항공사 고객 만족도 예측

논리적 근거에 기반한 데이터 분석

201713544 신익환

Index

1. 데이터 확인 및 고려사항 체크

Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	Gate location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling	Checkin service	Cleanliness	Online boarding	Departure Delay in Minutes	Arrival Delay in Minutes	target
Female	disloyal Customer	22	Business travel	Eco	1599	3	0	3	3	4	3	4	4	5	4	4	4	5	4	0	0.0	0
Female	Loyal Customer	37	Business travel	Business	2810	2	4	4	4	1	4	3	5	5	4	2	1	5	2	18	18.0	0
Male	Loyal Customer	46	Business travel	Business	2622	1	1	1	1	4	5	5	4	4	4	4	5	4	3	0	0.0	1
Female	disloyal Customer	24	Business travel	Eco	2348	3	3	3	3	3	3	3	3	2	4	5	3	4	3	10	2.0	0
Female	Loyal Customer	58	Business travel	Business	105	3	3	3	3	4	4	5	4	4	4	4	4	4	5	0	0.0	1

1.Target은 0&1의 binary로 표현
Logistic regression (활성함수 sigmoid)이 적합

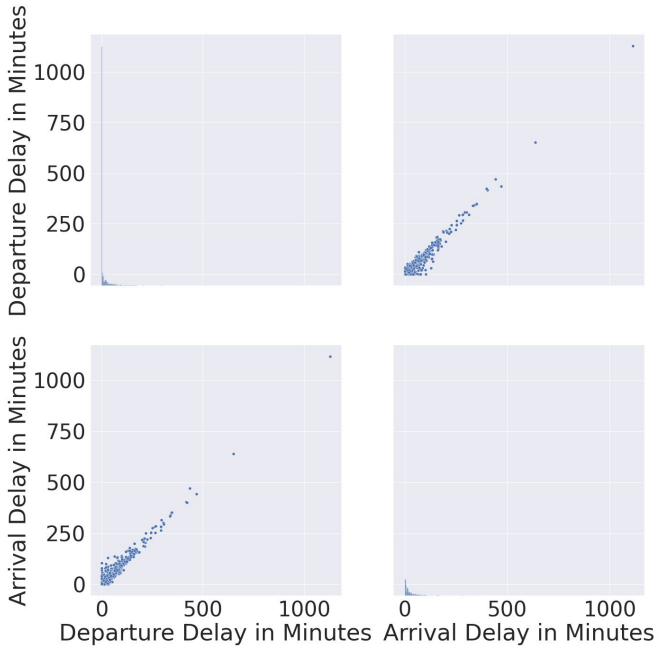
2.Feature(0~5 int)
null 값은 없음. 그러나, **0**값이 존재하는데 이에 대한 해석 및 처리가 필요

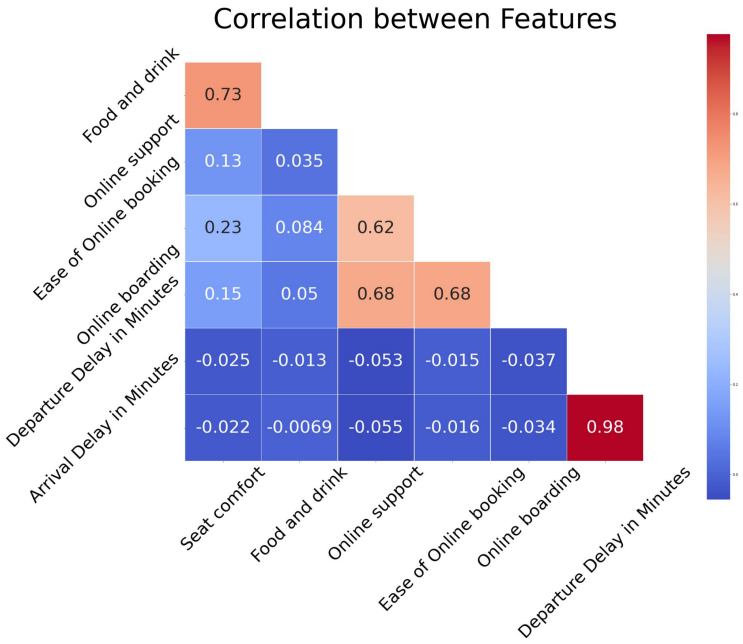
3. 기타 **Feature** 값 **(Delay minutes / Distance / Age)**
위의 지표와 다른 숫자 형태를 갖고 있으며 **Scale** 또한 달라 이에 대한 처리가 필요

