

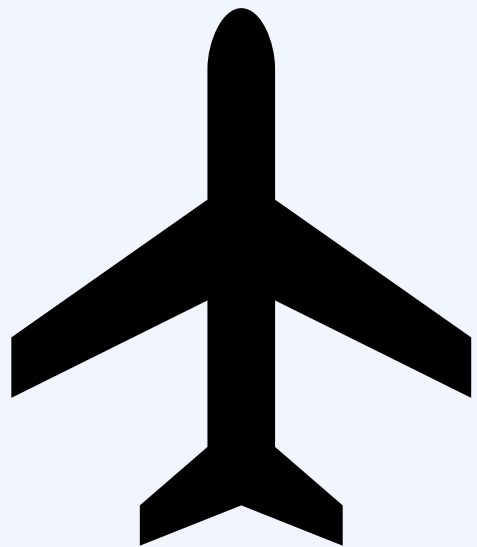


# Airline Passenger Satisfaction

16기 이은찬 | 기계공학과 2017170748

16기 정은미 | 통계학과 202015404

16기 윤지현 | 바이오의공학과 2020250046



- 데이터 개요
- EDA
- 데이터 전처리
- 모델링
- 성능 향상 방법

Feature 변수

Passenger 특징

각종 flight service

출발 및 도착 지연시간



Target 변수

satisfaction

범주형 변수

순서형 변수

**Class**

Eco,  
Eco Plus,  
Business

명목형 변수

**Gender**

Female,  
Male

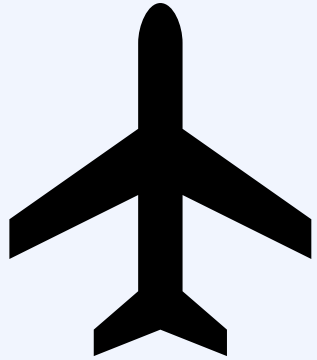
**Customer Type**

Loyal,  
disloyal

**Type of Travel**

Personal,  
Business

수치형 변수



가설1. **class** 변수는 원핫 인코딩보다 **라벨 인코딩**을 하는 것이 높은 성능을 보일 것이다.

