

데이터분석을 통한
항공여객기 연착 예측

AI07이기한

CONTENTS



Contents 01

데이터셋 선정 이유
기준모델



Contents 02

데이터 전처리



Contents 03

머신 러닝 적용



Contents 04

모델 해석

01

Contents 01

데이터셋 선정 이유

항공기 운항과 관련있는 여러 요소를 통해 항공기의 연착 유무를 예측.

직접적인 요인 : 비행거리, 항공기 수명, 항공기 좌석 수 등

간접적인 요인 : 날씨, 공항, 항공사 서비스 수준 등

어떤 요인이 얼마나 연착에 영향을 미치는지 확인하고 연착되지 않게 대비할 수 있음

기업의 비용 감소 효과를 기대할 수 있음

(고객유지비용 감소 등의 경제적 손실 감소)

01 Contents 01

기준모델

연착 0분 ~ 15분 : 0

연착 15분 이상 : 1

분류 학습의 기준 모델 : 최빈값

```
target='DEP_DEL15'  
  
df1[target].value_counts(normalize=True)
```

```
0    0.804612  
1    0.195388
```

불균형 클래스 평가지표는 정밀도, 재현율, f1, AUC 스코어까지 확인해볼 필요가 있음

02 Contents 02

데이터 전처리

data leakage 방지, 학습 시간 향상을 위해 '필요없는' 특성 삭제

DEPARTING_AIRPORT	LATITUDE	LONGITUDE	
Dallas Fort Worth Regional	32.894	-97.030	
Port Columbus International	39.991	-82.878	
Savannah/Hilton Head International	32.127	-81.202	
Douglas Municipal	35.219	-80.936	
Houston Intercontinental	29.983	-95.340	AIRPORT_FLIGHTS_MONTH: Avg Airport Flights per Month / 월별 공항 평균 비행거리
...	AIRLINE_FLIGHTS_MONTH: Avg Airline Flights per Month / 월별 항공사 평균 비행거리
LaGuardia	40.779	-73.876	AIRLINE_AIRPORT_FLIGHTS_MONTH: Avg Flights per month for Airline AND Airport / 월별 항공사 & 공항 평균 비행거리
Indianapolis Muni/Weir Cook	39.729	-86.282	
Logan International	42.364	-71.006	
San Jose International	37.363	-121.941	
Chicago O'Hare International	41.978	-87.906	

03 Contents 03

머신러닝 적용

Catboost 모델 적용 이유

1. 범주형 데이터를 숫자로 변환하는 과정(인코딩)이 필요하지 않음
2. 하이퍼파라미터 튜닝 과정이 간소화되어있음(모델 베스트 스코어 적용)
3. 연산속도가 매우 빠름

```
import numpy as np

cat_features = ['DEP_BLOCK', 'CARRIER_NAME', 'DEPARTING_AIRPORT', 'PREVIOUS_AIRPORT']
print(cat_features)

['DEP_BLOCK', 'CARRIER_NAME', 'DEPARTING_AIRPORT', 'PREVIOUS_AIRPORT']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=1234)
X_train, X_val, y_train, y_val = train_test_split(X_train, y_train, train_size=0.8, random_state=1234)

vc = y_train.value_counts().tolist() # sum(negative instances) / sum(positive instances)
ratio = float(vc[0]/vc[1])
vc, ratio # ratio = class_weight

([819297, 199116], 4.114671849575122)
```

```
from catboost import CatBoostClassifier
model = CatBoostClassifier(eval_metric='AUC',
                           use_best_model=True,
                           task_type="GPU",
                           scale_pos_weight=ratio
                           )
cb = model.fit(X_train, y_train, cat_features=cat_features, eval_set=(X_val, y_val))
```

```
X_test[cat_features] = X_test[cat_features].astype(str)
y_pred = model.predict(X_test)
y_pred

array([0, 1, 1, ..., 0, 1, 0])
```

04

Contents 04

모델 해석

Precision : 모델이 연착되지 않는다 예측한 것 중에 실제 예측이 맞은 경우

Recall : 실제 연착이 되지 않는다 중에 모델이 연착 되지 않는다고 예측한 경우

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.89	0.67	0.76	256173
1	0.32	0.64	0.43	62082
accuracy			0.67	318255
macro avg	0.60	0.66	0.60	318255
weighted avg	0.78	0.67	0.70	318255

04 Contents 04

모델 해석

Feature importance

특성이 예측에 얼마나 중요하게 작용했는지?

	feature_importance	feature_names
3	19.188661	DEP_BLOCK
16	11.597990	PRCP
0	10.294932	MONTH
4	9.908237	SEGMENT_NUMBER
7	7.293308	CARRIER_NAME
19	5.688217	TMAX
14	5.060389	DEPARTING_AIRPORT
1	4.551967	DAY_OF_WEEK
15	4.385929	PREVIOUS_AIRPORT
20	2.999889	AWND
6	2.918150	NUMBER_OF_SEATS
9	2.796486	Avg_MONTHLY_PASS_AIRPORT
2	2.497576	DISTANCE_GROUP
5	1.868091	CONCURRENT_FLIGHTS
8	1.852910	AIRLINE_AIRPORT_FLIGHTS_MONTH
13	1.570856	PLANE_AGE
10	1.500904	Avg_MONTHLY_PASS_AIRLINE
17	1.275643	SNOW
12	1.187625	GROUND_SERV_PER_PASS
11	0.858846	FLT_ATTENDANTS_PER_PASS
18	0.703395	SNWD

