항공서비스개선을 위한 항공사 이용객 설문조사와 고객 만족도 예측

CONTESTNS



1. 개요

- 프로젝트목표
- 분석할 데이터 소개



2.EDA 및 데이터전처리

- 데이터 컬럼 분석
- 데이터 전처리



3.가설 설정 및 검증

- 물리적 서비스
- 편의성서비스
- 정보제공서비스
- 인적서비스
- 그외요인들



5. 모델해석

- PDP
- SHAP



4. 모델 학습 및 검증

모델 학습

- RandomForest
- XGboost



6. 결론

1. 개요

프로젝트 목표 분석할 데이터 소개

프로젝트의 목표

문제 제시

항공사는 고객들이 원하는 욕구를 충족시켜 우리 항공사를 이용할 수 있도록 하고, 단골고객의 이탈 등의 방지를 위해 항공서비스에 대한 만족도를 항상 조사하고, 분석하여 개선해나가야 합니다.

목표

항공사를 이용한 고객들의 만족도 조사를 통해 '만족'과 '불만족'을 예측하여 개선해 나갈 서비스 파악합니다.

목표를 이루기 위해 분석해야 할 내용

인구통계적 속성,물리적 서비스,편의성 서비스,정보제공 서비스,인적서비스에 대한 만족도 점수와 항공사 만족도와의 관계

데이터 소개

항공사를 이용한 고객이 항공사에 만족하는지 여부를 알려줍니다. 여러가지 서비스에 대한 리커트척도와 몇가지 인적사항들로 이루어져 있습니다.

범주형 데이터

	count	unique	top
Gender	103904	2	Female
Customer Type	103904	2	Loyal Customer
Type of Travel	103904	2	Business travel
Class	103904	3	Business
satisfaction	103904	2	neutral or dissatisfied

'Gender': male, female

'Customer Type': Loyal Customer, disloyal Customer

'Type of Travel': Personal Travel, Business travel

'Class': Eco, Eco Plus, Business

'Satisfaction': neutral or dissatisfied, satisfied

target 컬럼인 'satisfaction'에서 **중립과 불만족**을 하나의 데이터로 묶었고,

이 데이터에서 이하 '불만족'이라고 부르겠습니다.

데이터 소개

항공사를 이용한 고객이 항공사에 만족하는지 여부를 알려줍니다. 여러가지 서비스에 대한 리커트척도와 몇가지 인적사항들로 이루어져 있습니다.

수치형 데이터

Age 1 Flight Distance 1	103904.0 103904.0 103904.0	64924.210502 39.379706	37463.812252 15.114964	1.0	32533.75	64856.5	97368 25	129880.0
Flight Distance 1			15.114964	7.0				125000.0
	103904.0			7.0	27.00	40.0	51.00	85.0
1.00.11.20.		1189.448375	997.147281	31.0	414.00	843.0	1743.00	4983.0
Inflight wifi service 1	103904.0	2.729683	1.327829	0.0	2.00	3.0	4.00	5.0
Departure/Arrival time convenient 1	103904.0	3.060296	1.525075	0.0	2.00	3.0	4.00	5.0
Ease of Online booking 1	103904.0	2.756901	1.398929	0.0	2.00	3.0	4.00	5.0
Gate location 1	103904.0	2.976883	1.277621	0.0	2,00	3.0	4.00	5.0
Food and drink	103904.0	3.202129	1.329533	0.0	2.00	3.0	4.00	5.0
Online boarding	103904.0	3.250375	1.349509	0.0	2.00	3.0	4.00	5.0
Seat comfort 1	103904.0	3.439396	1.319088	0.0	2.00	4.0	5.00	5.0
Inflight entertainment 1	103904.0	3.358158	1.332991	0.0	2.00	4.0	4.00	5.0
On-board service 1	103904.0	3.382363	1.288354	0.0	2.00	4.0	4.00	5.0
Leg room service 1	103904.0	3.351055	1.315605	0.0	2.00	4.0	4.00	5.0
Baggage handling 1	103904.0	3.631833	1.180903	1.0	3.00	4.0	5.00	5.0
Checkin service 1	103904.0	3.304290	1.265396	0.0	3.00	3.0	4.00	5.0
Inflight service 1	103904.0	3.640428	1.175663	0.0	3.00	4.0	5.00	5.0
Cleanliness 1	103904.0	3.286351	1.312273	0.0	2.00	3.0	4.00	5.0
Departure Delay in Minutes 1	103904.0	14.815618	38.230901	0.0	0.00	0.0	12.00	1592.0
Arrival Delay in Minutes	103594.0	15.178678	38.698682	0.0	0.00	0.0	13.00	1584.0

