항공기 지연 예측

항공기 지연 예측 보고서

TEAM KUDS

임한동, 허종국, 장민호, 윤예솔, 이혜원

목차

Part 1

데이터의 이해

다양한 항공편

국내선 항공편은 여객 운송(여객기), 화물 운송(화물기), 페리 비행 등 다양한 목적으로 운용된다.

실제로는 운용 목적에 따라 지연 사유 및 시간 등에 차이가 날 수 때문에 실제적인 모델을 만들기 위한 요인으로 고려하였다.

항공기 연결편

민간 상용 항공기는 정해진 스케줄에 따라 여러 노선을 연속적으로 운항한다.

그로 인해 이전 항공편에서 장시간 지연이 발생하거나 그라운드 타임(Ground Time)이 짧은 경우 항공편 지연이 발생하게 된다.

공항의 중요성

특정 공항은 수용 능력을 고려해 시간당 슬롯(Slot)을 배정함으로써 항공을 통제한다.

그러나 수용 능력의 정체로 인해 특정 시간 대 활주로가 포화되면서 항공기 지연이 빈번하게 발생하고 있다.

SFSNT와 실제 공항 스케줄을 비교하여 공항 및 항공사 코드를 확인

ARP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
공항	김포	김해	제주	대구	울산	청주	무안	광주	여수	양양	포항	사천	군산	횡성	인천

FLO	A	B	F	H	I	J	L	M
항공사	아시아나	에어부산	이스타항공	제주항공	진에어	대한항공	티웨이항공	코리아 익스프레스에어

지연시간

실제 시간과 계획 시각의 차이를 지연 시간으로 파악하였다. 지연 여부(DLY) 와의 관계를 보면 ATT-STT가 30 분 이상일 때,
즉 지연 시간이 30분 이상일 때 'DLY==Y' 로 표시됨을 알 수 있다.

비 정 기

제공받은 평가 데이터(AFSNT_DLY)에는 부정기편이 없었다.

결 항

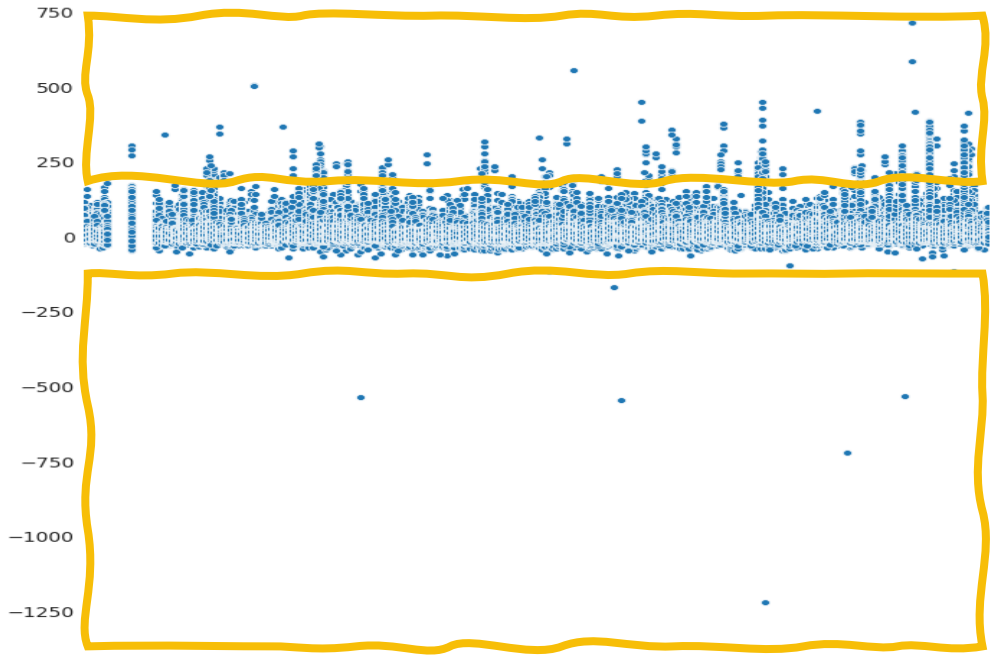
결항 항공편은 지연 시간에 영향을 주지 않는다. 그러므로 정확한 지연 모델을 만들기 위해 결항 항공편을 모두 제거하였다.

Part 2

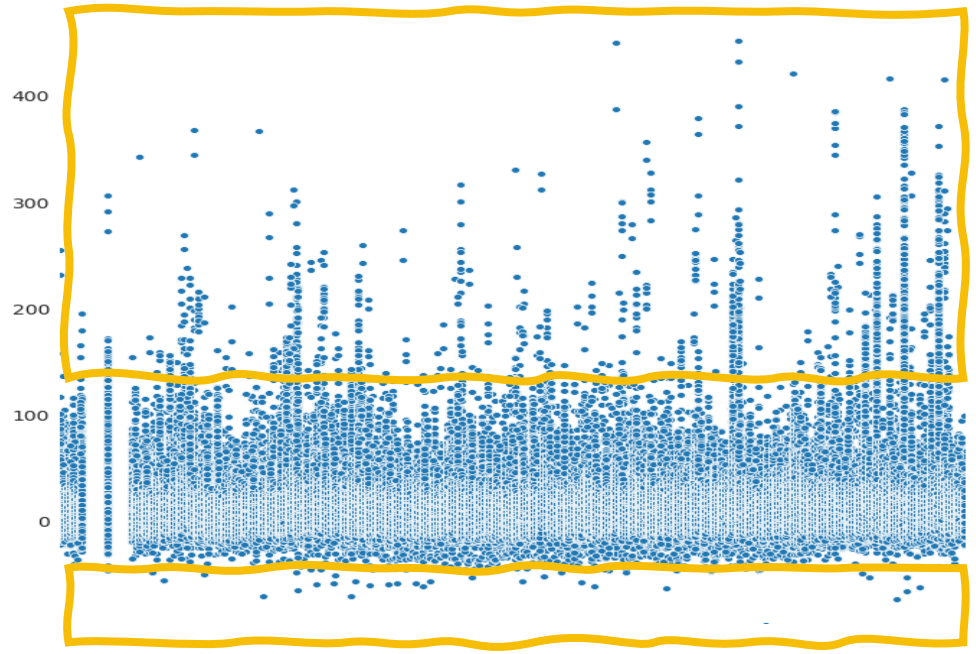
EDA

EDA는 시각적 방법을 통해 주어진 데이터 세트의 주요한 특징을 분석하고 요약하는 과정이다.
즉, 정규 모델링 혹은 가설(Hypothesis) 테스트 작업 전에 데이터가 말해줄 수 있는 부분들을
확인하는 것이다. 우리는 항공기 지연에 영향을 주는 변수들을 발견하기 위해 비정기편과 결항을 제외한
항공편을 가지고 지연 항공편 수, 지연 시간, 지연 비율을 파악하였다.

이상치 발견(Outlier Detection)



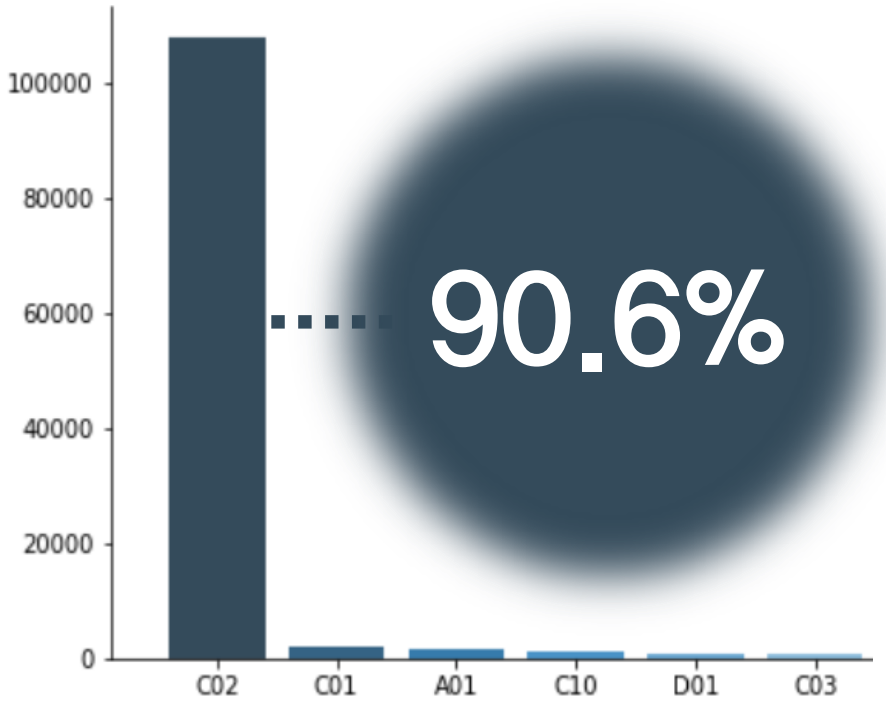
[실제 시각 – 계획 시각]

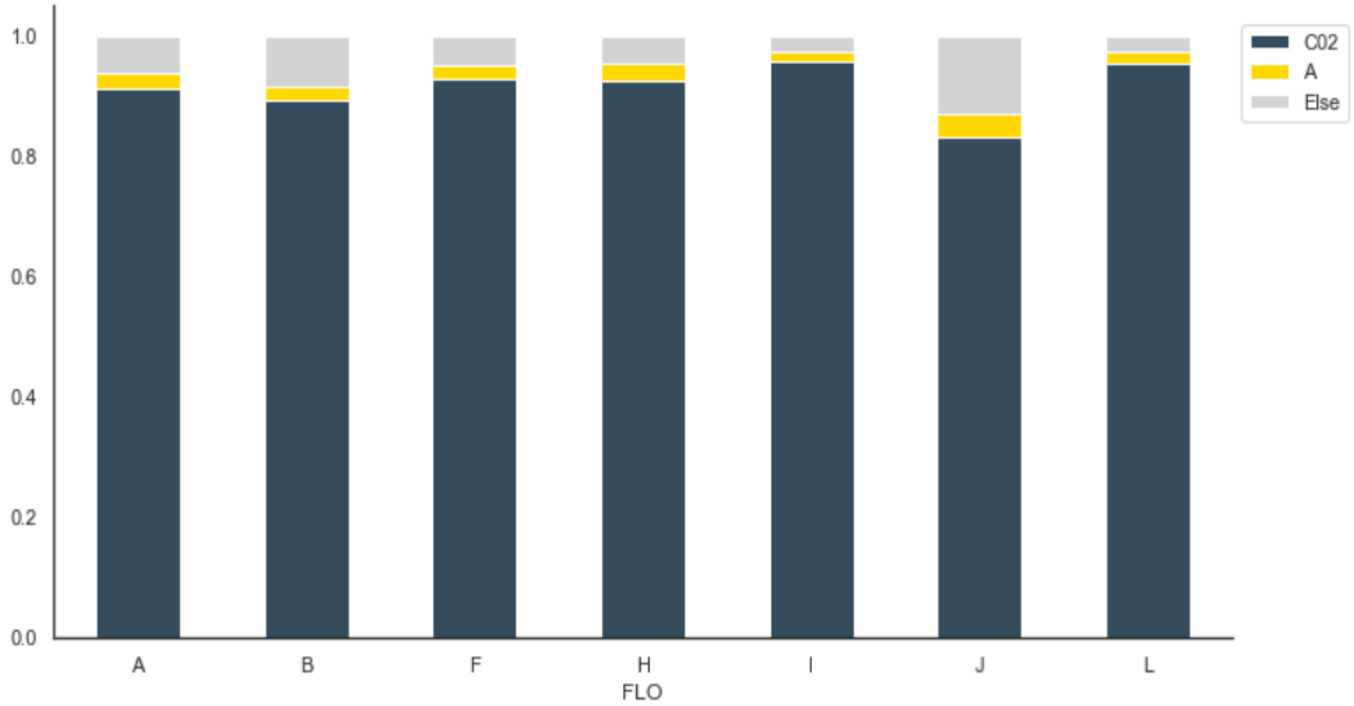


[지연 시간]

