빅 콘테스트

항공 운항 데이터를 활용하여 항공 지연을 예측하다

·시로 앤 마로 ·

박재우 · 안지민 · 이형선 · 정민영 · 주은혁

1 주제 및 데이터의 이해

2 EDA

3 Feature Engineering

4 Modeling & Ensemble

5 지연율 계산

Concept

주어진 데이터에서 의미 있는 변수를 찾아내고,
새로운 변수를 만들어 예측력 높은 모델을 통해

항공 지연율을 예측한다

주제 및 데이터 이해

2017년 1월 1일부터 2019년 6월 30일까지의 항공 데이터를 이용하여

2019년 9월 16일부터 2019년 9월 30일의 항공기 지연 여부를 예측

raw data

AFSNT_DLY: 2019.9.16 ~ 2019.9.30 까지 지연율 계산
AFSNT: 2017.1.1 ~ 2019. 6.30 까지의 운항 실적 데이터
SFSNT: 2019년 하계 스케줄 중 7월~9월이 포함된 시즌 데이터

주제 및 데이터 이해

SDT_YY, SDT_MM, SDT_DD, SDT_DY	연, 월, 일, 요일
ARP, ODP	공항, 상대 공항
FLO, FLT, REG	항공사, 편명, 등록 기호
AOD	출도착
IRR	부정기편
STT, ATT	계획 시각, 실제 시각
DLY, DRR	지연 여부, 지연 사유
CNL, CNR	결항 여부, 결항 사유

주제 및 데이터 이해

날씨 요인	A01	안개: 기온이 이슬점 이하일 때, 지표의 온도가 공기의 온도보다 낮아지면 발생
	A03	강우: 비가 내리는 것
	A07	운고: 구름의 높이. 구름 밑부분의 고도
공항 요인	B01	계류장 혼잡 → 계류장: 비행장 내에서 항공기에 승객을 탑승시키거나 우편 또는 화물을 적재, 급유·주기 및 정비하기 위하여 지정된 구역
	B03	활주로 사정 → 활주로: 항공기 이/착륙을 위하여 국토교통부령으로 정하는 크기로 이루어지는 공항 또는 비행장에 설정된 구역
기체 요인	C01	A/C 정비: 항공기 정비
	C02	A/C 접속: 전편 항공기의 지연 및 결항으로 다음 연결편에 영향

EDA (Exploratory Data Analysis)

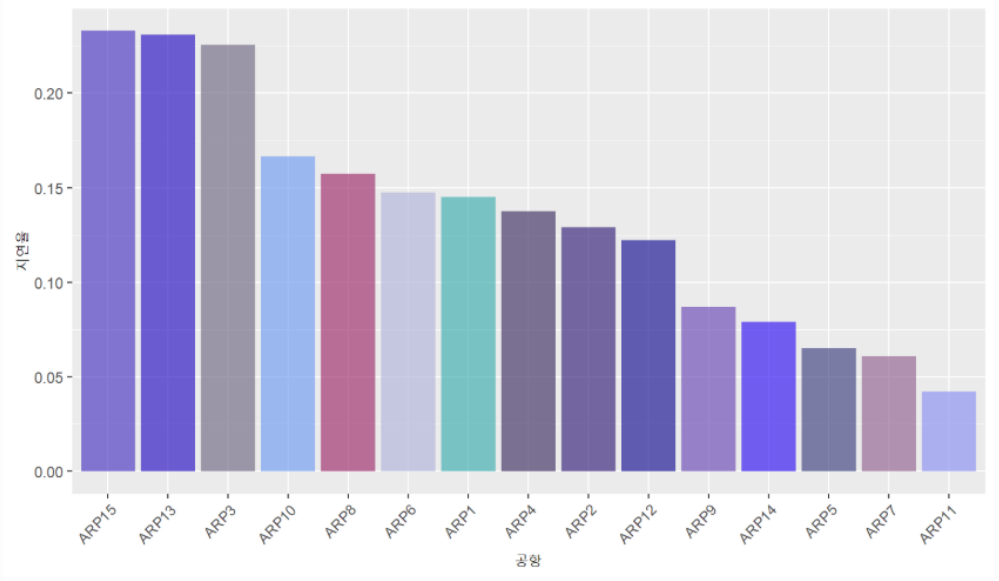
AFSNT의 전체 데이터 중 **항공편 지연(DLY)** 개수 확인

▶ 지연 사유 코드(DRR) 확인

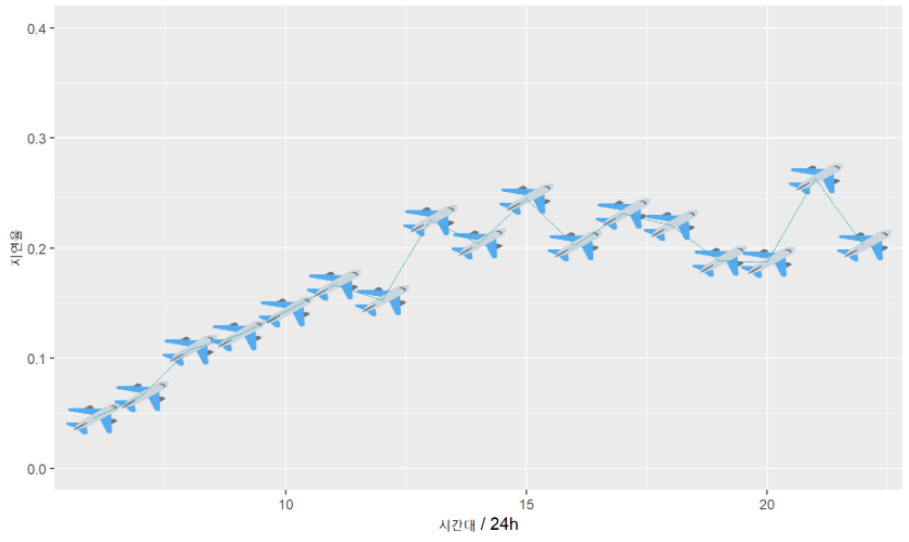
▶ 지연 사유 코드 분류 : 날씨 요인(A) / 공항 요인(B) / 기체 요인(C) / 기타

