- 해당 그래프들은 해당 시간 요일에 어떤 config가 많이 발생했는지를 나타내는 누적 통계 차트

요일 별 config 비율



요일마다 주로 활성화 되는 Config가 다름

시간에 따른 config 비율



시간대에 따라 활성화 되는 Config가 다른 패턴을 보임

요일, 시간에 따라 활성화 되는 config가 다른 패턴을 보이기 때문에 이를 가공하여 파생변수로 사용

파생 변수 생성 - 항공 교통 량 활용

시간 당 항공 교통량에 따른 Config변화

교통량 적을 때
(시간 당 0 ~ 20대)



시간당 교통량이 적을 때 : 기상 환경에 따라 이착륙에 사용되는 활주로가 다양함
Ex : 교통량이 적을 때 1~2개의 활주로만 사용되기도 함

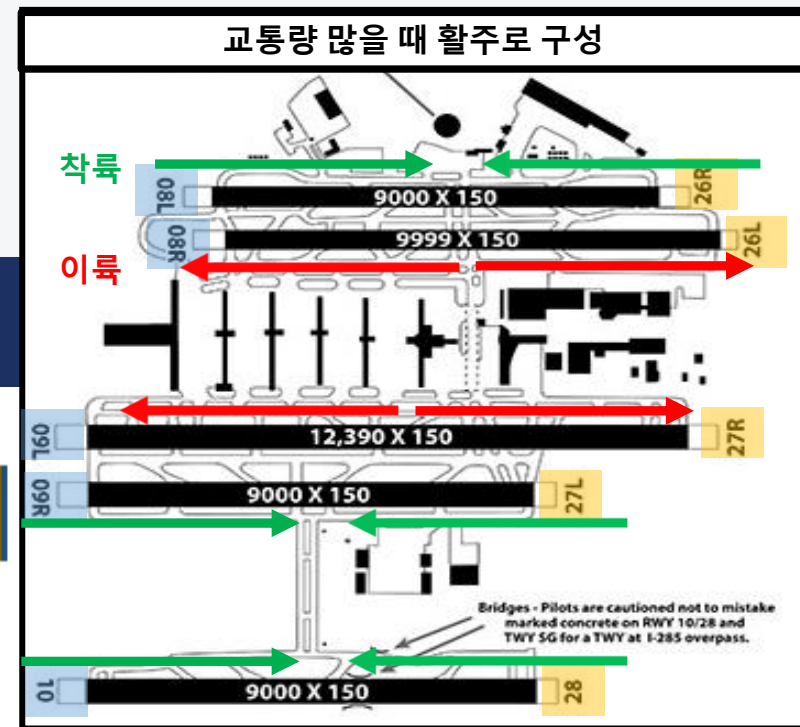
교통량 많을 때
(시간 당 100 ~ 120대)



D_26L_27R_A_26R_27L_28

D_8R_9L_A_10_8L_9R

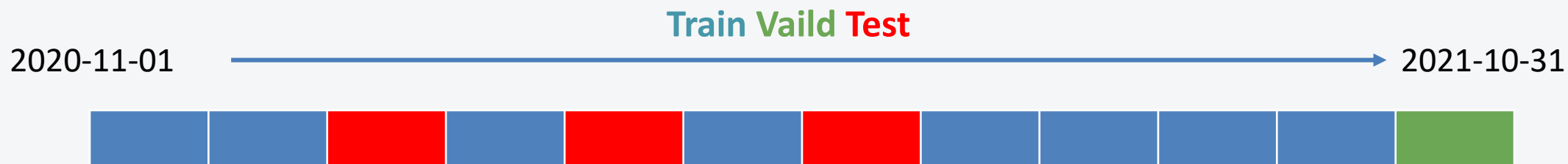
시간당 교통량이 많을 때 : 많은 교통량을 수용하기 위해 특정 config만 사용



↓

교통량이 많을 때 모든 활주로에서
일정한 방향으로 이륙 또는 착륙 진행

데이터 나누기 Train – Valid – Test으로 나누어 검증



- 1년간의 데이터를 이용해 학습 진행
- Test 데이터는 주어진 1년간의 기간 중 일정 기간의 데이터로 주어짐
- Valid 데이터는 학습 데이터의 가장 마지막 시점 10%로 사용