다양한 원인으로 인해 계획 시각과 실제 시각 간의 차이가 발생하게 된다. 지연 시간 데이터는 대체로 밀집되어 있으나 데이터들이 넓게 분포하기 때문에 이상치 처리의 필요성이 제기된다.

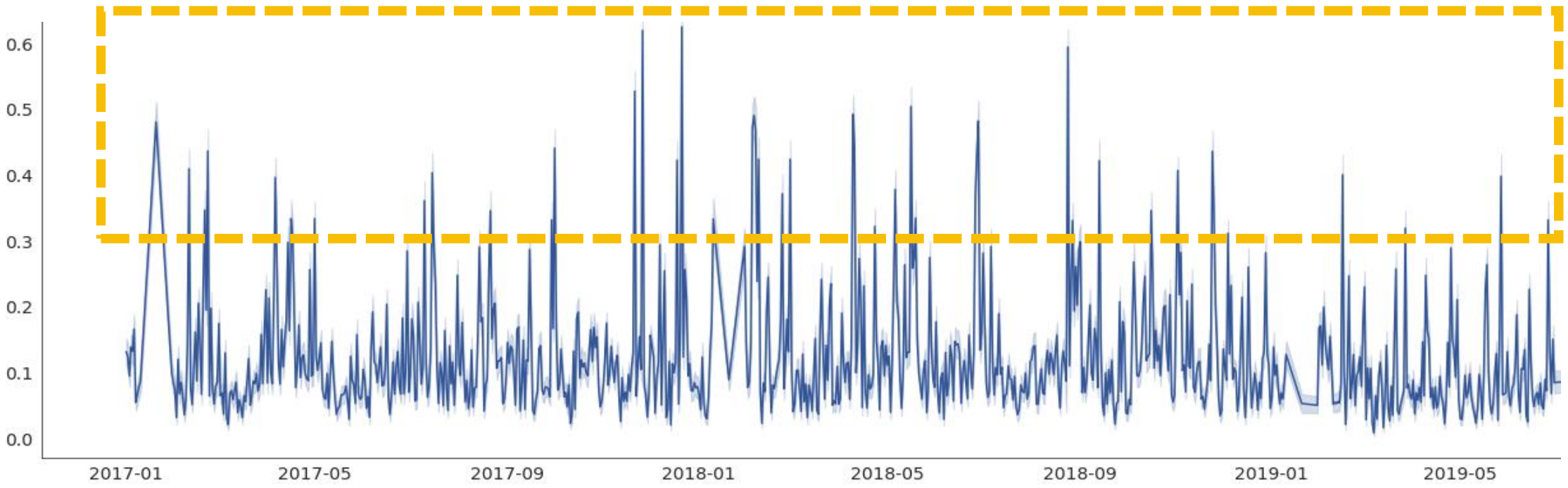
지연 요인 별 항공편 비교





전체 항공기 지연 요인 중 C02(AC접속)는 90% 이상을 차지한다. 항공사 별로 지연 요인 비율의 경우도 유사하다.
날씨로 인한 지연은 전체의 1% 정도를 차지한다. 그래서 C02로 인한 항공기 지연 원인을 중점적으로 분석하였다.

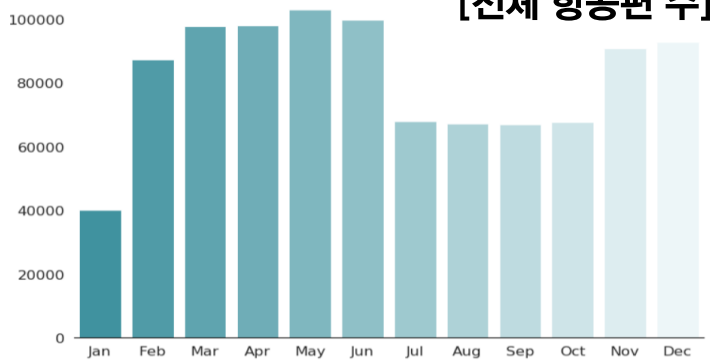
날짜 별 지연 비율



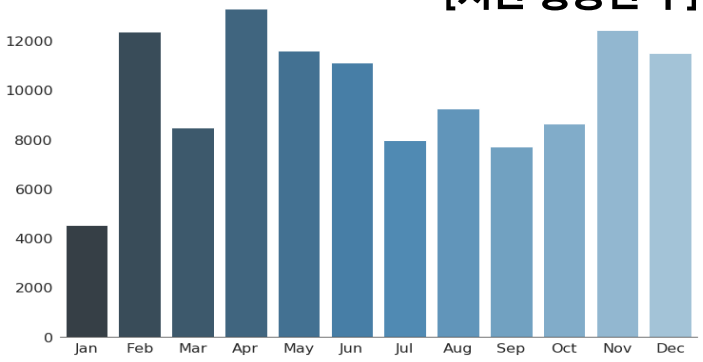
통시적으로 자료를 분석했을 때 특정 구간에서 지연율이 급증하거나 급감하는 현상이 발생하였다. 일반적으로 항공기 지연이 날씨 혹은 활주로 혼잡으로 인해 발생한다는 것을 고려했을 때, 해당 결과가 시계열적인 요소를 반영하는 것처럼 보인다.

월(月) 별 지연 비율

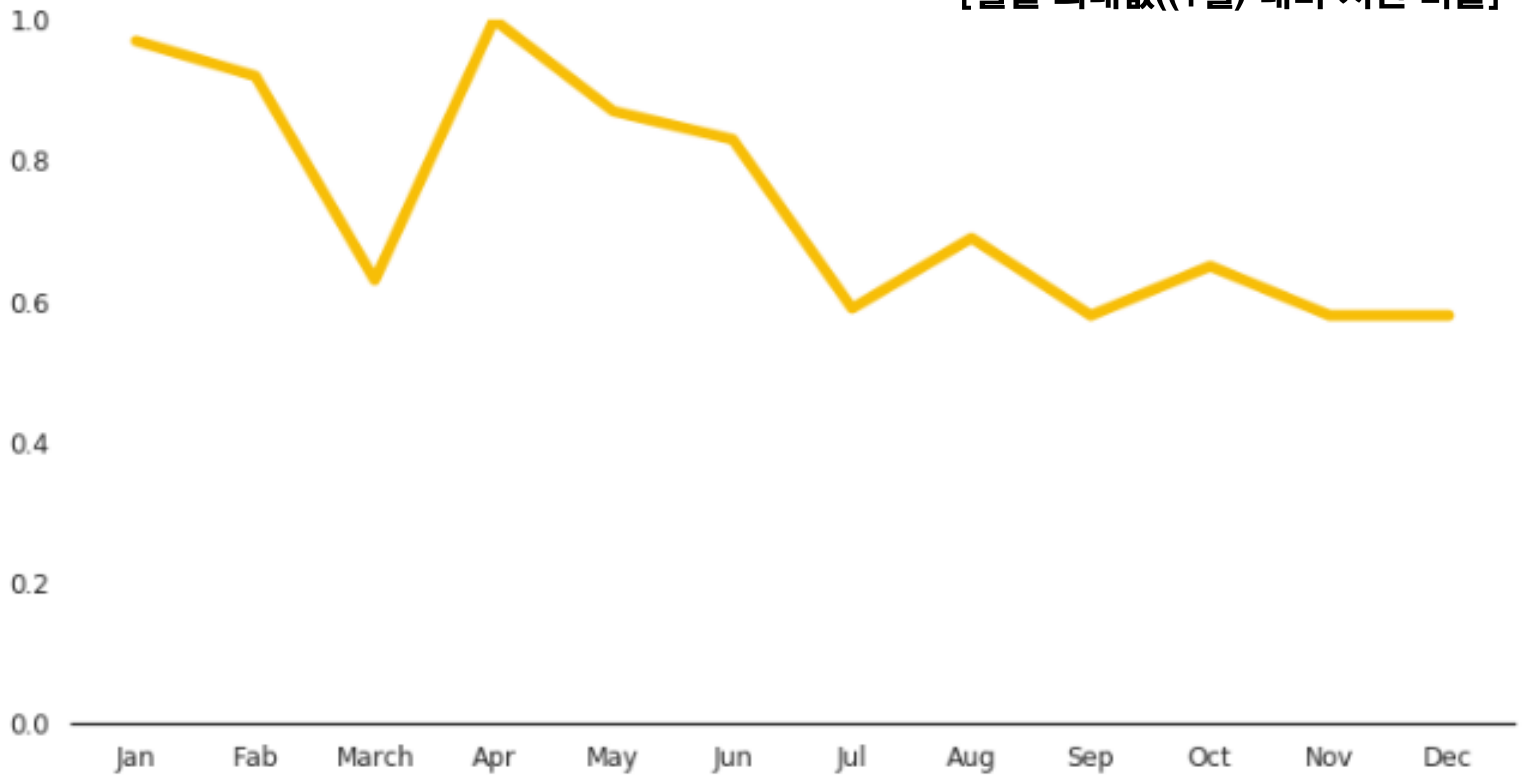
[전체 항공편 수]



[지연 항공편 수]



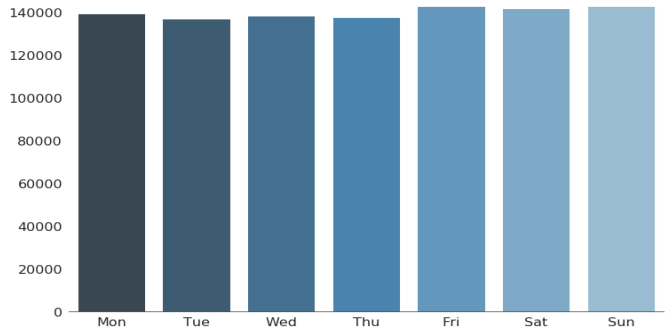
[월별 최대값((1월) 대비 지연 비율]



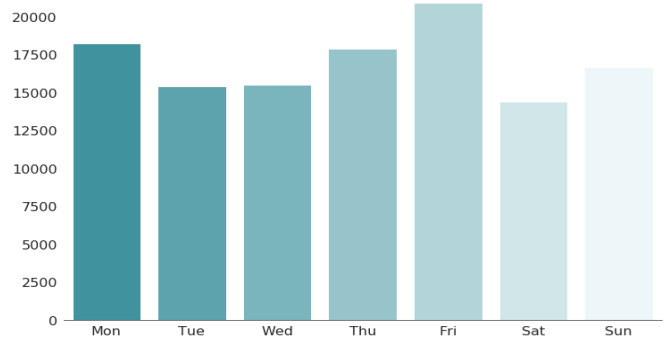
월 별 지연 항공편 수를 분석했을 때, 특정 월에 따라 지연 비율에 편차가 존재한다.