4. Object Feature (Gender, Customer Type, Type of Travel, Class)
점수가 아닌 **Grouping**의 기준이 되는 값으로 이를 어떻게 고려/활용 할 것인가에 대한 논의가 필요

2. Logistic Regression을 위한 변수 확인/처리

2.1 다중공선성 제거





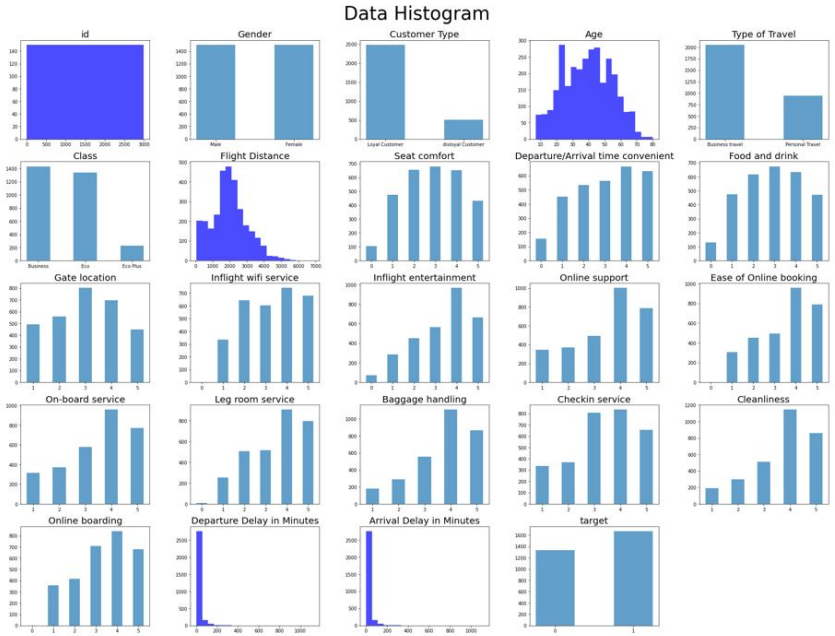
Departure Delay in Minutes ~~& Arrival Delay in Minutes~~(0.98)

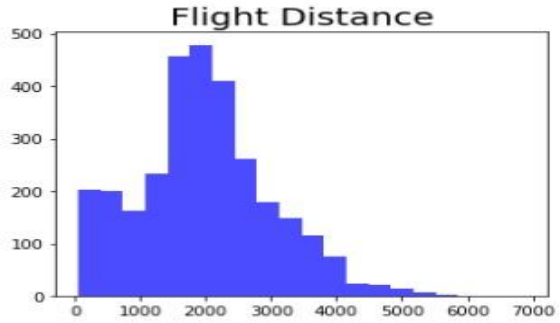
Seat Comfort ~~& Food and drink~~-(0.73)

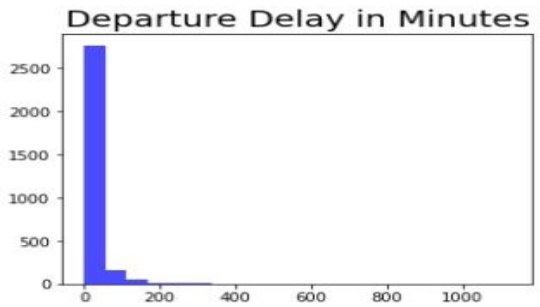
Online Boarding ~~& Online Support & Ease of Online Booking~~ (0.68)

