5주차 모델링 심화

김부현

1. 과제 목표

- **주제 -** 항공사 승객 만족도 요인분석 및 점수 예측
- **과제 목표** - 항공사에 대한 승객의 만족도를 높이는 요인을 알아보고 점수를 예측해보고자 함.
- **분석결과 활용 프로세스** - 고객 관리, 서비스 품질 개선, 수익성 유지

2. 분석 주요 내용

1. EDA

- 결측값 확인
- 고객 및 항공 서비스 정보 분석
- 변수간 상관관계 파악 (Correlation coefficient, p-value, Cronbach-alpha 등)
- 변수 처리 및 변환 – 리코딩, 원 핫 인코딩, 파생변수 추가 등

2. 모델링 및 결과 도출

- 적합한 모델 선정 – Tree-based, Boosting 등
- 평가기준 – Accuracy, F1-Score, AUC 등의 다양한 평가지표 활용
- 데이터/그래프로 시각화
- 하이퍼파라미터 튜닝 후 최종 점수 예측
- 향후 비즈니스 활용도 분석

3. 분석 데이터셋

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

1. 인구사회학적 요인

Gender: Gender of the passengers (Female, Male)
Age: The actual age of the passengers

2. 승객 유형 분류

Customer Type: The customer type (Loyal customer, disloyal customer)
Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance: The flight distance of this journey

3. 분석 데이터셋

3. 개별 서비스 만족도
Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
Ease of Online booking: Satisfaction level of online booking
Gate location: Satisfaction level of Gate location
Food and drink: Satisfaction level of Food and drink
Online boarding: Satisfaction level of online boarding
Seat comfort: Satisfaction level of Seat comfort
Inflight entertainment: Satisfaction level of inflight entertainment
On-board service: Satisfaction level of On-board service
Leg room service: Satisfaction level of Leg room service
Baggage handling: Satisfaction level of baggage handling
Check-in service: Satisfaction level of Check-in service
Inflight service: Satisfaction level of inflight service
Cleanliness: Satisfaction level of Cleanliness

4. 지연시간
Departure Delay in Minutes: Minutes delayed when departure
Arrival Delay in Minutes: Minutes delayed when Arrival

5. 항공 만족도 (결과변수 - 이진형)
Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

4. 초기 분석

*Training Dataset 80%, Test Dataset 20%
Stratified 10-fold Cross Validation

Method	Model	Accuracy	AUC	Recall	Precision	F1
Boosting	CatBoost	0.9632	0.9949	0.9415	0.9715	0.9563
	LGBM	0.9627	0.9944	0.9369	0.9750	0.9556
	XGBoost	0.9621	0.9946	0.9406	0.9699	0.9550
	Gradient Boosting	0.9406	0.9871	0.9151	0.9444	0.9295
	AdaBoost	0.9264	0.9767	0.9075	0.9193	0.9134
Tree-based	Random Forest	0.9607	0.9933	0.9354	0.9716	0.9532
	Extra Trees	0.9588	0.9928	0.9323	0.9701	0.9508
	Decision Tree	0.9432	0.9423	0.9362	0.9314	0.9338
Discriminant Analysis	LDA	0.8937	0.9565	0.8797	0.8727	0.8762
	QDA	0.8655	0.9462	0.8310	0.8512	0.8410
Regression	Logistic Regression	0.8919	0.9556	0.8814	0.8679	0.8746
KNN	K Nearest Neighbor	0.7520	0.8140	0.6914	0.7184	0.7046

Catboost가 1위,

전반적으로 Boosting과 Tree-based 기반의 방법이 비교적 점수가 높은 경향을 보임.