요일 별 지연 비율

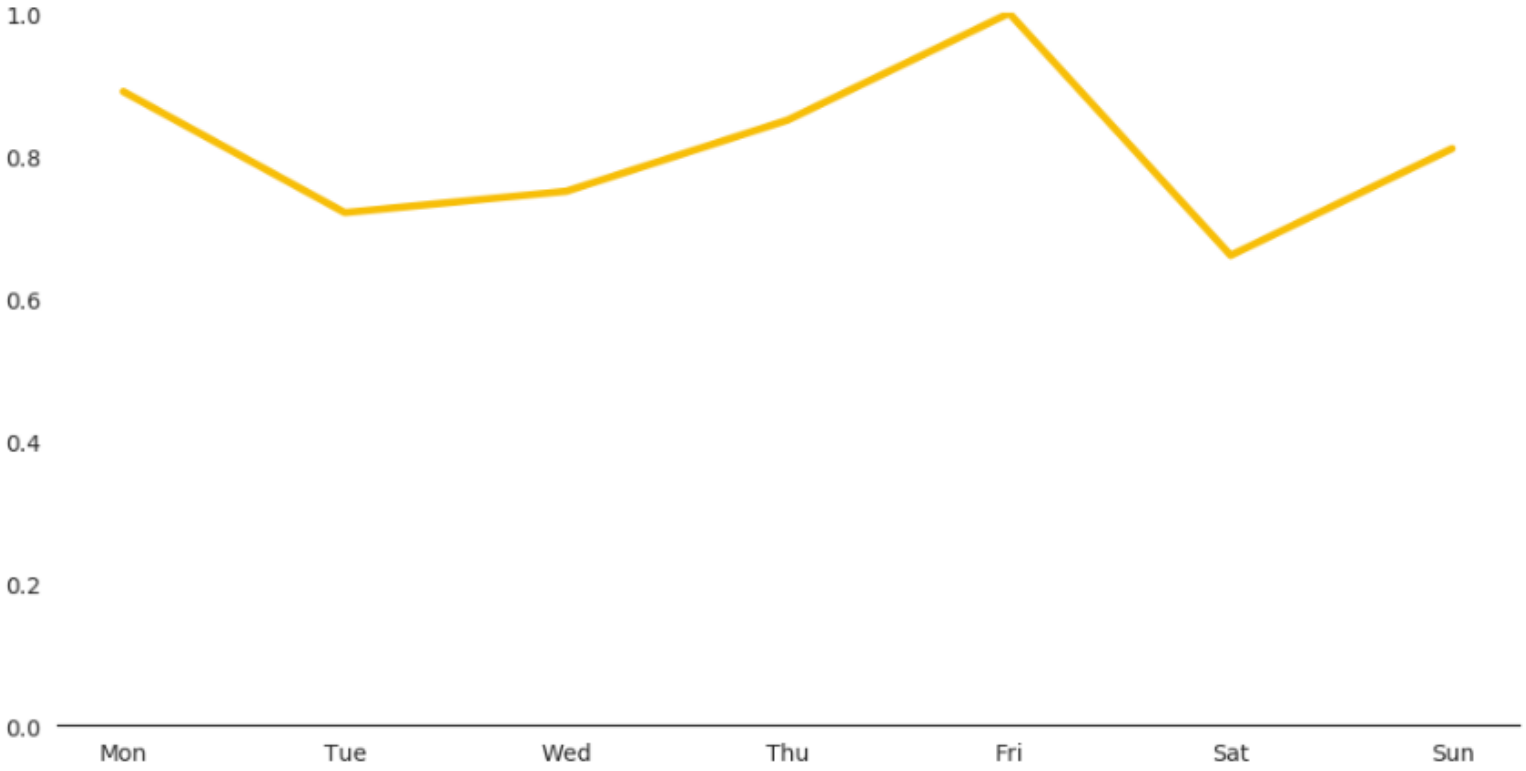
[전체 항공편 수]



[지연 항공편 수]



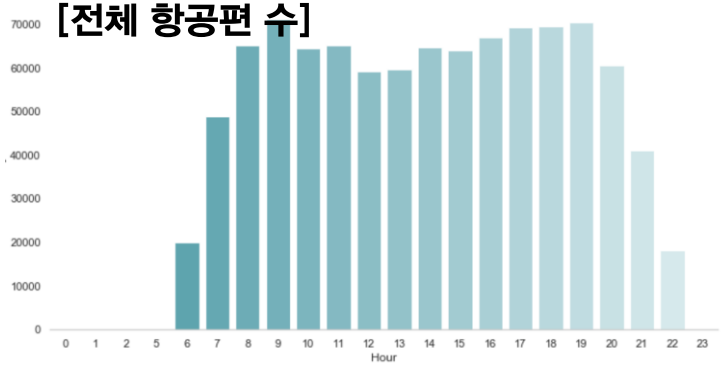
[요일 별 최대값(금요일) 대비 항공편 지연 비율]



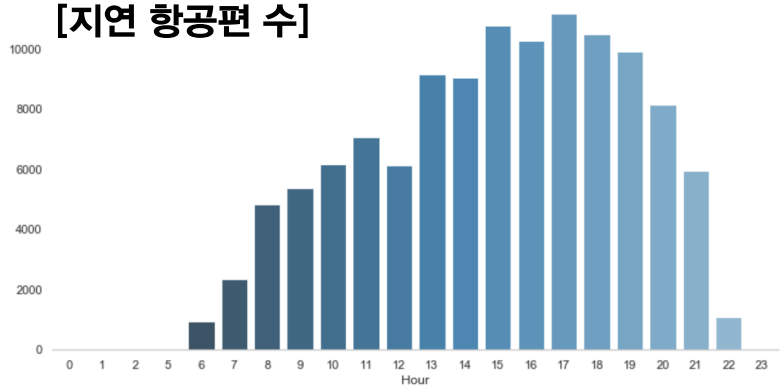
운항 항공편 수는 요일 별로 비슷하고 지연 항공편 수의 차이 또한 크지 않다. 그래서 월 별 지연 비율과 비교했을 때 상대적으로 변동률이 낮게 나타난다.

시간 별 지연 비율

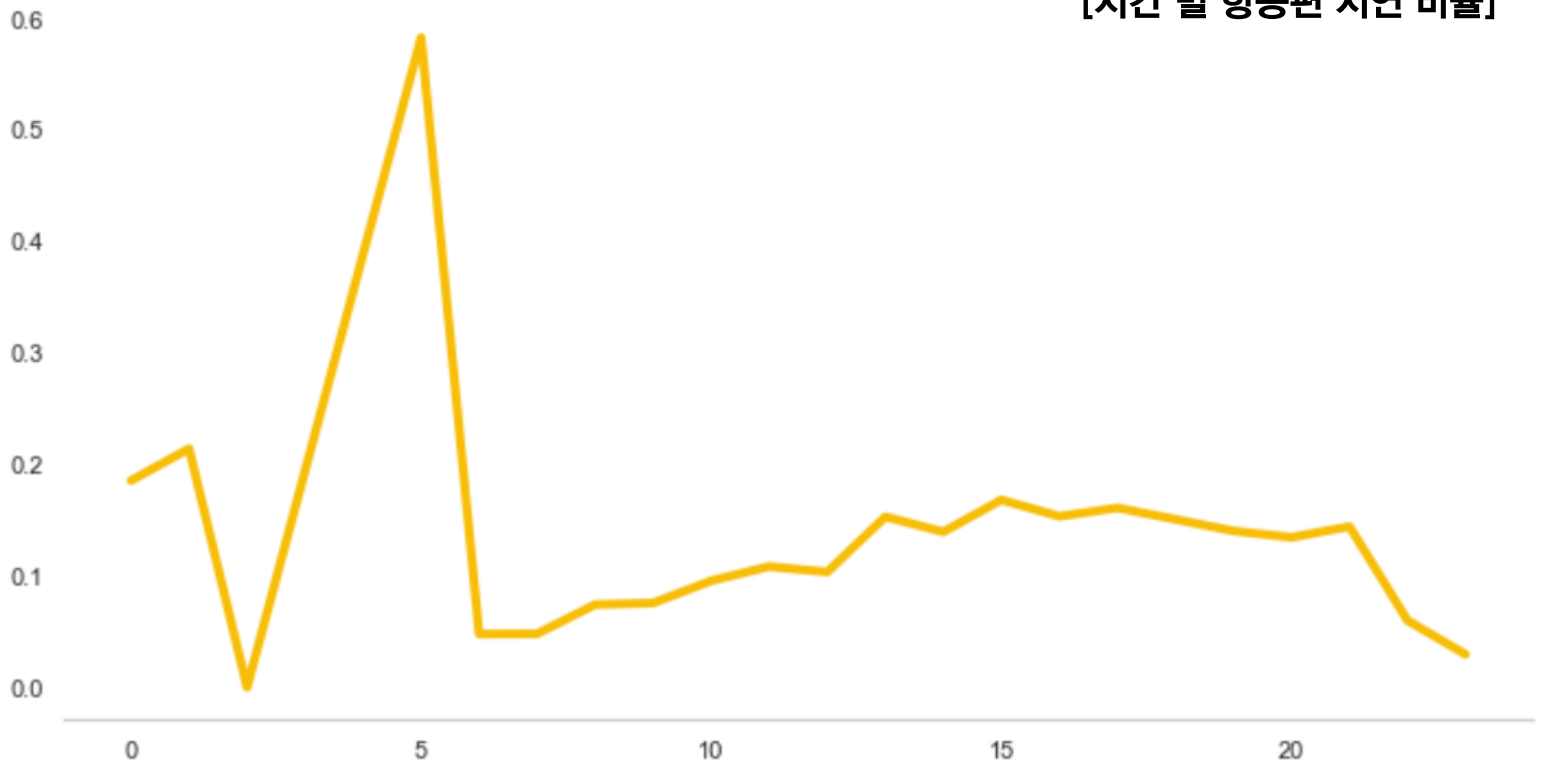
[전체 항공편 수]