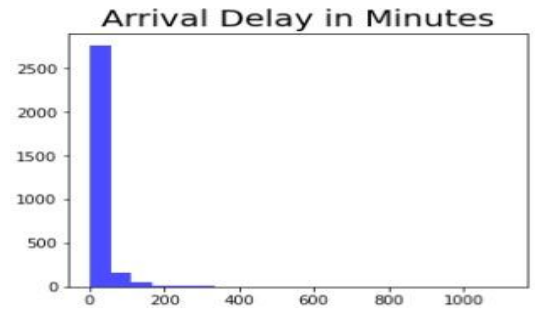
2. Logistic Regression을 위한 변수 확인/처리

2.2 데이터 편중 여부 및 스케일 확인









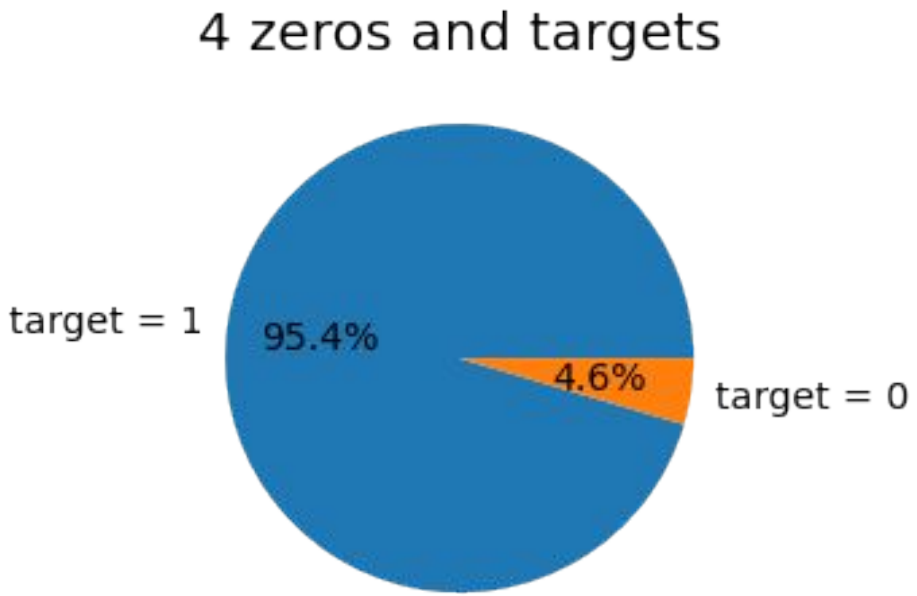
Flight Distance, Delay in Minutes 값들은 좌편향 되어있으며,
리커트척도 값(**0~5**점 표현)들과 **scale**이 크게 다르다. 따라서, 평가지표로 반영할 경우 스케일링 및 정규화가 필요할 것이다.
(그러나, 뒤에 이어질 **4-1**항목의 논거에 의해 최종 평가지표로 반영하지 않을 것이므로 변환하지 않았다)

3. ‘0’값에 대한 해석 및 처리

3.1 ‘0’값을 ‘Null’값으로 판단한 근거

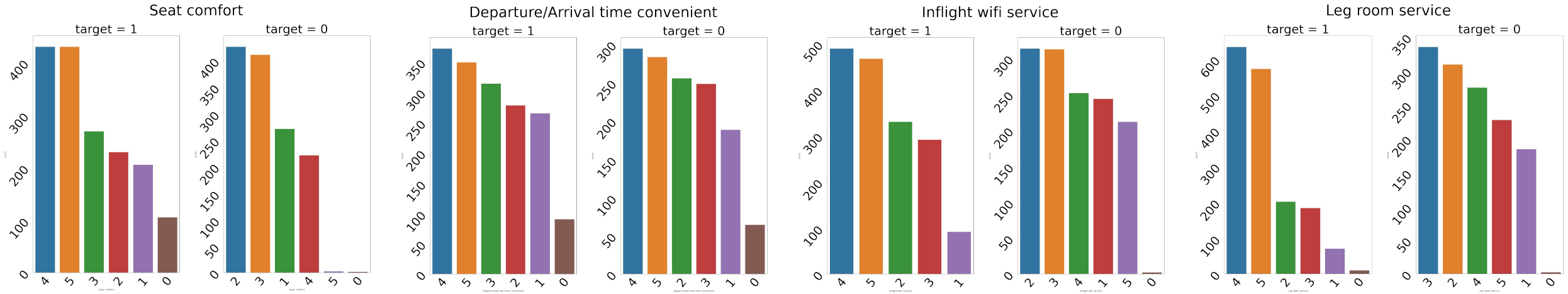
- **일반적으로 만족도 조사에 사용되는 리커트 척도는 1~ 5 또는 1~ 7점으로 조사한다.**
- **점수상 0값은 critical 한 점수이다. 그러나 3~4개 항목에 0점을 주고도 최종 만족 응답인 경우가 상당 수 존재**

