3주차 모델링

김부현

0. 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

0. 분석 데이터셋

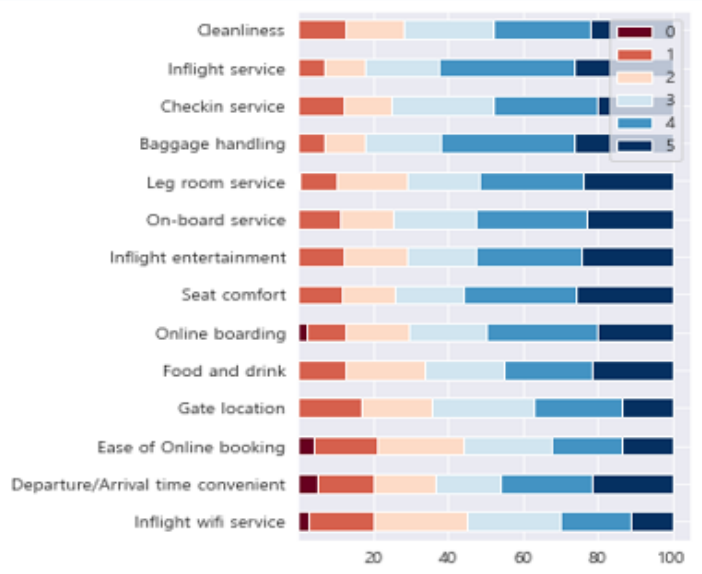
3. 개별 서비스 만족도
Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
Ease of Online booking: Satisfaction level of online booking
Gate location: Satisfaction level of Gate 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)

1. 추가적인 데이터 전처리

5. 개별 서비스 만족도 (5점 척도)

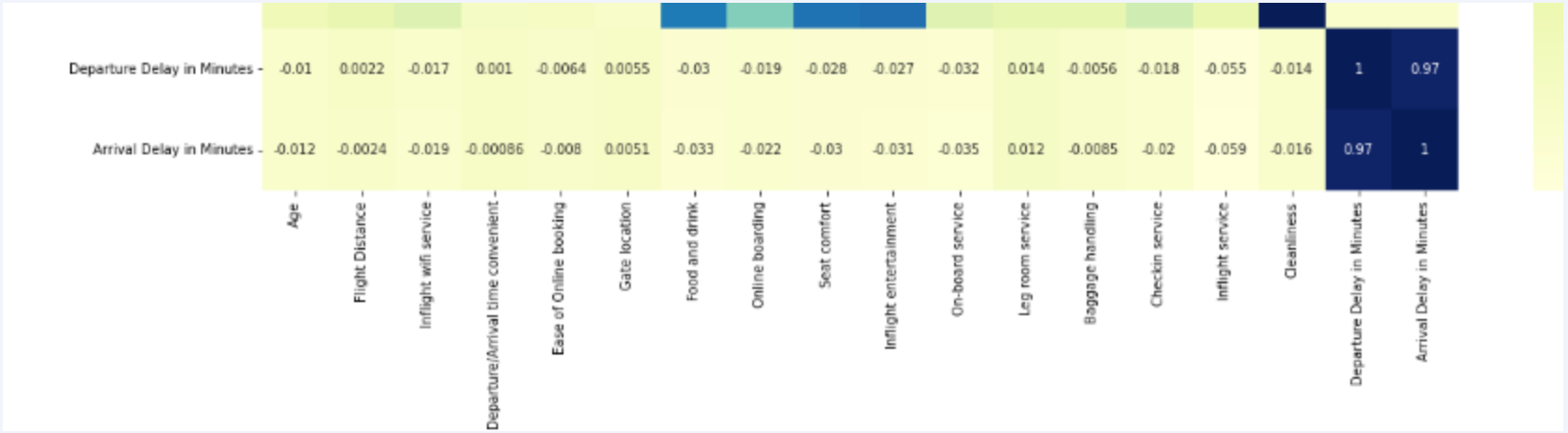


Service	mean	std
Inflight wifi service	2.73	1.33
Departure/Arrival time convenient	3.06	1.53
Ease of Online booking	2.76	1.40
Gate location	2.98	1.28
Food and drink	3.20	1.33
Online boarding	3.25	1.35
Seat comfort	3.44	1.32
Inflight entertainment	3.36	1.33
On-board service	3.38	1.29
Leg room service	3.35	1.32
Baggage handling	3.63	1.18
Checkin service	3.30	1.27
Inflight service	3.64	1.18
Cleanliness	3.29	1.31

5점제 척도의 변수 타입: Numerical, Categorical 중 어떻게 처리할까 고민하다가 결국 둘 다 해보기로 결정.