EDA (Exploratory Data Analysis)

공항별 지연율 그래프



시간대별 지연율 그래프



공항 별로 / 시간대 별로 지연율의 차이가 나타남

EDA (Exploratory Data Analysis)

공항별 / 시간대별 지연 그래프

공항과, 시간대에 따라 지연율의 차이가 큰 것을 확인

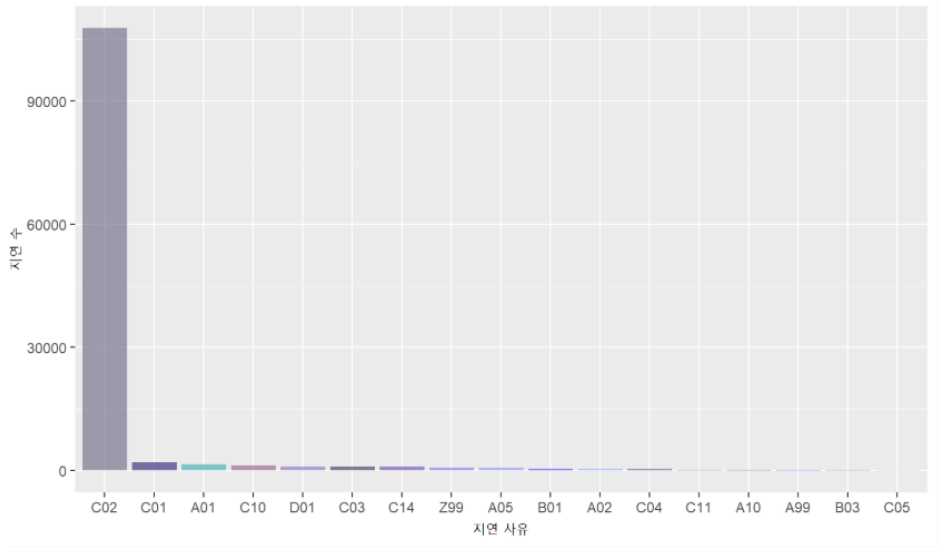
→ ARP15(인천), ARP13(군산), ARP3(제주) 세 개 공항의 지연율이 높음

→ 오전보다 오후 비행의 지연율이 높음

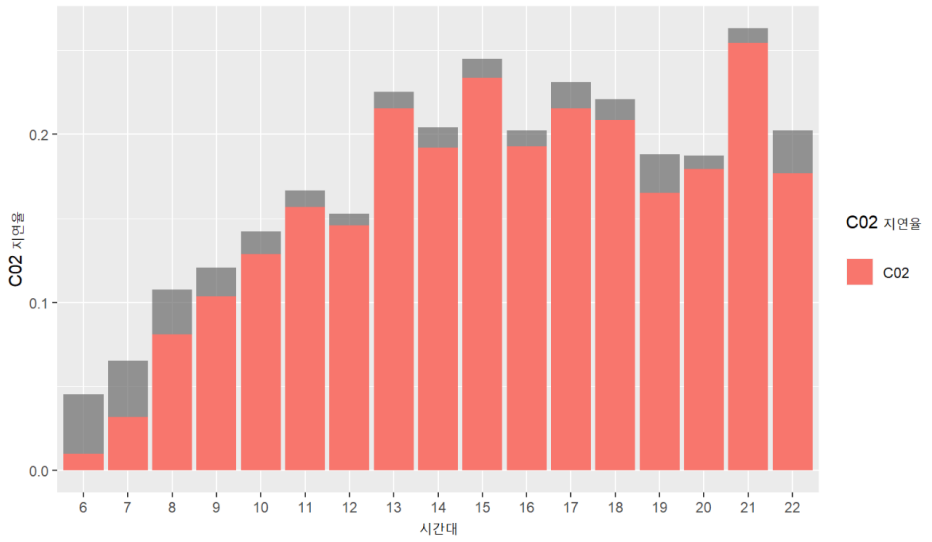
⇒ 이 후 변수를 생성할 때 공항 별, 시간대 별로 분류하여 처리할 예정

EDA (Exploratory Data Analysis)

지연 사유 그래프



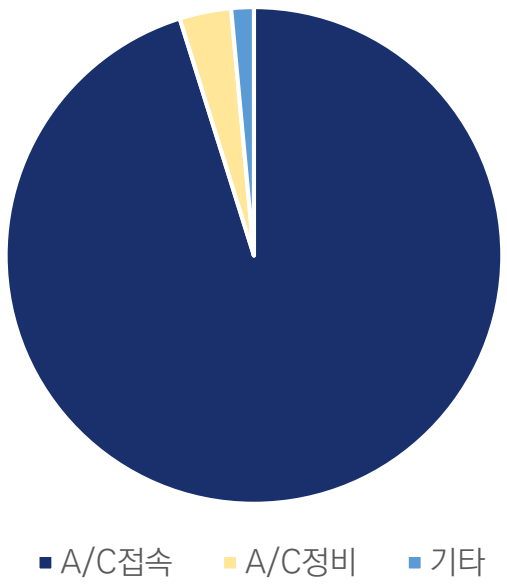
C02 지연율 그래프



지연의 약 90%가 C02 (A/C접속) → C02가 지연 예측에 대한 가장 중요한 키워드

EDA (Exploratory Data Analysis)

가장 중요한 요인: C02 (A/C 접속)



동일 항공기체 운항 시,앞 운항의 지연이 다음 운항 지연에 미치는 영향

- 앞 운항이 지연 되었을 때
  Odds = 1.1662
- 앞 운항이 지연되지 않았을 때
  Odds = 0.0677
- 앞 운항의 지연에 따른
  Odds Ratio = 17.230

⇒ **“연쇄적 지연이 발생할 것이다.”**

EDA (Exploratory Data Analysis)

AFSNT :: REG

AFSNT_DLY :: REG

SFSNT를 통해 REG(기체)별 일정의 반복성을 파악,
AFSNT의 REG를 AFSNT_DLY에 적용

EDA (Exploratory Data Analysis)

By 연관 분석 (Association Analysis)

