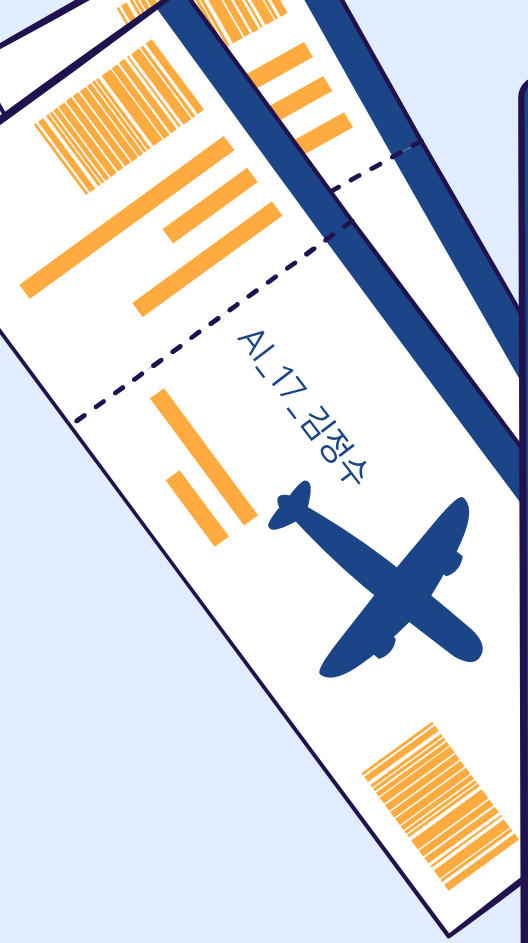


항공사 만족도 예측 모델

Here is where your presentation begins

contents

01

프로젝트 & 데이터 소개

💡 신규 LCC 합류로 소비자들의 선택 폭은 넓어질 수 있지만 공급과잉과 과당경쟁에 따른 수익성 악화로 이어질 수 있다는 시각이 우세하다.

💡 그렇다면 소비자들이 우리 항공사를 선택할 수 있도록 고객 만족도 예측을 통해 파악해보려 한다.

고려해야할 부분

어떤 서비스를 주력으로 제공해야할까?

만족도와 직결되는 요소가 무엇일까?

01

프로젝트 & 데이터 소개

Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service
Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3
Male	disloyal Customer	25	Business travel	Business	235	3
Female	Loyal Customer	26	Business travel	Business	1142	2
Female	Loyal Customer	25	Business travel	Business	562	2
Male	Loyal Customer	61	Business travel	Business	214	3

129487 rows × 23 columns

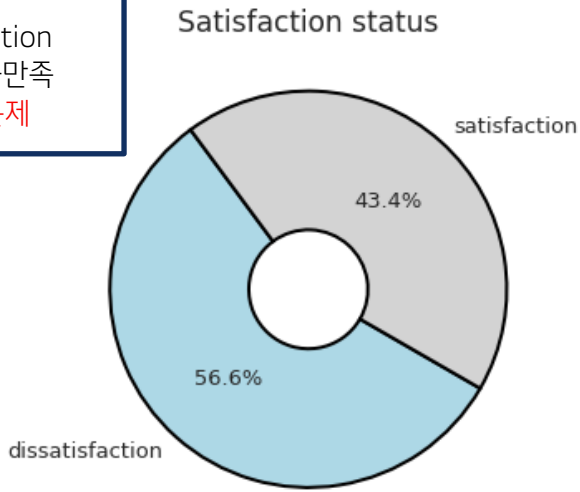
#	Column	Non-Null Count	Dtype
0	Gender	129487 non-null	object
1	Customer Type	129487 non-null	object
2	Age	129487 non-null	int64
3	Type of Travel	129487 non-null	object
4	Class	129487 non-null	object
5	Flight Distance	129487 non-null	int64
6	Inflight wifi service	129487 non-null	int64
7	Departure/Arrival time convenient	129487 non-null	int64
8	Ease of Online booking	129487 non-null	int64
9	Gate location	129487 non-null	int64
10	Food and drink	129487 non-null	int64
11	Online boarding	129487 non-null	int64
12	Seat comfort	129487 non-null	int64
13	Inflight entertainment	129487 non-null	int64
14	On-board service	129487 non-null	int64
15	Leg room service	129487 non-null	int64
16	Baggage handling	129487 non-null	int64
17	Checkin service	129487 non-null	int64
18	Inflight service	129487 non-null	int64
19	Cleanliness	129487 non-null	int64
20	Departure Delay in Minutes	129487 non-null	int64
21	Arrival Delay in Minutes	129487 non-null	float64
22	satisfaction	129487 non-null	int64

02

Target, 가설 설정, EDA

Target

Satisfaction
만족 / 불만족
분류 문제



가설

1. 이코노미 좌석일수록 만족도가 낮다.
2. 연령대가 높을수록 만족도가 낮다.
3. 비행거리가 짧을 수록 만족도가 낮다.
4. 충성 고객은 오히려 만족도가 더 낮다.