'id': 설문조사에 응한 고객 고유번호

'Age': 설문조사에 응한 고객 나이

'Flight Distance': 비행거리

'Departure Delay in Minutes', 'Arrival Delay in Minutes':

출발/도착 지연시간

나머지 특성:항공사서비스에 대한 리커트척도로 나타낸 점수 (0~5점)



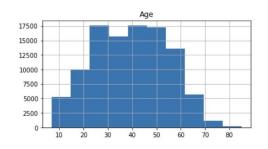
리커트(Likert) 척도란?

'전혀 아니다', '아니다', '보통이다', '그렇다', '매우그렇다' 의 정도를 표시하도록하는 측정방법 이 데이터에서 0점은 평가없음. 1점 → '전혀 아니다', 5점 → '매우 좋다' 로 나타내어졌습니다.

2. EDA 및 데이터 전처리

데이터 컬럼 분석 데이터 전처리

수치형데이터 분포 확인



Departure Delay in Minutes

1000

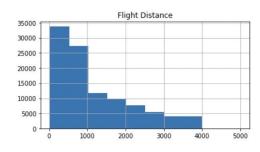
1200 1400

80000

60000

40000

20000





1000 1200

- ▷ 나이는 7세부터 85세까지 분포20대부터 50대에 가장 많이 분포
- ▷ 비행거리의 최단거리는 31km, 최장거리는 약 5000km

 출발지연시간으로 인해 도착지연시간이 결정되기때문에 같은 데이터를 가진 컬럼이라고 볼 수 있음

수치형데이터 분포 확인

satisfaction	Arrival Delay in Minutes	Departure Delay in Minutes
neutral or dissatisfied	1584.0	1592
satisfied	1280.0	1305
satisfied	1011.0	1017
neutral or dissatisfied	970.0	978
satisfied	920.0	933
neutral or dissatisfied	952.0	930
neutral or dissatisfied	924.0	921
satisfied	860.0	859
neutral or dissatisfied	823.0	853
satisfied	729.0	750

출발/도착지연시간 상위 10개 데이터 -> 800분 이상 지연 긴 시간이 지연되었지만 '만족'을 준 케이스도 있기때문에 이상치라 생각하지 않고 같이 분석

실제로 상위 두번째 데이터에서 도착시간이 1280분(21시간) 이 지연 된 기사

https://www.sedaily.com/NewsVlew/22P3NH20R8

獨서 접촉사고 난 대한항공 항공기…21시간 운항지연

머니투데이 이건희 기자	
2019.11.17 09:52	(기) (의견남기기)
정상 착륙 후 대기 중 발생한 사고…인명 피해 없어	

데이터 전처리

	missing_number	missing_percentage
Arrival Delay in Minutes	310	0.002984
id	0	0.000000
Gender	0	0.000000
Departure Delay in Minutes	0	0.000000
Cleanliness	0	0.000000
Inflight service	0	0.000000

satisfaction	Arrival Delay in Minutes	Departure Delay in Minutes
neutral or dissatisfied	1584.0	1592
satisfied	1280.0	1305
satisfied	1011.0	1017
neutral or dissatisfied	970.0	978
satisfied	920.0	933
neutral or dissatisfied	952.0	930

- ▷ 결측치는 도착지연시간(분)에서만 존재
 도착지연시간은 출발지연시간이 원인이기때문에 비슷한 값을 가지므로 출발지연시간와 같은 값으로 채워줌
- ▷ 'id'컬럼은 고객의 고유번호이기때문에 만족도 예측에 영향X -> 삭제

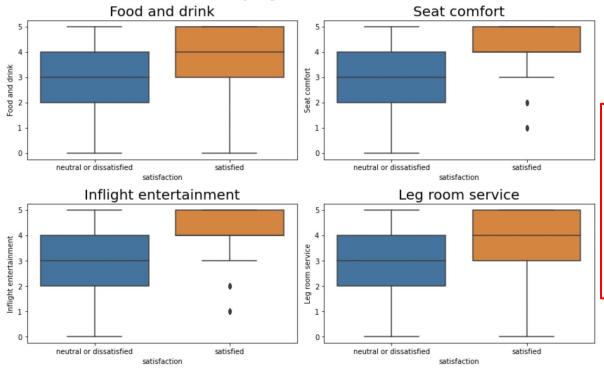


3. 가설 검증

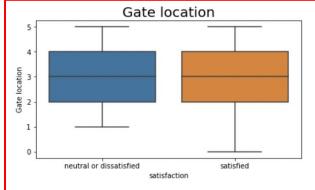
데이터 시각화를 통해 가설 검증

가설 1. 물리적 서비스가 만족도에 영향을 준다.