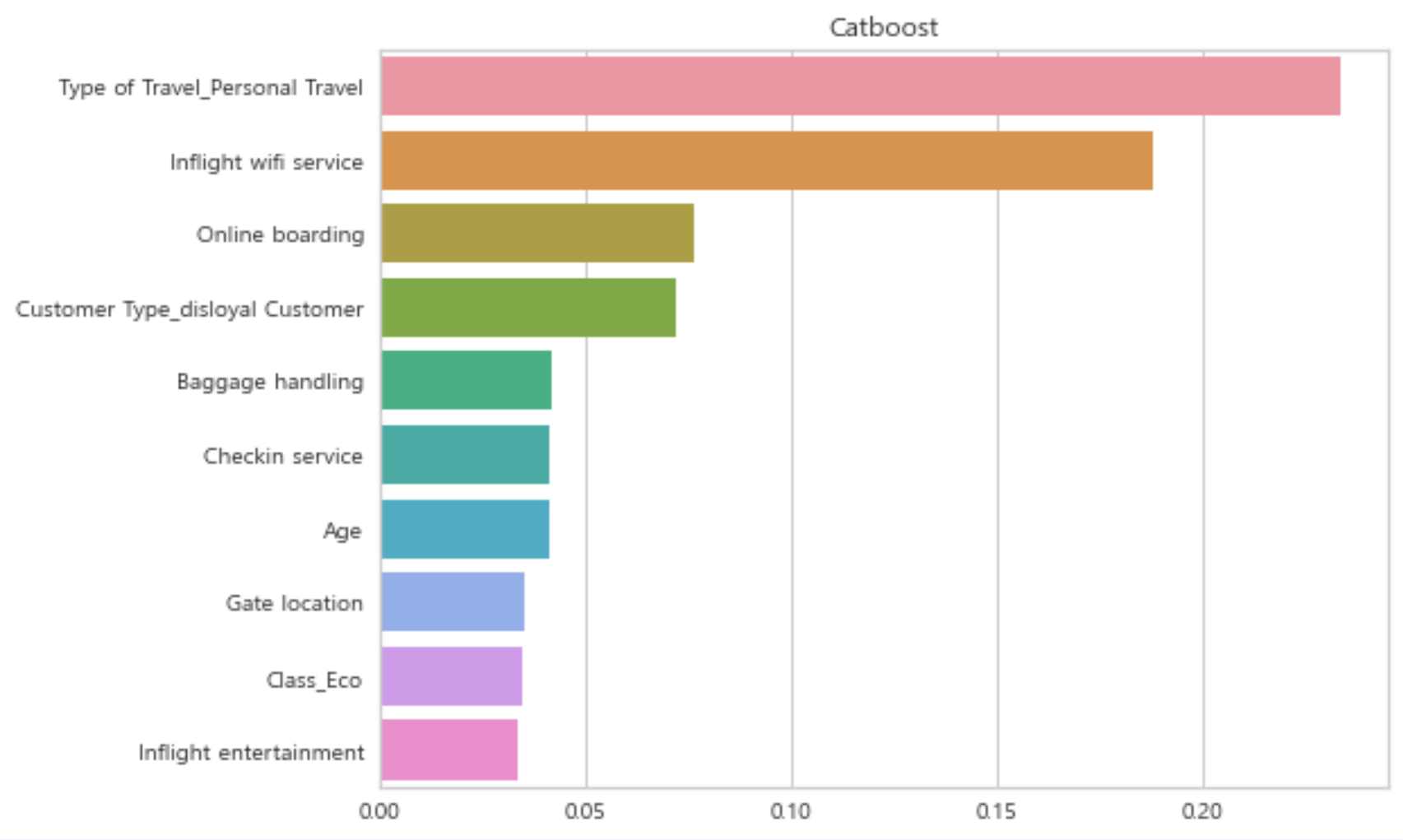
5. Catboost

- 2017년 러시아 회사 Yandex에서 출시.
- Gradient Boosting Machines 알고리즘 기반의 모델.
- GBM의 단점인 학습 시간이 오래 걸린다는 점과, 과적합 문제를 개선.
- 별도의 파라미터 조절 없이도 효율적인 성능을 보여주는 것으로 알려져 있음.