02

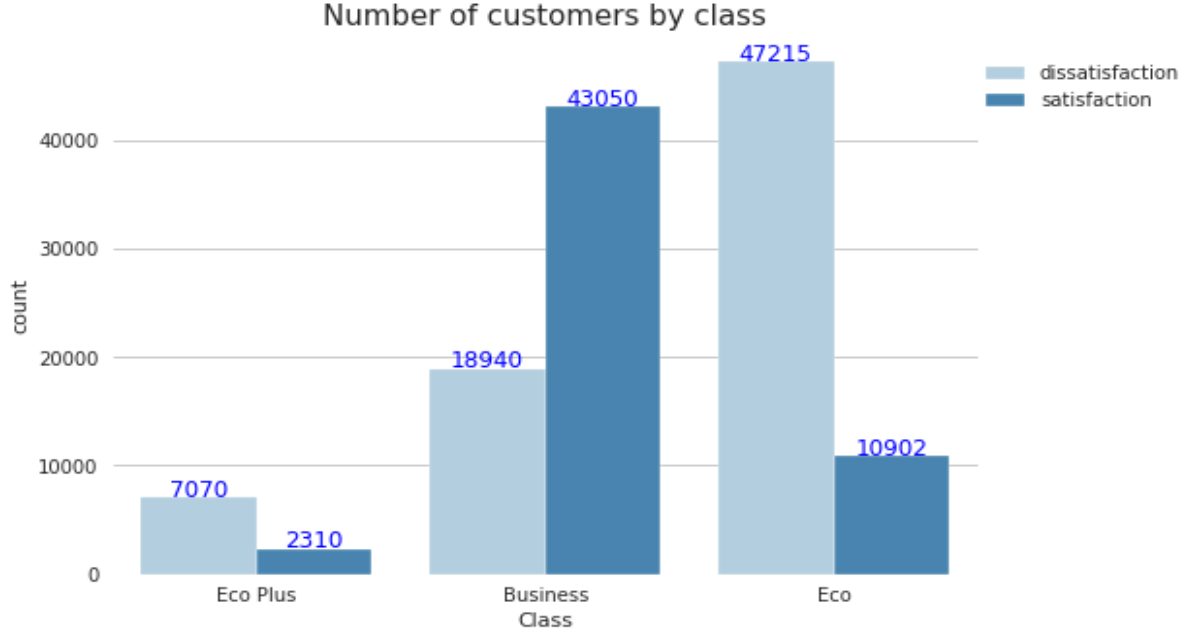
Target, 가설 설정, EDA

데이터 전처리

1. 결측치 : 항공기 지연 시간에 대한 결측치 총 80개 삭제(샘플 데이터의 수가 충분하다고 판단)
2. 타켓(satisfaction) : 불만족(0), 만족(1) 인 int 값으로 대체.
3. 불필요한 컬럼 삭제 : id
4. 특성이 범주형인 경우 카디널리티가 높지 않아 그대로 유지.

02

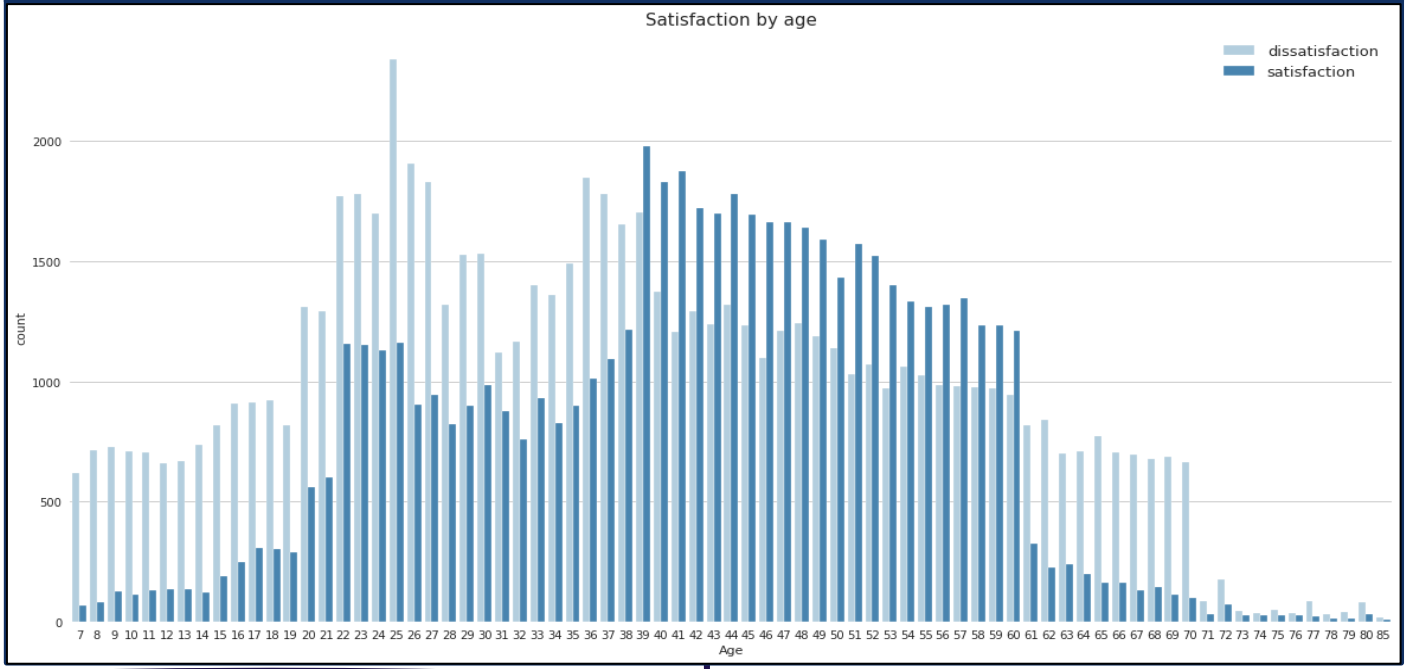
가설 : 이코노미 좌석일수록 만족도가 낮다.
이코노미와 이코노미 플러스를 이용한 고객들의 불만족도가 만족에 비해 3~4배 높다.