Correlation between physical service and satisfaction



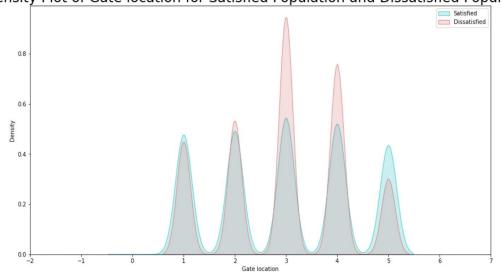
▷ 기내식, 편안한 좌석, 기내엔터테인먼트, 레그룸: 설문조사 점수가 높을수록 '만족'



게이트위치는 설문조사 점수와 관계없이 '불만족', '만족' 같은 비율

가설 1-1. 게이트위치와 만족도는 상관관계가 없다.

Density Plot of Gate location for Satisfied Population and Dissatisfied Population

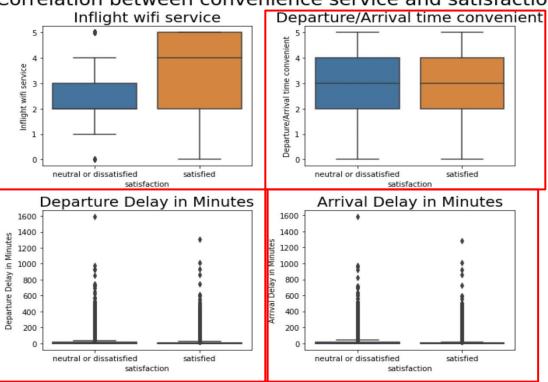


 ▷ 게이트위치의 점수가 '매우 좋다' 인 5점일 때는 만족도가 더 높지만 3, 4점에서는 불만족이 2배 많은 것으로 보아 무의미한 관계로 보임.

결론:물리적 서비스 중 게이트위치는 만족도에 영향을 미치지 않는다.

가설2. 편의성 서비스가 만족도에 영향을 준다.

Correlation between convenience service and satisfaction

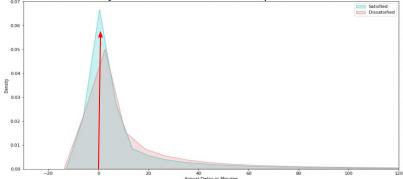


▷ 기내 와이파이서비스는 만족도 점수가 높을수록 '만족'

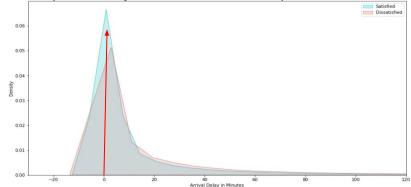
● 편리한 출발/도착시간, 출발/도착지연시간과 만족도는 영향이 없는것으로 보임.

가설 2-1. 출발/도착지연시간과 만족도는 상관관계가 없다.

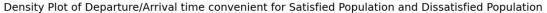
Density Plot of Arrival Delay in Minutes for Satisfied Population and Dissatisfied Population

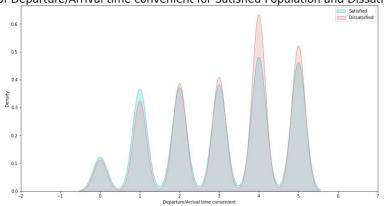


Density Plot of Departure Delay in Minutes for Satisfied Population and Dissatisfied Population



▷ 지연시간이 없을때는 확실히 만족도가 높다 하지만 delay가 존재할때에는 '만족', '불만족'차이가 크게 없다. 가설 2-2. 편리한 출발/도착시간과 만족도는 상관관계가 없다.





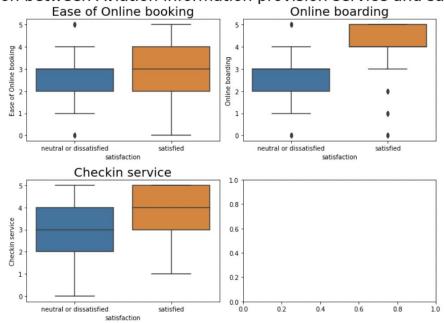
편리한 출발/도착시간 만족도 점수에서 높은 4,5점을 주었지만 '불만족'이 더 높게 나옴.