## 라벨 인코딩

Class	Label
Eco	0
Eco Plus	1
Business	2

## 원핫 인코딩

Eco	Eco Plus	Business
1	0	0
0	1	0
0	0	1

## 데이터 개요 • Feature 변수

### 범주형 변수

### 수치형 변수

14개의 세부적인  
서비스 척도 변수

Age

Flight Distance

Cleanliness

Gate location

Seat comfort

Food and drink

Online boarding

Inflight service

Arrival Delay in Minutes

Departure Delay in Minutes

On-board service

Leg room service

Baggage handling

Check-in service

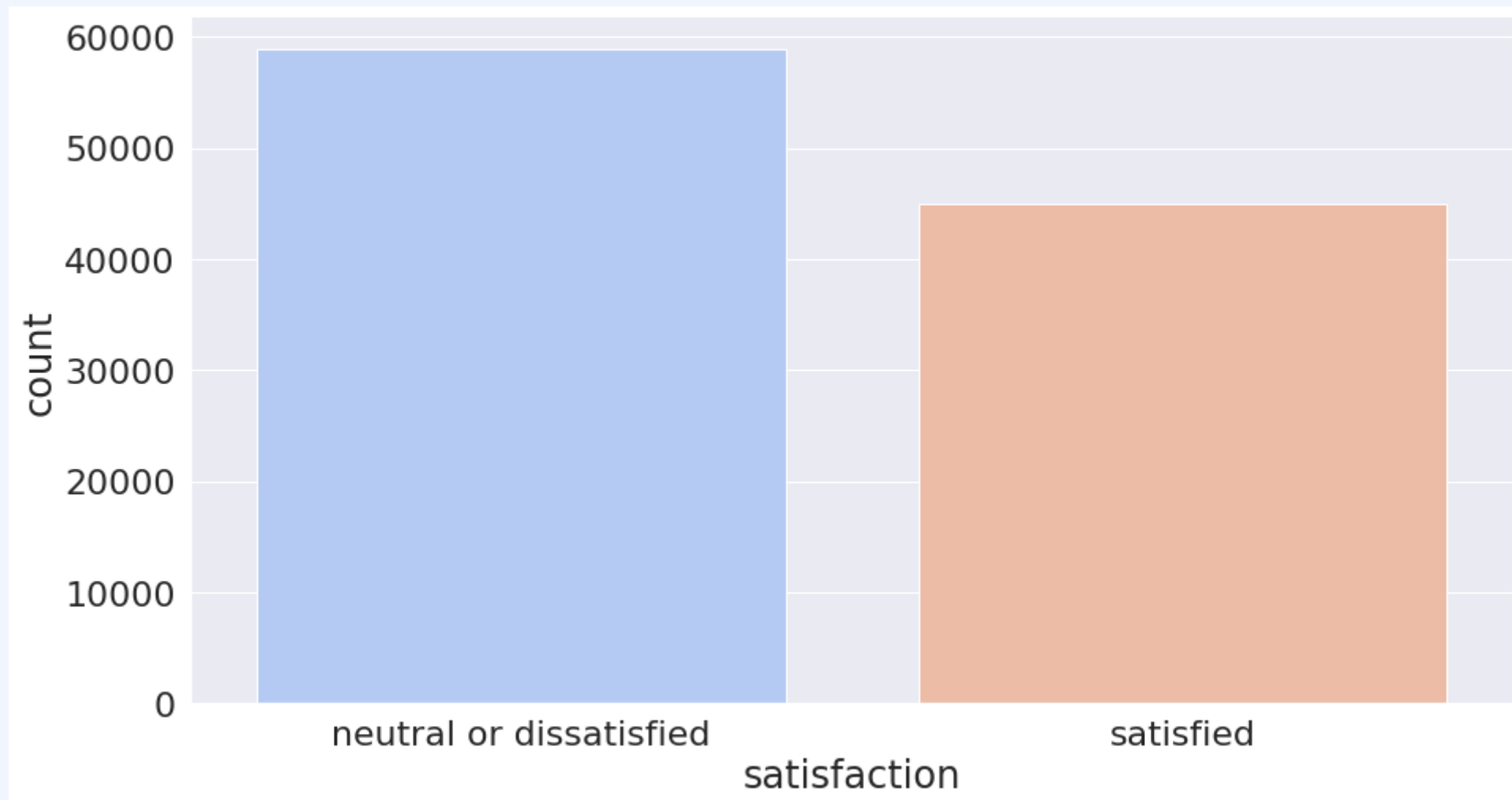
Inflight wifi service