02

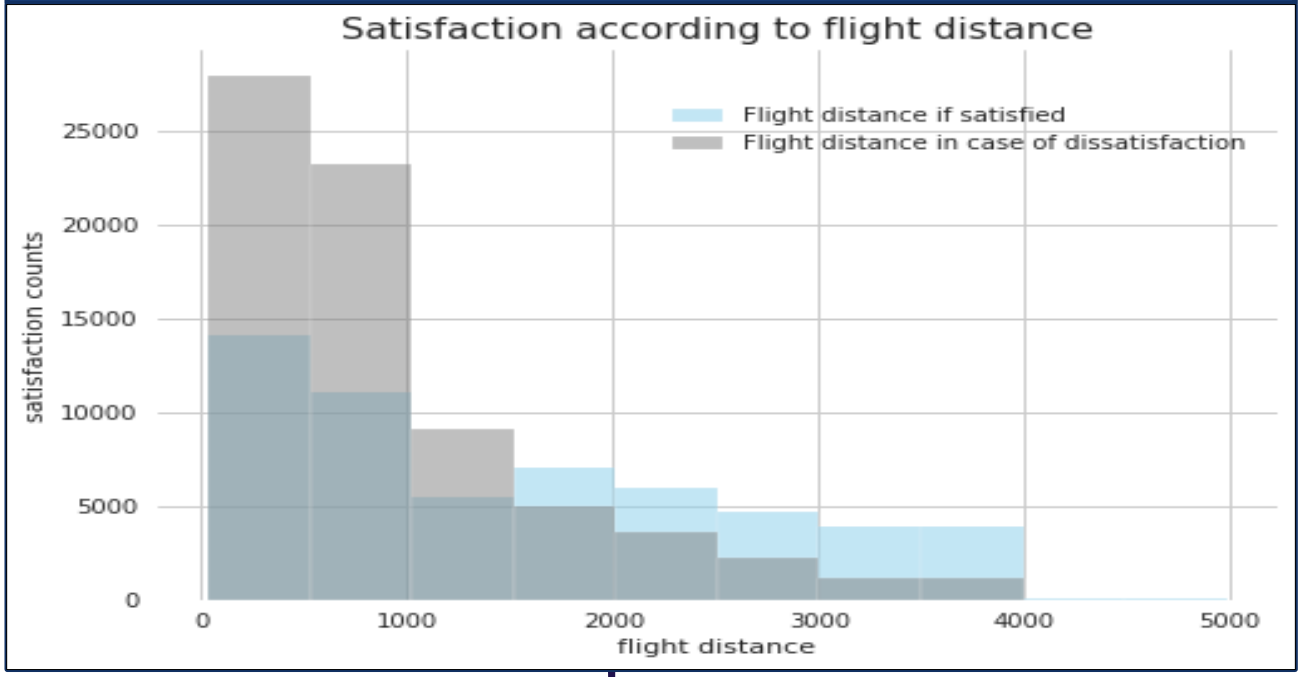
가설 : 연령대가 높을수록 만족도가 낮다

38세까지는 대체적으로 불만족도가 높고, 39세~60세까지는 만족도가 높다가 그 이후에는 다시 불만족도가 높다. 즉, 연령대가 높을수록 만족도가 무조건 낮다고 할 수 없다.



