머신러닝을 활용한 승객 만족도 개선 프로젝트

Project for Predict and improve passenger satisfaction

발표자 : 김수영

FLOW OF CONTENTS

문제 정의와 가설 설정

The first step in any project is defining your problem. You can use the most powerful and shiniest algorithms available, but the results will be meaningless if you are solving the wrong problem.

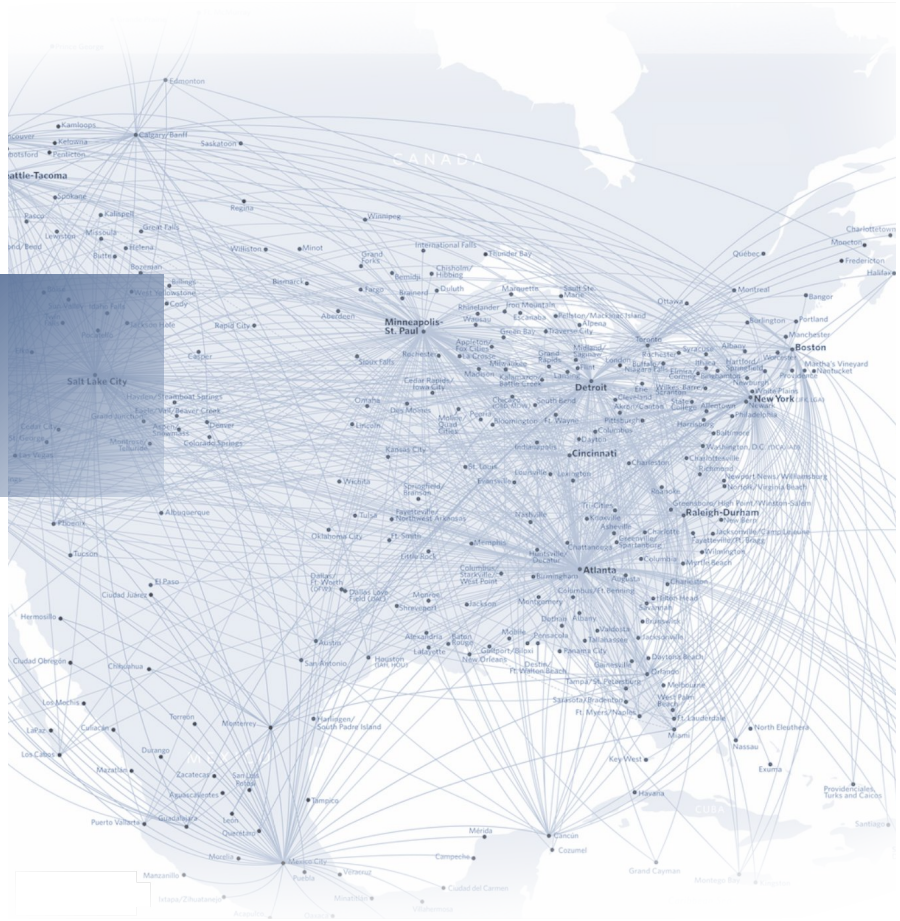
01

문제 정의

프로젝트 목표:
승객 만족도를 예측하는 분류 모델을 설계하고
모델해석을 통해 항공사의 고객 만족도 개선을 위한 인사이트 도출

데이터 분석 및 머신러닝 모델해석을 통해 다음과 같은 문제들을 해결하고자 합니다.

1. 승객 만족도 개선에 가장 중요한 특성은 무엇인가요?
2. 비행 지연은 승객의 만족도에 어떤 영향을 주나요? 비행지연이 발생한 경우에는 어떻게 대처해야 할까요?
3. 나이, 비행 목적, 좌석 등급 등 승객의 조건에 따라 비행 만족도에 영향을 주는 특성들이 어떻게 달라질까요?



ABOUT DATA SET

미국의 한 항공사에서 실시한 설문조사 자료

10만 개 이상의 학습 데이터
2만 개 이상의 평가 데이터

주요 Column (총 24개 Column)

Type of Travel: 승객의 여행 목적 (개인 여행/비즈니스 목적)
Class: 비행기 좌석 등급
Flight distance: 비행 거리(miles)
Inflight wifi service: 기내 와이파이 서비스 만족도
Seat comfort: 시트의 착석감
Baggage handling: 수하물 관리 만족도
Check-in service: 탑승수속 만족도
Departure Delay in Minutes: 출발 지연 시간(분)
Arrival Delay in Minutes: 도착 지연 시간(분)

Satisfaction(target): 비행 만족도 [만족/ 불만족]