편리한 출발/도착시간은 항공사만족도에 영향을 주지 않는 것으로 보임

결론:편의성 서비스 중 편리한 출발/도착시간, 출발/도착지연시간은 만족도에 영향을 미치지 않는다.

가설 3. 정보제공 서비스가 만족도에 영향을 준다.

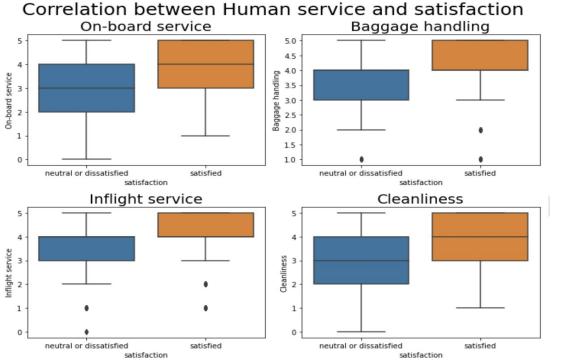
Correlation between Aviation information provision service and satisfaction



▷ 쉬운 온라인 예약, 온라인탑승정보, 체크인서비스 모두 만족도 점수가 높을수록 '만족'이 높다.

결론:정보제공서비스모두만족도에 영향을 미친다.

가설 4. 인적 서비스가 만족도에 영향을 준다.

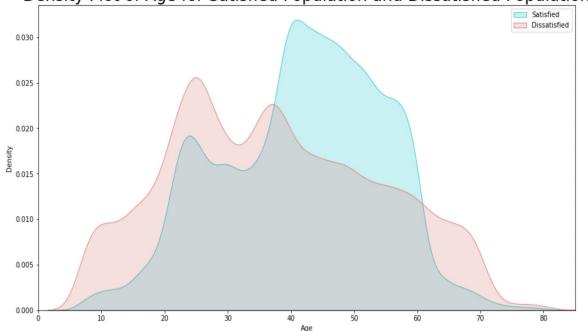


○ On-board서비스, 수하물취급, 기내서비스, 청결도 모두 만족도에 영향을 미침.

결론: 인적 서비스 모두 만족도에 영향을 미친다.

가설 5. 나이와 만족도에 상관관계가 있다.



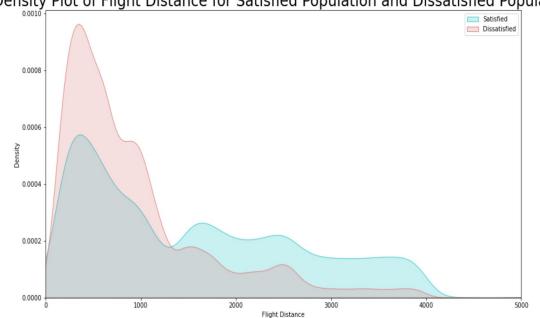


- ▷ '만족'의 경우 30대 후반에서 50대후반까지 가장 많음
- ▷ '불만족'의 경우 **20**대, **30**대에서 가장 많음

결론:나이와 항공사 만족도가 유의미한 관계라고 볼 수 있다.

가설 6. 비행거리와 만족도에 상관관계가 있다.

Density Plot of Flight Distance for Satisfied Population and Dissatisfied Population



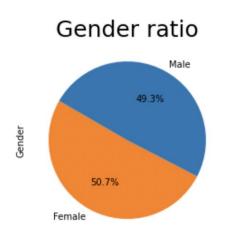
비행거리가 길수록 '불만족'이 높을거라고 생각했지만
 1300km이하인 거리에서 '불만족'이 높게 나왔고, 그 이상일때 '만족'이 더 높게 나옴 확실한 차이가 보이므로

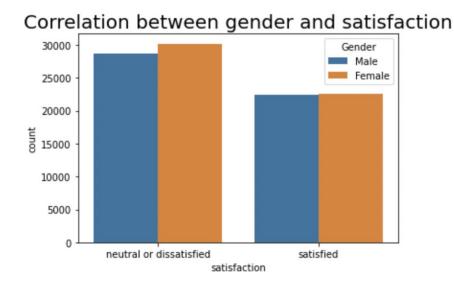
유의미한 변수라고 생각

결론: 비행거리와 항공사 만족도가 유의미한 관계라고 볼 수

21

가설 7. 성별과 만족도에 상관관계가 있다.