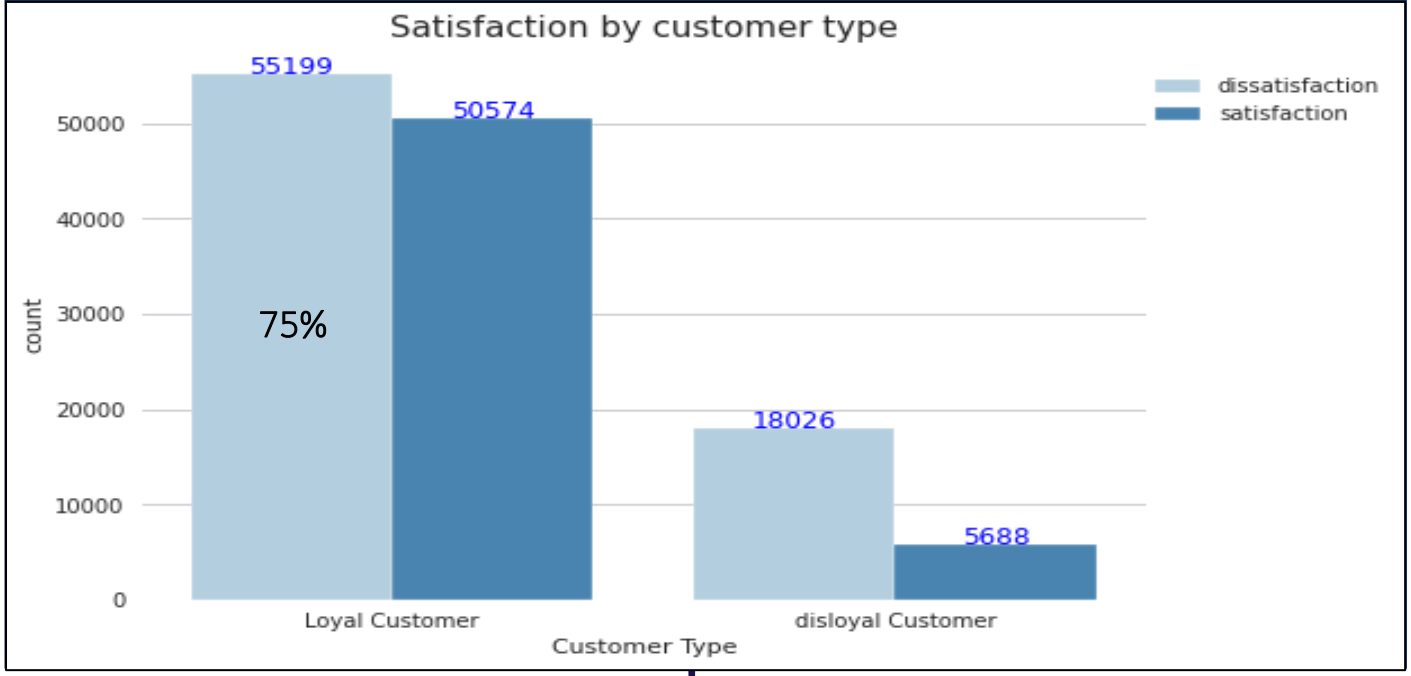
02

가설: 비행거리가 짧을 수록 만족도가 낮다
비행거리가 짧은 경우 만족도가 더 낮고, 비행거리가 길수록 만족도가 높다.



02

가설 : 충성 고객의 만족도가 더 낮다.
불만족인 고객 중 충성고객(Loyal Customer)이 75%로 충성 고객의 만족도가 더 낮았다



03

모델 학습 & 모델 평가

Hold out 기법을 통해
train / val / test 총 3가지로 분리

X_train : (72515, 22),y_train : (72515,)
X_val : (31079, 22),y_val : (31079,)
X_test : (25893, 22),y_test : (25893,)

타켓 비율이 완전 불균형하지 않기때문에
정확도와 f1 score 확인.

1. LogisticRegression (기준모델)
- 회귀를 이용한 분류 모델

2. DecisionTree
- 훈련데이터에 대한 제약사항이 없어서 유연한 모델이다. 즉 과적합 위험 가능성이 높다.

3. RandomForest
- DecisionTree 의 과적합 문제를 해소해주는 모델.
- 배깅을 대표하는 모델로 각각의 기본 모델이 랜덤으로 예측하는 성능보다 좋을 경우 합치는 과정에서 에러가 상쇄되어 더 정확한 예측 가능

4. XGBoost
-틀린 데이터에 집중하기 위해 가중 샘플링을 하는 대신 잔차를 학습한다.
- 잔차가 큰 관측치를 더 학습하도록 하는 효과가 있고, 이전 모델이 틀린 만큼을 직접 학습하며 이전 모델을 순차적으로 보완한다.

03 LogisticRegression (기준모델)

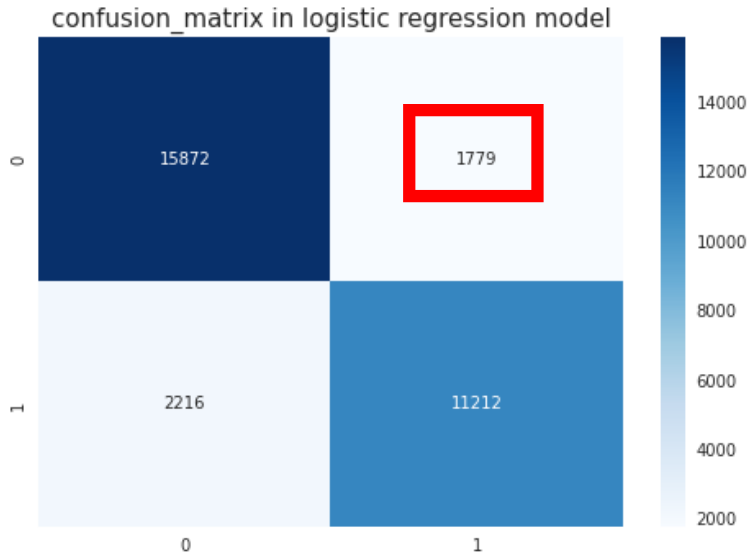
LogisticRegression은 L2 패널티가 적용되기 때문에 스케일링 필수

----------train----------

	precision	recall	f1-score	support
0	0.88	0.91	0.89	41046
1	0.87	0.83	0.85	31469
accuracy			0.87	72515
macro avg	0.87	0.87	0.87	72515
weighted avg	0.87	0.87	0.87	72515