  (4개항목 기준 21명이 존재하며 특히, 4개의 항목에서 0점을 주었던 고객 중 95.3%가 최종 만족을 나타내었다.)



3. ‘0’값에 대한 해석 및 처리

3.1 ‘0’값을 ‘Null’값으로 판단한 근거



0값이 평가 점수라면 극단적인 부정의 표현으로 만족(**Target =1**)보다 불만족(**Target=0**)에서 더 많이 나타나는 것이 타당하나, 오히려 **만족(Target =1)에서 더 많이 나타나는 것으로 보아 부정을 의미하는 지표가 아닌 ‘Null’값을 의미하는 것으로 추론할 수 있다.**

3.2 Hot Deck (KNN method) 를 활용한 ‘Null’값 처리

	항목1	항목2	항목3	항목4	항목5
A	1	2	3	4	5
B	Null	3	5	2	4
C	4	1	2	4	1
D	3	3	5	1	4
E	5	4	3	2	1

다른 평가 항목에서 유사한 점수를 주었다

⇩

평가 성향이 비슷하다

⇩

Null 값에 대한 평가도 비슷할 가능성이 높다.

3. ‘0’값에 대한 해석 및 처리

3.2 Hot Deck (KNN method) 를 활용한 ‘Null’값 처리

```
1 from sklearn.impute import KNNImputer
2 data = df.replace(0, np.NaN)
3 imputer = KNNImputer(n_neighbors=1)
4
5 KNN_data = imputer.fit_transform(data.iloc[:,7:16])
6 df.iloc[:,7:16] = KNN_data
7 df.head()
```

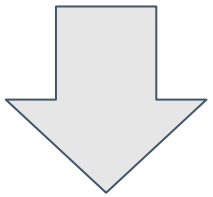
	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	Gate location	Inflight wifi service	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Cleanliness	Online boarding	Delay in Minutes	target
0	1	Female	disloyal Customer	22	Business travel	Eco	1599	3.0	4.0	3.0	4.0	3.0	5.0	4.0	4.0	4.0	5	4	0.0	0
1	2	Female	Loyal Customer	37	Business travel	Business	2810	2.0	4.0	4.0	1.0	4.0	5.0	4.0	2.0	1.0	5	2	18.0	0
2	3	Male	Loyal Customer	46	Business travel	Business	2622	1.0	1.0	1.0	4.0	5.0	4.0	4.0	4.0	5.0	4	3	0.0	1
3	4	Female	disloyal Customer	24	Business travel	Eco	2348	3.0	3.0	3.0	3.0	3.0	2.0	4.0	5.0	3.0	4	3	2.0	0
4	5	Female	Loyal Customer	58	Business travel	Business	105	3.0	3.0	3.0	4.0	4.0	4.0	4.0	4.0	4.0	4	5	0.0	1