Numerical로 처리한 경우는 0점 값을 전부 결측치로 보고 제거함. (총 데이터 중에서 약 9% 정도의 손실.)
Categorical로 처리한 경우는 0점부터 시작하는 경우는 0점을, 1점부터 시작하는 경우는 1점을 더미변수로 두고 One-Hot encoding을 함.

1. 추가적인 데이터 전처리



이륙 지연시간과 착륙 지연시간과의 상관계수는 0.97로,

Multicollinearity Problem의 위험성으로 인해 해당 변수를 drop하고 파생변수를 만들어 줌.

파생변수(Delay): Departure Delay와 Arrival Delay의 합

2. Confusion Matrix

Confusion Matrix		실제	
		Positive	Negative
예측	Positive	TP	FP
	Negative	FN	TN

1. Accuracy (정확도)

전제 예측 중 얼마나 잘 맞추었는지 나타내는 지표.

(TP+TN)/All로 정의된다.

2. Recall (재현율)

실제 Positive를 얼마나 잘 예측했는지 나타내는 지표.

TP/(TP+FN)로 정의된다.

3. Precision (정밀도)

Positive로 예측한 것을 얼마나 잘 맞추었는지 나타내는 지표.

TP/(TP+FP)로 정의된다.

4. Specificity (특이도)

실제 Negative를 얼마나 잘 예측했는지 나타내는 지표.

TN/(FP+TN)로 정의된다.

5. F1 Score

Recall과 Precision의 조화 평균.

6. AUC

ROC는 x축은 1- Specificity, y축은 Recall로 나타내는 곡선이며, AUC는 ROC Curve의 밑면적을 계산한 값

3. 초기 분석 (5점 척도 - 범주형)

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

Method	Model	Accuracy	AUC	Recall	Precision	F1
Boosting	CatBoost	0.9646	0.9953	0.9449	0.9727	0.9586
	LGBM	0.9638	0.9950	0.9399	0.9758	0.9575
	XGBoost	0.9633	0.9950	0.9442	0.9703	0.9571
	Gradient Boosting	0.9454	0.9882	0.9214	0.9512	0.9360
	AdaBoost	0.9278	0.9776	0.9096	0.9230	0.9162
Tree-based	Extra Trees	0.9602	0.9929	0.9378	0.9695	0.9534
	Random Forest	0.9594	0.9933	0.9338	0.9714	0.9522
	Decision Tree	0.9465	0.9458	0.9399	0.9371	0.9385
Discriminant Analysis	LDA	0.9310	0.9755	0.9077	0.9315	0.9195
	QDA	0.5162	0.5630	0.9171	0.4712	0.6220
Regression	Logistic Regression	0.9321	0.9782	0.9120	0.9301	0.9209
KNN	K Nearest Neighbor	0.6968	0.7430	0.6355	0.6553	0.6452

4. 초기 분석 (5점 척도 - 연속형)