----------val----------

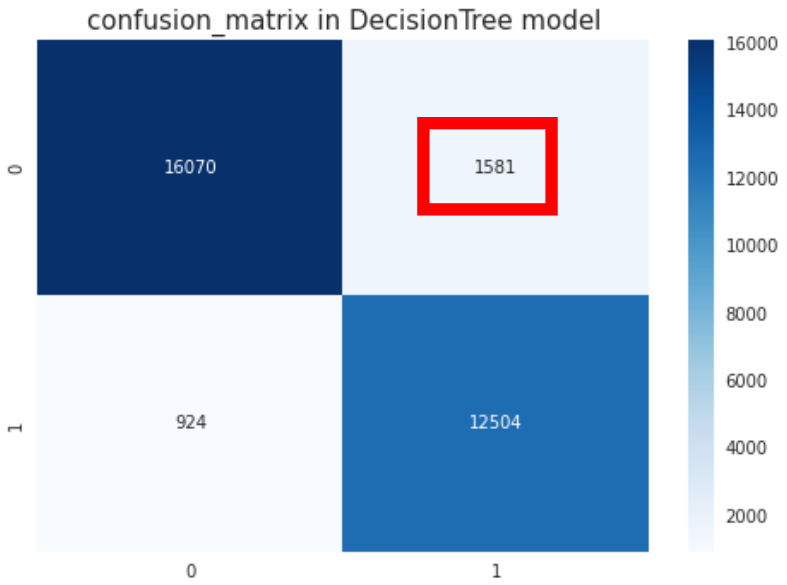
	precision	recall	f1-score	support
0	0.88	0.90	0.89	17651
1	0.86	0.83	0.85	13428
accuracy			0.87	31079
macro avg	0.87	0.87	0.87	31079
weighted avg	0.87	0.87	0.87	31079



OrdinalEncoder(),
StandardScaler(),
LogisticRegression(random_state=2)

03 DecisionTree

```
------------------------train-------------------------
              precision    recall  f1-score   support

           0       0.94      0.92      0.93     41046
           1       0.89      0.93      0.91     31469

    accuracy                           0.92     72515
   macro avg       0.92      0.92      0.92     72515
weighted avg       0.92      0.92      0.92     72515

--------------------------val-------------------------
              precision    recall  f1-score   support

           0       0.95      0.91      0.93     17651
           1       0.89      0.93      0.91     13428

    accuracy                           0.92     31079
   macro avg       0.92      0.92      0.92     31079
weighted avg       0.92      0.92      0.92     31079
```



OrdinalEncoder(),
DecisionTreeClassifier
(random_state=2, criterion="entropy",max_depth=6)

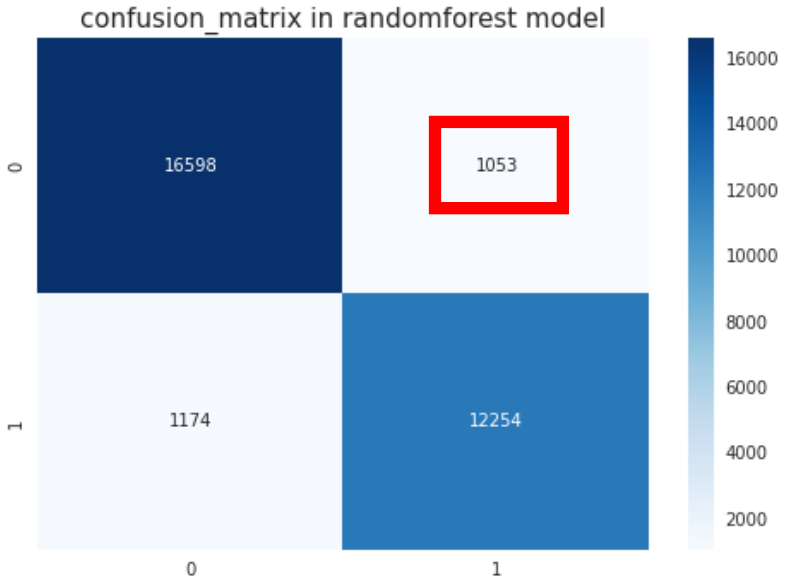
03 RandomForest

----------train----------

	precision	recall	f1-score	support
0	0.93	0.94	0.94	41046
1	0.93	0.91	0.92	31469
accuracy			0.93	72515
macro avg	0.93	0.93	0.93	72515
weighted avg	0.93	0.93	0.93	72515

----------val----------

	precision	recall	f1-score	support
0	0.93	0.94	0.94	17651
1	0.92	0.91	0.92	13428
accuracy			0.93	31079
macro avg	0.93	0.93	0.93	31079
weighted avg	0.93	0.93	0.93	31079



OrdinalEncoder(),
RandomForestClassifier
(random_state=2, oob_score=True,
n_jobs=-1,max_depth=6)