[지연 항공편 수]



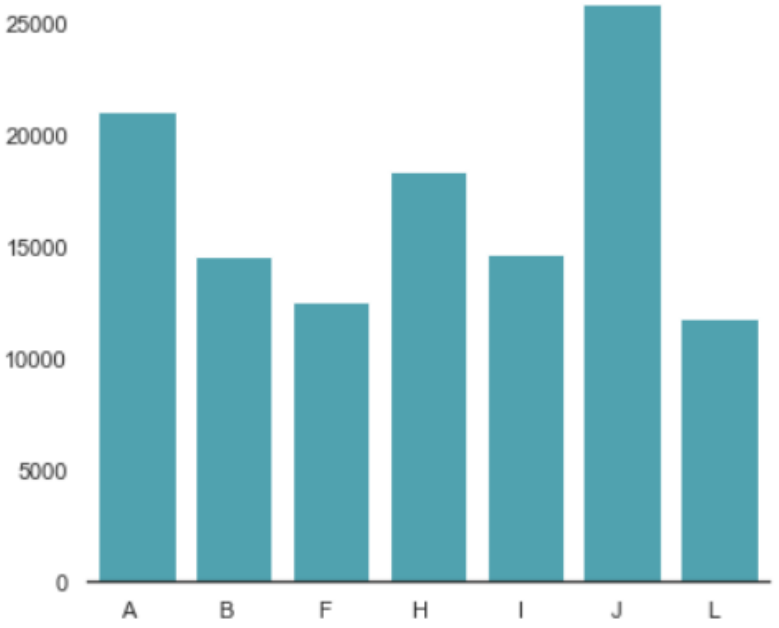
[시간 별 항공편 지연 비율]



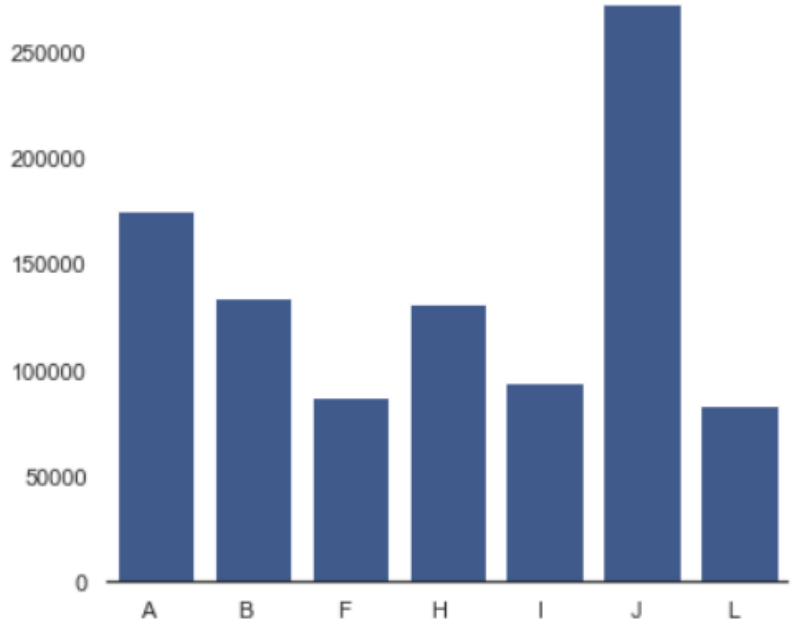
계획 시간 대 별로 항공기 지연을 분석한 결과, 특정 시간대(오후 15 ~ 18시)에서 지연율이 높게 나타났다.
새벽 시간대(0 ~ 5시)에서 지연율이 급격히 높아지고 있으나 운항 항공편 수에서 차이가 많이 나고 비정기편이 많았다.

항공사 별 지연 비율

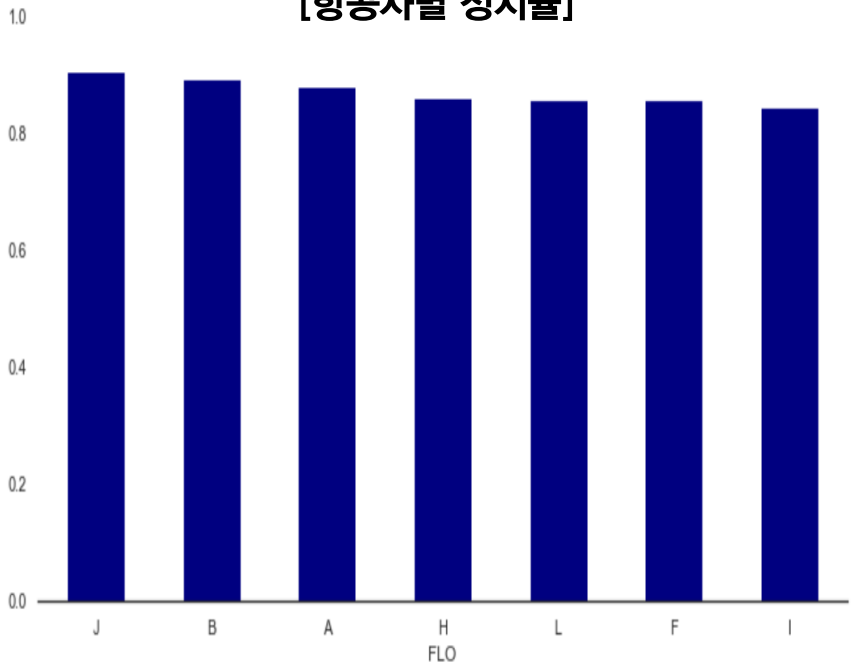
[지연 항공편 수]



[전체 항공편 수]



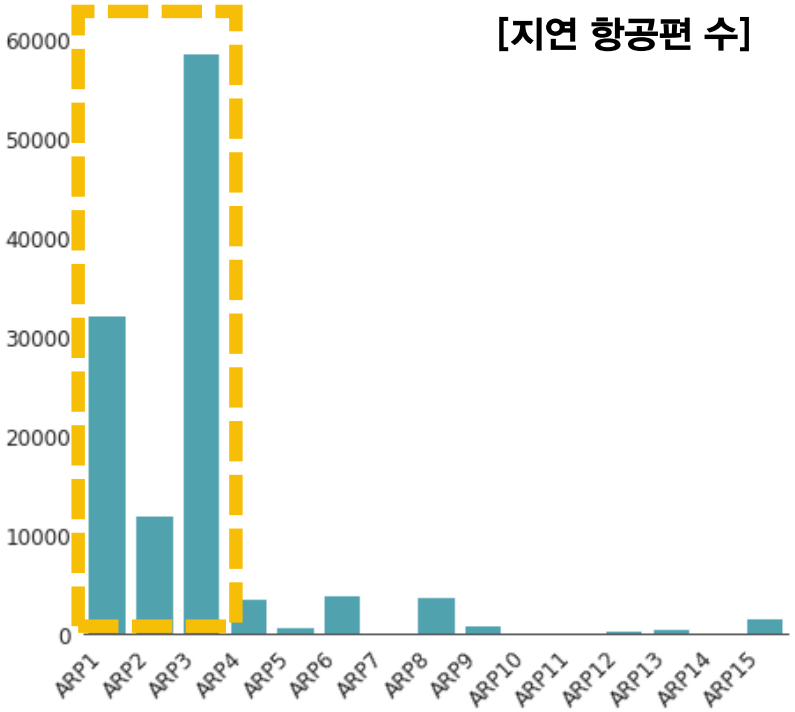
[항공사별 정시율]



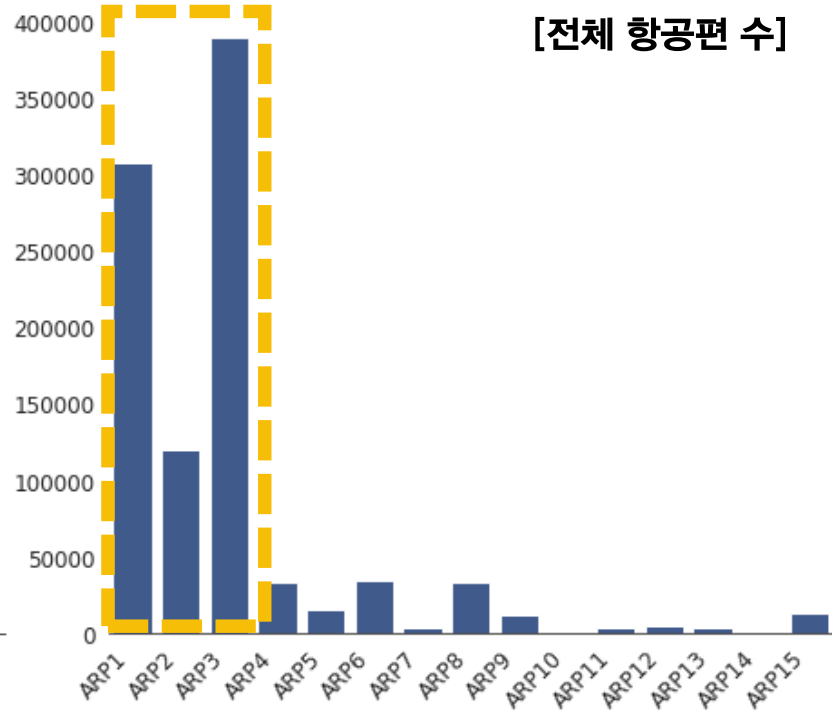
대체적으로 항공사별로 전체 항공편 수가 많을 수록 지연되는 항공편 수 또한 많아진다. 지연 비율로 보았을 때에는 큰 차이가 없었다.