Departure/Arrival time convenient

Ease of Online booking

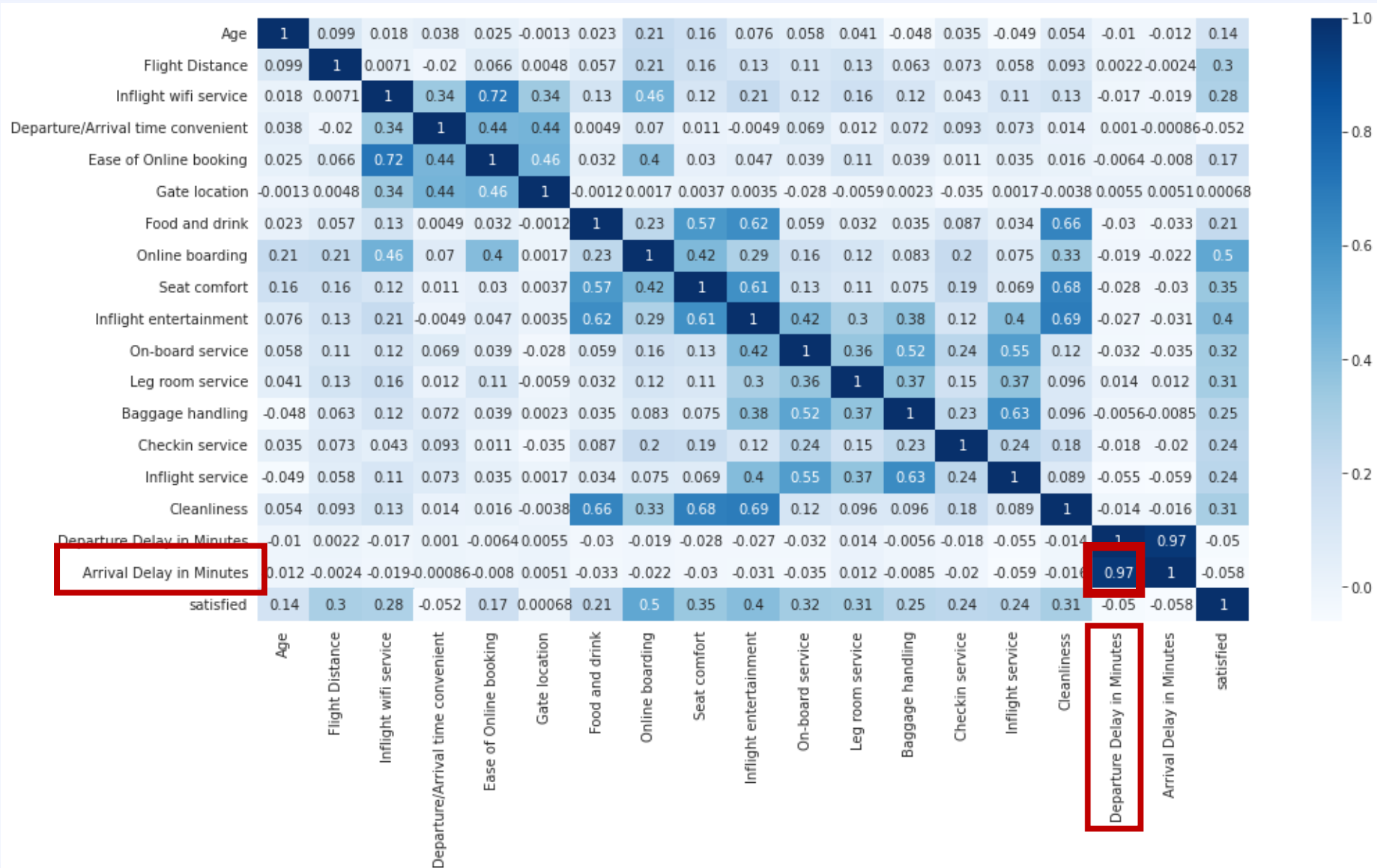
Inflight entertainment

## EDA • Y의 균형 확인

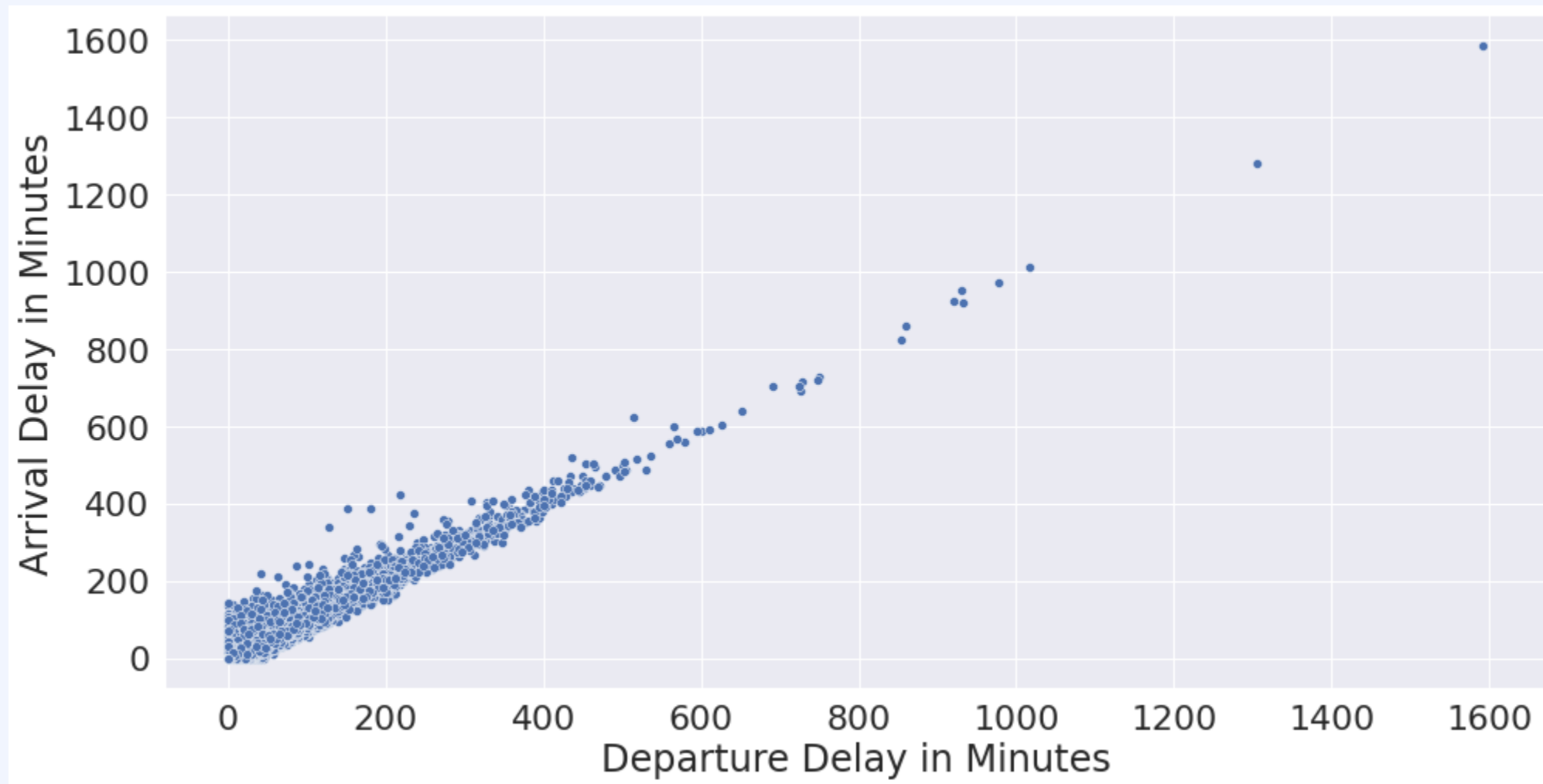


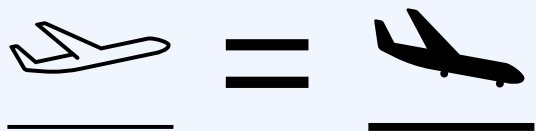
# EDA

## • 수치형 변수 사이의 상관계수 확인









상관계수 = 0.97

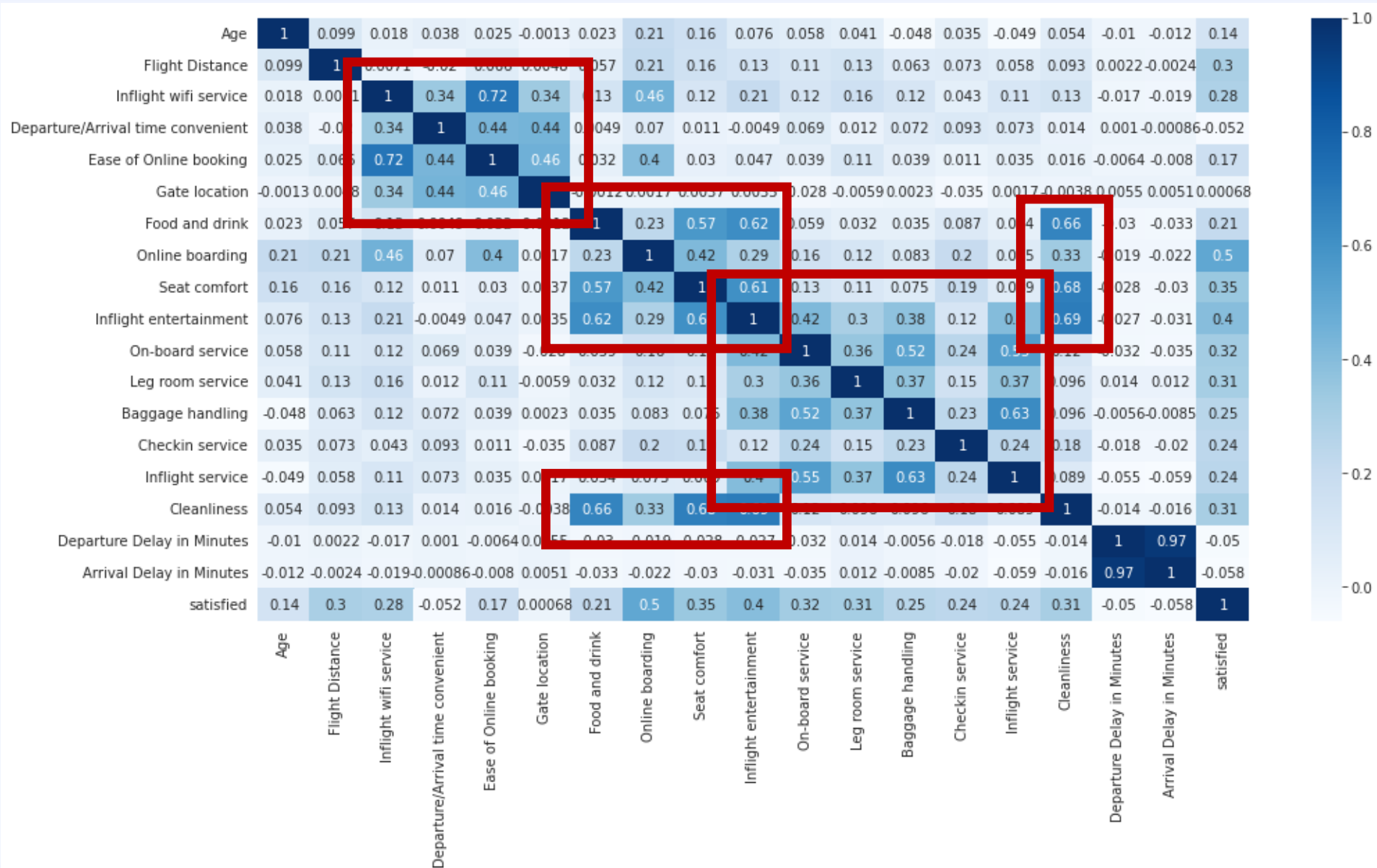
다중공산성 방지를 위해 하나의 column 제거

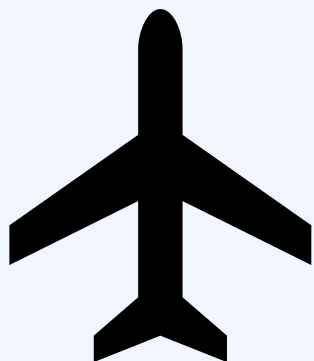
Gender	0
Customer Type	0
Age	0
Type of Travel	0
Class	0
Flight Distance	0
Inflight wifi service	0