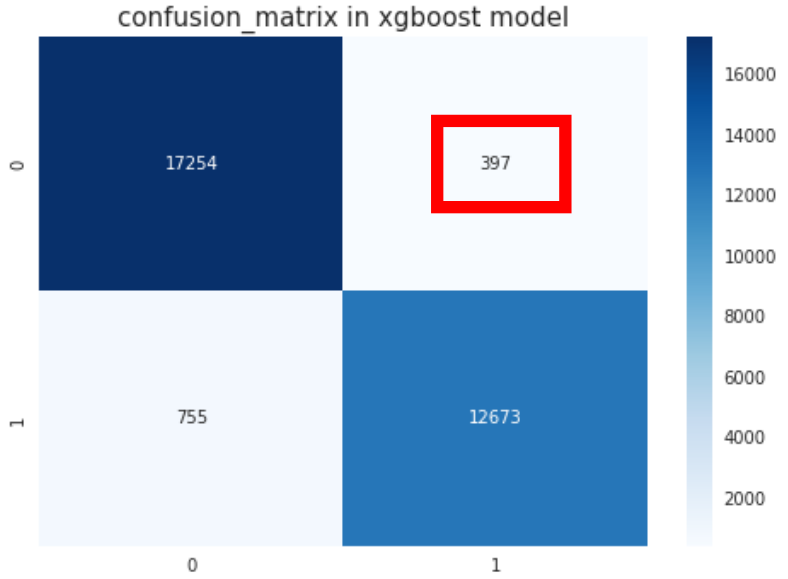
03 XGBoost

----------train----------

	precision	recall	f1-score	support
0	0.97	0.99	0.98	41046
1	0.99	0.96	0.97	31469
accuracy			0.98	72515
macro avg	0.98	0.98	0.98	72515
weighted avg	0.98	0.98	0.98	72515

----------val----------

	precision	recall	f1-score	support
0	0.96	0.98	0.97	17651
1	0.97	0.94	0.96	13428
accuracy			0.96	31079
macro avg	0.96	0.96	0.96	31079
weighted avg	0.96	0.96	0.96	31079



OrdinalEncoder(),
XGBClassifier(random_state=2)

GridSearchCV를 사용하여 하이퍼 파라미터 튜닝.
max_depth 최적의 값 : 6

04

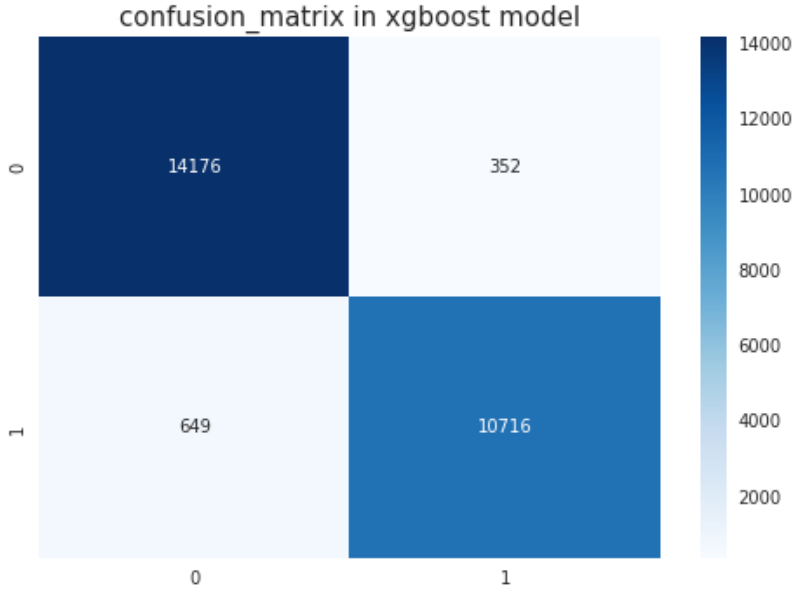
최종 모델 평가 및 모델 해석

최종 모델 : XGBoost

선택 이유 :
정확도와 f1 score 가 가장 높기도 했지만, 해당 문제는 만족하지 않았는데 만족했다고 예측된 ERROR가 더 중요하다. 즉, 불만족인데 만족이라고 예측한 수가 가장 적다.

테스트 셋 성능

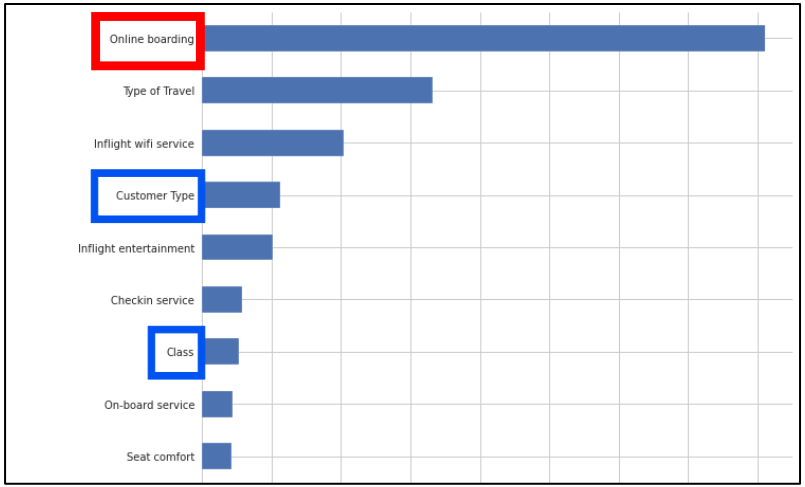
	precision	recall	f1-score	support
0	0.96	0.98	0.97	14528
1	0.97	0.94	0.96	11365
accuracy			0.96	25893
macro avg	0.96	0.96	0.96	25893
weighted avg	0.96	0.96	0.96	25893