공항 별 지연 비율

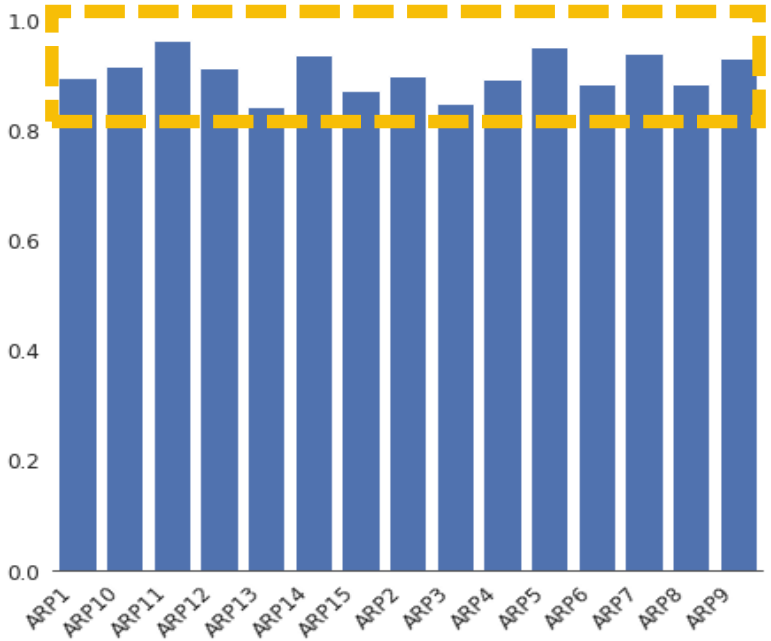
[지연 항공편 수]



[전체 항공편 수]



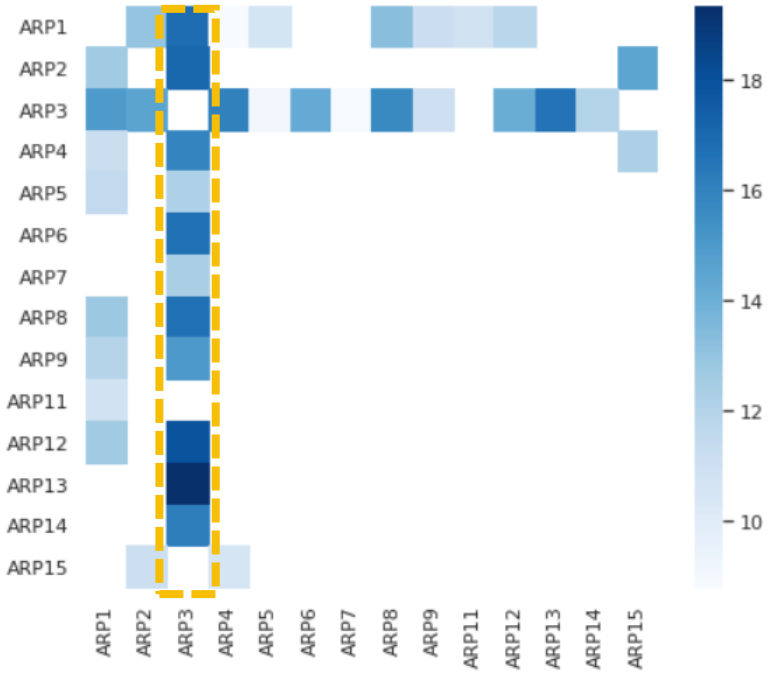
[항공편의 정시율]



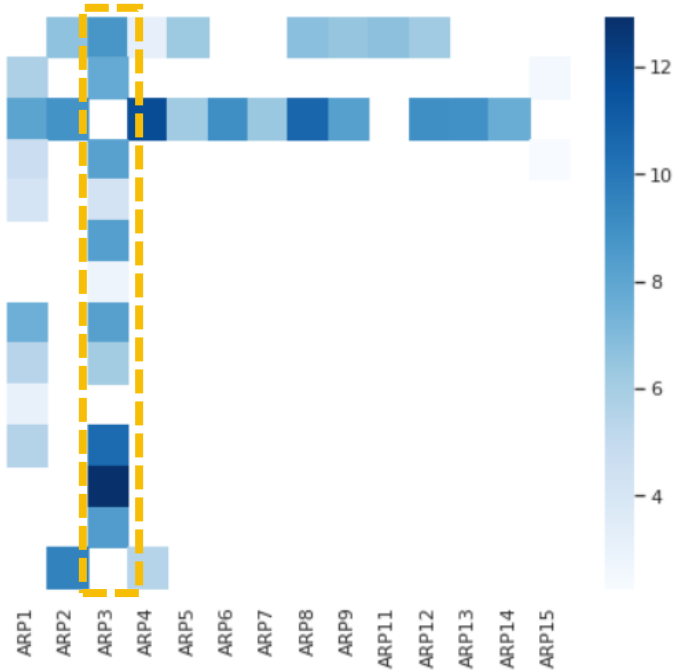
지연 항공편은 공항별로 큰 차이를 보인다. 전체 15 개의 공항 중 3 개의 공항에서 항공기 지연이 집중되는 한편,
지연 항공편 수는 기간 내 운항한 전체 항공편 수에 비례하는 양상을 띄고 있다.
공항 별 정시율은 대체로 유사하나 지연이 집중되는 공항들의 경우 유독 낮게 나타났다.

출/도착 공항 별 평균 지연 시간

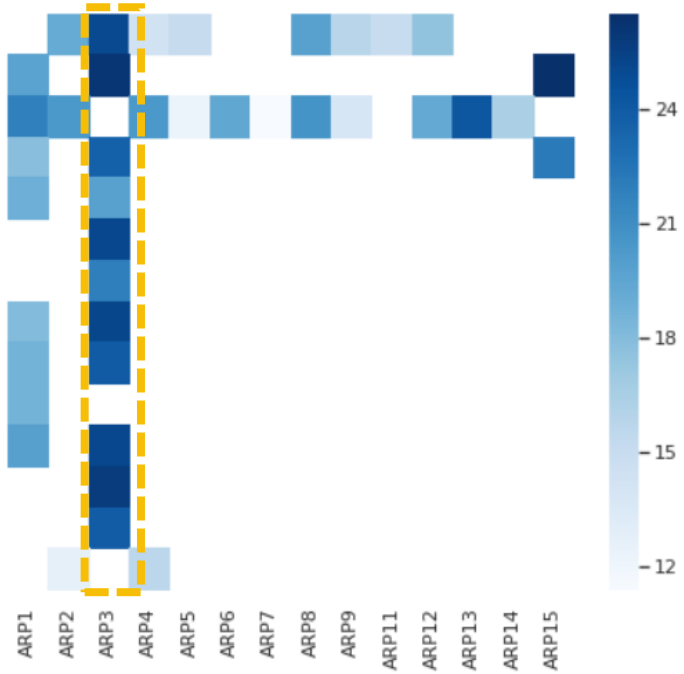
[출발 항공편]



[도착 항공편]



[전체 항공편]



평균적으로 도착 항공편보다 출발 항공편에서 지연 시간이 길게 나타난다. 특히, 히트맵에서 ARP3 공항에서 출발하거나 ARP3 공항으로 도착하는 항공편들이 비교적 더 길게 지연되는 경향을 확인할 수 있었다.

EDA 결과 요약

- C02 (AC접속)이 지연 요인 데이터 중에서 압도적으로 많았고 다른 요인들은 비슷한 양상을 보였기 때문에 분류 모델이 이를 잘 학습하게 만드는 것이 중요하다고 생각하였다.

- 날씨(Weather) 데이터의 경우 일부 공항에 적용가능한 METAR 값이 결여되어 있다. 평가 기간의 METAR 데이터가 사전에 주어지지 않았기 때문에 모델에 입력 변수로 줄 수 없다고 생각했다.

- ARP의 경우 EDA를 통해 각 항공사 혹은 공항에서 전체 항공편 수가 많으면 지연 항공편 수도 많아진다는 것을 볼 수 있었다. 또한 날짜(Date) 데이터는 월, 요일, 계획 시간에 따라 항공기 지연 양상이 뚜렷이 구분되었다. 이러한 결과를 바탕으로 같은 Date에 운항하고 같은 공항(ARP)을 사용하는 항공편 수를 묶어 피처로 만들었다.

- 계획 시각(STT)에 따라 항공기 지연 비율은 변동폭이 크다는 것을 EDA로부터 확인할 수 있었다. 또한 계획 시각(STT)와 실제 시각(ATT)의 차이가 30분 이상이면 항공기를 지연으로 판단한다는 점으로부터 STT는 항공기 지연 예측에 중요하게 작용하는 변수라는 결론을 내렸다.

Part 3

Feature Engineering

AFSNT.csv

STT 가공

STT 에서 시간과 분을 분리하고 분 단위의 실수로 변환하였다.

INT 에서 백분위 99% (91분 이상) 은 91분으로 통일하고 음수값은 모두 0으로 처리했다.

값이 너무 클 경우 예측 성능이 떨어지므로 60(분)으로 나눠서 Re-scaling

(시간은 정수부분으로, 분 단위는 소수부분으로 들어간다.)

Flights & Slots

공항의 하루 총항공편 개수를 flights라 했다. 같은 Date와 ARP 내에 해당하는 항공편 수를 더하여 계산했다.

공항의 수용 능력을 고려한 슬롯 데이터가 모든 공항에 대하여 존재하지 않는다. Slot의 의미를 살려 같은 Date와 ARP에 내에서 STT를 20등분으로 나눈 것을 slots라 했다

AFSNT.csv

지연 여부 & 출도착 (DLY & AOD)	지연된 경우 '1', 지연되지 않은 경우 '0'으로 환산하여 스케일링 출발의 경우 '1', 도착의 경우 '0'으로 환산하여 스케일링
출발 & 도착공항 (ARP & ODP)	각각의 공항들을 One hot encoding을 사용하여 Numerical Value로 변환
월(Month) 요일(Days of Week)	월의 계절성과 요일의 주기성을 반영하기 위해서 월과 요일은 One-hot Encoding 하지 않았다.

Part 4

Modeling

Gradient Boost Machine

부스팅 알고리즘(GBM)은 약한 학습기(Weak Learner)를 순차적으로 학습–예측하면서 잘못 예측한 데이터에 가중치 부여를 통해 오류를 개선해 나가는 방식이다. GBM의 가장 큰 특징은 가중치 업데이트를 경사하강법 (Gradient Descent)을 이용하는 것이다.

의사 결정나무에 기반한 Ensemble 기법은 안정적인 성능을 내고, 과적합될 가능성이 적다. 그 중, Boosting 기법은 클래스 비율의 차이로 인한 편향을 줄일 수 있다.

Model Trial

1. Boosting 계열을 활용한 모델
   - XGBoost Model
   - Light GBM Model
   - CatBoost Model
2. 신경망 구조를 활용한 예측 모델
   - RNN

Boosting 계열

모델 선택 이유

AFSNT 데이터 전처리 과정에서 카테고리형 변수들이 다수 생성되었다.
다양한 머신러닝 알고리즘 중 카테고리형 변수를 분석하는 데 성능이 좋은 Boosting 계열을 선택했다.
LGBM, XGBoost, CatBoost 등을 시도해본 후, 최종적으로 조금 더 좋은 성능을 내는 CatBoost를 사용하기로 했다.