- 군집분석에 의해 분류된 cluster를 대상으로
  그룹에 대한 특성을 분석
- AFSNT의 REG에 따른 FLT의 연관성 분석 시행
- 도출된 REG와 FLT의 연관성을 AFSNT_DLY에 적용

앞 운항 지연 예측의 핵심인 REG를
AFSNT_DLY에서 재확인 후 변수로 추가

EDA (Exploratory Data Analysis)

By 연관 분석 (Association Analysis)

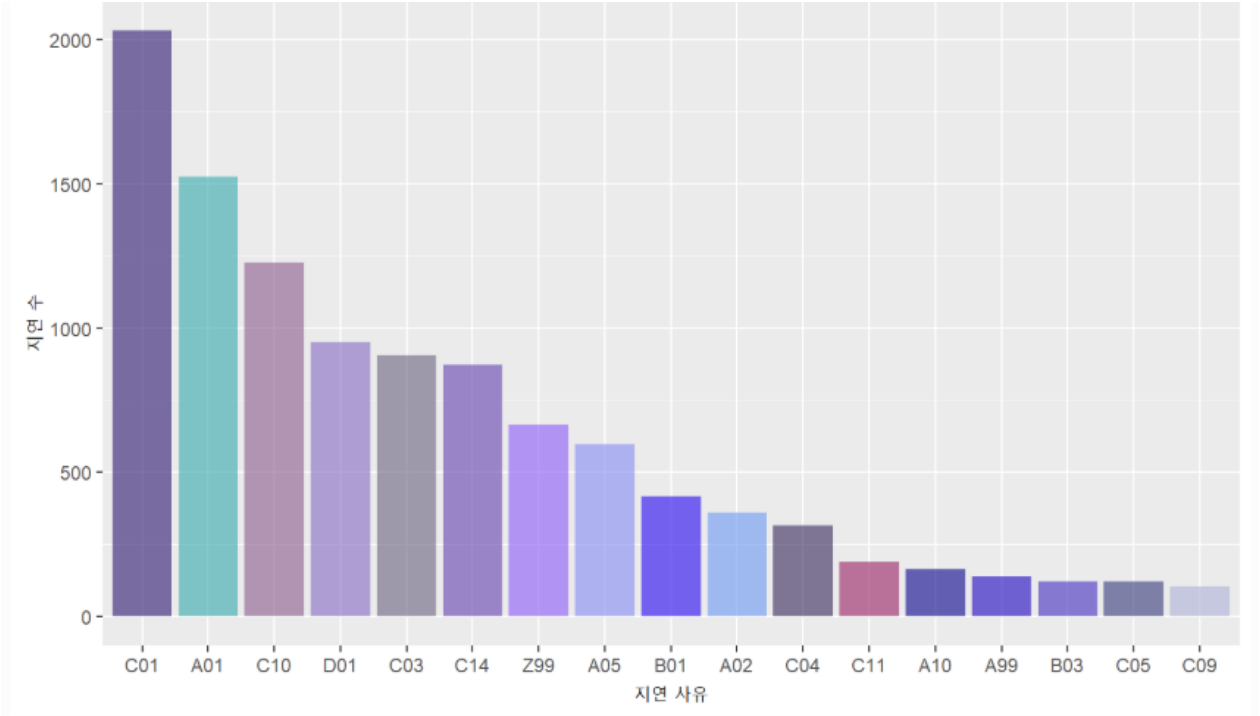
다른 요인에 대해서도 확인해보자

- 군집분석에 의해 분류된 cluster를 대상으로
그룹에 대한 특성을 분석
- AFSNT의 REG에 따른 FLT의 연관성 분석 시행
- 도출된 REG와 FLT의 연관성을 AFSNT_DLY에 적용

앞 운항 지연 예측의 핵심인 REG를
AFSNT_DLY에서 재확인 후 변수로 추가

EDA (Exploratory Data Analysis)

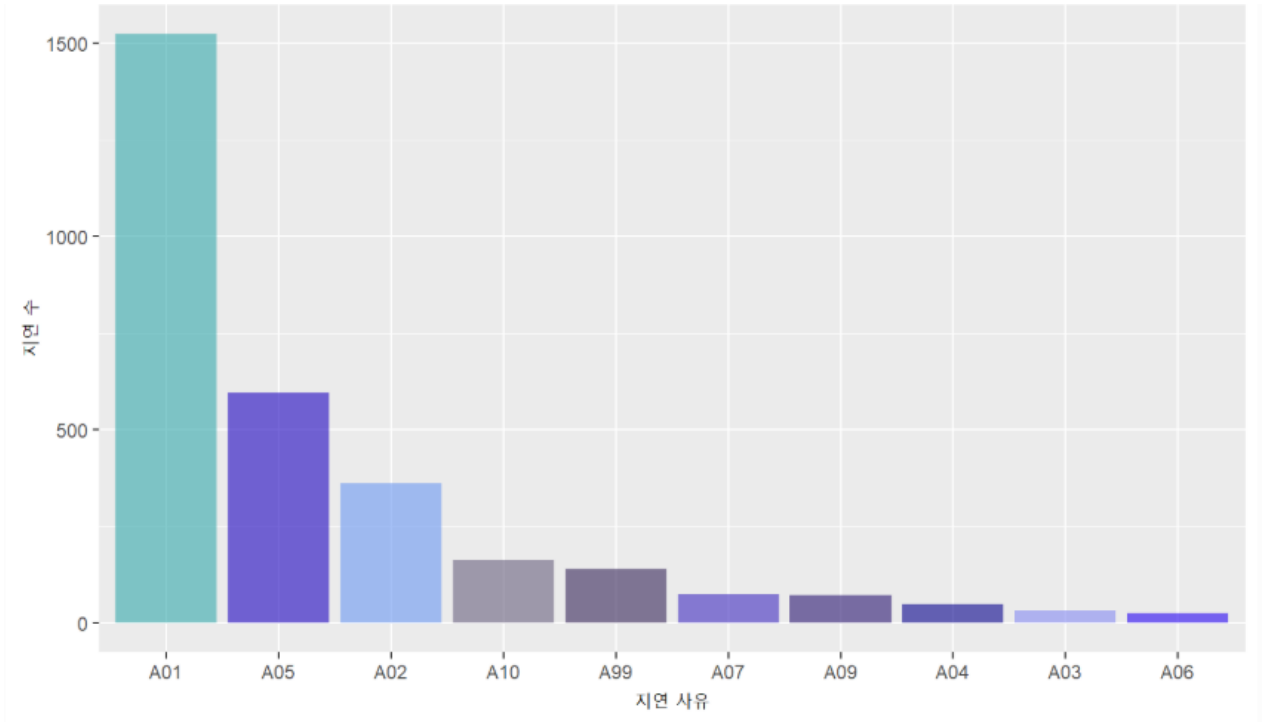
상위 지연 사유 그래프



C02로 인한 지연을 제외하면, A B C 로 인한 지연이 골고루 발생함

EDA (Exploratory Data Analysis)