Departure/Arrival time convenient	0
Ease of Online booking	0
Gate location	0
Food and drink	0
Online boarding	0
Seat comfort	0
Inflight entertainment	0
On-board service	0
Leg room service	0
Baggage handling	0
Checkin service	0
Inflight service	0
Cleanliness	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	310
satisfaction	0
satisfied	0
dtype: int64	

→ 결측치 가지는 Arrival 변수 제거

# EDA

## • 수치형 변수 사이의 상관계수 확인

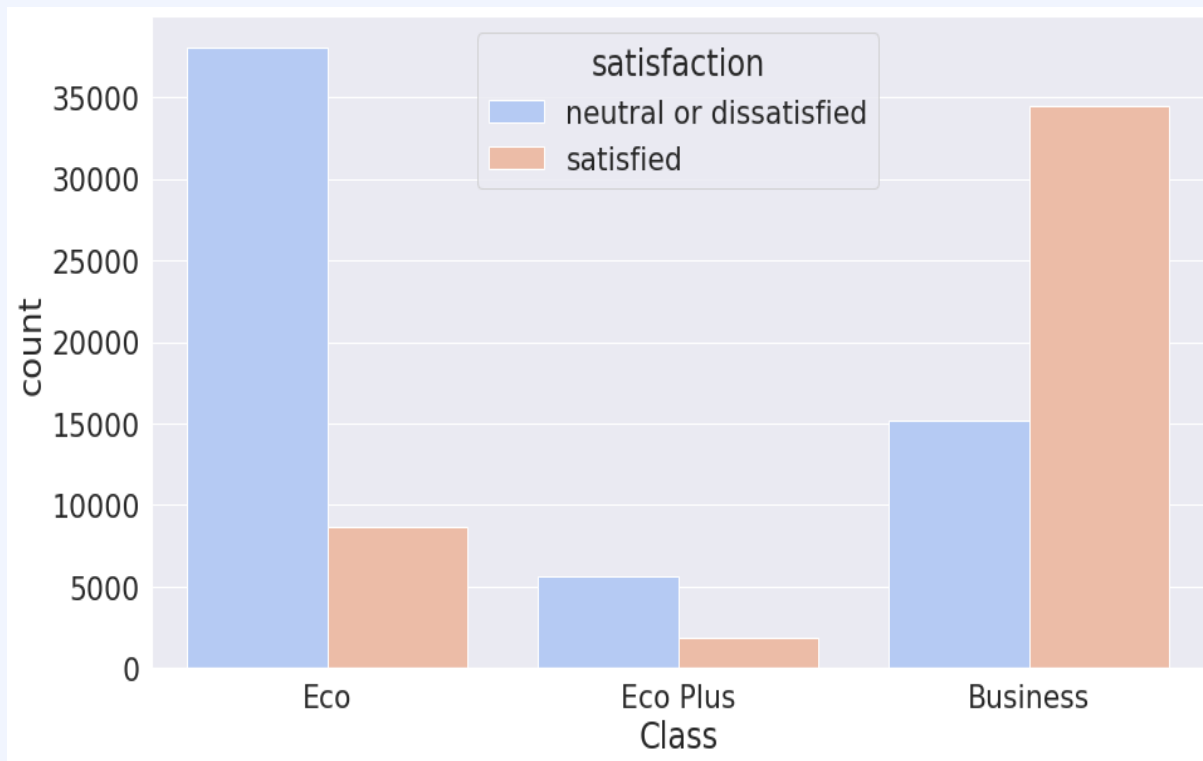




가설2. PCA를 진행하는 것이 높은 성능을 보일 것이다.