한편, EDA를 통해 C02가 항공기 지연 사유 중 가장 큰 비율을 차지했었다. 따라서 LGBM에서 DRR이 C02인 데이터들을 추출하여(AFSNT_C02.csv) 피처 분석을 진행했다.

How to apply LGBM & XGBoost

Step 1: train, test, validation set 나누기 & 가중치 설정

2018년 9월 15일~2018년 10월 1일까지의 데이터를 test set(X_test)으로, 나머지 데이터를 train set(X_train) 으로 만들었다.

x_train, x_test에서 AOD는 D인 데이터는 1로, A는 0으로 설정했다. 그리고 ARP, ODP, FLO를 One-hot Encoding 하여 카테고리형 변수로 만들었다.

y_train, y_test에서 DLY는 Binary 로 처리하였다. DLY의 데이터 비율이 매우 불균형 하기 때문에 DLY 가 Y인 데이터에 **가중치(weight_2 : num(DLY==N)/num(DLY==Y))**를 주었다.

Step 2: Stratified K fold 진행 (Cross Validation)

3번으로 나누어서 Stratified K-fold를 진행한다. 각 iteration 에 x_train, y_train 의 1/3 만큼 사용하고 이를 validation set으로 활용한다. Data Imbalance 문제를 해결하기 위해 **stratify = y_train set 내에서 3 fold data set을 설정**하여 모든 class가 일정한 비율을 갖게 만든다.

이 단계를 통해 lgbmodel_final_{} 모델을 생성한다.

Step 3: 예측 값(y_pred) 찾기

3번의 iteration 을 통해 얻은 lgbmodel_final_{}을 합쳐 model_list를 생성한다. 그리고 함수 predict_proba를 이용해 model_list. 각 피처의 확률값(probability)을 추출하고 array로 만든 다음, 확률값이 가장 큰 피처의 index를 뽑아서 y_pred 를 생성한다. 이후 y_test와 y_pred를 이용해 **Confusion Matrix 를 생성**한다.

How to apply CatBoost

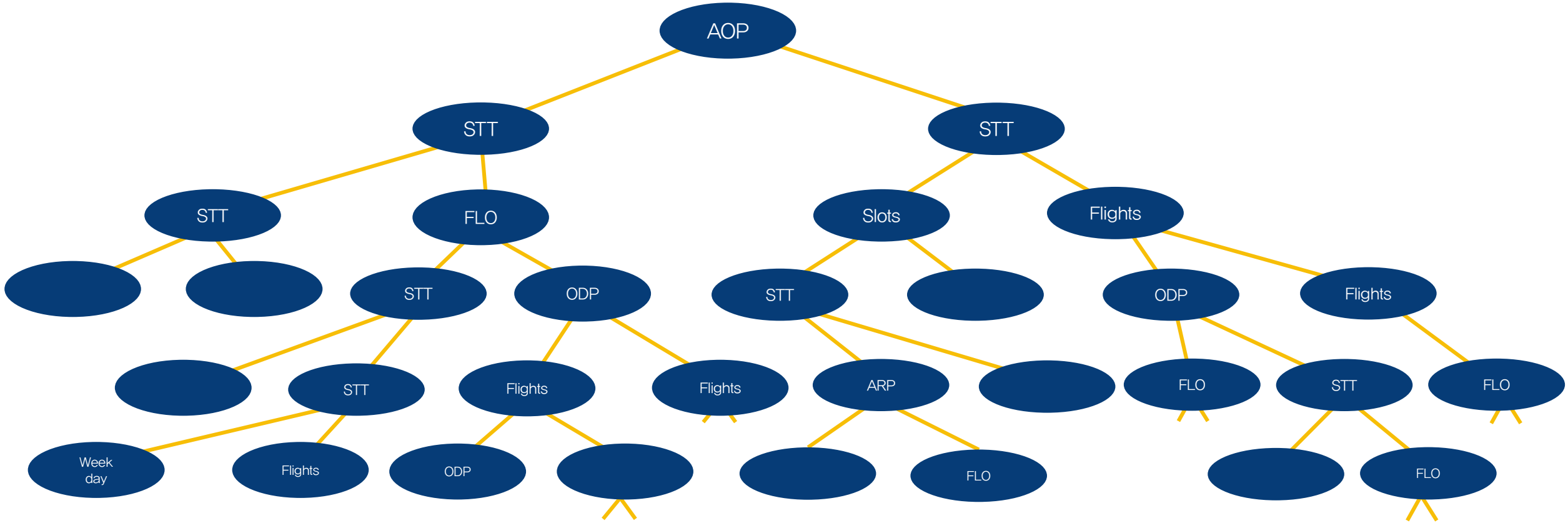
Step 1: Preprocessing

CatBoost는 **자동적으로 카테고리형 변수를 인식**하기 때문에 이전 모델에서 인코딩한 전처리 파일이 아닌, 원본 파일을 불러와 간단히 전처리한 뒤 학습하였다. 전처리에는 날짜 열 생성, 모델이 학습하지 않는 데이터(결항, 등록기호, 소수 공항 및 편명 등)와 이상치(백분위 상하 1%)를 제거하였다. 그리고 학습에 사용할 변수(ARP, ODP, FLO, AOD, IRR, Month, Weekdays)를 문자형으로 바꾸었다. 단, 계획 시간의 경우 학습을 위해 실수형 변수로 치환하였다.

Step 2: train, test, validation set 나누기 & 가중치 설정

Train_set과 Test_set을 나눈 후 validation set으로 한 번 더 Train_set을 나눠주었다. 이전 모델과 마찬가지로 Data Imbalance를 해결하기 위해 동일한 가중치를 두고 학습하였다. 학습에 사용된 estimator는 10000 개, **Early_Stopping_Round를 500으로 두어** 모델의 성능이 더 이상 나아지지 않는 경우, 학습을 조기 종료한 뒤 Confusion Matrix 와 ROC Curve를 추출하였다.

모델 도식화


AOP
STT
STT
STT
FLO
Slots
Flights
STT
ODP
STT
ODP
Flights
STT
Flights
Flights
ARP
FLO
STT
FLO
Week day
Flights
ODP
FLO
FLO

모델 학습 결과

모 델 명	Accuracy	Recall	AUC
Light GBM Model	74.75%	70.10%	72.62%
CatBoost Model	70.60%	69.75%	77.22%
XGBoost Model	72.80%	69.01%	71.06%

신경망 구조 (RNN)

모델 선택 이유

RNN은 신경망 구조의 한 형태로서, 시계열 (Sequence)의 형태를 가지는 데이터를 활용하여 은닉층(Hidden layer)의 값을 전달하고 이를 기억한다. 그리고 다음 단계의 네트워크가 저장된 정보를 사용하여 결과값을 예측한다.
LightGBM은 categorical variable을 처리하는 데는 용이하지만, Sequential variable을 처리할 때는 한계가 있다.
반면 항공데이터에서 각각의 항공편은 날짜 및 노선에 따라 달리 구성된다. 따라서 우리는 이를 연결편으로 묶어 파악하고자 하였으므로 RNN 모델을 대조군으로 설정하고 시도해 보았다.

How to apply RNN

Step 1: 연결편 추출

추출 과정에 날짜(Date), 편명(FLT), 등록기호(REG) 등을 사용해 같은 항공편들을 분류하고 출도착(AOD), 예정시각(STT), 실제시각(ATT) 데이터를 가지고 정규표현식을 활용해 총 58 가지의 연결편으로 분류하였다.

Date	ARP	FLT	REG	IRR	STT	ATT
2017 – 01– 01	ARP3	J1955	SEw3NzE4	N	10:05	10:32
2017 – 01– 01	ARP6	J1955	SEw3NzE4	N	11:10	11:18
2017 – 01– 01	ARP3	J1956	SEw3NzE4	N	11:45	12:11

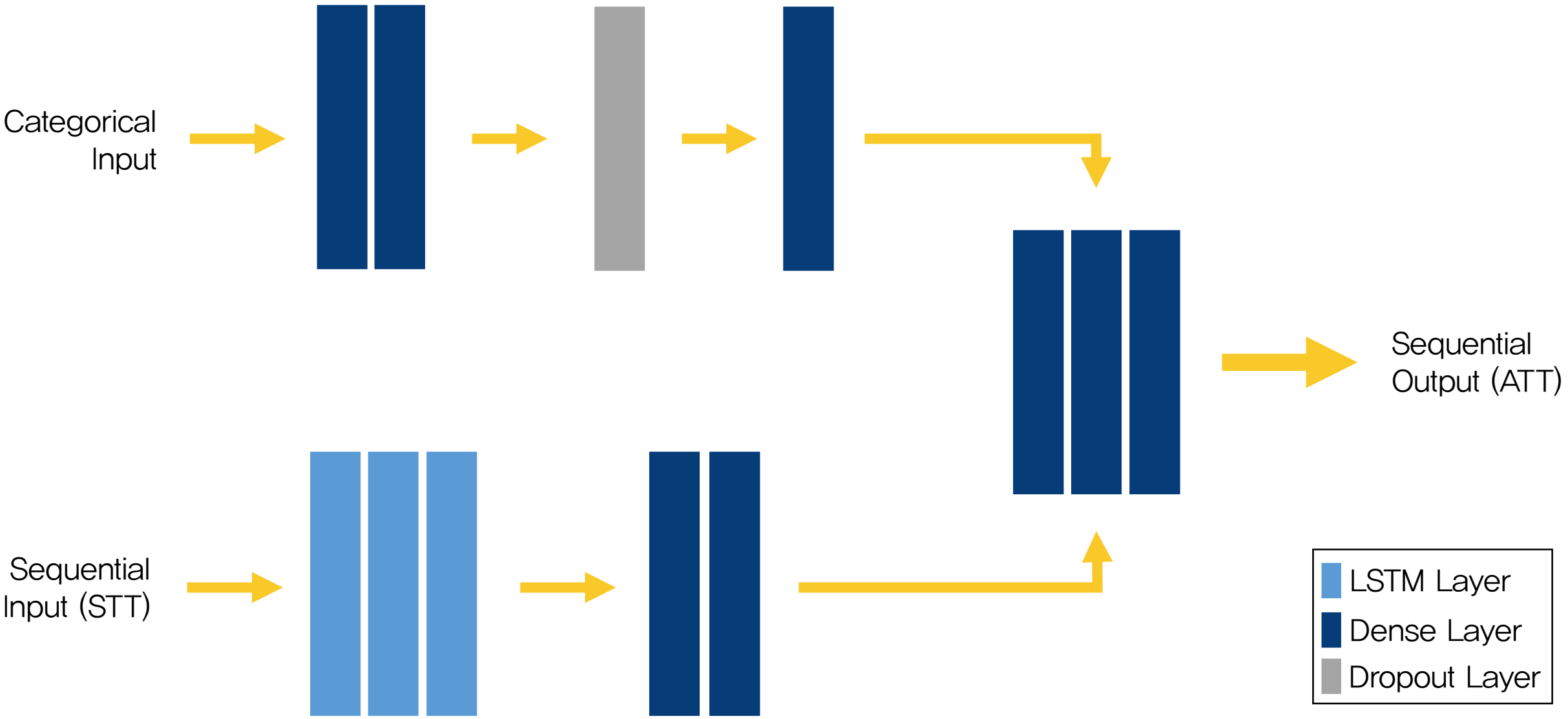
Step 2: 데이터를 Input & Output Sequential 데이터 형태로 변환한 뒤 모델 학습

Input 데이터로는 STT Sequence(float format)와 Feature Sequence(One–hot encoded features), Output 데이터에는 ATT Sequence(float format)와 Predict_label과 비교할 실제 레이블 값인 Label Sequence를 생성해 모델을 학습시킨다.