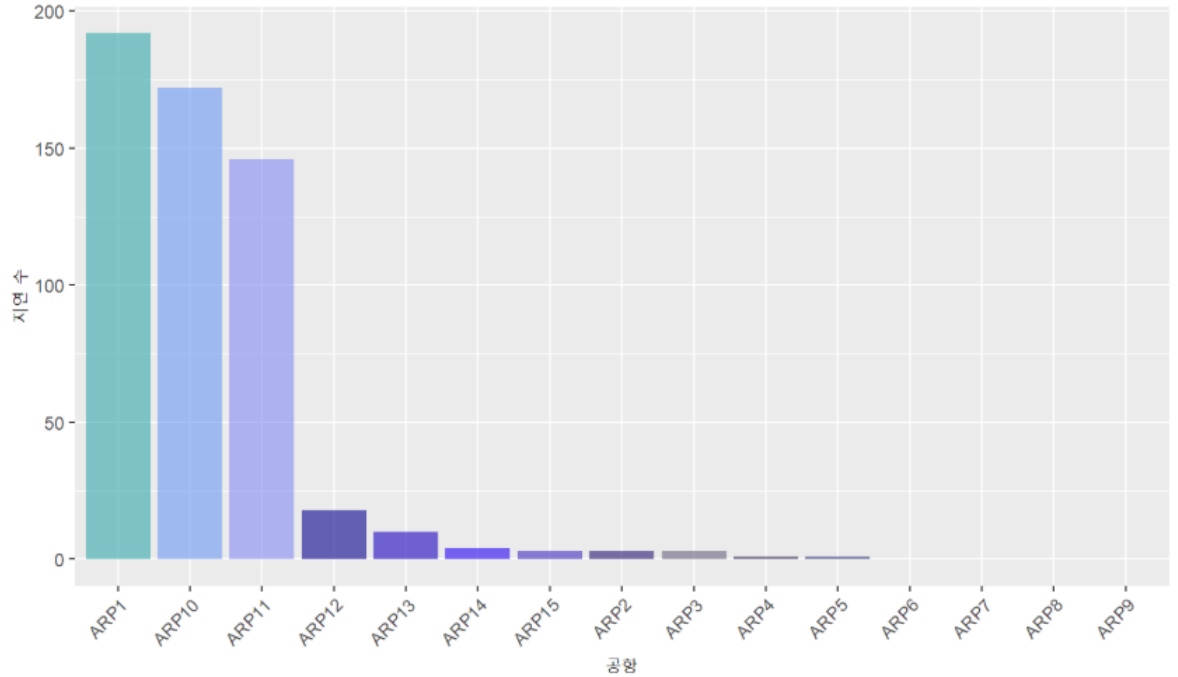
날씨 요인 지연 그래프



A01(안개)으로 인한 지연이 가장 많은 비중을 차지함 → 안개 예측이 중요

EDA (Exploratory Data Analysis)

공항 요인 지연 그래프



B(공항 요인)로 인한 지연 수의 차이가 공항 별로 상이하게 나타남
→ 공항 환경이 지연에 유의하게 영향을 미칠 것

EDA (Exploratory Data Analysis)

날씨 & 공항 요인 변수

- PS_AVG : 평균 해면 기압(hpa)
- TMP_AVG, TMP_MAX, TMP_MNM : 평균 / 최대 / 최저 기온 (℃)
- TD_AVG : 평균 이슬점(℃)
- HM_AVG, HM_MNM : 평균 / 최저 상대 습도(%)
- WSPD_AVG, WSPD_MAX, WSPD_INS : 평균 / 최대 / 최대순간 풍속 (knot)
- CA_TOT_AVG, CLA : 평균 / 최저운고 운량(1/8)
- RN_SUM : 강수량(mm)
- VIS_MNM : 최단 시정(10m)
- APRON_SIZE : 계류장 면적
- CRAFT_SIZE : 최대 동시 주기 수

Feature Engineering

: 머신 러닝 알고리즘을 작동하기 위해 데이터에 대한

도메인 지식을 활용하여 **특징(Feature)** 를 만들어내는 과정

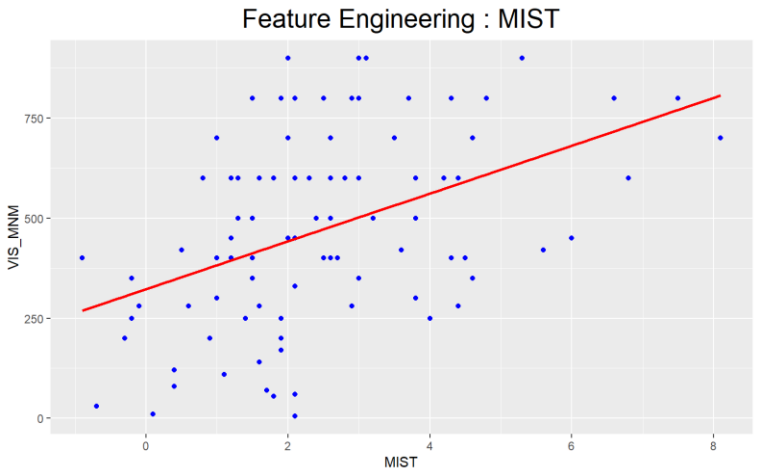
Feature Engineering

날씨 (Mist & HM_INV)