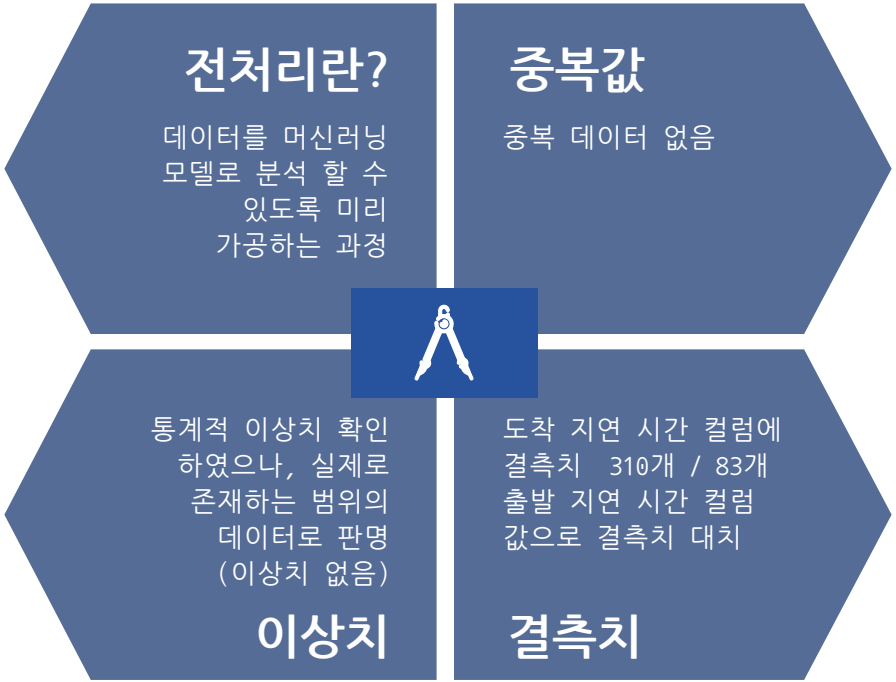
가설 설정

1. 좌석 클래스가 높을수록 만족도가 높을것이다.
2. 좌석의 편안함이 만족도에 가장 많은 영향을 줄 것이다.
3. 지연시간이 30분 이내라면 고객의 만족도에 큰 영향이 없을 것이다.
4. 비즈니스 목적의 승객이, 지연시간에 더 민감하게 반응할 것이다.
5. 연령별로 만족도에 영향을 주는 요인들이 다를것이다.