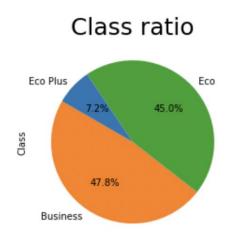


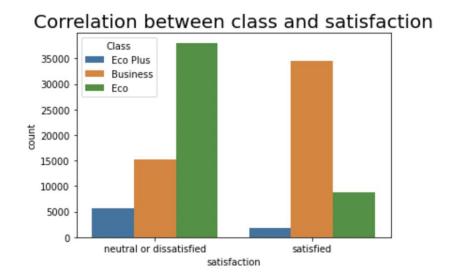


▷ 데이터에서 **남자**와 <mark>여자</mark>의 비율이 같고,'만족','불만족'의 비율도 비슷하게 나옴

결론: 성별과 항공사 만족도에는 상관관계가 없다.

가설 8. 좌석(class)와 만족도에 상관관계가 있다.



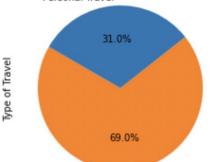


- ▷ **이코노미, 이코노미플러스**: 만족 < **불만족 / 비즈니스**: **만족** > 불만족
- ▷ 편안한 좌석이 이코노미 < 이코노미 플러스 < 비즈니스 이기때문에 비즈니스 이용고객의 만족도가 높고, 이코노미를 이용한 고객은 불만족이 높은 것으로 보임

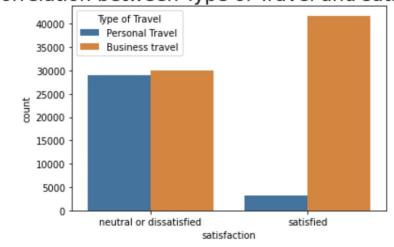
결론: 좌석(class)과 항공사 만족도에는 상관관계가 있다.

가설 9. 여행유형과 만족도에 상관관계가 있다.





Correlation between Type of Travel and satisfaction

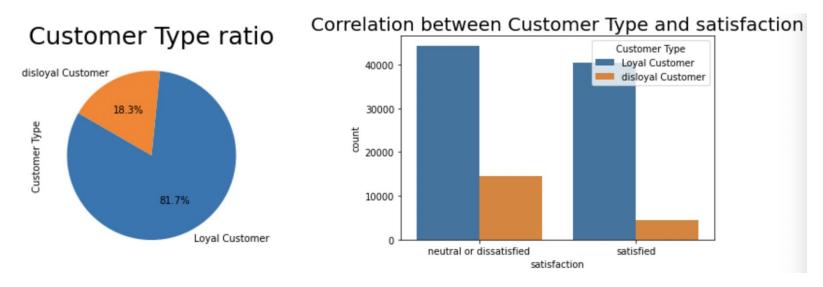


- ▷ 출장이 70%, 개인여행이 30%로 출장이 더 많은데 출장인 경우에는 만족이 더 높고, 개인여행인 경우 불만족이 훨씬 높음
- ▷ 개인여행일 때 만족도에 대한 기준이 더 높은 것으로 보임

Business travel

결론: 여행유형과 항공사 만족도에는 상관관계가 있다.

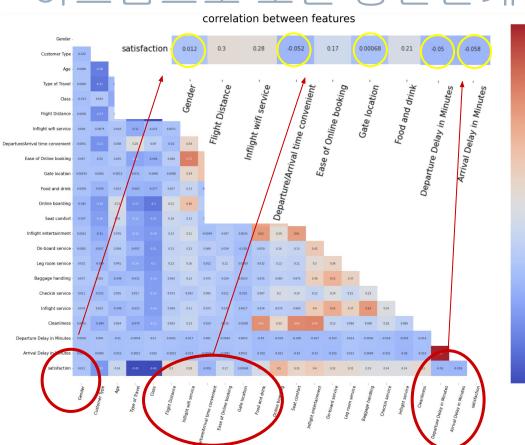
가설 10. 고객유형과 만족도에 상관관계가 있다.



- ▷ **단골고객**은 '만족'과 '불만족'이 비슷한 비율이고, <mark>단골이 아닌 고객</mark>은 **불만족**이 만족의 3배
- ▷ 단골이 아닌 고객을 잡기 위해서는 항공사 서비스개선이 꼭 필요함

결론: 고객유형과 항공사 만족도에는 상관관계가 있다.