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

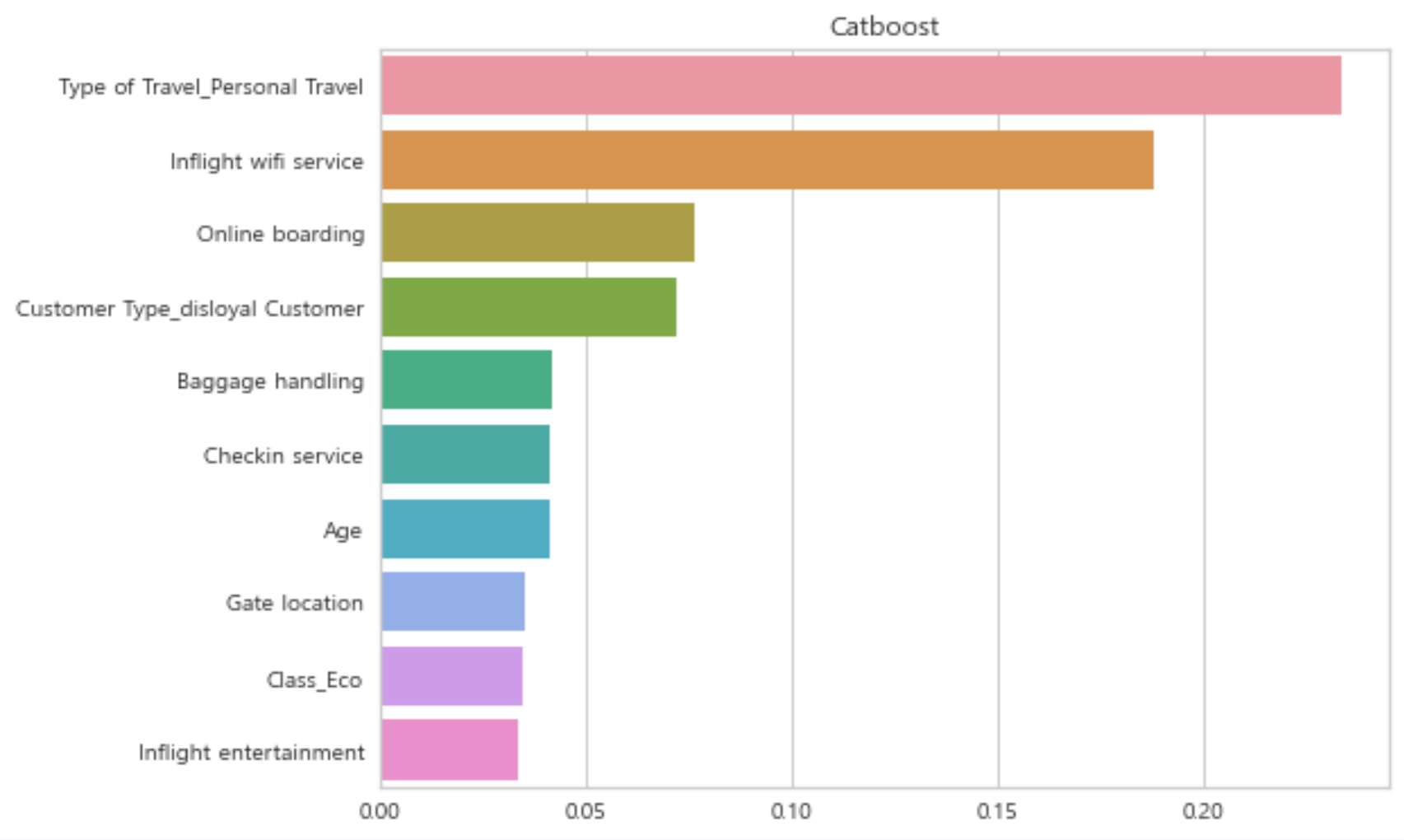
Method	Model	Accuracy	AUC	Recall	Precision	F1
Boosting	CatBoost	0.9632	0.9949	0.9415	0.9715	0.9563
	LGBM	0.9627	0.9944	0.9369	0.9750	0.9556
	XGBoost	0.9621	0.9946	0.9406	0.9699	0.9550
	Gradient Boosting	0.9406	0.9871	0.9151	0.9444	0.9295
	AdaBoost	0.9264	0.9767	0.9075	0.9193	0.9134
Tree-based	Random 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.