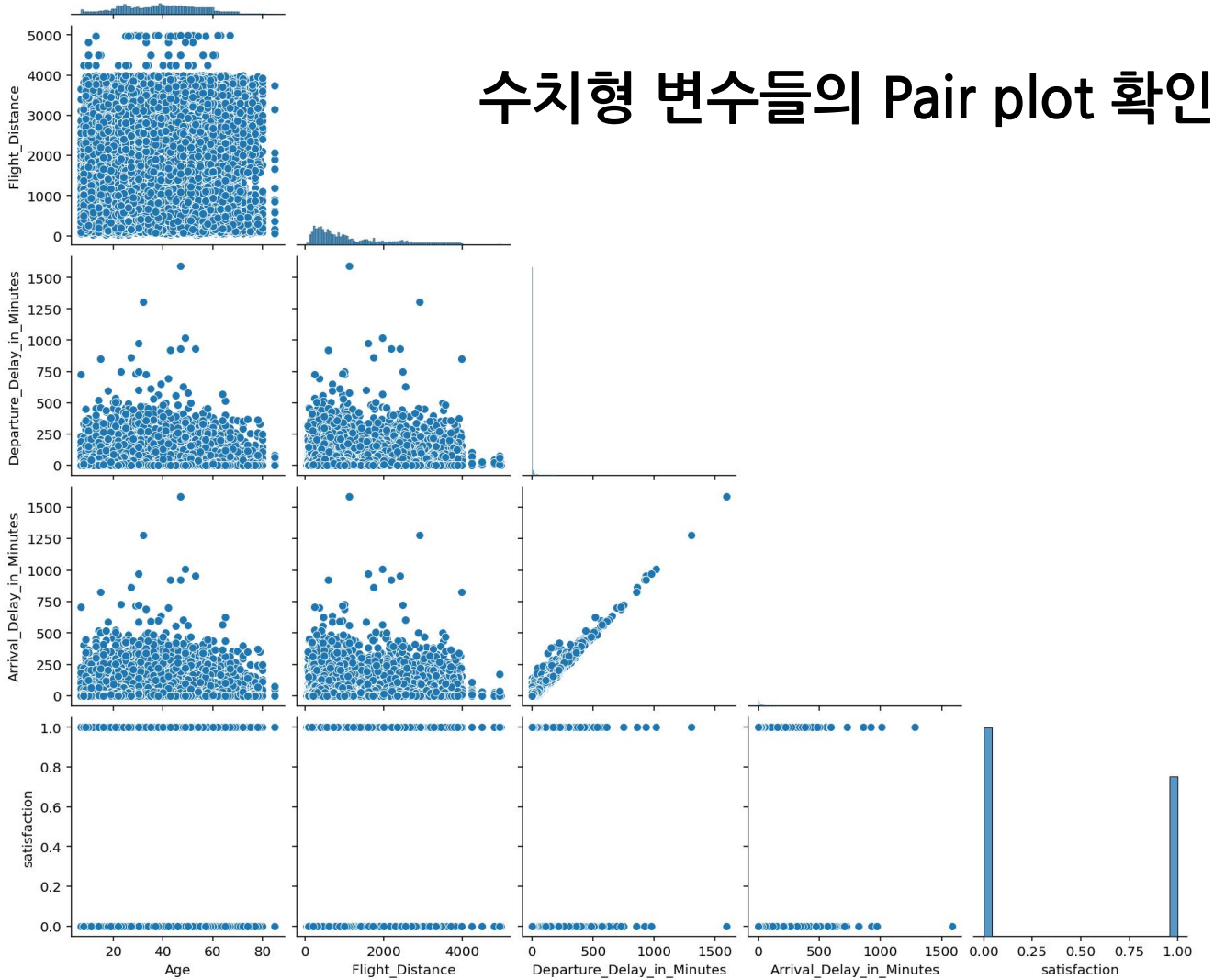
EDA & 전처리

02

DATA PREPROCESSING



수치형 변수들의 Pair plot 확인

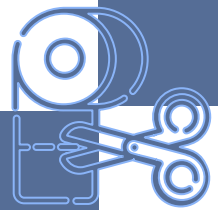


FEATURE ENGINEERING

‘ID’ 컬럼 삭제

‘Overall_Rating’ 특성 추가

0 또는 1~5의 값을 가지는 설문조사 특성들의 평균값



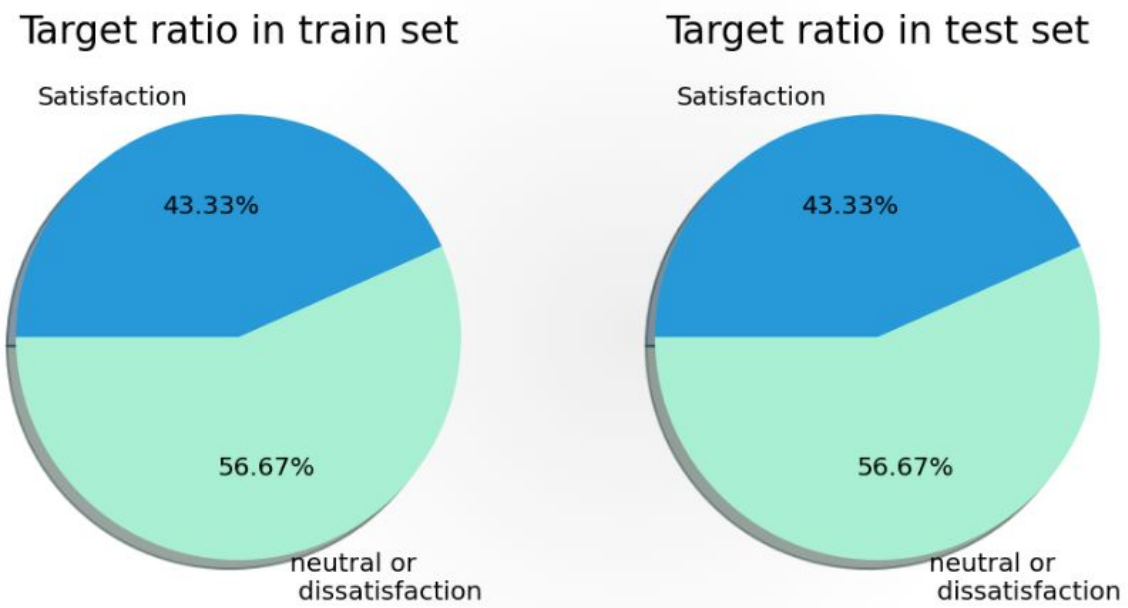
'Average_delay' 특성 추가

지연시간 컬럼 두개의 평균을 값으로 가지는
평균지연시간 특성을 추가

'Departure_Delay_in_Minutes', 'Arrival_Delay_in_Minutes' 삭제

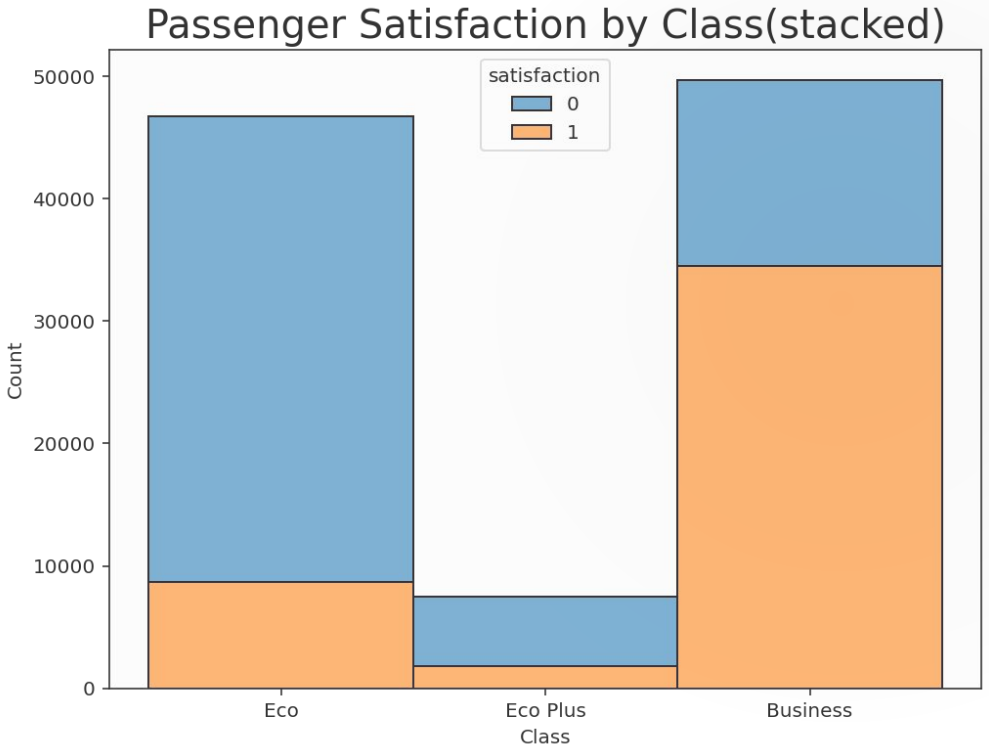
높은 상관관계를 가지는 특성을 삭제하여 모델의 일반화 성능 향상, 모델 해석의 정확도 상승