히트맵으로 보는 상관관계



특성들과 만족도의 상관계수

성별:0.012

편리한 출발/도착시간: -0.052

게이트위치: 0.00068

출발지연시간: -0.05

도착지연시간: -0.058

0에 가까운 상관계수를 가졌고, 만족도에 영향을 미치지 않는다는 것을 한 번 더 확인.

→ 가설 결론

게이트 위치, 편리한 출발/도착시간, 출발/도착지연시간, 성별은 만족도에 영향을 주지 않습니다.

그 외 다른 서비스는 만족도점수가 높은만큼 항공사 '만족'으로 이어지고, 나이, 비행거리, Class, 여행유형, 고객유형이 항공사 만족도에 영향을 미친다고 분석할 수 있습니다. 4. 모델 학습 및 검증

> 기준모델 모델학습

모델링



문제유형선정 '만족', '불만족'을 예측하는 분류문제



분류문제란?

- -> 지도학습: 입력과 대응하는 미리알려진 결과(데이터에 존재)를 학습 크게 분류와 회귀문제로 나뉨
 - 분류: 결과 출력이(예, 아니요)(참, 거짓)(0, 1)과 같은 범주형을 예측하는문제
 - 회귀: 연속인 수치형을 예측하는 문제

분류문제의 가장 간단한 기준모델은 최빈값으로 예측.

Baseline모델

'만족', '불만족'을 예측하는 분류문제 분류문제의 가장 간단한 기준모델은 최빈값으로 예측.

▷ 학습셋에서 '불만족'이 **0.56**이라는 비율로 최빈값은 '불만족'

```
# 기준모델
from sklearn.metrics import accuracy_score

major = y_train.mode()[0]
y_pred = [major] * len(y_train)
print("학습 정확도: ", accuracy_score(y_train, y_pred))
학습 정확도: 0.5668106300301962
```

최빈값으로 예측한다면 정확도는 0.56매우 낮은 수준.

순열중요도/Feature Engineering

위에서 살펴본결과대로'만족도'에 영향을 미치지 않는 특성 (게이트위치, 편리한 출발/도착시간, 출발/도착지연시간, 성별) 제거 할 예정

순열중요도를 통해 중요도가 낮은 특성도 제외

순열 중요도는 각각의 특성을 무작위로 배열하여 제 기능을 하지 못하게 만든 뒤 결과를 비교하여 각 특성의 중요도를 계산

중요도가 -인 특성이나 0.001이하인 특성을 제외해도 성능은 거의 영향이 없으며, 모델학습 속도는 개선됩니다.

- 중요도가 낮은 특성: 중요도가 0.001이하인 하위 6개

Weight	Feature
0.1286 ± 0.0049	Inflight wifi service
0.1198 ± 0.0053	Type of Travel
0.0504 ± 0.0015	Customer Type
0.0346 ± 0.0011	Online boarding
0.0223 ± 0.0017	Checkin service
0.0215 ± 0.0012	Class
0.0174 ± 0.0016	Baggage handling
0.0169 ± 0.0026	Seat comfort
0.0168 ± 0.0024	Cleanliness
0.0148 ± 0.0010	Inflight service
0.0049 ± 0.0009	On-board service
0.0045 ± 0.0014	Leg room service
0.0038 ± 0.0012	Inflight entertainment
0.0030 ± 0.0009	Age
0.0019 ± 0.0005	Gate location
0.0016 ± 0.0008	Departure/Arrival time convenient
0.0008 ± 0.0007	Ease of Online booking
0.0004 ± 0.0005	Flight Distance
0.0004 ± 0.0004	Food and drink
0.0002 ± 0.0008	Arrival Delay in Minutes
0.0000 ± 0.0003	Departure Delay in Minutes
-0.0005 ± 0.0003	Gender

Feature Engineering

- '만족도'에 영향을 미치지 않는 특성

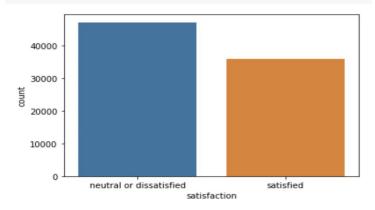
(게이트위치, 편리한 출발/도착시간, 출발/도착지연시간, 성별) 제거

- 중요도가 낮은 특성 (기내식, 비행거리, 쉬운 온라인예약) 제거

분류문제를 풀기전에 먼저 타겟 범주가 어떤 비율을 가지고 있는지 확인

'불만족'과 '만족'이 학습셋, 검증셋, 테스트셋 모두 0.56, 0.43으로 균형을 이룸, 그대로 진행

sns.countplot(x=y_train); # 타겟이 균형을 이루고 있다.