- 날씨 요인 중 안개(A01)로 인한 지연율이 높은 것을 확인
- 안개가 생성되기 위해서는 기온이 이슬점에 도달하고, 습도가 높아야 함
- 안개가 생긴다면 시정거리가 짧아진다는 가정 하에, 이를 표현할 수 있는 새로운 변수 생성

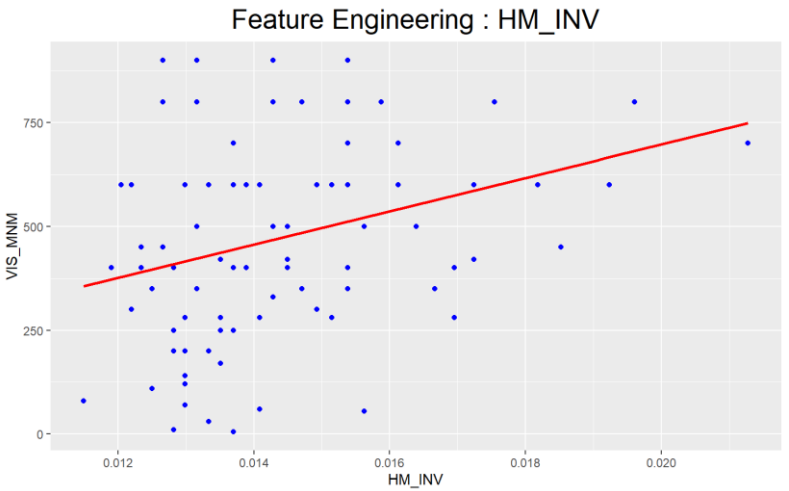
Feature Engineering

날씨 (Mist)



$$VIS\_MNM$$
$$\propto (TMP\_MNM - TD\_AVG) \Rightarrow MIST$$

날씨 (HM_INV)



$$VIS\_MNM$$
$$\propto \frac{1}{HM\_AVG} \Rightarrow HM\_INV$$

$*$ VIS_MNM : 최단 시정 | TMP_MNM : 최저 온도 | TD_AVG : 평균 이슬점 | HM_AVG : 평균 습도

Feature Engineering

공항 교통 혼잡도 (COMPLEXITY)

특정 항공기의 출발 예정 시간(STT)을 기준으로 1시간 이전부터 예정되어 있는 운항 수를 계산

→ 단위 시간 내의 비행 수가 증가하면 공항이 혼잡해 진다는 가정을 하고,
이것이 비행 지연에 영향을 끼칠 것으로 판단

Modeling & Ensemble

변수 설명

교통 혼잡도 *COMPLEXITY*	기준 출발 시간의 1시간 이전부터 예정된 운항 수
계류장 면적 *APRON_SIZE*	공항 별 계류장 면적 (m^2)
주기장 능력 *CRAFT_STAND*	공항 별 주기장의 최대 동시 주기 수
기준 공항 *ARP*	기준 공항(ARP)에 따라 생성한 더미변수 (ARP10제외)
상대 공항 *ODP*	상대 공항(ODP)에 따라 생성한 더미변수

Modeling & Ensemble

변수 설명

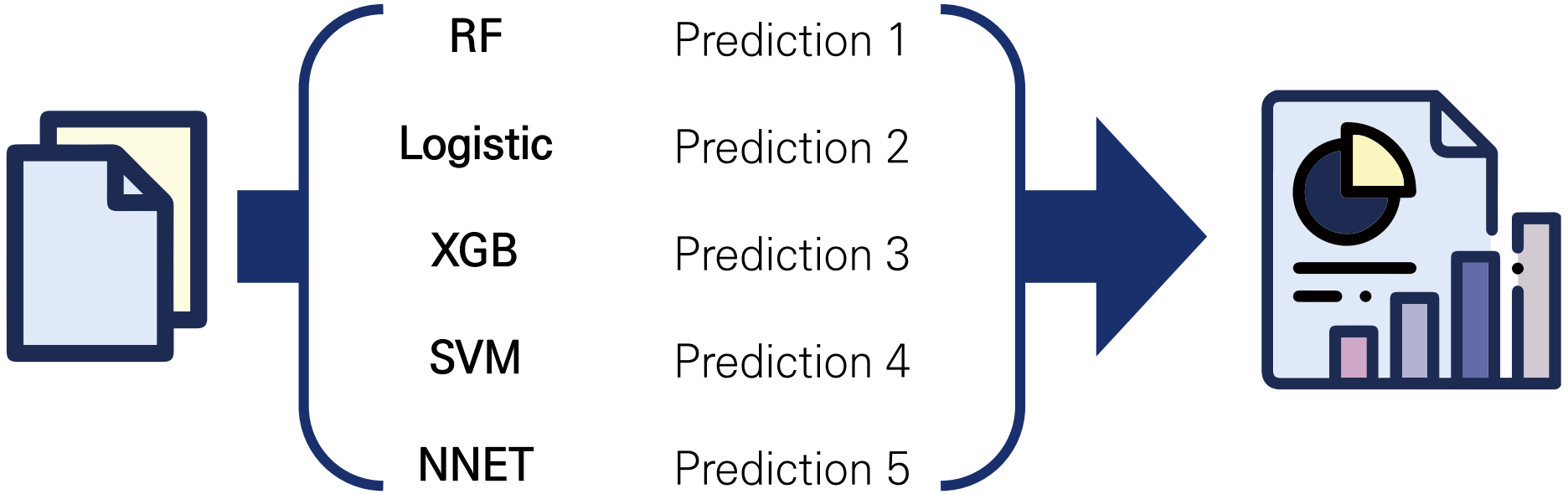
해면 기압 *PS_AVG*	관측 지점의 평균해면에서 관측된 기압(*hpa*)
기온 *WSPD_AVG,MAX,INS*	대기의 온도(℃)
운량 *CA_TOT_AVG, CLA*	구름이 하늘을 덮고 있는 정도
강우량 *RN_SUM*	일정 기간 동안 일정한 곳에 내린 비의 분량