타겟 분포 확인



비교적 균형적인 타겟 분포를 보임

가설1 좌석 클래스가 높을수록 만족도가 높을것이다.



가설 확인
높은 클래스의 좌석일 수록 승객
만족도가 높은것을 확인 할 수 있다.

머신러닝 모델 학습 및 평가

04

머신러닝 모델

기준모델

최소한의 성능을 나타내는
기준이 되는 모델.
타겟의 최빈클래스(0)를
기준모델로 선정했다.

Logistic Regression

선형회귀에 sigmoid함수를 씌워서
분류 문제에 사용

RandomForest

배깅(bagging) 방식 앙상블
트리기반 알고리즘

LightGBM

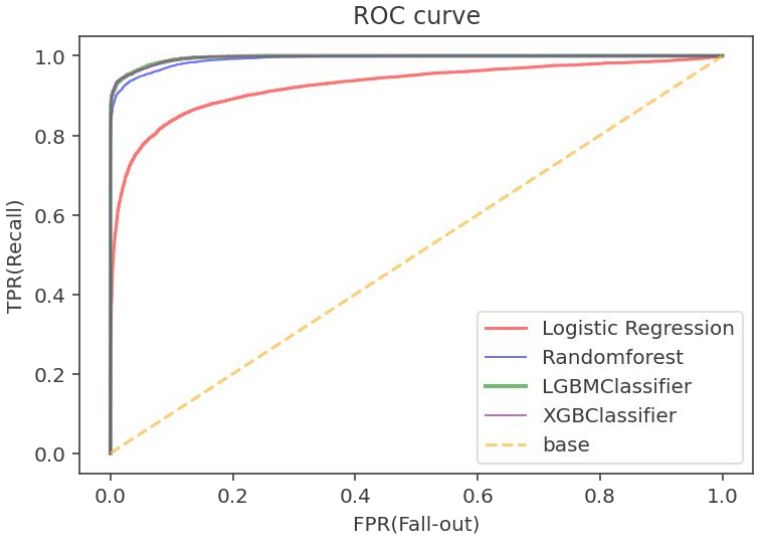
Gradient Boosting 기법의
Tree 기반 학습 알고리즘
빠르다

XGBoost

Extreme Gradient Boosting
Tree 기반 학습 알고리즘

모델 성능 비교

	Base	Logistic	Randomforest	LGBM	XGBoost
accuracy	0.567	0.873	0.954	0.965	0.965
precision	0.000	0.873	0.960	0.976	0.976
recall	0.000	0.829	0.932	0.941	0.942
f1	0.000	0.850	0.946	0.958	0.959
auc	0.500	0.926	0.992	0.996	0.996



주요 평가지표로 f1 사용

성능은 XGB와 거의 비슷하지만 속도면에서 뛰어난 모습을 보이는 LGBMClassifier를 최종모델로 선정

모델 튜닝 및 최종 성능

LGBMClassifier 모델

하이퍼 파라미터 튜닝 결과

colsample_bytree = 0.97

subsample =0.82

learning_rate = 0.08,

max_depth = 12,

min_child_weight = 4,

n_estimators=350,

최종 모델의 교차 검증(cross validation)
f1 -score는 0.96

Test set에 대한 평가지표를 확인하여 일반화 성능을 검증하자.