4. 기타 Feature 값들에 대한 해석 및 처리

(Delay in Minutes) 반영시 **scaling** 및 **1~5**점 표준화 필요

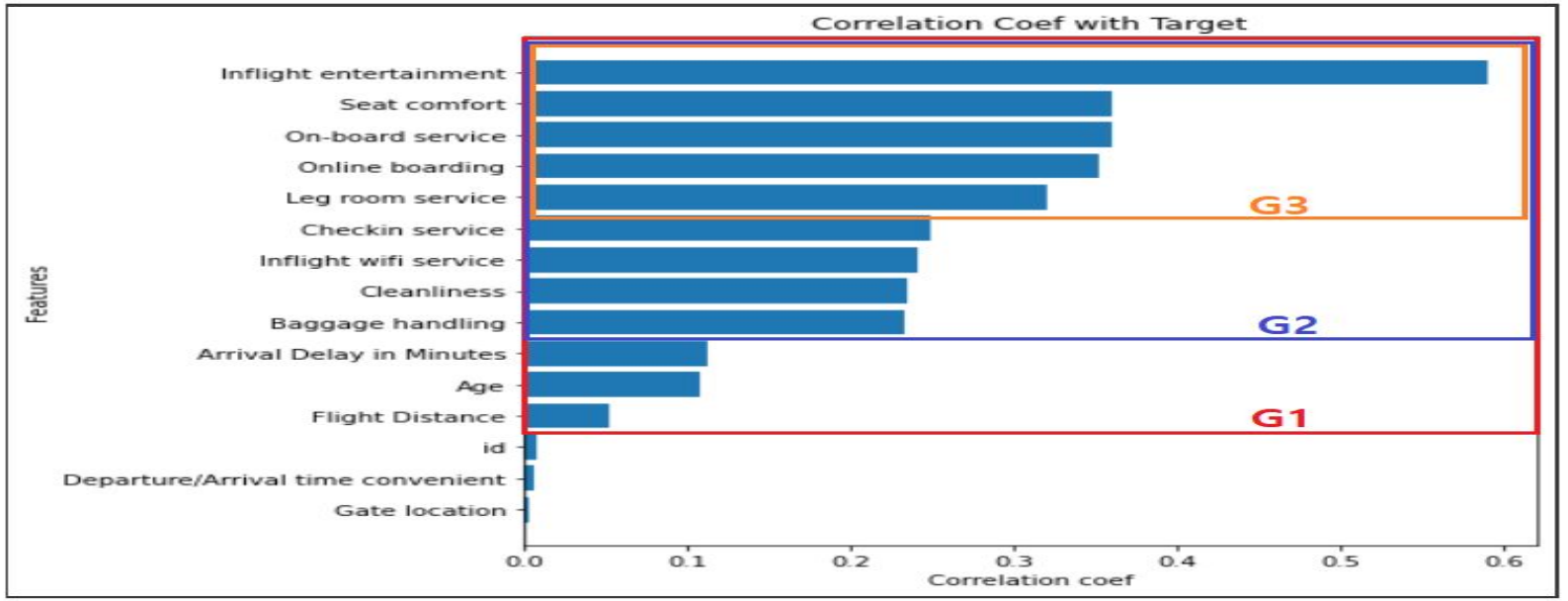
(Flight Distance) 고객이 사전 인지하는 요소이기 때문에 평가지표로서 크게 반영되지는 않을 것.
(단, 장거리 일수록 **seat comfort** 등의 요소와 상관관계가 있을 수 있을 것이다. → 위 **2-1**의 **Heat map**에서 기각)

(Age) Box Plot으로 성별에 따른 **target** 분포 차이가 나는지 확인



데이터 전처리 전, 상관도(가중치)가 너무 작은 항목들은 분석 지표에서 누락시켜도 무방하므로 **3**번까지 전처리한 데이터를 사용하여 먼저 가중치 분석을 하였음.