## EDA • 유의미해보이는 Feature

### Y 와 좌석등급

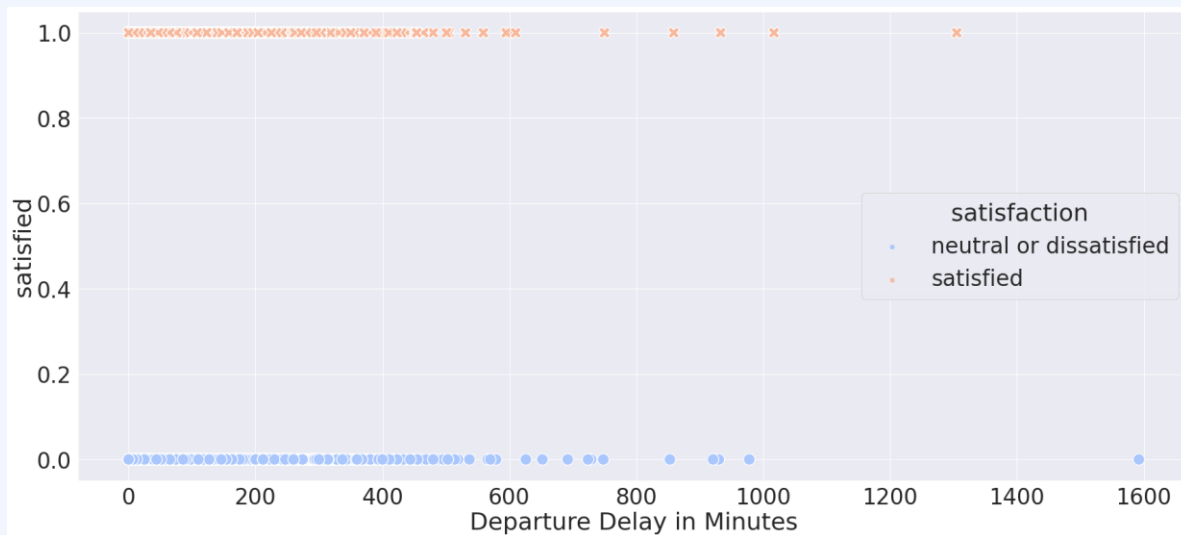


### Y 와 여행목적

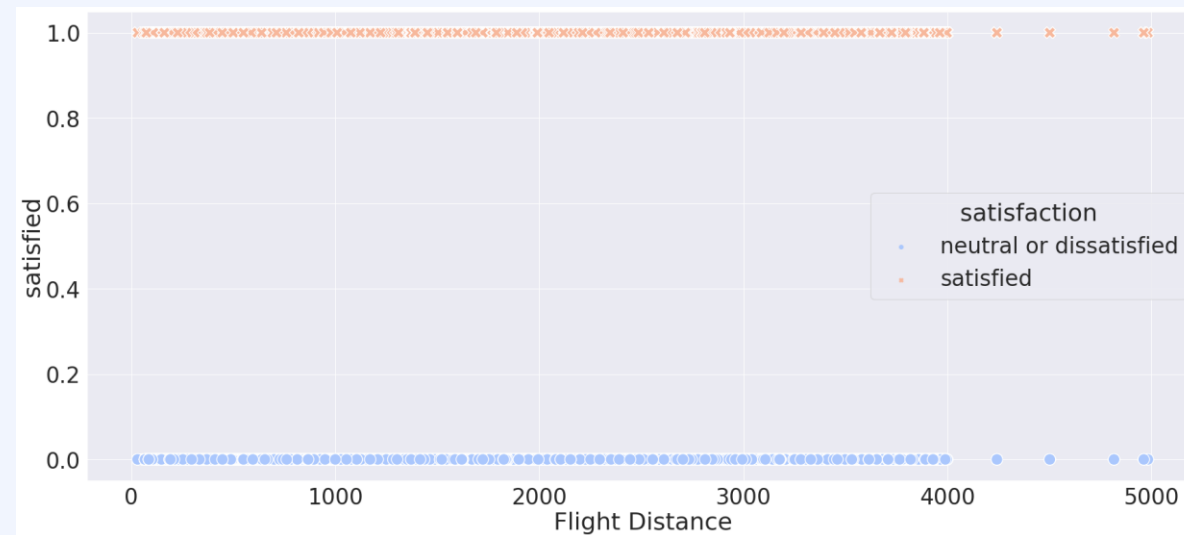


# EDA • 무의미해보이는 Feature

## Y와 출발연기시간

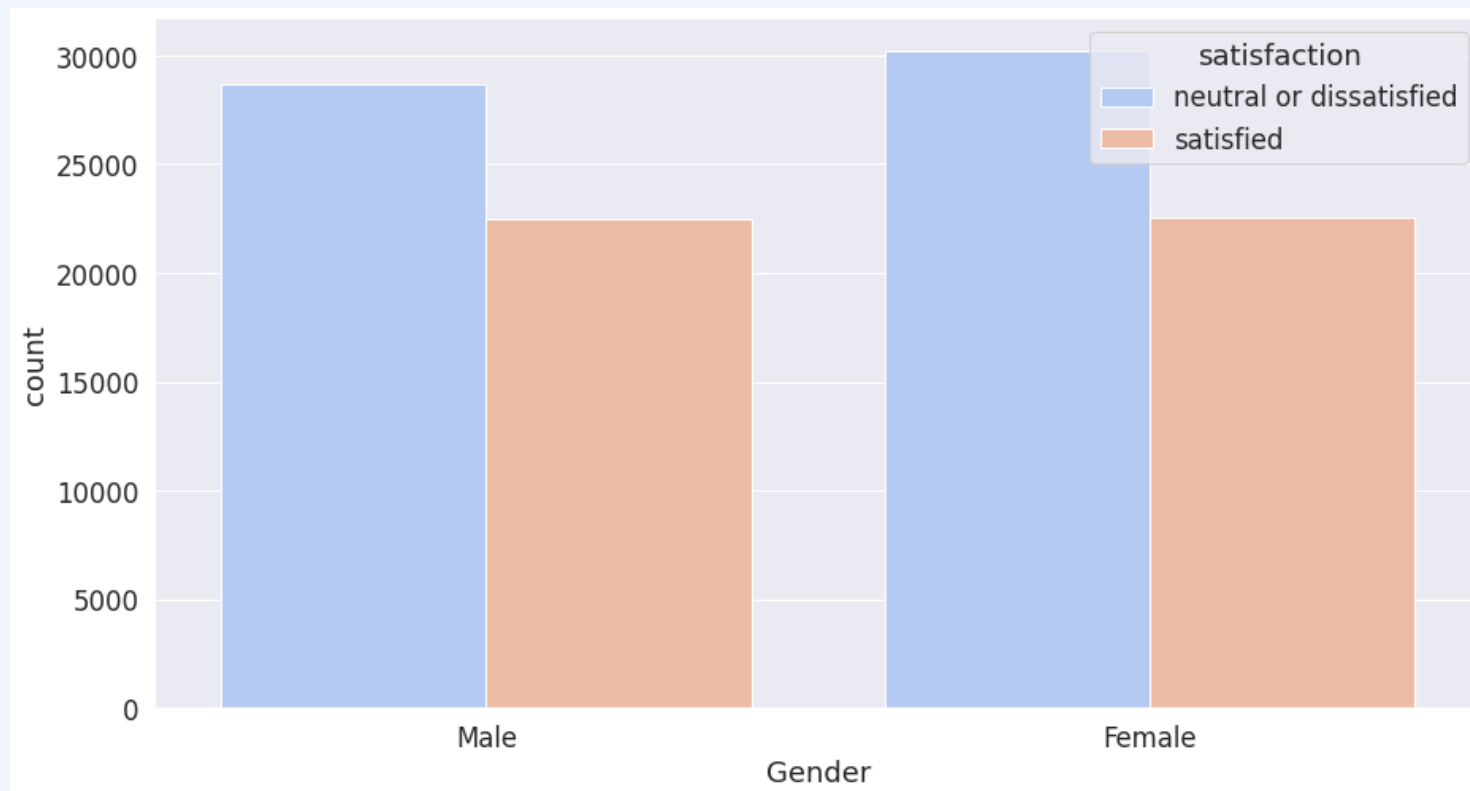


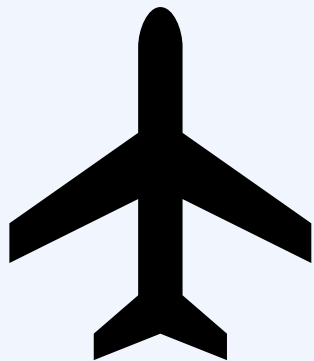
## Y와 비행거리



## EDA • 무의미해보이는 Feature

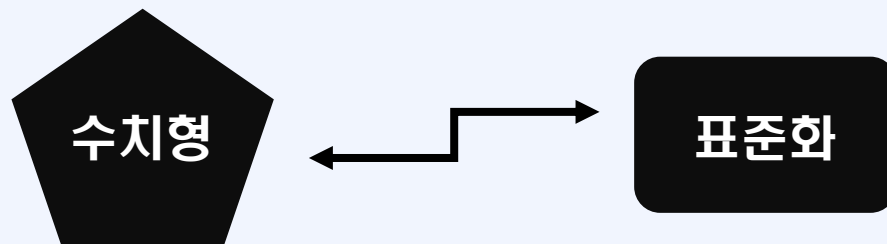
### Y 와 성별





가설3. 시각적으로 의미 없어 보이는 변수들을 제거  
하는 것이 높은 성능을 보일 것이다.





가설4. 수치형 변수를 표준화하는 것이  
높은 성능을 보일 것이다.

## 데이터 전처리 방향 설정 by 가설

가설1. class 변수는 원핫 인코딩보다 라벨 인코딩을 하는 것이

→ Class 변수 라벨인코딩? 원핫인코딩?

가설 2. PCA를 진행하는 것이

→ PCA yes? no?

가설3. 시각적으로 의미없어 보이는 변수들을 제거하는 것이

→ 변수 3개 제거 yes? no?

가설 4. 수치형 변수를 표준화하는 것이

→ 표준화 yes? no?

높은 성능을 보일 것이다.



각 방안을 4가지 모델로 성능평가 후  
방안 사용 유무 결정

## Data Preprocessing

가설1. **class** 변수는 원핫 인코딩보다 **라벨 인코딩**을 하는 것이 높은 성능을 보일 것이다.

→ Class 변수 라벨인코딩? 원핫인코딩?

Model: default parameters  
No validation data

Metric : Acc	LR	NB	DT	RF
Class One-hot encoding	82.98	84.64	94.66	96.30
X	<b>84.04</b>	84.63	94.63	96.42