val['satisfaction'].value_counts(normalize=True)

neutral or dissatisfied 0.566094 satisfied 0.433906 Name: satisfaction, dtype: float64

test['satisfaction'].value_counts(normalize=True)

neutral or dissatisfied 0.561018 satisfied 0.438982 Name: satisfaction, dtype: float64

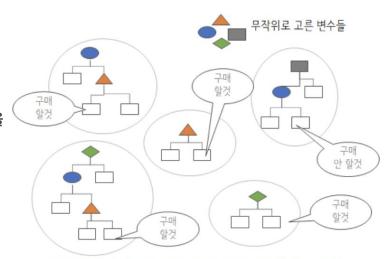
모델학습 1. 랜덤포레스트

랜덤포레스트란?

"무작위 숲" 랜덤 포레스트는 훈련을 통해 구성해놓은 다수의 의사결정나무들로부터 분류 결과를 취합해서 결론을 얻는,

일종의 인기 투표

분류문제에서는 의사결정나무를 무작위로 많이 만들어서 최빈값에 투표하는 것



[랜덤포레스트] 4:1로 구매할 것이라고 예측

모델학습 1. 랜덤포레스트

RandomizedSearchCV 이용 하이퍼파라미터 튜닝

- max_depth : 트리의 최대 깊이 (깊이가 깊어지면 과적합될 수 있으므로 적절히 제어 필요)
- n_estimators : 결정트리의 갯수를 지정(무작정 트리 갯수를 늘리면 성능이 좋아지는 것 대비 시간이 걸릴수 있음)
- max features : 최적의 분할을 위해 고려할 최대 feature 개수

검증세트 정확도 0.9636206149					
	precision	recall	f1-score	support	
neutral or dissatisfied	0.96	0.98	0.97	11764	
satisfied	0.97	0.94	0.96	9017	
accuracy			0.96	20781	
macro avg	0.96	0.96	0.96	20781	
weighted avg	0.96	0.96	0.96	20781	

한습세트 정확도 0.9868267507188143

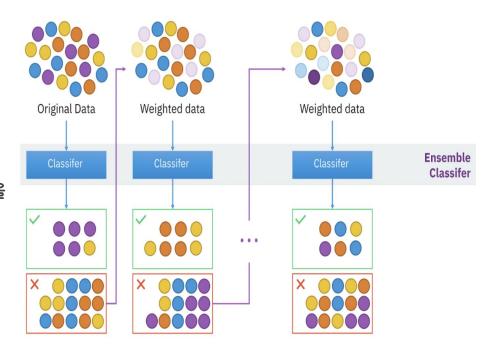
- 최적 하이퍼파라미터:{'n_estimators': 150, 'max_features': 10, 'max_depth': 17}
- ▷ 학습셋의 정확도와 검증셋의 정확도가 크게 차이나지 않는 것으로 과적합 방지

모델학습 2, XGBoost

XGBoost란? Extreme Gradient Boosting

Boosting: 약한 모델들의 **학습 에러에 가중치**를 두고, 순차적으로 다음 학습모델에 반영하여 하나의 강한 모델을 만드는 과정

Gradient Boosting은 시간이 오래걸리는 모델인데 이 모델의 단점을 보완하여 병렬연산을 지원하므로써 계산속도를 향상 시킨 모델



모델학습 2, XGBoost

RandomizedSearchCV 이용 하이퍼파라미터 튜닝

- max_depth : 트리의 깊이
- min_child_weight(기본값1): leaf node에 포함되는 최소 관측 수, 작은값을 가질수록 과적합 발생 가능성높음
- colsample bytree(기본값1): 트리 생성에 필요한 피처의 샘플링에 사용
- n estimators : 약한 학습기의 개수. 약한 학습기가 순차적으로 오류를 보정하므로 개수가 많을수록 예측