6. Catboost의 주요 하이퍼파라미터 튜닝

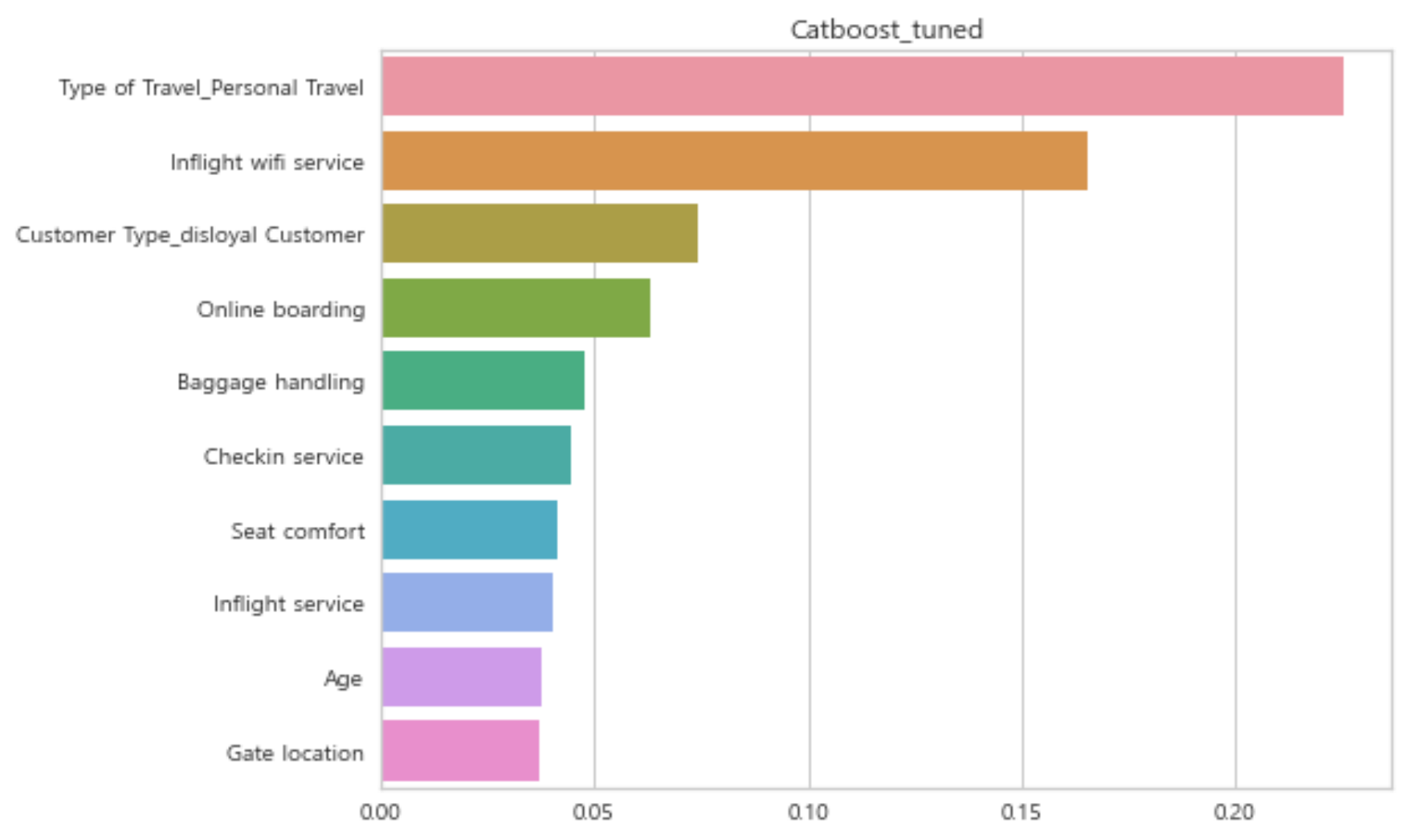
Hyperparameter	Default	Tuned	Description
iterations	1000	260	반복횟수, 의사결정나무 가지의 최대 개수 (=n_estimators)
learning_rate	0.065	0.05	학습률
depth	6	11	깊이
max_leaves	64	2048	최대 잎 개수
l2_leaf_reg	3	2	L2 규제, 손실함수에 가중치의 제곱을 더한다. 전체적으로 가중치를 작아지게 함으로 과적합을 방지하는 파라미터.
random_strength	1	0.6	트리 구조를 선택할 때 분할 평가 요소에 무작위 점수를 더함. 과적합을 방지하는 파라미터.