- Labeling Encoding을 적용했을 때, LR에서 성능 개선이 있는 것으로 볼 수 있음

→ Class 변수 라벨인코딩 적용!

## Data Preprocessing

가설2. PCA를 진행하는 것이 높은 성능을 보일 것이다.

→ PCA yes? no?

Metric : Acc	LR	NB	DT	RF
PCA(n=10)	86.93	83.79	93.37	95.62
X	87.06	<b>86.41</b>	<b>94.54</b>	<b>96.37</b>

- PCA 진행하였을 때, 성능 감소 확인

→ PCA 적용 X !

## Data Preprocessing

가설3. 무의미해 보이는 변수들을 제거하는 것이 높은 성능을 보일 것이다.

→ 변수 3개 제거 yes? no?

제거 변수: Male, Departure Delay in Hours, Flight Distance

Metric : Acc	LR	NB	DT	RF
변수 제거	<b>84.41</b>	<b>86.41</b>	94.42	96.33
변수 유지	84.04	84.63	94.63	96.42

- 변수 제거 시, 성능 개선 확인

→ 일부 변수 제거!

## Data Preprocessing

가설4. 수치형 변수를 표준화하는 것이 높은 성능을 보일 것이다.

→ 표준화 yes? no?

척도 변수: 수치형으로 가정하고 PCA 진행

Metric : Acc	LR	NB	DT	RF
표준화	<b>87.06</b>	86.41	94.54	96.37
X	84.41	86.41	94.42	96.33

- 표준화 진행하였을 때, 성능 개선 확인

→ 표준화 진행!

## 데이터 전처리 방향 확장 by 가설

가설1. class 변수는 원핫 인코딩보다 라벨 인코딩을 하는 것이

→ Class 변수 라벨인코딩 적용!

가설 2. PCA를 진행하는 것이

→ PCA 적용 X !

가설3. 시각적으로 의미없어 보이는 변수들을 제거하는 것이

→ 3개 변수 제거!