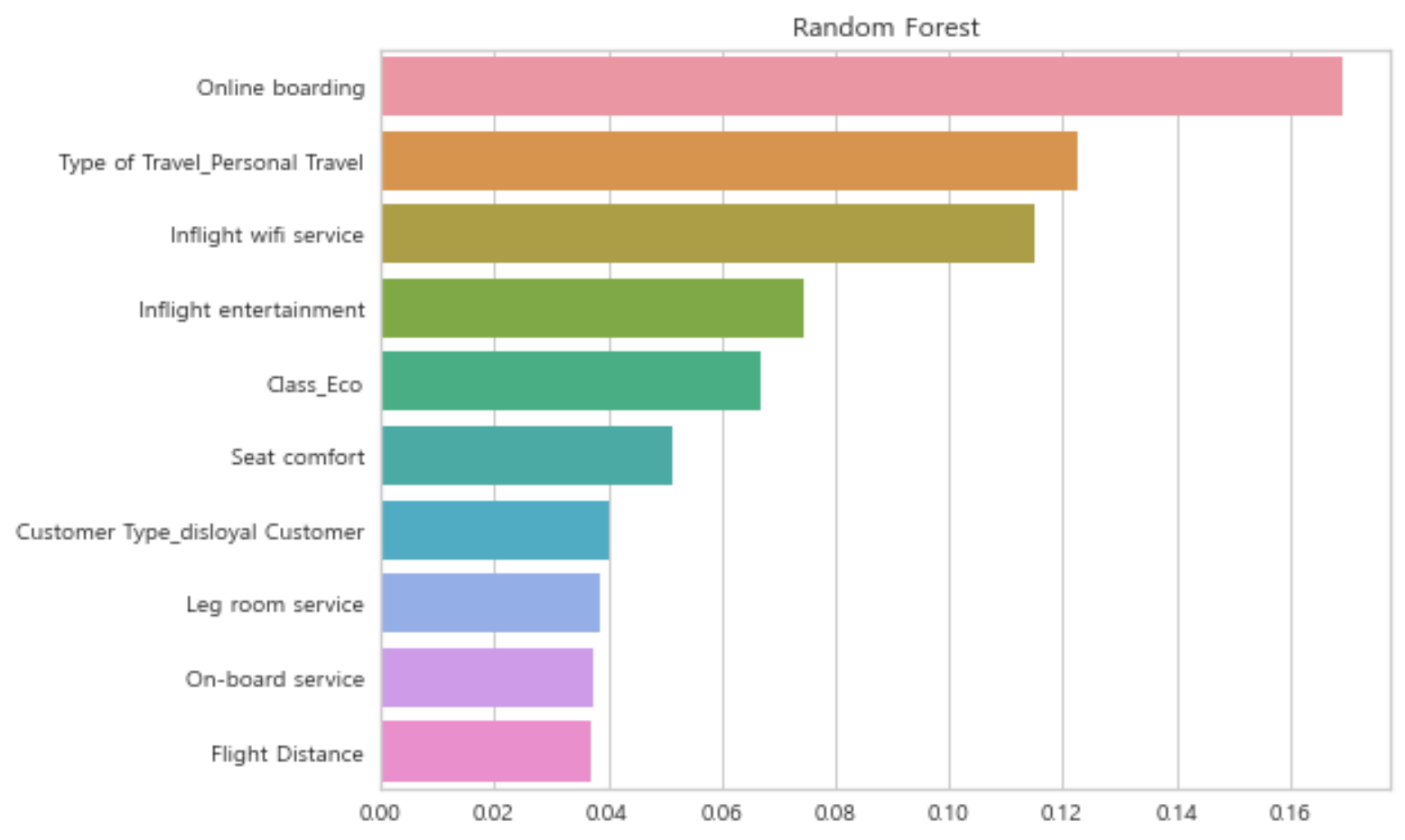
5. Feature Importance Plot - Catboost

*이하의 자료는 전부 5점제 척도를 Numerical로 처리한 기준으로 함.