Modeling & Ensemble

변수 설명

시정 *VIS_MNM*	목표물을 명확하게 식별할 수 있는 거리(10m)로, 대기의 혼탁도를 나타내는 척도
고도 *BASE*	운고에 영향을 미치는 구름의 최저 고도
안개 *MIST*	[최저기온(TMP_MNM) – 이슬점(TD_AVG)]으로 생성한 안개 변수

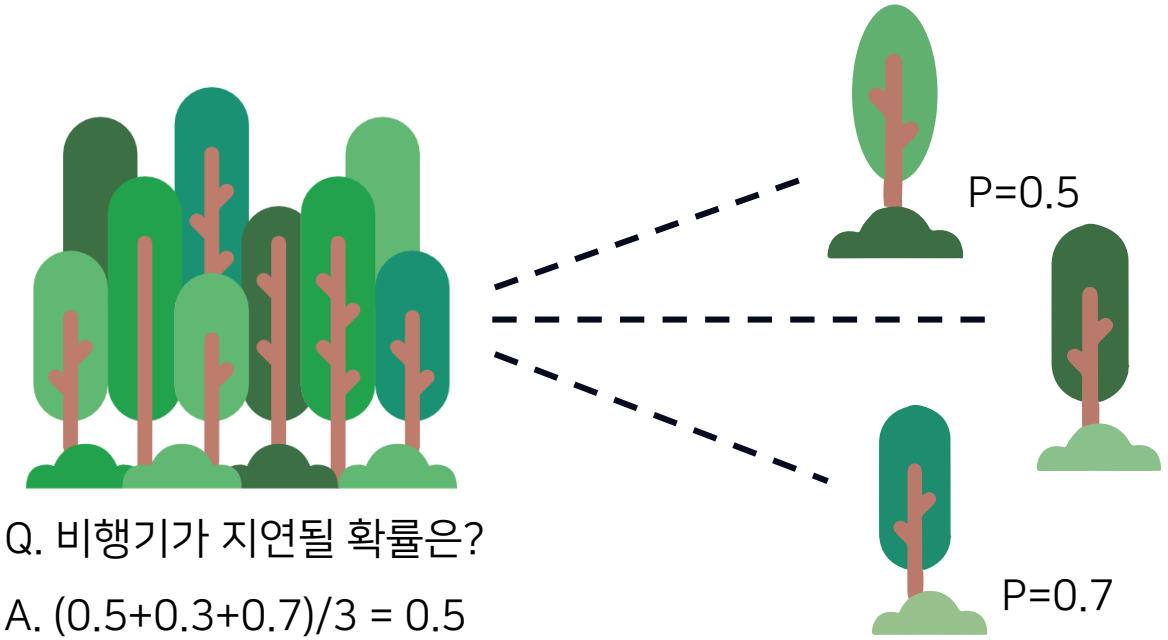
Modeling & Ensemble



5개의 학습법으로 앙상블을 사용하여 과적합을 방지하고 예측력을 향상시키고자 함

Modeling & Ensemble

Random Forest



Q. 비행기가 지연될 확률은?

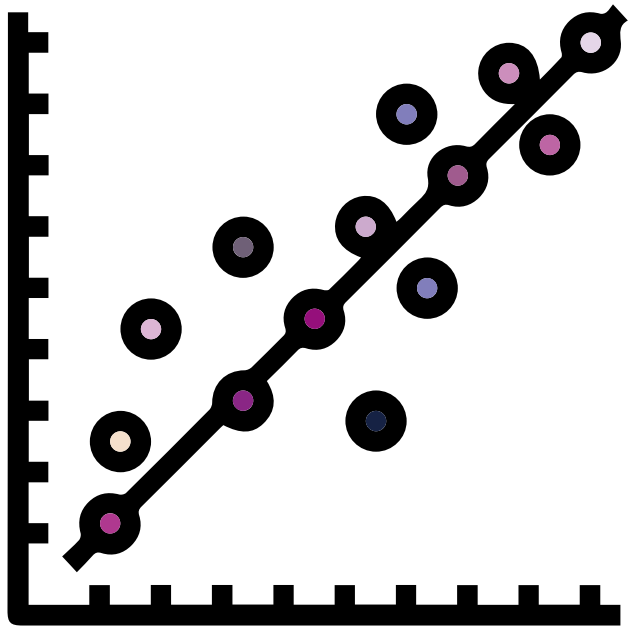
A. (0.5+0.3+0.7)/3 = 0.5

랜덤 포레스트는 여러 개의 의사결정나무를 이용하는 학습법으로 평균 예측값을 출력함

과적합을 방지하고 다수의 변수를 일괄적으로 다룰 수 있다는 장점이 있음

Modeling & Ensemble

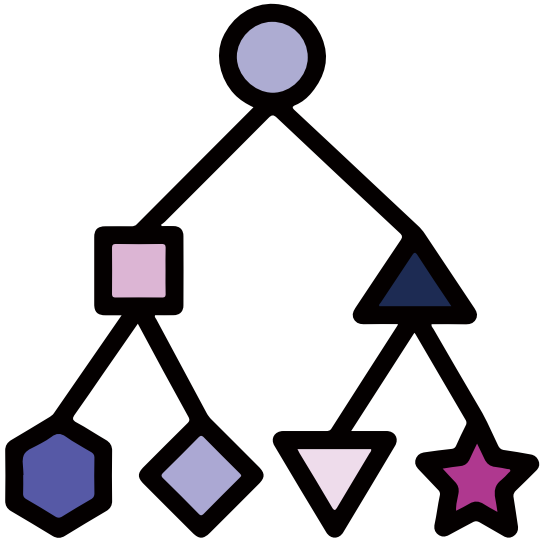
Logistic Regression



로지스틱 회귀 모델은 가장 일반적인 분류 모델로 시그모이드 함수를 사용하여 출력 값을 0과 1 사이의 확률로 도출함

Modeling & Ensemble

eXtreme Gradient Boosting

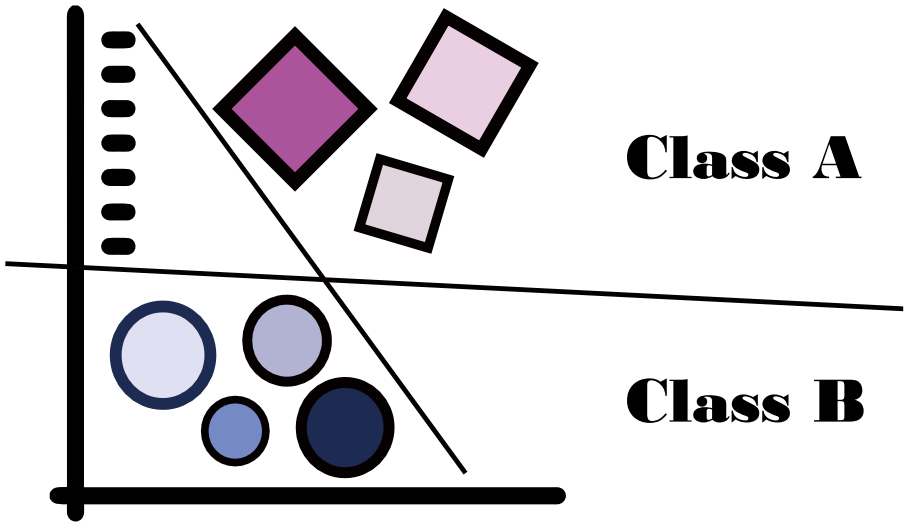


XGBoost는 높은 예측률을 보이는 학습법으로 알려졌으나 과적합의 위험도 가지고 있음