III. 모델링

- 모델 비교

모델 비교

- 3개의 각기 다른 모델을 통해 예측 및 결과를 도출 함

룰 기반

확률 파생 변수

사용 변수 : 확률 기반 5개 변수

균등분포, 누적확률분포, 현재 구성에 대한 가중치, 1시간 이내에 활성화 되었던 가중치

① ② ③ ④ ⑤

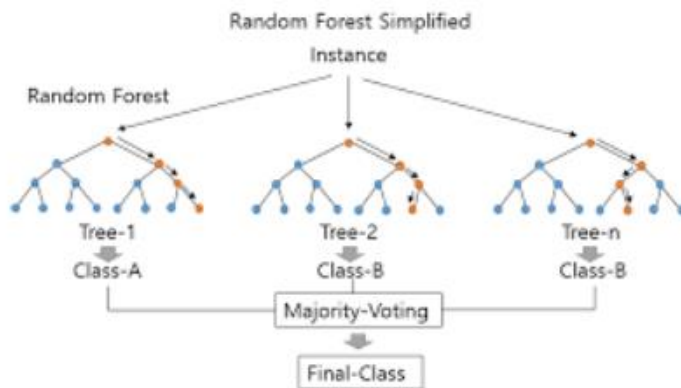
	uniform	config_dist	current	possible	mixture
katl:D_26L_27R_A_27L_28	0.037037	0.000000	0.0	0.0	0.014245
katl:D_26L_28_A_26L_28	0.037037	0.000000	0.0	0.0	0.014245
katl:D_26L_28_A_26R_27L_28	0.037037	0.000000	0.0	0.0	0.014245
katl:D_26L_28_A_26R_28	0.037037	0.000000	0.0	0.0	0.014245
katl:D_26R_27R_A_26R_27L_28	0.037037	0.019544	0.0	0.0	0.021762
katl:D_26R_28_A_26R_28	0.037037	0.000000	0.0	0.0	0.014245
katl:D_8L_9L_A_10_8L_9R	0.037037	0.117264	0.0	0.0	0.059346
katl:D_8R_9L_A_10_8L_9R	0.037037	0.332248	0.6	0.0	0.372802
katl:D_8R_9L_A_10_8R_9R	0.037037	0.026059	0.0	0.0	0.024268
katl:D_8R_9L_A_10_9R	0.037037	0.000000	0.0	0.0	0.014245
katl:D_8R_9L_A_8L_9R	0.037037	0.000000	0.0	0.0	0.014245
katl:D_8R_9L_A_8R_9L	0.037037	0.000000	0.0	0.0	0.014245
katl:D_8R_9R_A_10_8L_9R	0.037037	0.000000	0.0	0.0	0.014245
katl:D_9L_A_9R	0.037037	0.000000	0.0	0.0	0.014245
katl:other	0.037037	0.094463	0.0	0.0	0.050577