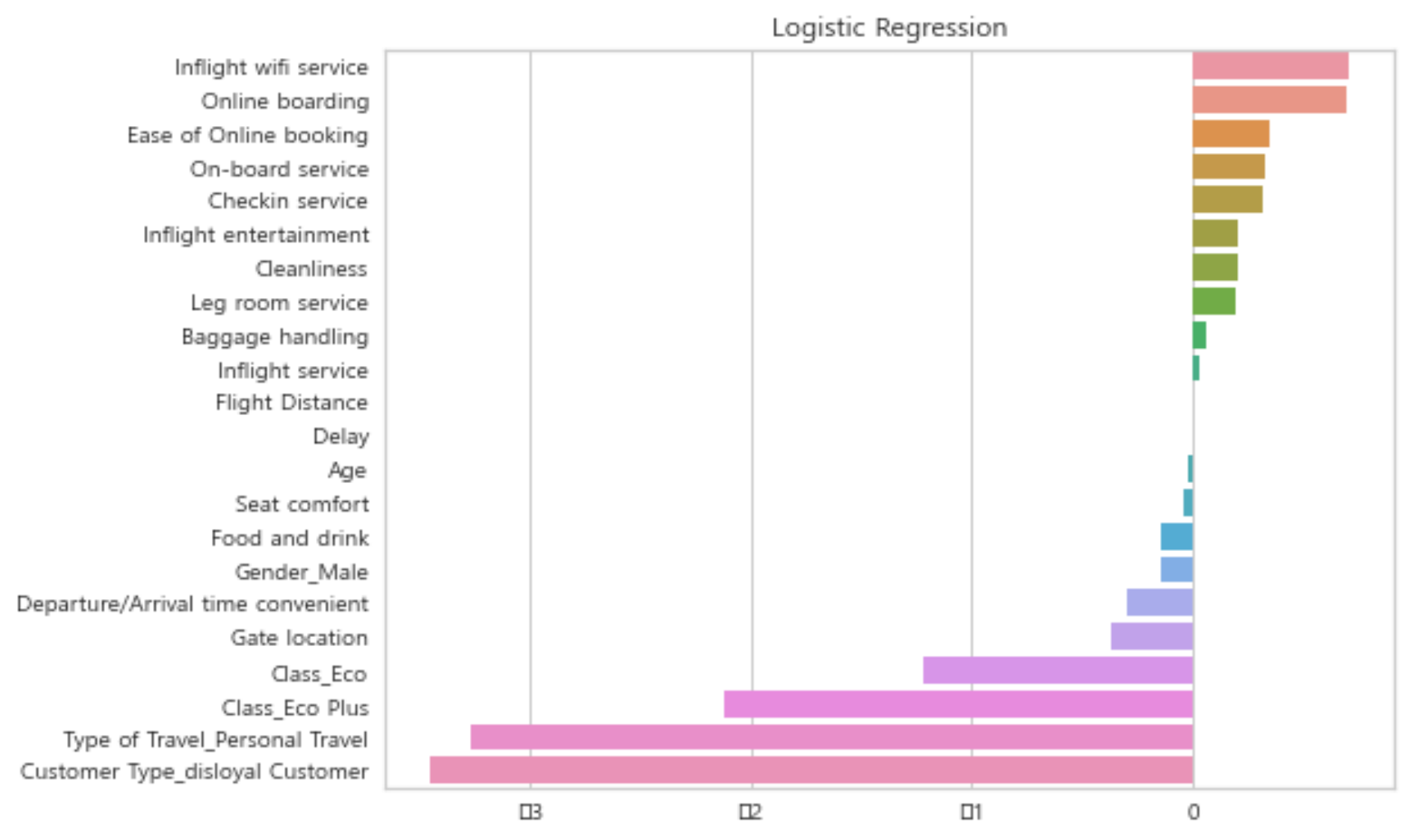
여행타입 (개인목적), 와이파이 서비스 편의성, 온라인 보딩 순으로 결과변수에 영향을 끼침.