Step 3: Interval Predict를 활용해 레이블 예측

Interval Predict(ATT Test의 예측값 – STT Test)가 0.5를 넘어가면 Label test가 1, 그렇지 않으면 0으로 예측한다. 다음으로 실제 레이블과 예측 레이블을 비교하여 평가한다.

모델 구조화


Categorical Input
Sequential Input (STT)
Sequential Output (ATT)
LSTM Layer
Dense Layer
Dropout Layer

최종 학습 결과

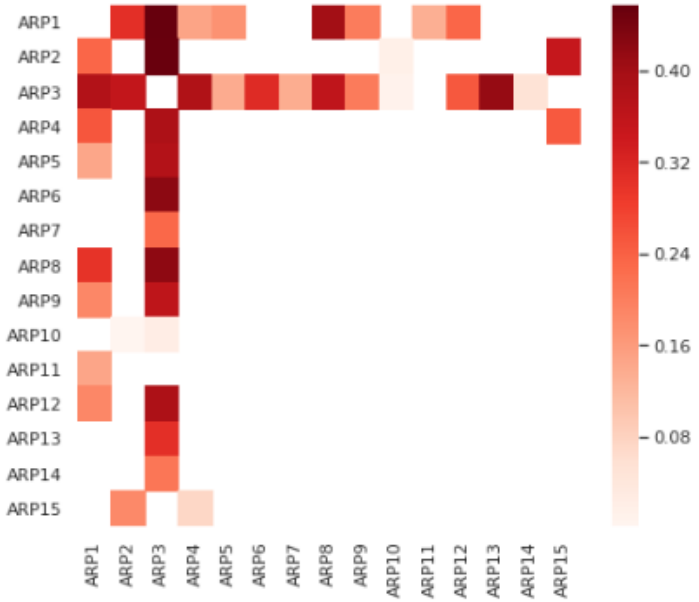
모 델 명	Accuracy	Recall	AUC
Light GBM Model	74.75%	70.10%	72.62%
CatBoost Model	**70.60%**	**69.75%**	**77.22%**
XGBoost Model	72.80%	69.01%	71.06%
RNN Model	72.65%	50.36%	–

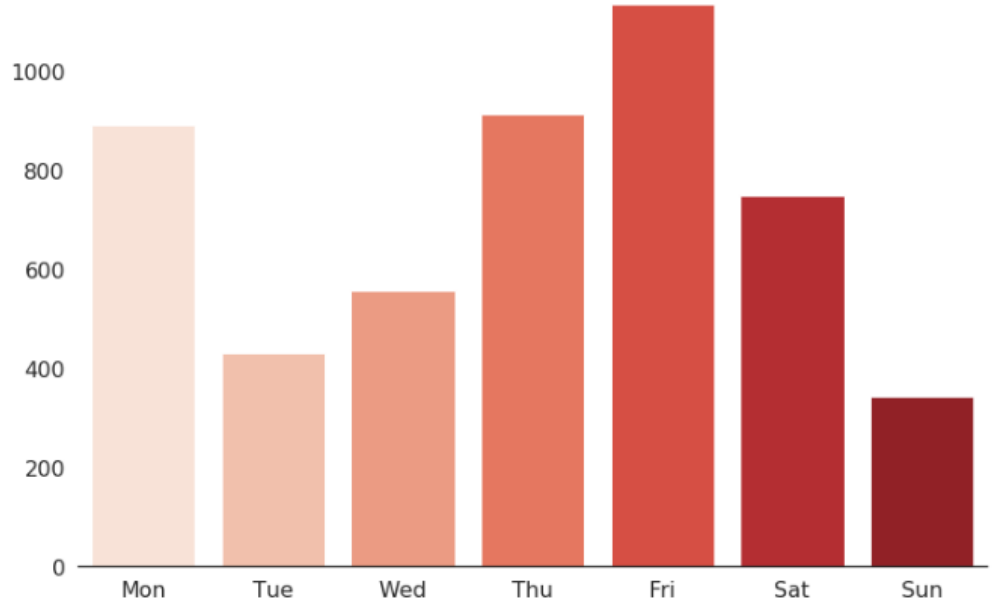
Part 5

Conclusion

평가 데이터 리뷰 (1)

ARP3(제주공항)에서 항공편이 지연될 가능성이 다른 공항에 비해 높게 나타난다. 또한, 우리 모델은 금요일에 더 많은 항공편이 지연될 것이라 예측하였다.

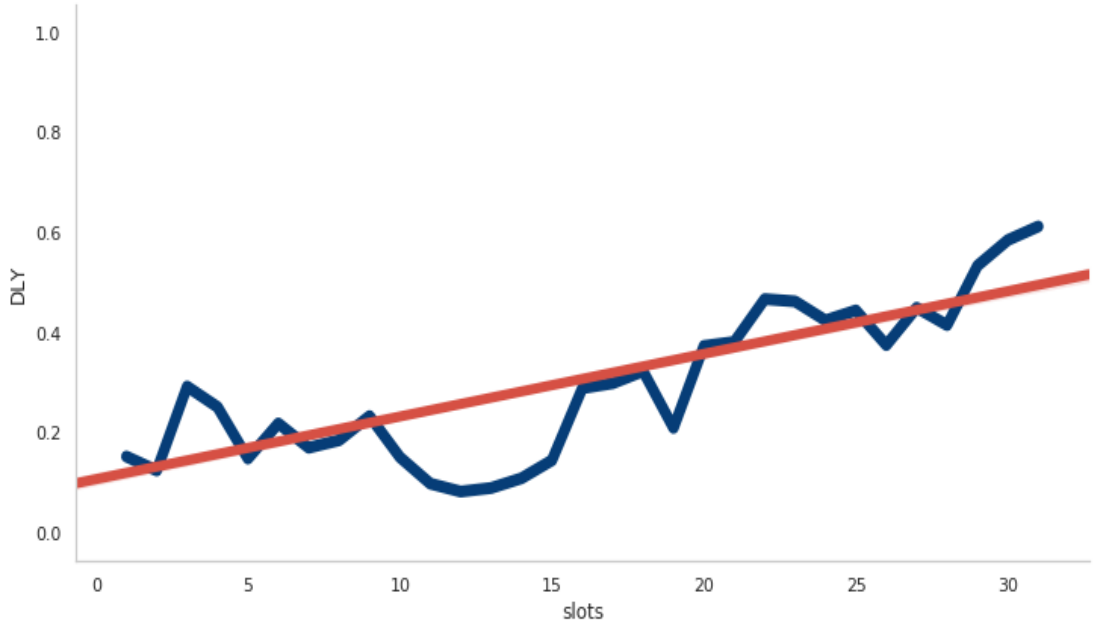


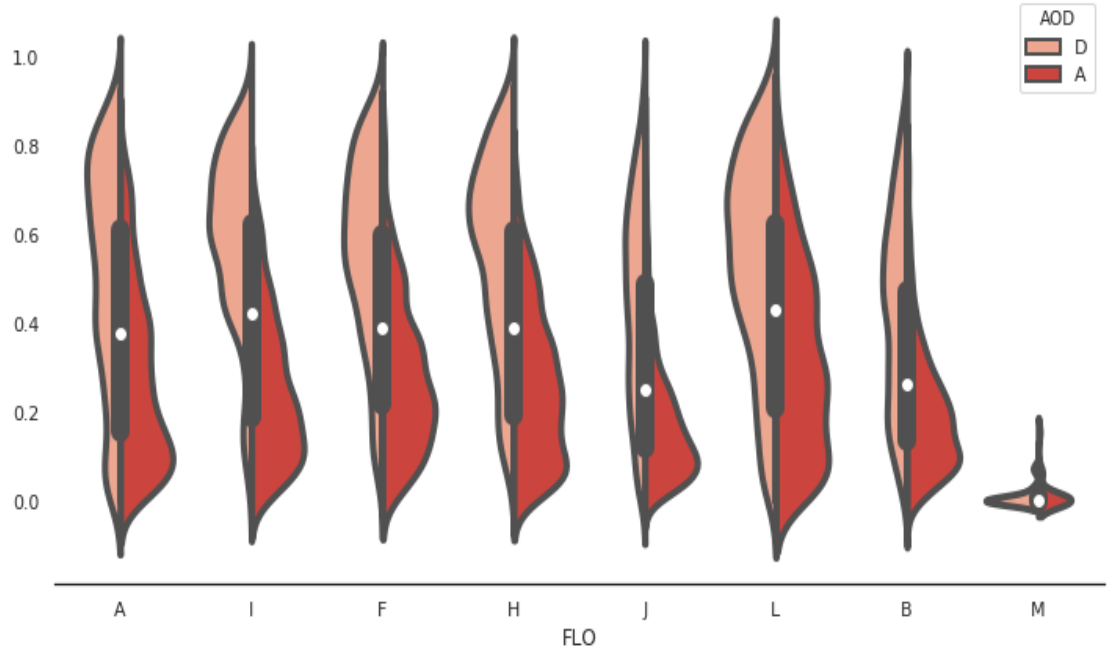


평가 데이터 리뷰 (2)

시간 당 출/도착 항공편 수가 증가할 수록 지연 가능성이 높아지는 경향을 보인다. 항공사에 따른 차이는 크지 않았으나 일부 항공사들(대한항공, 에어부산)은 출발보다 도착할 때 더 많이 지연될 것이라 예상된다.

전체적으로 우리 모델은 EDA에서 본 것과 같이 출발 항공편에서 더 많이 지연될 것이라 판단하였다.