Evaluation Metrics of Test set
accuracy : 0.965
precision : 0.974
recall : 0.945
f1 : 0.96
auc score : 0.995

보지 않은 데이터인 Test set에 대해서도 높은 평가지표를 나타내므로 일반화 성능이 높은 모델이라고 할수 있음

모델 해석

05

Weight	Feature
0.1839 ± 0.0054	Type_of_Travel
0.1650 ± 0.0036	Inflight_wifi_service
0.0884 ± 0.0015	Customer_Type
0.0315 ± 0.0020	Online_boarding
0.0306 ± 0.0012	Baggage_handling
0.0261 ± 0.0017	Inflight_service
0.0238 ± 0.0019	Checkin_service
0.0235 ± 0.0016	Seat_comfort
0.0180 ± 0.0012	Cleanliness
0.0094 ± 0.0009	Class
0.0064 ± 0.0011	Gate_location
0.0061 ± 0.0012	Age
0.0048 ± 0.0007	On-board_service
0.0034 ± 0.0013	Inflight_entertainment
0.0026 ± 0.0011	Ease_of_Online_booking
0.0022 ± 0.0009	Flight_Distance
0.0009 ± 0.0007	Average_delay
0.0007 ± 0.0007	Overall_Rating
0.0006 ± 0.0004	Leg_room_service
0.0005 ± 0.0005	Departure/Arrival_time_convenient
0.0001 ± 0.0004	Food_and_drink
-0.0000 ± 0.0002	Gender

Permutation Importance

순열 중요도를 확인한 결과

Type_of_Travel

Inflight_wifi_service

Customer_Type, Online_boarding

Baggage_handling

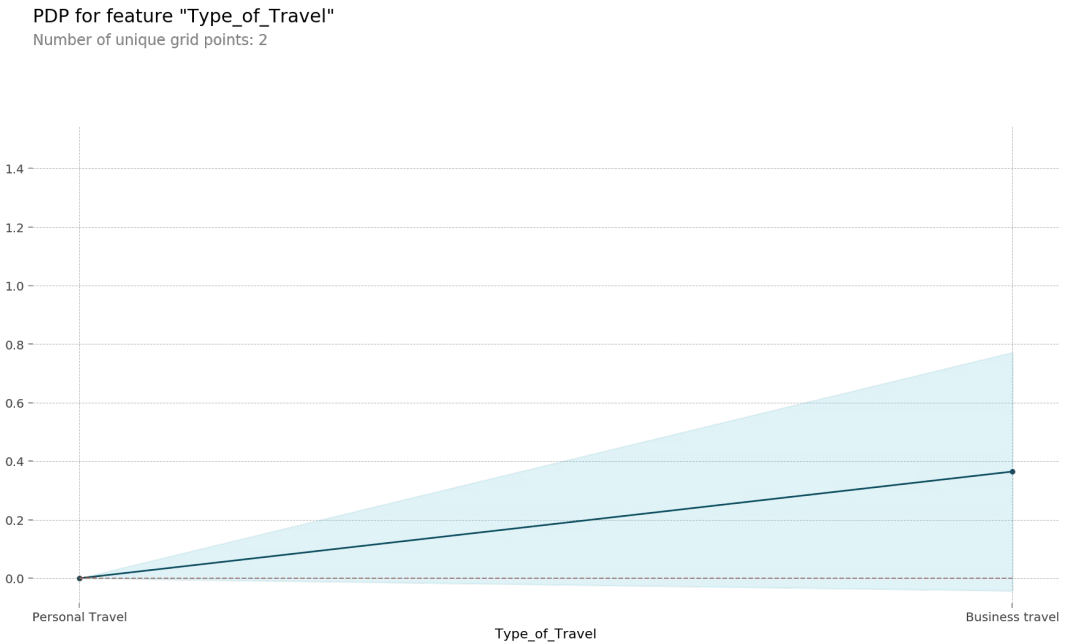
순으로 중요도가 높은 것을 확인했다.

가설2 데이터 특성 중 Seat comfort가 만족도에 가장 많은 영향을 줄 것이다?

순열중요도를 통해 확인한 결과, 비행 만족도에 가장 많은 영향을 주는 요인은 Seat comfort가 아닌 Type_of_Travel임을 알 수 있었다.

가설 기각

Type_of_Travel 특성의 PDP 확인

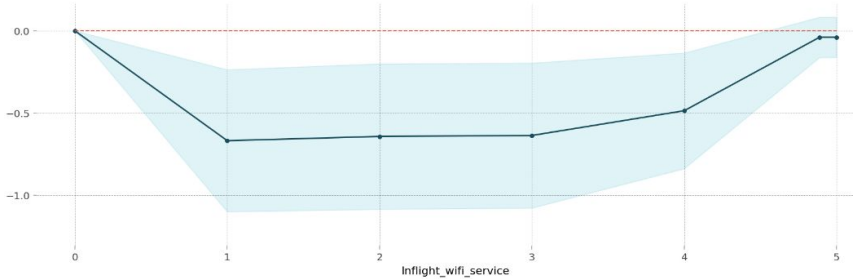


비즈니스 목적인 경우에 개인 여행 목적의 승객보다 만족도가 높은 경향을 보입니다.