5. Feature Importance Plot – Random Forest



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

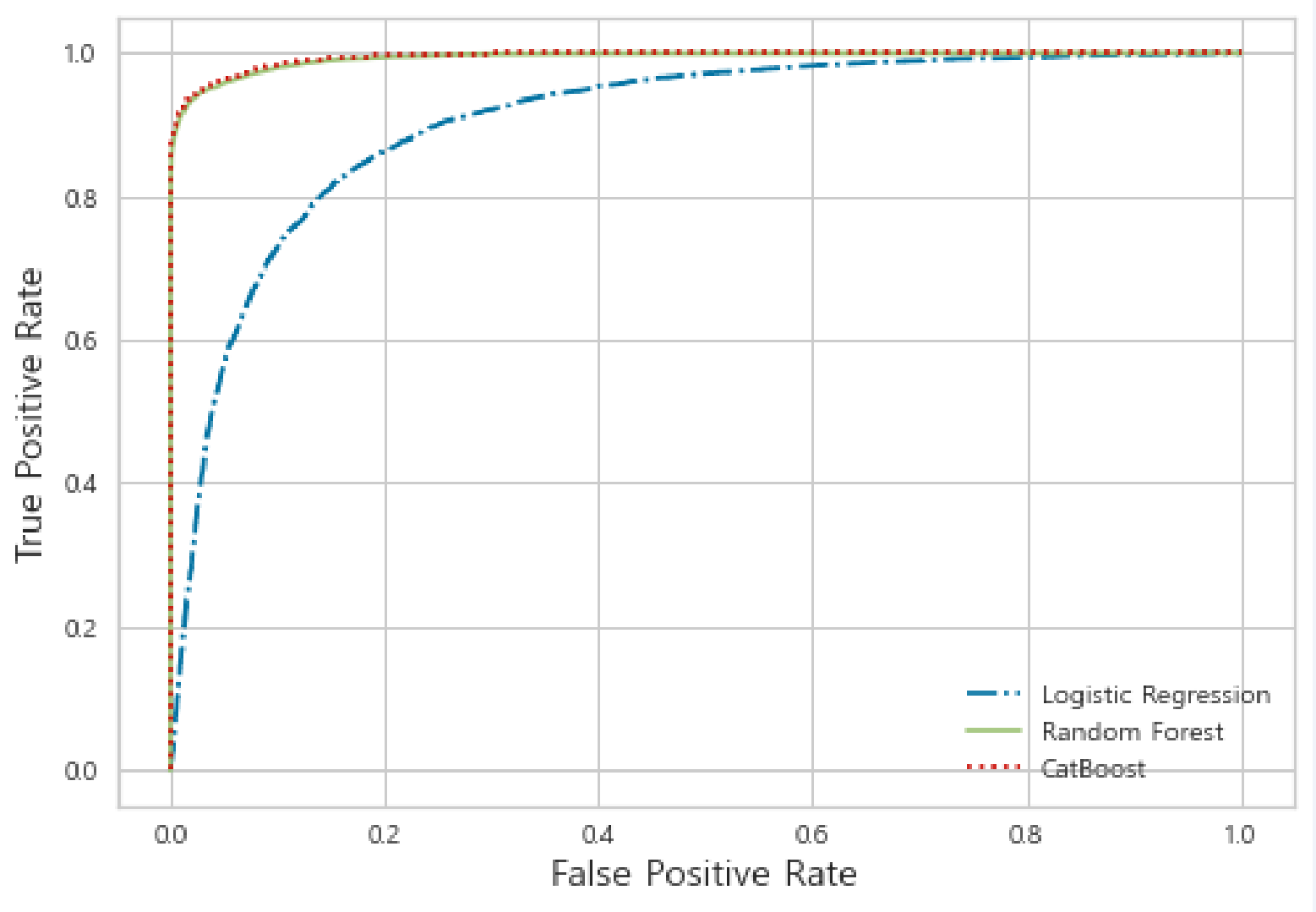
6. Coefficient Plot – Logistic Regression



와이파이 서비스 편의성, 온라인 보딩 순으로 양의 관계를 가지고,

비충성 고객, 여행 타입 (개인 목적) 순으로 음의 관계를 가짐.

7. 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

8. 결론

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

2. Catboost, Random Forest, Logistic Regression 공통적으로 3가지 모델에서
기내 와이파이 편의성과 온라인 보딩 시스템 만족도가 큰 승객은, 주로 최종 항공 서비스 만족도에 긍정적인 영향을 끼침.

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

9. 앞으로의 개선점

1. 하이퍼파라미터 튜닝 전/후 비교.
2. Test dataset으로 최종 점수까지 예측.