7. Feature Importance Plot - Catboost



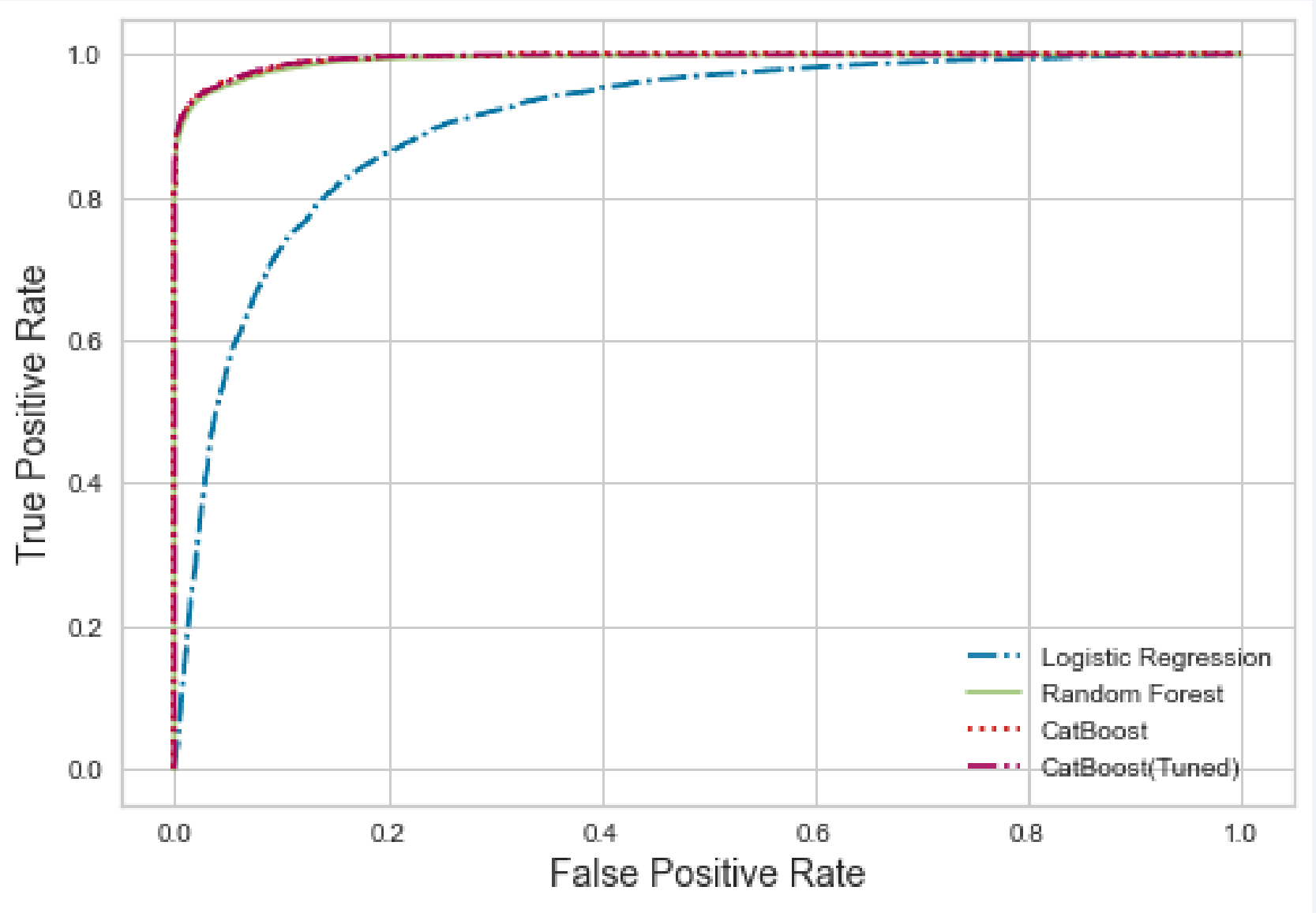
여행타입 (개인목적), 와이파이 서비스 편의성, 온라인 보딩 순으로 결과변수에 영향을 끼침.

8. Feature Importance Plot – Catboost (Hyperparameter Tuning)



여행타입 (개인목적), 와이파이 서비스 편의성, 고객타입 (비충성고객) 순으로
결과변수에 영향을 끼침.

9. ROC Curve


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
Random Forest
CatBoost
CatBoost(Tuned)

10. 최종 점수

* Accuracy 중심 Hyperparameter 튜닝

Train					
Hyperparameter	Accuracy	AUC	Recall	Precision	F1
Default	0.9632	0.9949	0.9415	0.9715	0.9563
Tuned	0.9640	0.9949	0.9400	0.9749	0.9572
Test					
Hyperparameter	Accuracy	AUC	Recall	Precision	F1
Default	0.9629	0.9946	0.9392	0.9724	0.9555
Tuned	0.9627	0.9947	0.9362	0.9747	0.9550

튜닝 후에 Accuracy 점수를 살펴보면, Train Dataset에서 상승했으나, Test Dataset에서는 오히려 감소함.

11. 예상 원인

1. Dataset 분포가 양극화되어 Accuracy가 이미 매우 높은 편이라 추가로 점수 지표를 끌어올리기 어려움.
2. 과적합에 연관성이 있는 l2_leaf_reg, random_strength 파라미터가 보수적으로 조정됐을 가능성이 있음.

12. 결론

1. 전반적으로 Boosting과 Tree-based 기반의 방법이 비교적 점수가 높은 경향을 보임.

2. 기내 와이파이 편의성과 온라인 보딩 시스템 만족도가 큰 승객은, 주로 최종 항공 서비스 만족도에 긍정적인 영향을 끼침. 반면 기내 와이파이 편의성에서 1, 2점을 준 고객은 불만족의 비율이 큼.

→ 기내 와이파이 서비스를 재정비하고, 수요조사를 실시하는 등의 필요성이 보여짐.

고객의 주 사용처가 유튜브, 넷플릭스와 같은 OTT 서비스라면 최소한 동영상 스트리밍이 원활할 정도의 환경을 구축하는 것이 필요함.

온라인 보딩 시스템 (사전 탑승 수속)에 대해서는 문제점 파악을 위해 전반적으로 고객 대상으로 설문조사 등의 필요성이 보여짐.

탑승 수속의 대기시간 문제 개선, 사전결제 방법의 편리성, 더 쉬운 예약을 위한 웹페이지 UI 개선 등.

3. 개인적인 목적으로 여행을 가는 승객은, 주로 최종 항공 서비스 만족도에 부정적인 영향을 끼침.

→ 개인 목적으로 여행을 가는 고객들의 만족도를 상승시킬 방안이 필요함.

개인고객 대상으로 SNS 홍보, 부가적인 혜택 마련 등.