지연 요인 종합

공항의 최대 수용 능력은 정해져 있는 반면 이용객들의 항공편 수요는 시간에 따라 유동적으로 변한다.
수요가 많은 노선의 경우 항공사들이 항공기를 증편하게 되는데, 항공편의 수가 공항의 수용 능력을 초과할 때 항로 혼잡으로 인해 항공기 지연 문제가 빈번히 발생한다고 볼 수 있다. 또한 보딩 타임(Boarding Time) 등의 이유로 출발할 때보다 도착할 때가 지연 가능성이 높은 것으로 나타났다. 우리 모델은 **금요일에 제주공항에서 시간 당 항공편 수가 많아져 C02(AC접속)으로 인한 지연이 많이 발생**하는 것이라 예상하였다.

지연율을 낮추려면

실제로 제주 공항에서 항공기 지연은 오래 전부터 공론화된 문제였다. 국토교통부에서 발간한 2018년 항로교통 서비스보고서에 따르면 제주공항의 2018년 지연율은 16.1%로 17년 대비 2.3%p 증가하여 전체 공항 중 가장 높은 지연율 상승을 보였다.

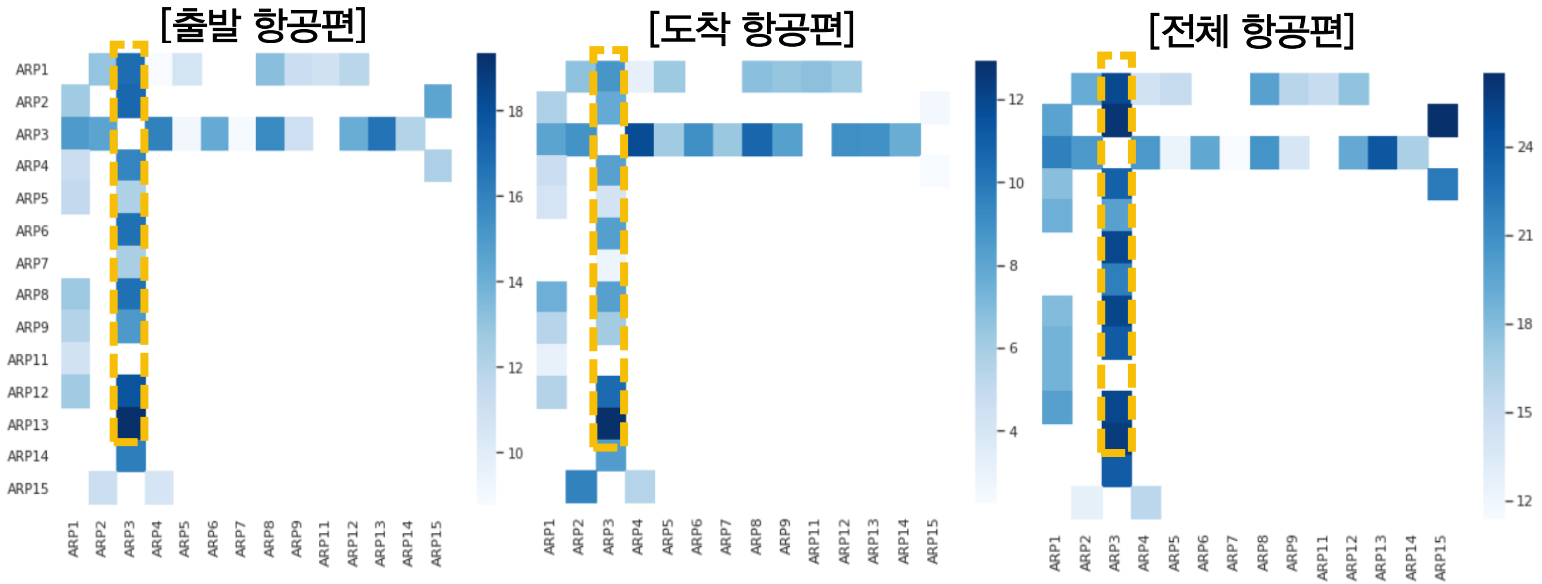
항공기 지연으로 인한 사회적 비용이 크다는 점에서 정시율을 높이는 방안이 시급하다. 지연율을 낮추기 위해서는 **제주공항의 활주로를 신설하여 공항의 수용 능력을 증가시키고, 각 항공사들과 연계하여 혼잡한 노선의 항공사 스케줄을 조정하는 대안이 필요하다.**

Part 6

Appendix

Appendix: 지연 원인(1) 공항

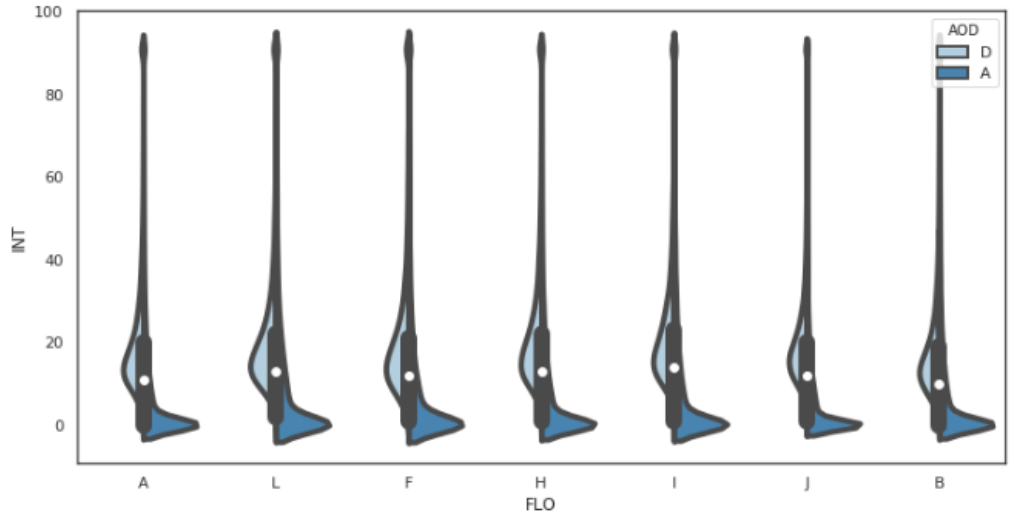
[EDA]



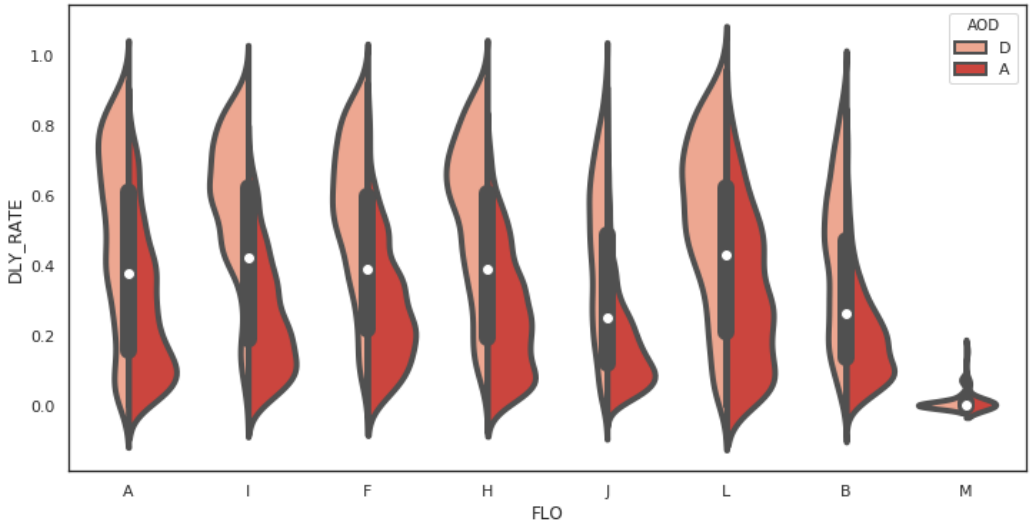
[평가파일]

Appendix: 지연 원인(2) AOD

[EDA]

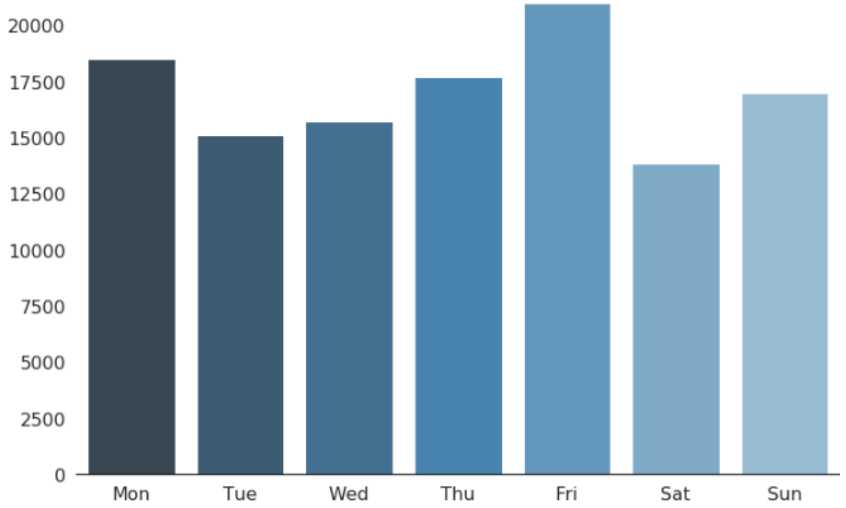


[평가파일]

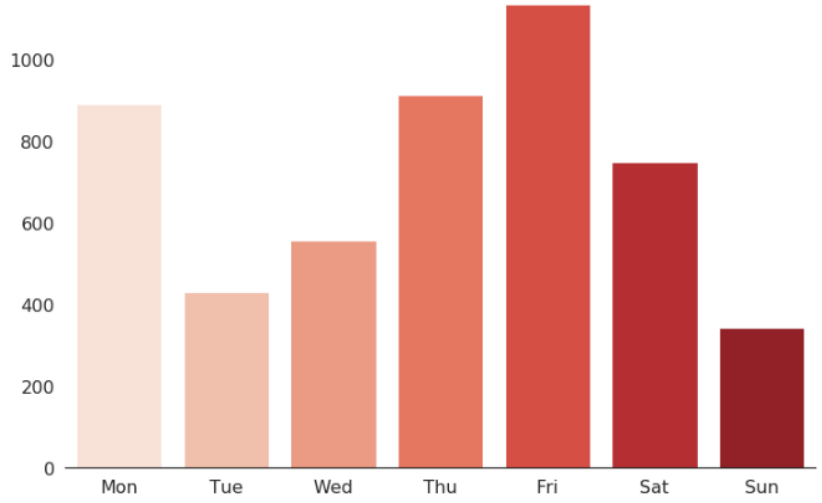


Appendix: 지연 원인(3) 요일

[EDA]



[평가파일]



Reference

Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016, September). A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)* (pp. 1-6). IEEE.


감사합니다