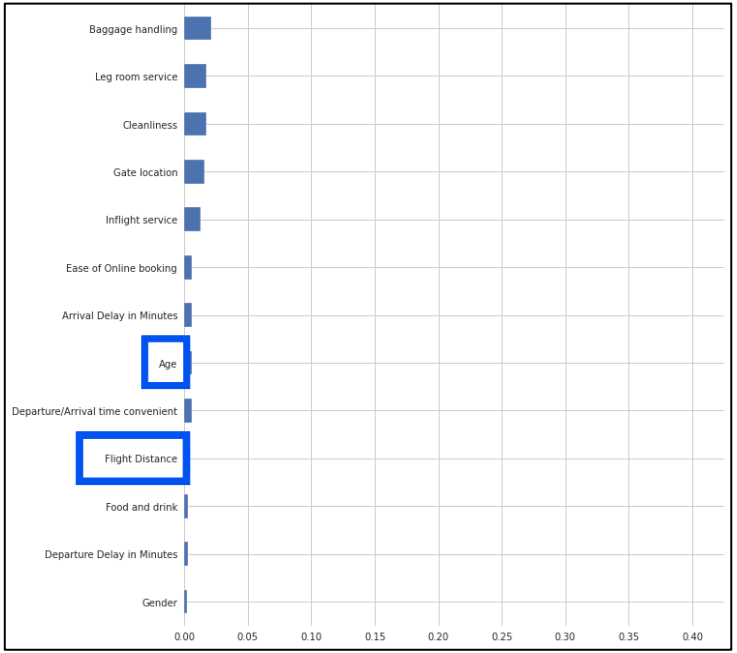


04

최종 모델 평가 및 모델 해석

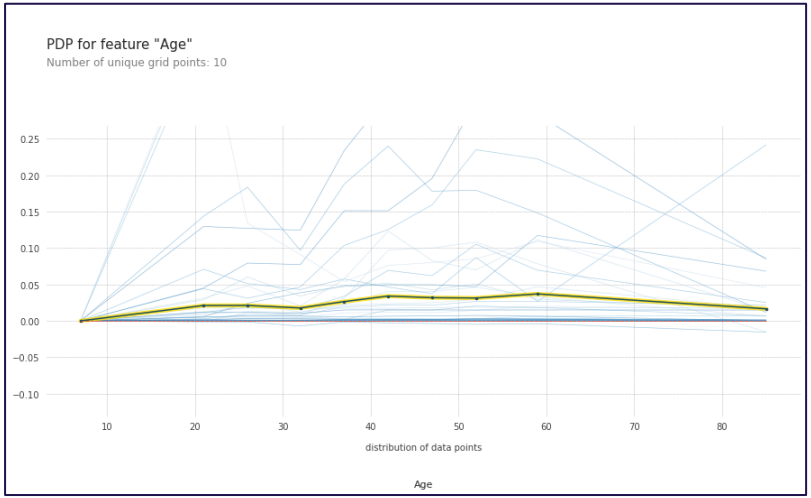
특성 중요도 (MDI)