순열중요도 높은 PDP 확인

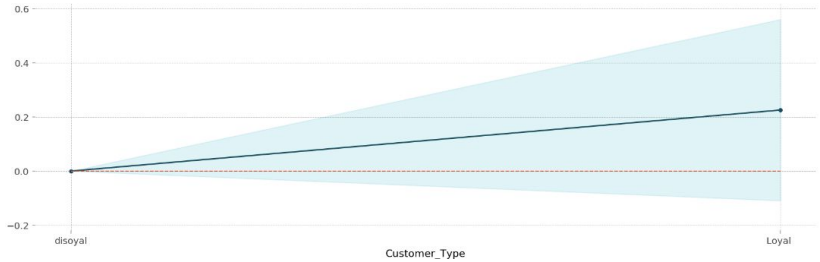
PDP for feature "Inflight_wifi_service"

Number of unique grid points: 7



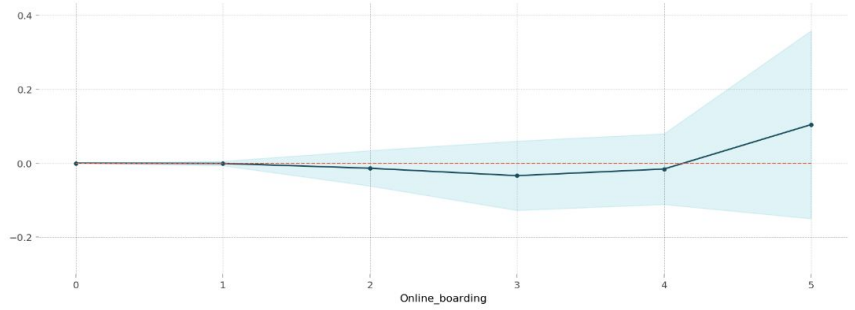
PDP for feature "Customer_Type"

Number of unique grid points: 2



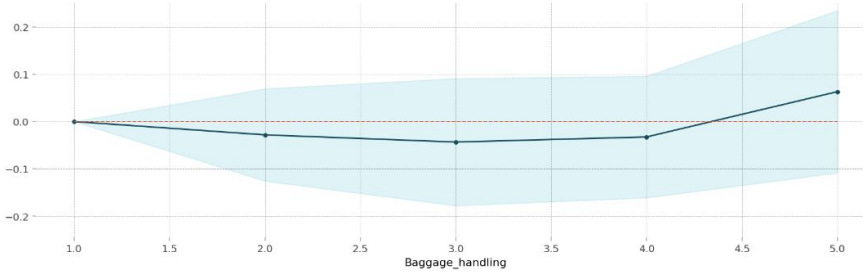
PDP for feature "Online_boarding"