가설 4. 수치형 변수를 표준화하는 것이

→ 표준화 진행!

높은 성능을 보일 것이다.

# Modeling

## 사용한 Metric

- Accuracy score
- precision score
- recall score
- AUROC
- AUPRC



## 결과 수치

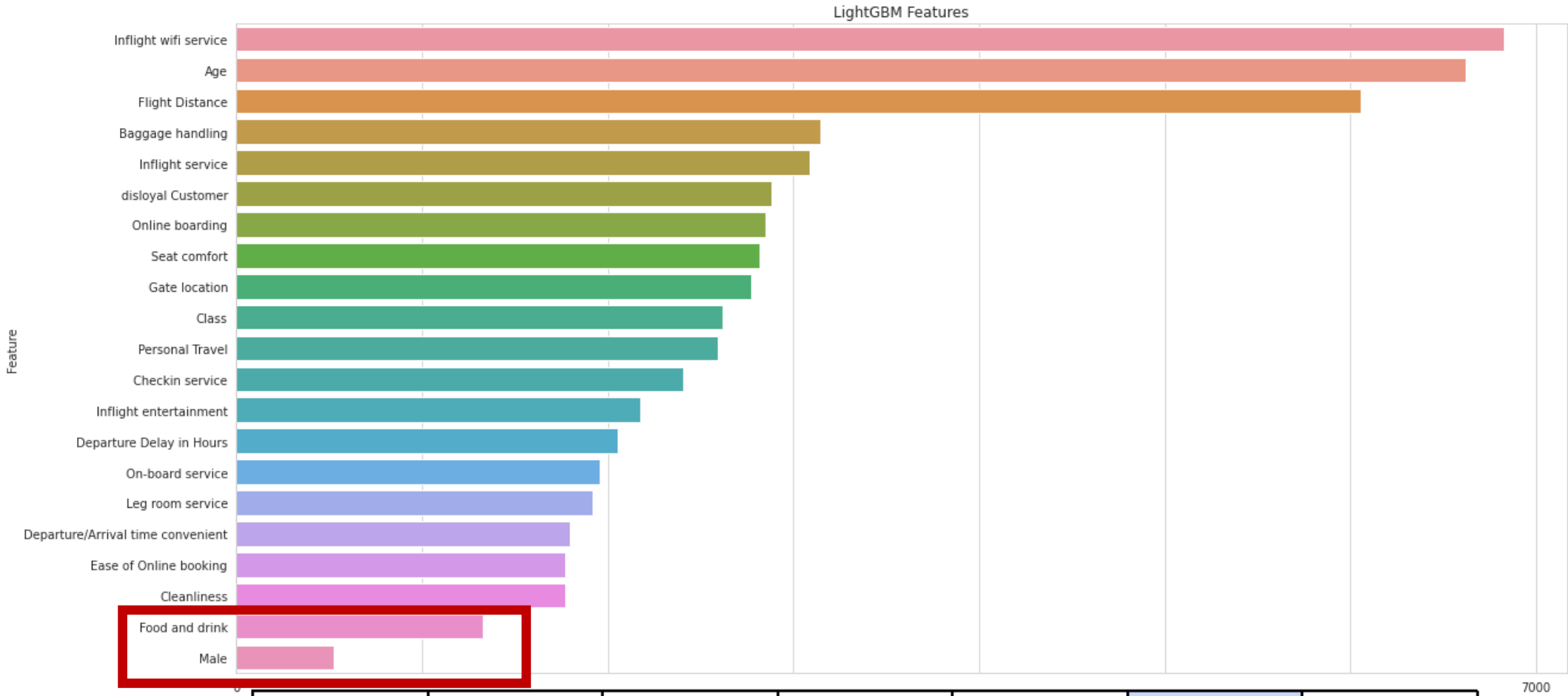
- Logistic Regression
- GaussianNB
- Decision Tree
- RandomForest
- LGBM
- XGB

LR	NB	DT	RF	LGBM	XGB
87.12	86.43	94.64	96.35	96.47	96.39

→ LGBM 모델 선택



# 성능 향상 방법



	LR	NB	DT	RF	LGBM	XGB
기존	87.12	86.43	94.64	96.35	96.47	96.39
FI 하위 2개 제외	87.11	86.33	94.71	96.40	96.49	96.43