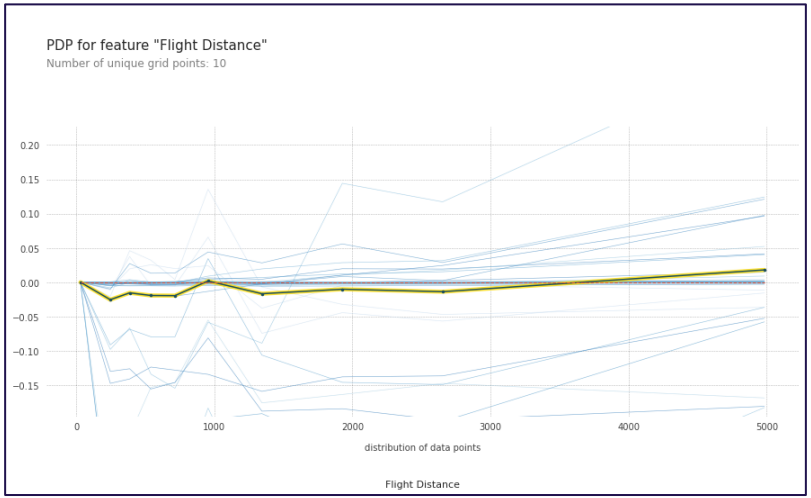




04

최종 모델 평가 및 모델 해석 _ PDP

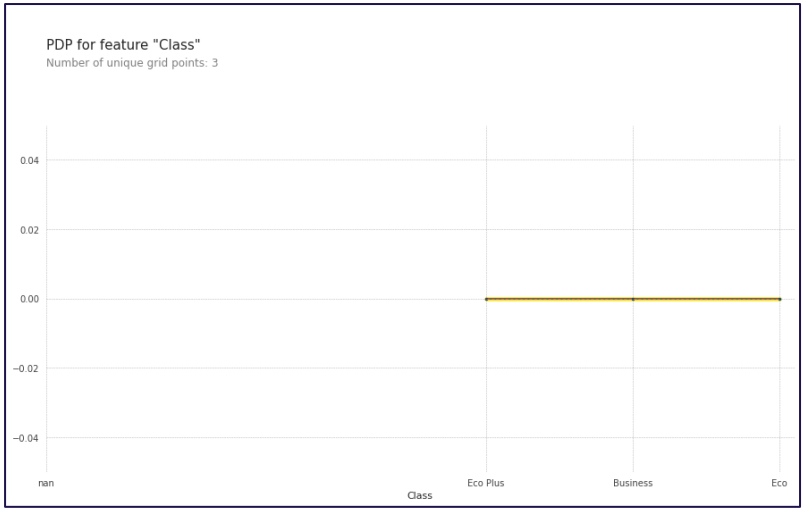


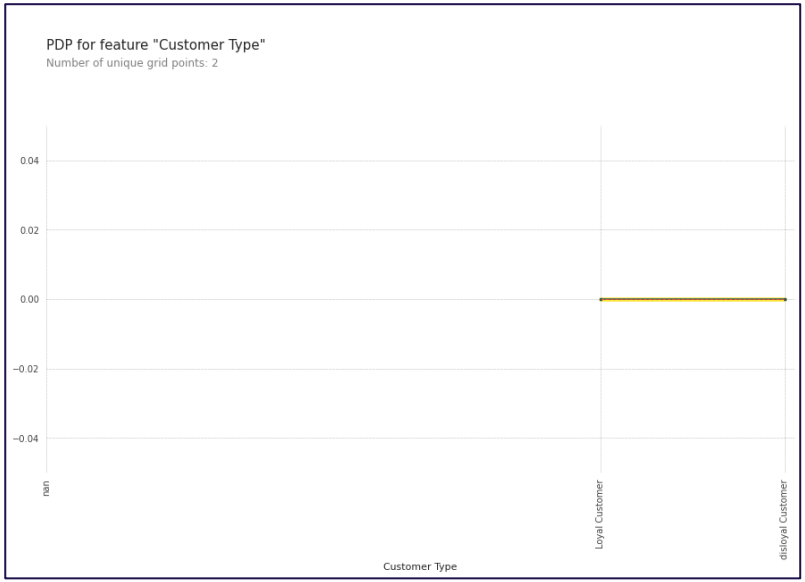


Age의 경우 나이가 들수록 만족도가 증 가하다가 감소하는 경향이 미세하게 보였지만, 영향력이 커보이진 않았고, Flight Distance의 경우 비행거리가 아주 먼 경우에만 만족도가 조금 높은 것으로 판단.

04

최종 모델 평가 및 모델 해석 _ PDP





Class나 Customer Type의 경우 만족도를 예측하는데 의미가 없다고 판단.

05

결론 _ 만족도를 높이기 위해서는?

1. 만족도에는 Online boarding의 특성 영향력이 가장 크다. 보통 종이탑승권을 발권하기 위해서는 카운터나 KIOSK를 이용하기 위해 직접 방문해야한다. 하지만 모바일 탑승권을 발행하면 굳이 방문하지 않아도 바로 비행기 탑승이 가능하다.

2. 즉, 비행기 탑승까지의 과정이 축소된다면 고객의 만족도를 높일 수 있다고 생각한다.

3. 또한, 그 외에도 "Inflight wifi service", "Inflight entertainment"의 특성 영향력도 높았다. 기내 안에서 즐길 거리가 다양하다면 고객의 만족도를 높일 수 있다고 생각한다.

06

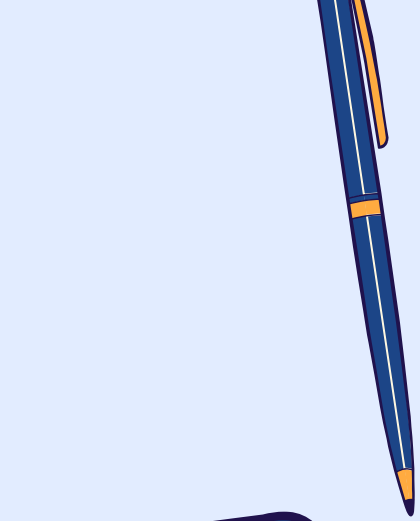
개선사항, 한계점

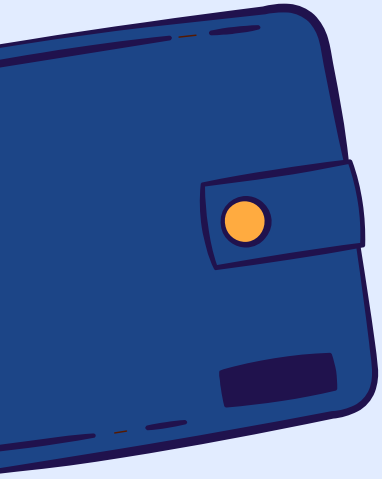
- Data Set 자체로 봤을 때, 데이터 정제가 거의 필요 없는 상태라서 다양한 전처리, Feature Engineering을 해보지 못한 아쉬움.
추후에 Kaggle Data 아닌 웹 크롤링 등을 이용해 직접 수집하고 다양한 전처리와 Feature Engineering을 시도해 볼 것.

- 모든 특성에 대해 알아보지 못한 점.

- 추후에 특성 중요도에 따라서 불필요한 컬럼을 추가로 삭제 후 성능 확인.

- PDP 해석하는데 조금 어려움이 있어서 다양한 사례를 보면서 해석해 볼 것.





thank you
for
listening