Loss

0.097

머신러닝 기반

RandomForest



사용변수 :
시간대, 요일, 과거 Config

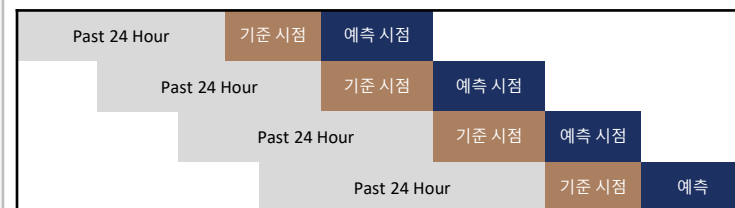
Loss

0.106

딥러닝 기반

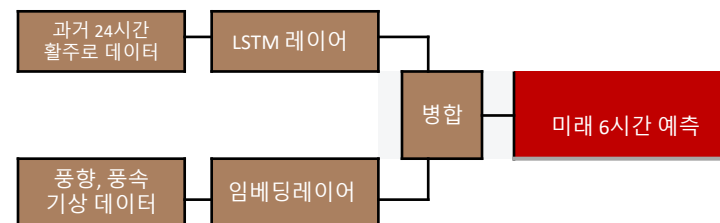
LSTM

사용변수 : 과거 config, 풍향 풍속



Input 데이터 기준 시점 예측 시점 : 미래 6시간

모델 구조

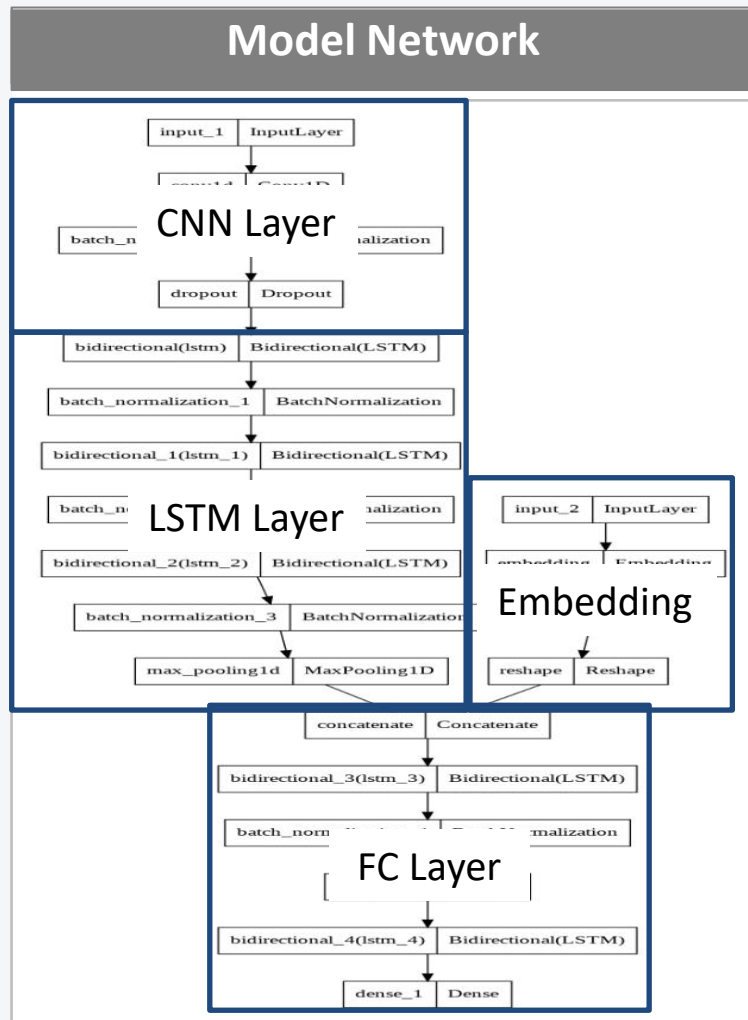


Loss

0.0893

※ 스코어는 loss 기반으로 책정 되며 낮을 수록 높은 예측률을 의미 함

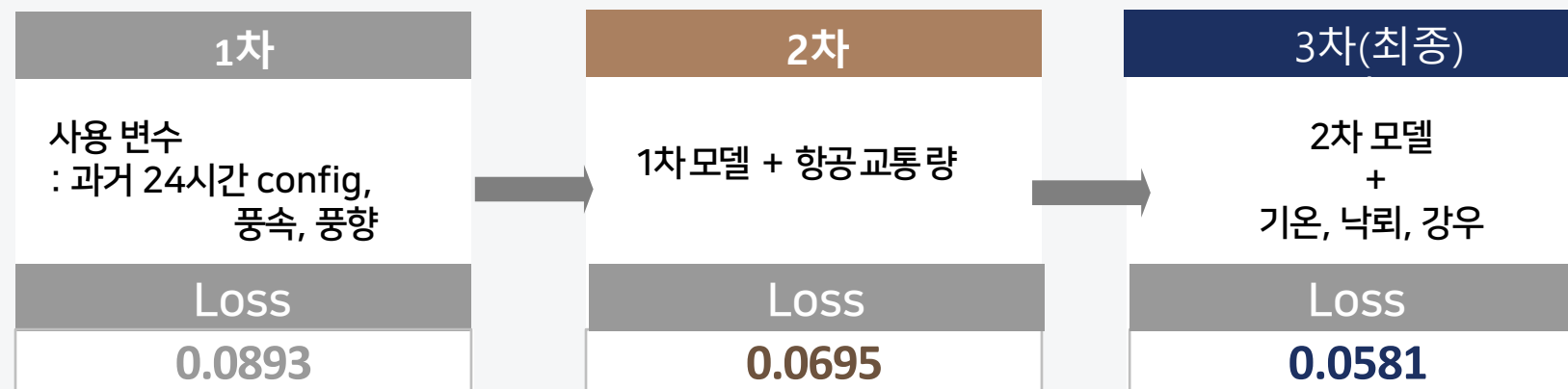
모델 Develop - 딥러닝 모델 개선



parameter

Optimizer : Rectified Adam
Learning rate Scheduler : Exponential Decay
Learning rate : 0.0005
Batch size : 64