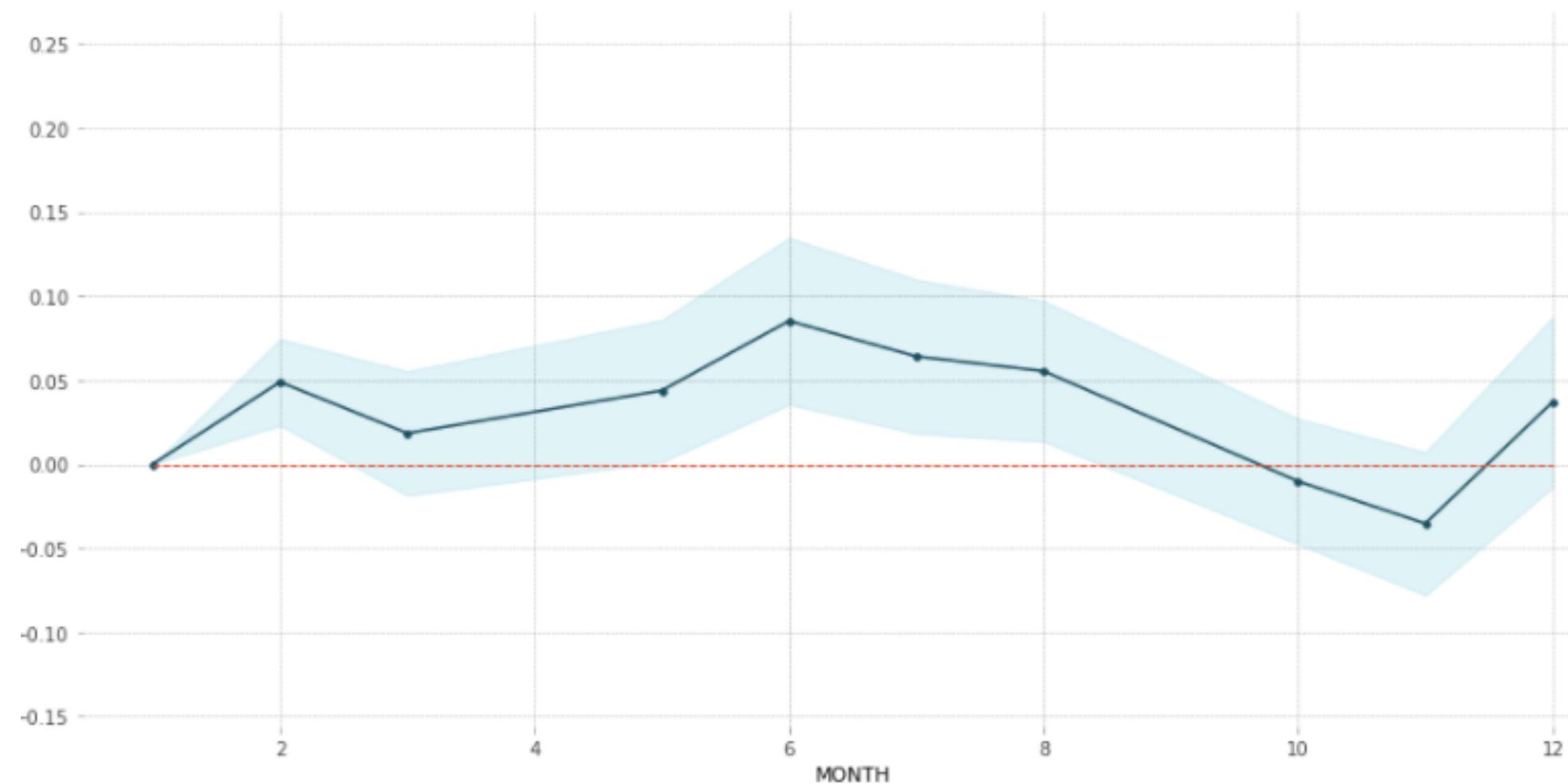
04 Contents 04

모델 해석

PDP

특성이 예측에 어떻게 작용했는지?

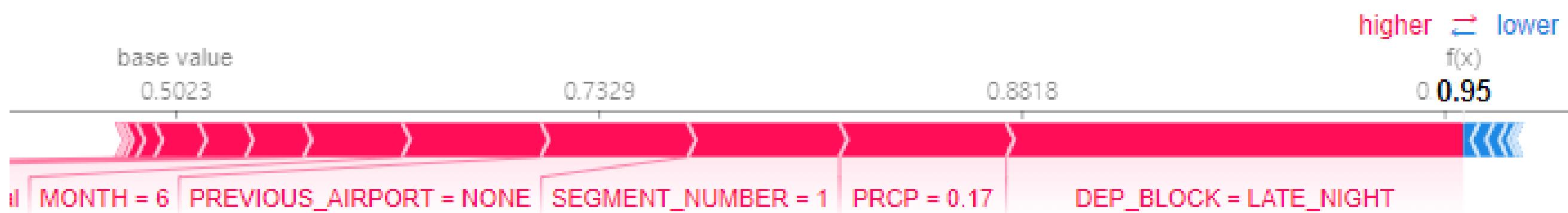
PDP for feature "MONTH"
Number of unique grid points: 10



04 Contents 04

모델 해석

하나의 샘플에 특성이 예측에 어떻게 작용했는지?



04 Contents 04

모델 해석

결론

출발 시간대, 강수량, 출발날짜(월) 순으로 항공기 연착에 중요한 영향을 미친다.

출발날짜(월)의 경우 연착 유무 예측에 중요한 영향을 끼치지만 수시로 연착률을 변동 시킨다.

Thank you!

감사합니다.

끝까지 함께해 주셔서 감사합니다.