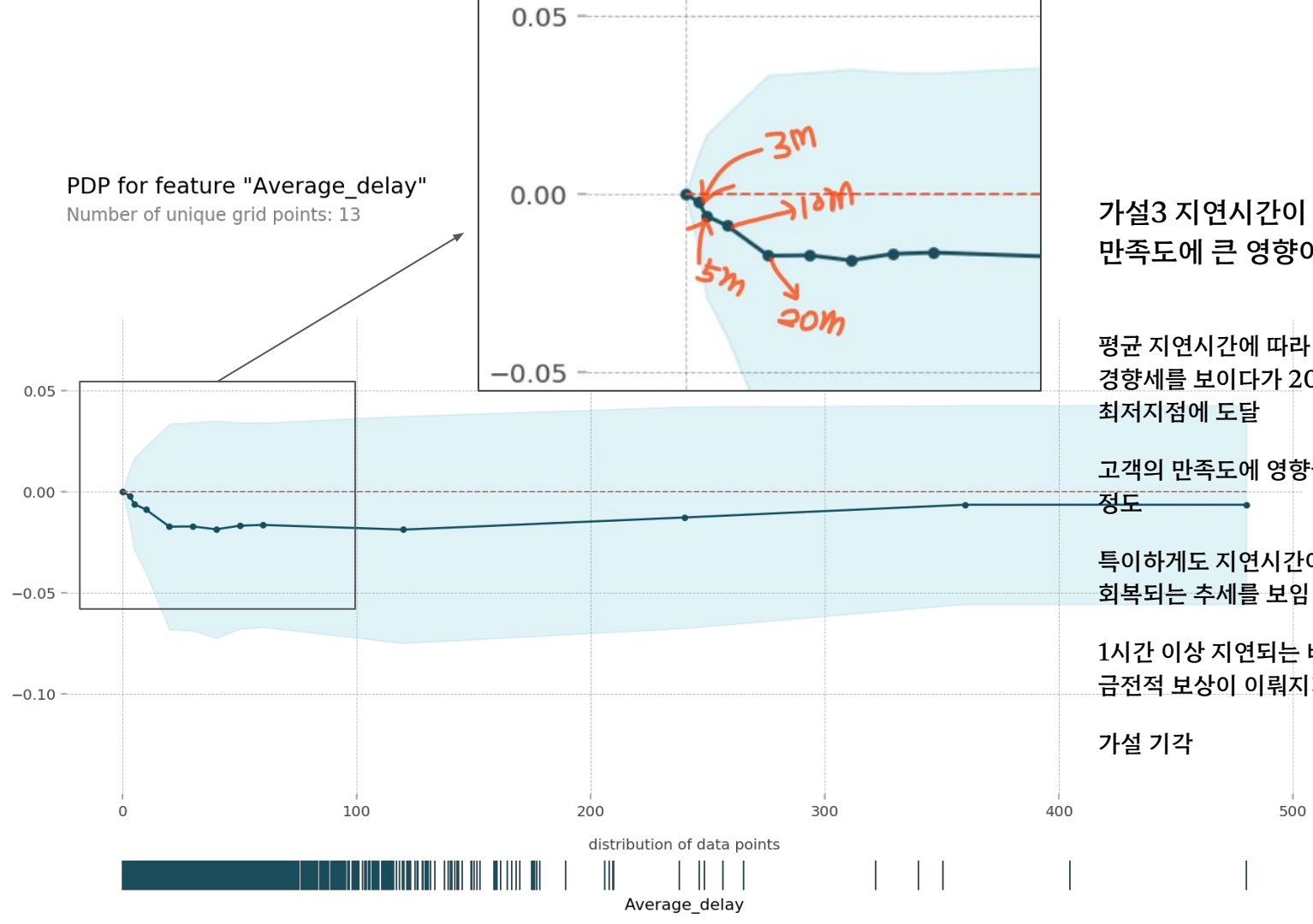
Number of unique grid points: 6



PDP for feature "Baggage_handling"

Number of unique grid points: 5





가설3 지연시간이 30분 이내라면 고객의 만족도에 큰 영향이 없을 것이다.

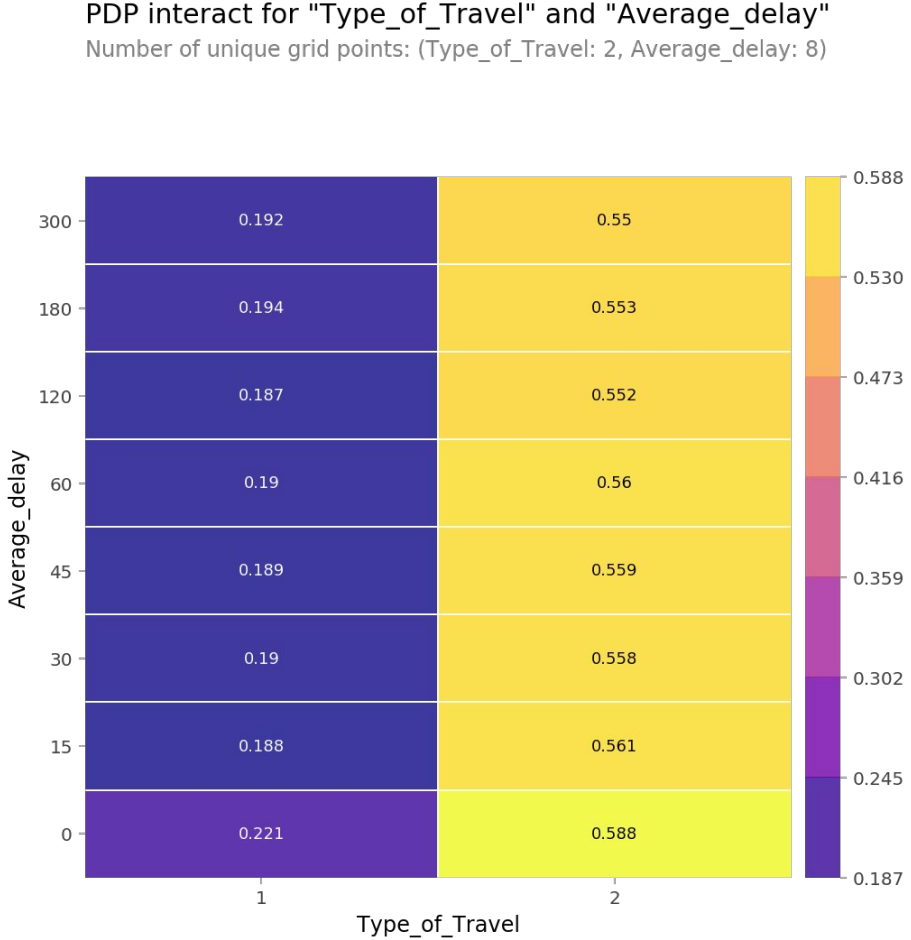
평균 지연시간에 따라 타겟 확률이 이미 하락하는 경향세를 보이다가 20분이 되면 하락세가 거의 최저지점에 도달

고객의 만족도에 영향을 주는 지연시간은 3~20분 정도

특이하게도 지연시간이 길어질수록 만족도가 다시 회복되는 추세를 보임

1시간 이상 지연되는 비행에 대해서는 규정상 금전적 보상이 이뤄지기 때문이 아닌가 추측

가설 기각



가설4 비즈니스 목적의 승객의 경우,
지연시간에 더 민감하게 반응할 것이다.