파생 변수들을 추가하며 성능 개선 도모



※ 스코어는 loss 기반으로 책정되며 낮을 수록 높은 예측률을 의미 함

V. 결론

결과 및 결론 - 보완점

가설 검정

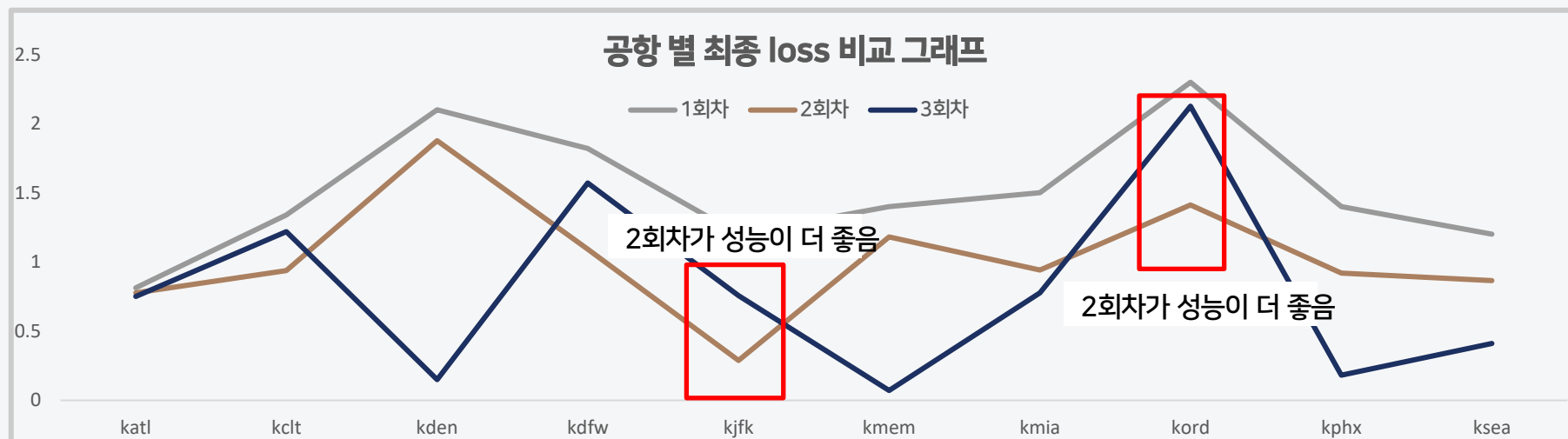
- 날씨, 시간, 요일 파생 변수를 사용했을 때 Loss가 더 감소 함
- 설정한 가설이 유효하다고 판단

공항 별 변수 영향도

- 초기 모델에서 가장 loss가 높음
- 최종 모델의 loss가 모든 공항에서 낮은 것은 아님
- 각 공항별로 영향을 미치는 변수가 다름
- 따라서 공항 별로 변수의 영향도가 다르며 더 좋은 성능을 위해선 공항 별로 각기 다른 변수 설정 필요
















K Fold

- 공항 별로 10개의 모델 생성 -> **학습시간 오래 소요 (Kfold 시 200시간 이상 소요)**
- Generalization gap을 완화하기 위한 전략으로 K-Fold가 시행되어야 하지만 물리적인 시간 문제로 시행되지 못 함



결과 및 결론

1등

User or team		Best public ↓ Mean Agg Log Loss ⓘ	Timestamp ⓘ	Trend (last 10)	# Entries
	HunHun_Jieun	1 0.0581	2022-03-30 03:41:10		13
	Stuytown	2 0.0695	2022-03-18 22:28:02		15
	pennswood	3 0.0910	2022-03-19 16:30:43		5
	mathurin	4 0.0969	2022-03-17 21:54:53		4
	dihoon	5 0.0970	2022-03-29 07:49:45		8
	Benchmark: Recency-weighted historical forecast	0.0980			
	agnim25	6 0.0980	2022-03-22 21:31:06		1
	Benchmark: No change forecast	0.1278			
	Nononing	7 0.1795	2022-03-19 17:01:55		6

2022 기업 연계 프로젝트

감사합니다. 이상 탐지 & 미래 예측

TEAM HunHun Jieun ho~!
임훈, 정지훈, 정지은, 이현호(멘토)