다른 모델과 결합하여 단점을 보완하도록 하였고 교차검증시 정확도를 향상하는 목적으로 사용

Modeling & Ensemble

Supporting Vector Machine

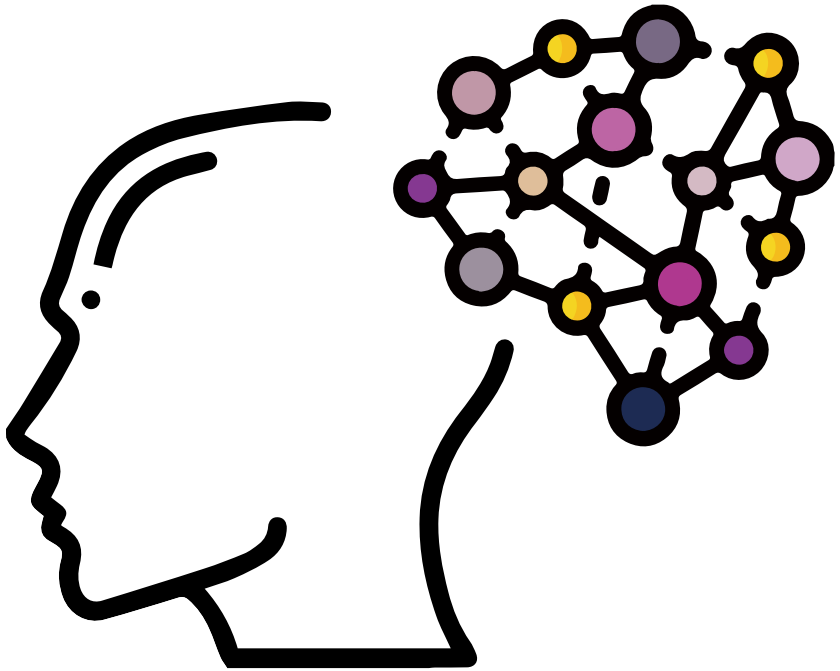


서포트 벡터 머신은 다차원의 데이터를 다루는 것에 유리한 학습법

변수가 50개를 넘는 고차원의 데이터를 학습해야 하므로 사용함

Modeling & Ensemble

Neural Network



인공 신경망은 문제 해결 능력을 가지는 모델 전반을 가리키는데,

입력 신호와 출력 신호 사이의 관계를 학습하며 인간의 뇌와 비슷하게 움직임

Modeling & Ensemble

모델링 구조

1. Bootstrapping

지연 여부가 약 1:7인 불균형 자료를 그대로 사용할 경우 예측률이 떨어진다.
이를 보완하여 균형 있는 샘플링을 위해 로즈 기법을 사용한다.

2. Searching Parameter

5개의 테스트 데이터를 사용하여 각 학습법에서의 MSE의 평균을 최소화하는
적절한 Parameter를 탐색한다.

Modeling & Ensemble

모델링 구조

3. Within Ensemble

훈련 데이터로부터 10,000개의 표본을 추출하여 모델에 적합하는 과정을 5회 반복한다. 얻어진 5개 모델의 평균 예측값을 학습의 결과 예측값으로 설정한다. 각 학습법마다 위 과정을 진행한다.

4. Between Ensemble

5가지 학습법으로 도출한 각 결과 예측값의 평균을 최종 예측값으로 설정한다. 최종 예측값을 이진 분류하여 지연 여부를 예측한다.