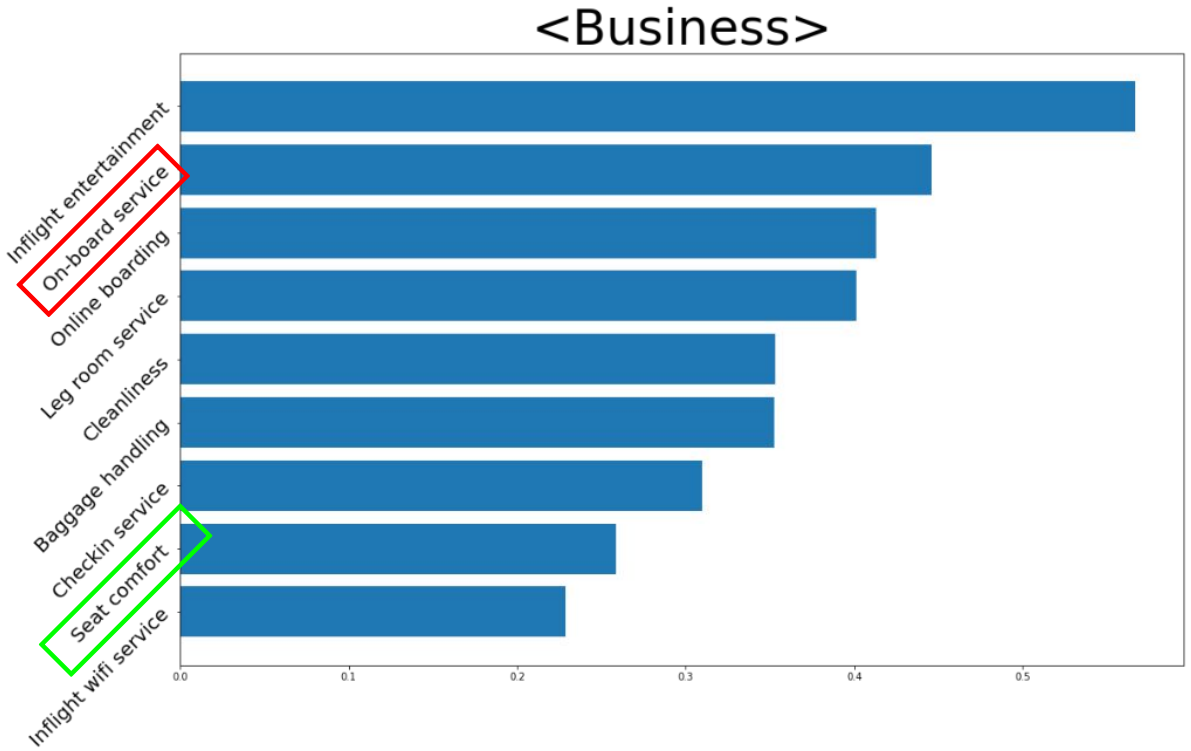
4. 기타 Feature 값들에 대한 해석 및 처리

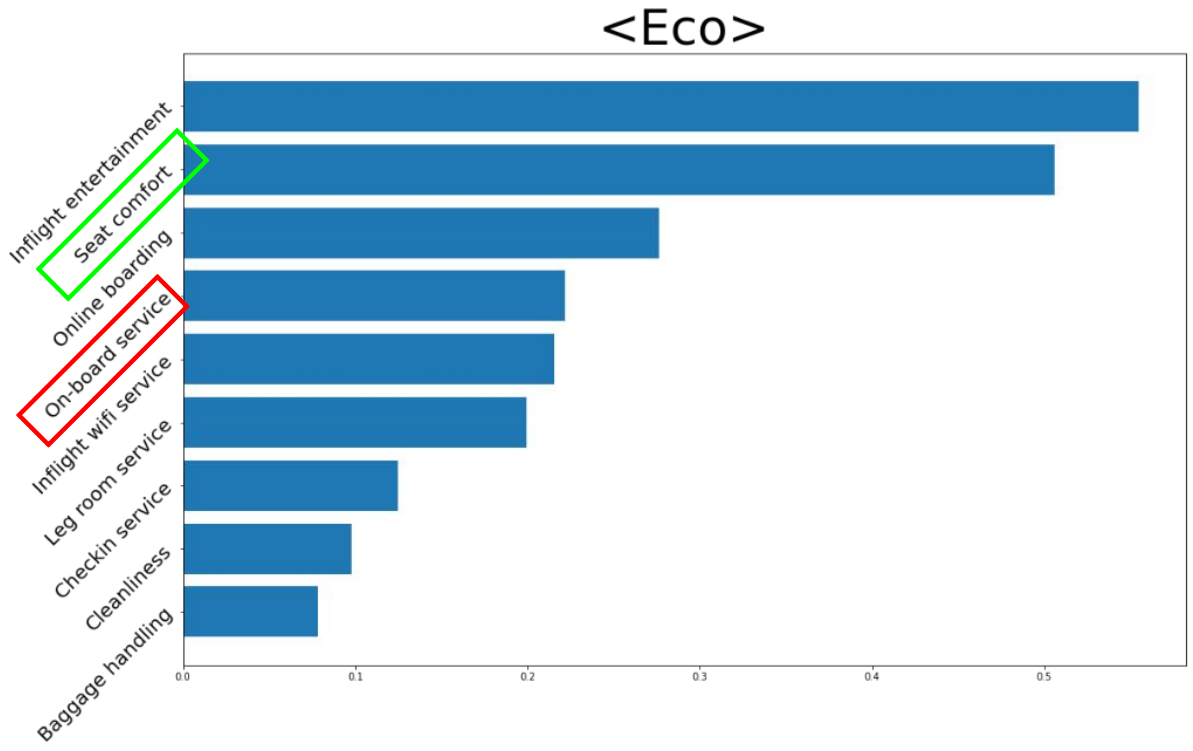


Flight Distance/Age/Arrival Delay in minutes 3항목의 **feature** 들은 가장 상관도가 낮은 항목(**0.2**이하)이며, 각 그룹간의 성능 차이가 매우 적으므로 데이터 분석 효율을 위해 **feature** 값에서 제외하기로 한다.

→ **G2** 채택 (**G3**가 아닌 **G2**를 채택한 이유에 대해서는 뒤의 **5.Object** 항목에서 다루기로 한다)