눈에 띄는 큰 차이는 보이지 않음

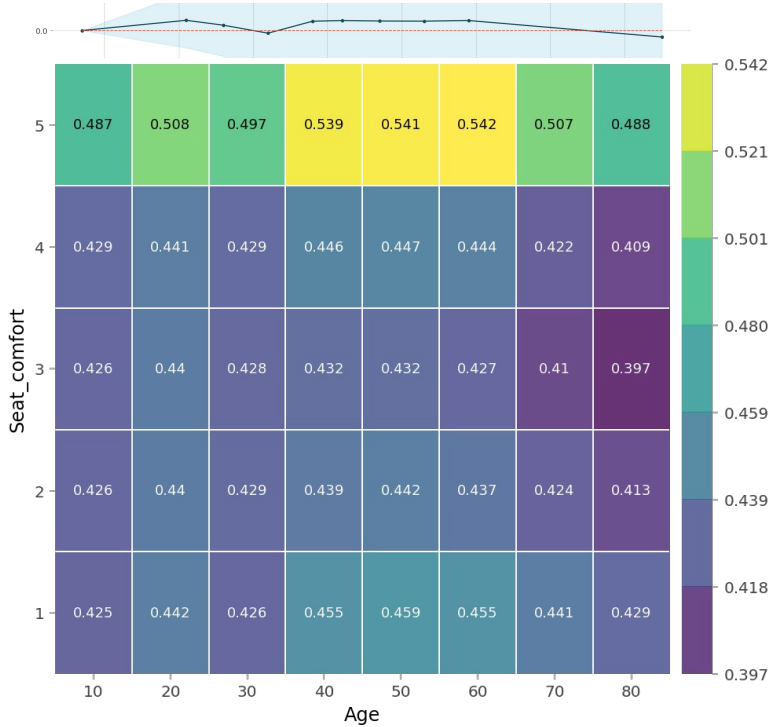
지연시간이 발생하면 만족도가 떨어지는 것은 확실하다

하지만 고객의 여행 목적에 따라 그 정도가 크게 다르지 않았다.

가설 기각

PDP interact for "Age" and "Seat_comfort"

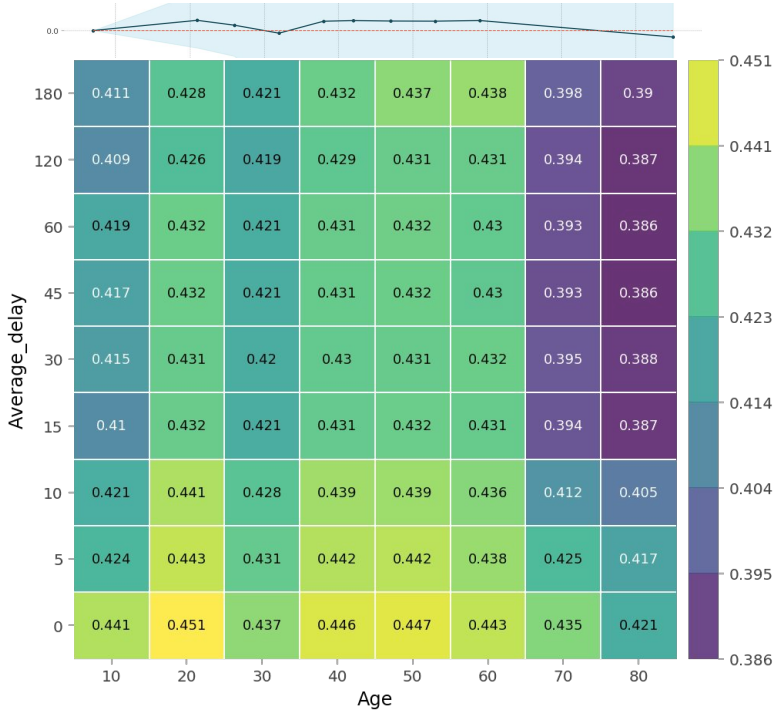
Number of unique grid points: (Age: 8, Seat_comfort: 5)



40대~60대의 경우 좌석의 편안함이 만족도에 영향을 주는 정도가 다른 연령대 보다 높게 나타났습니다.

PDP interact for "Age" and "Average_delay"

Number of unique grid points: (Age: 8, Average_delay: 9)



70대이상의 고령층에게는 평균 지연시간이 10분 이상 생겼을 때, 만족도에 주는 부정적 영향이 다른 연령대보다 더 크게 나타나는것을 확인할 수 있습니다.

가설5. 연령별로 만족도에 영향을 주는 요인들이 다를것이다.

가설과 동일하게 연령별로 비행 만족도에 영향을 주는 요인들이 다른것을 확인 할 수 있었습니다.

결론 및 인사이트 도출

서비스 개선

서비스 개선의 관점에서
기내 와이파이 환경과
온라인 탑승수속 서비스
수하물 관리 시스템을
중점으로 고객 서비스를
개선

비행 지연이 발생한다면?

70세 이상의 고령층에 집중해서
컴플레인 핸들링
1시간 이내의 항공기 지연에도
적당한 보상을 드리는 것이
장기적인 고객 만족도 관리에
도움

좌석 업그레이드

높은 클래스의 좌석일수록
비행 만족도가 높은 경향 보임
출항 전 고급석이 비어있다면
좌석 업그레이드를 제공하는
것이 좋다

시트 개량

중장년층 고객에게 좌석의
착석감에 대한 피드백을
주기적으로 받아서 항공기
시트의 편안함을 개량


감사합니다
PARIS
MEXICO