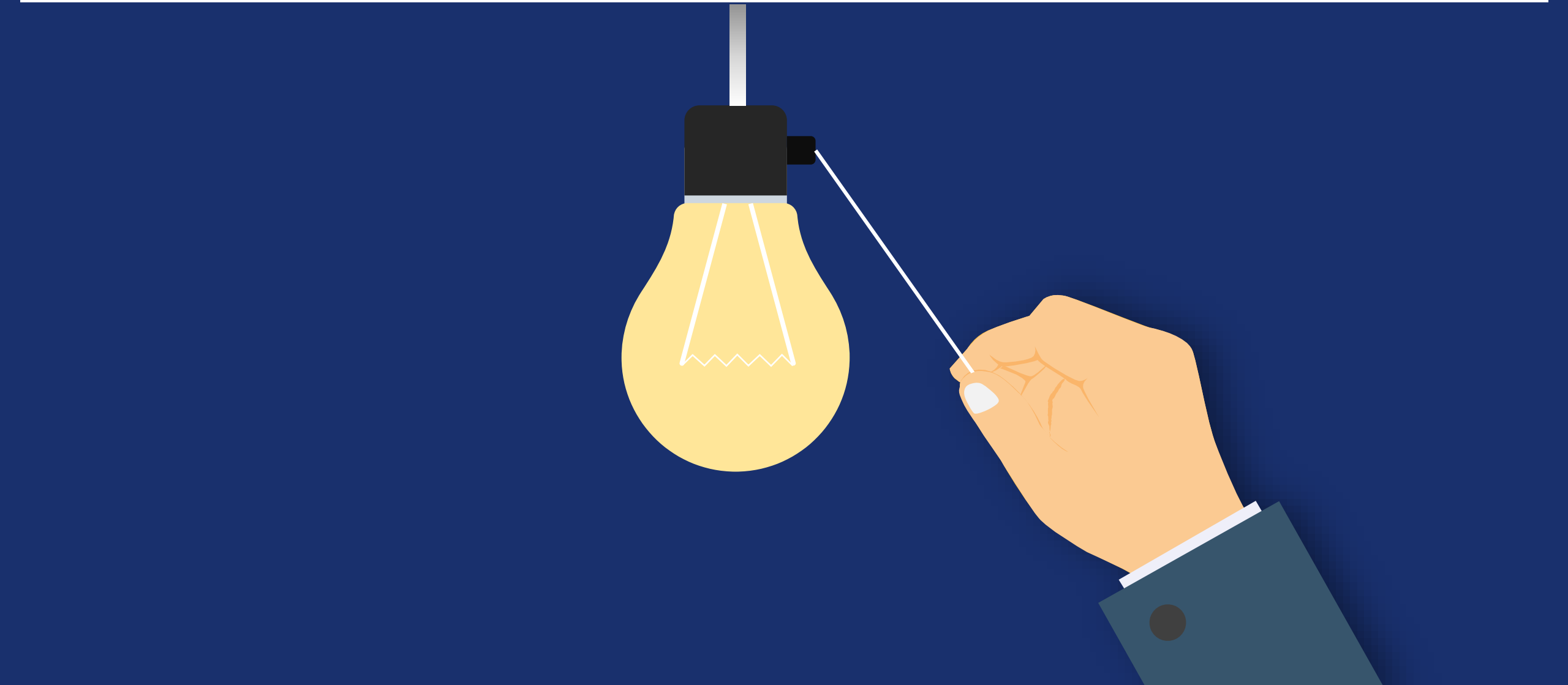
Modeling & Ensemble

모델링 구조

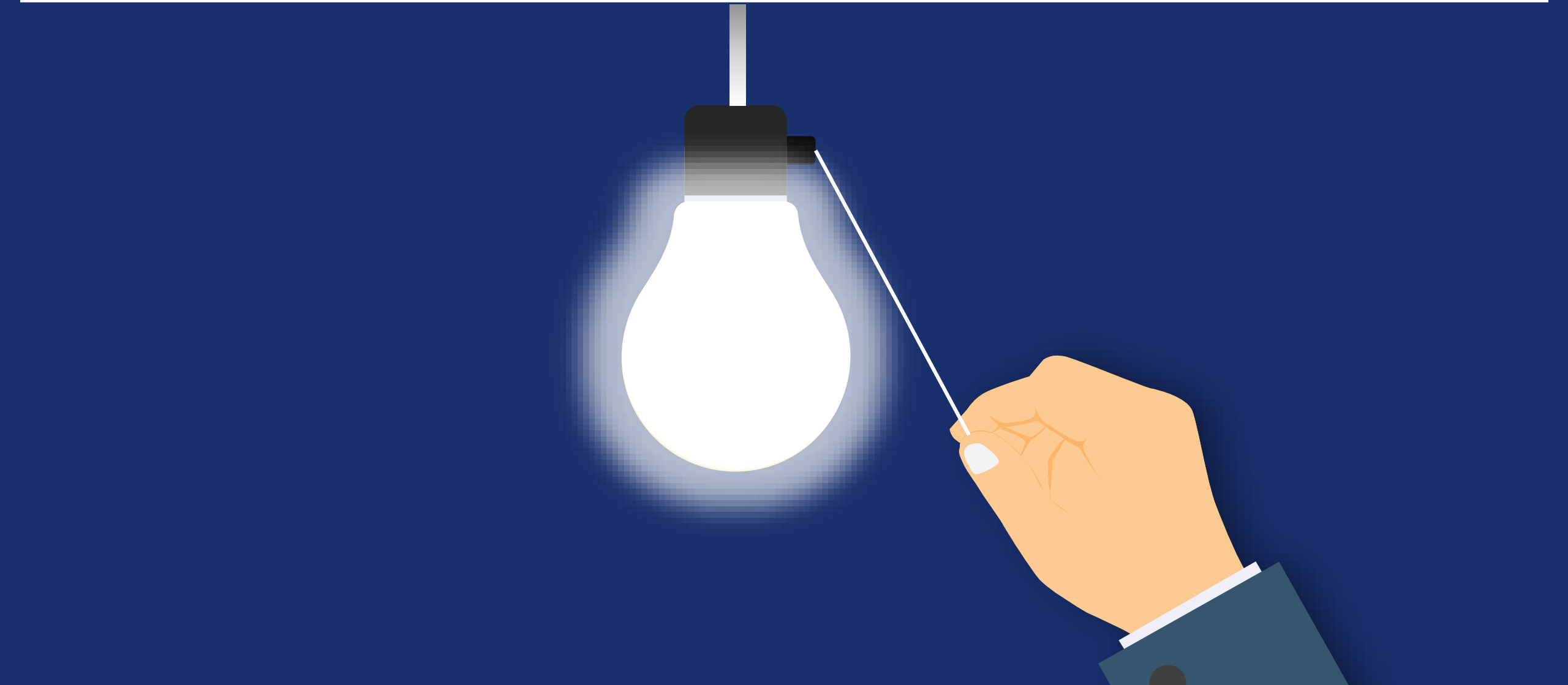
5. Predicting

한 비행이 지연된다면 동일 기체로 운항하는 다음 비행이 A/C접속으로 지연될 확률이 높은 것을 확인함에 따라, 같은 기체(REG)로 그룹핑하여 예측을 진행한다.

지연율 계산



지연율 계산




Thank you
We are shiro and maro
START