## 성능 향상 방법

- **Hyperparameter Tuning**
  - 여러 번의 실험 결과 최적의 hyperparameter 찾음

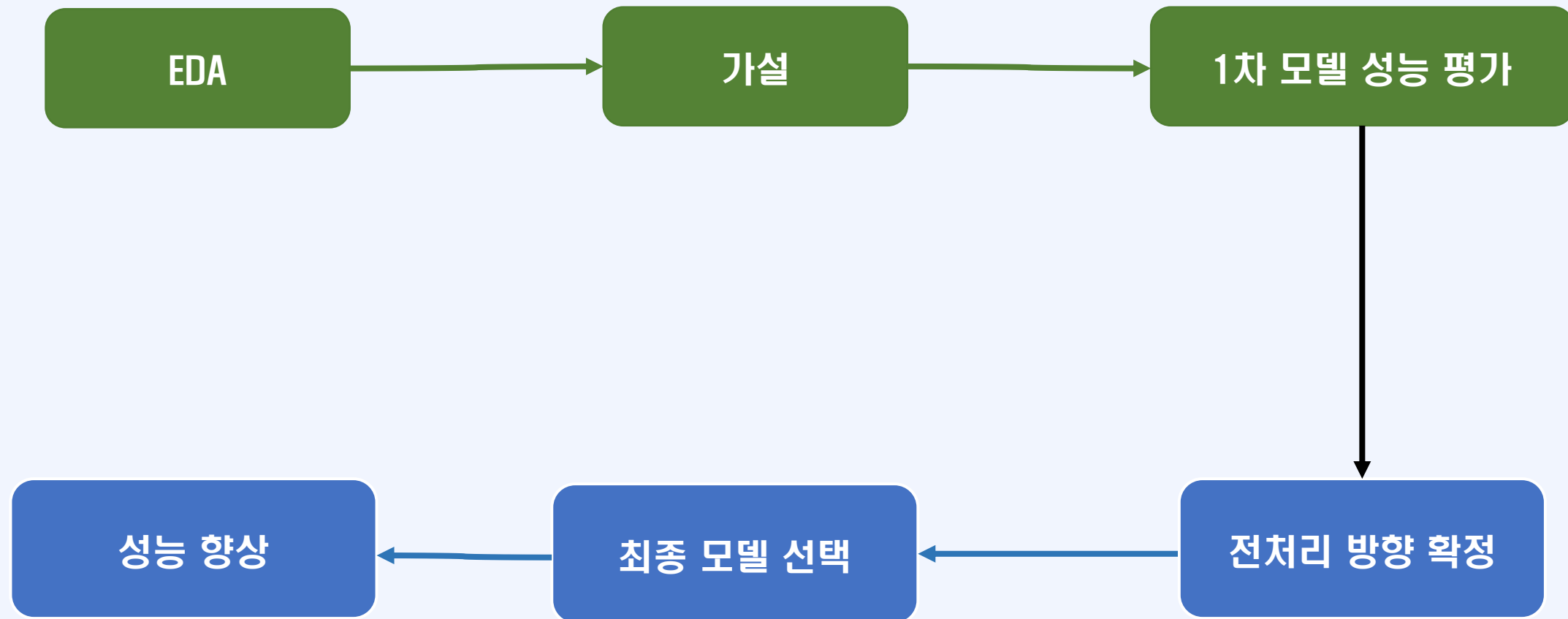
### **LGBM**

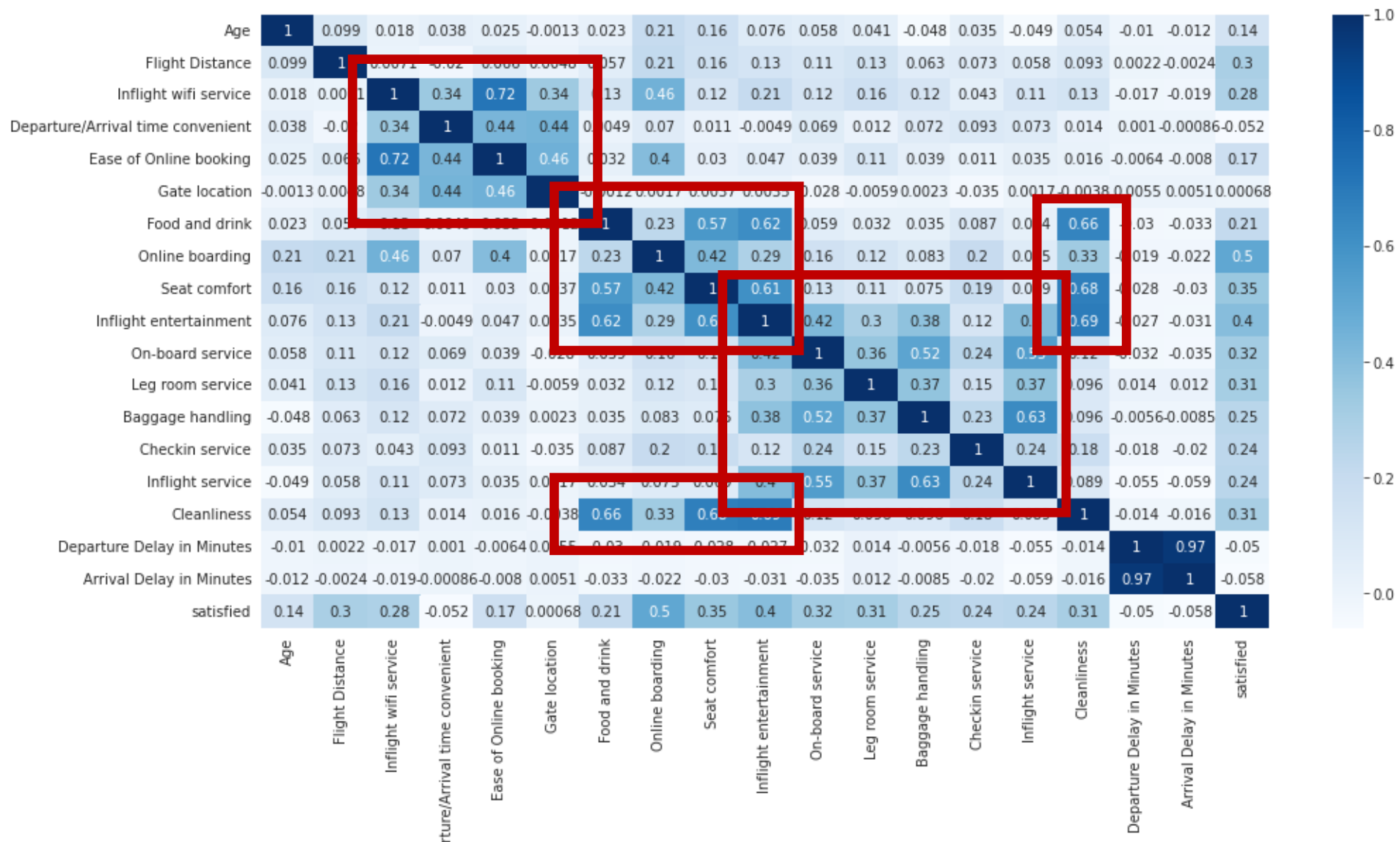
```
boosting_type = "gbdt", n_estimators = 2000, learning_rate = 0.01
```

### **XGB**

```
n_estimators = 1000, eta = 0.005, max_depth = 5, max_leaves = 24
```

## 요약





14개의 세부적인 서비스 척도 변수의 합 => Total service score



감 사 합 니 다