- 11 63011	iators. $\neg =$	コロハー	게ㅜ. ㄱ 근	コロハハ	느 시 ㄱ -	ㅗ ㅗㅠᆯ ㅗㅇ이ㅡㅗ 게ㅜ기 ᆭᆯㅜㅋ 에ㄱ
학습세트 정확도 0.9784295561					\triangleright	최적 하이퍼파라미터:'n_estimators': 400,
검증세트 정확도 0.9633318897	7069438					<pre>'min_child_weight': 10, 'max_depth': 7,</pre>
	precision	recall	f1-score	support		<pre>'colsample bytree': 0.8}</pre>
						_
neutral or dissatisfied	0.96	0.98	0.97	11764		
satisfied	0.97	0.95	0.96	9017	\triangleright	학습셋의 정확도와 검증셋의 정확도가 크게
						차이나지 않는 것으로 과적합 방지
accuracy			0.96	20781		시아니시 ほこ スーエ ガーロ こん
macro avg	0.96	0.96	0.96	20781		
weighted avg	0.96	0.96	0.96	20781		

모델학습 2. 랜덤포레스트와 XGBoost 비교

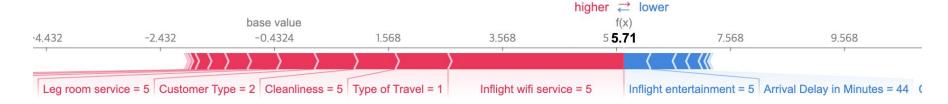
학습세트 정확되 0.9844086474 검증세트 정확되 0.9619845050	767528		61	학습세트 정확도 0.9757 <mark>227241</mark> 검증세트 정확 <mark>도</mark> 0.963 <mark>7</mark> 800105		recall	f1-score
	precision	recall	f1-score		precision	recarr	11-50016
neutral or dissatisfied	0.95	0.98	0.97	neutral or dissatisfied	0.96	0.98	0.97
satisfied	0.97	0.94	0.96	satisfied	0.97	0.94	0.96
accuracy			0.96	accuracy			0.96
macro avg	0.96	0.96	0.96	macro avg	0.96	0.96	0.96
weighted avg	0.96	0.96	0.96	weighted avg	0.96	0.96	0.96

- ▷ XGBoost에서 학습셋의 정확도와 검증셋의 정확도가 비슷하게 나옴.
- ▷ 랜덤포레스트에서 조금 더 과적합이 발생
- ▶ Precision, recall, f1-score 모두 값의 차이가 없음
- ▷ 랜덤포레스트 XGBoost 모두 좋은 성능을 가진 모델
- ▶ XGBoost로 학습하겠습니다.

일반화성능 제시

정확도 0.9633507853403142	precision	recall	f1-score	support	
neutral or dissatisfied satisfied	0.96 0.97	0.98	0.97	14573 11403	모델 학습과 검증에서 높은 정확도를 보여주었고, 테스트셋에
accuracy	0.06	0.06	0.96	25976	대해서도 높은 정확도가 나왔습니다.
macro avg weighted avg	0.96 0.96	0.96 0.96	0.96 0.96	25976 25976	학습이 잘 되었다고 볼 수 있습니다.

SHAP 이용 모델 해석

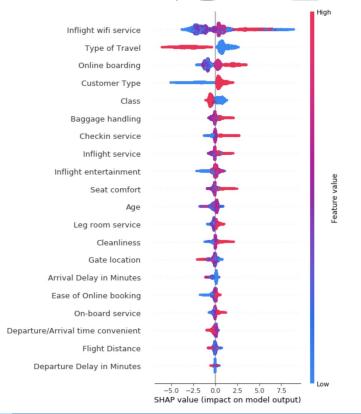


'만족'예측에 긍정적인 영향을 끼친 것은 빨간색

기내 와이파이 서비스, 여행유형, 청결도 순으로 '만족'에 긍정적인 영향 부정적 영향을 끼친 것은 파란색

기내엔터테인먼트에서 부정적인 영향을 준 것으로 볼 수 있습니다.

SHAP 이용 모델 해석



y축은 각 특성, x축은 Shapely value

빨간색: feature값이 높다, 파란색: feature값이 낮다.

예측에 미치는 영향력(=중요도)에 따라 정렬

x축 0 기준, 음의 영역 : 부정적요인 / 양의 영역 : 긍정적요인

→ 결론

기내 와이파이서비스, 온라인탑승정보, 수화물관리가 만족도에 큰 영향을 주기 때문에 우선순위를 두고 관리해야합니다.

그리고 '불만족'인 고객들의 낮은 서비스 점수들을 분석하여 해당 서비스를 개선할 수 있도록 조언하겠습니다.

감사합니다.