5. Object Feature 값들에 대한 해석 및 처리





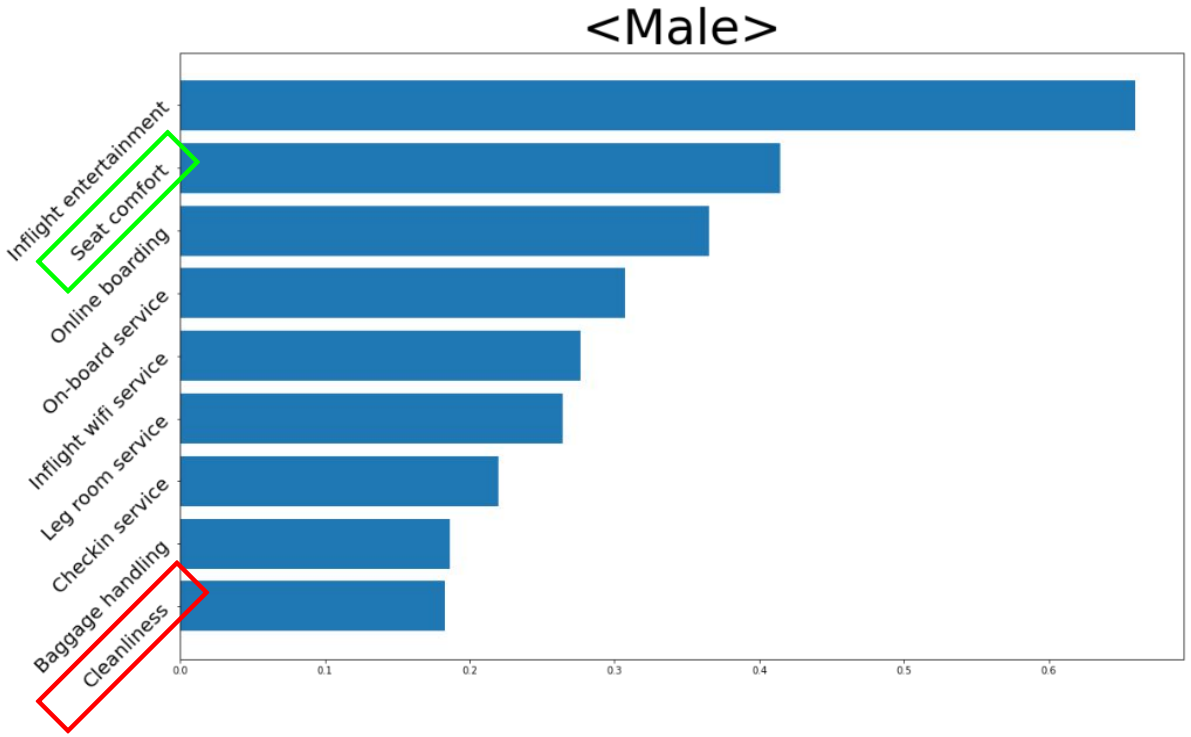
<해석 및 추론>

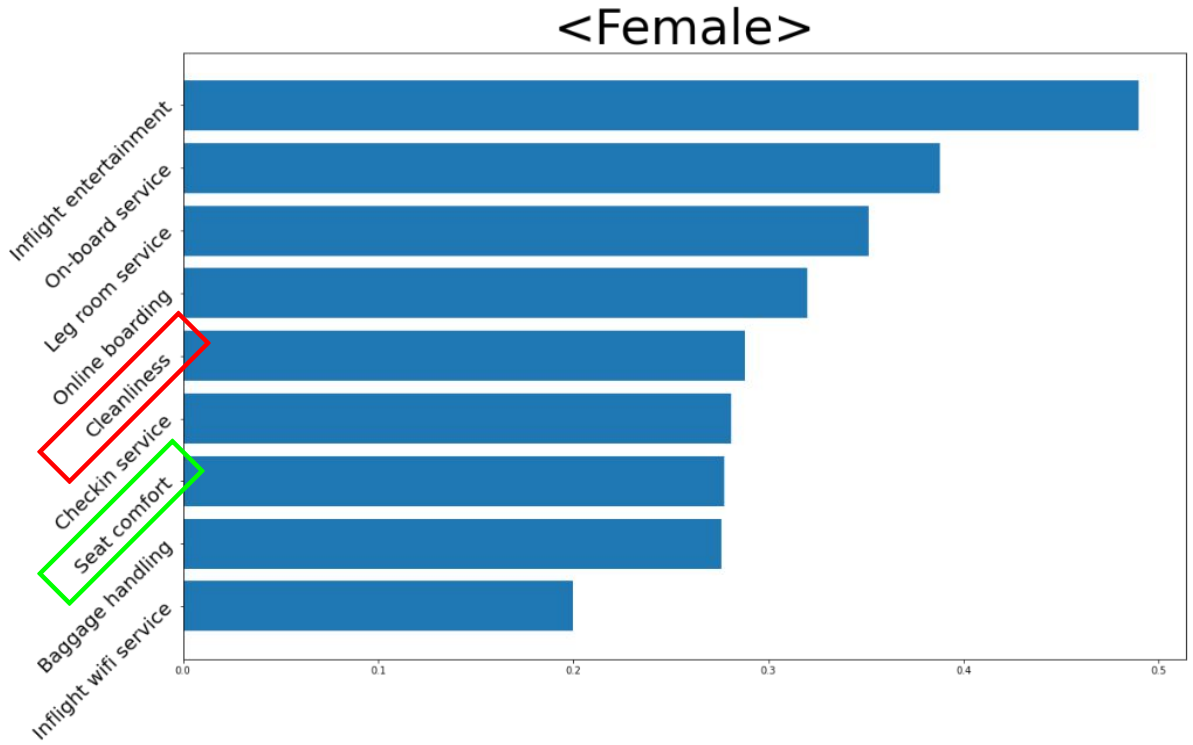
Business그룹은 **Seat comfort**의 상관도가 낮은반면 **Eco** 그룹은 상관도가 높게 나왔다.

이는 **Business** 좌석의 경우 기본적으로 **Seat comfort** 이 좋아 이에 대한 평가 **variation**이 적기 때문이라고 추론 가능.

또한 **On-board service** 항목의 상관도가 높은 것으로 보아 서비스 측면의 대우를 받는 것을 중시한다고 추론 가능

5. Object Feature 값들에 대한 해석 및 처리





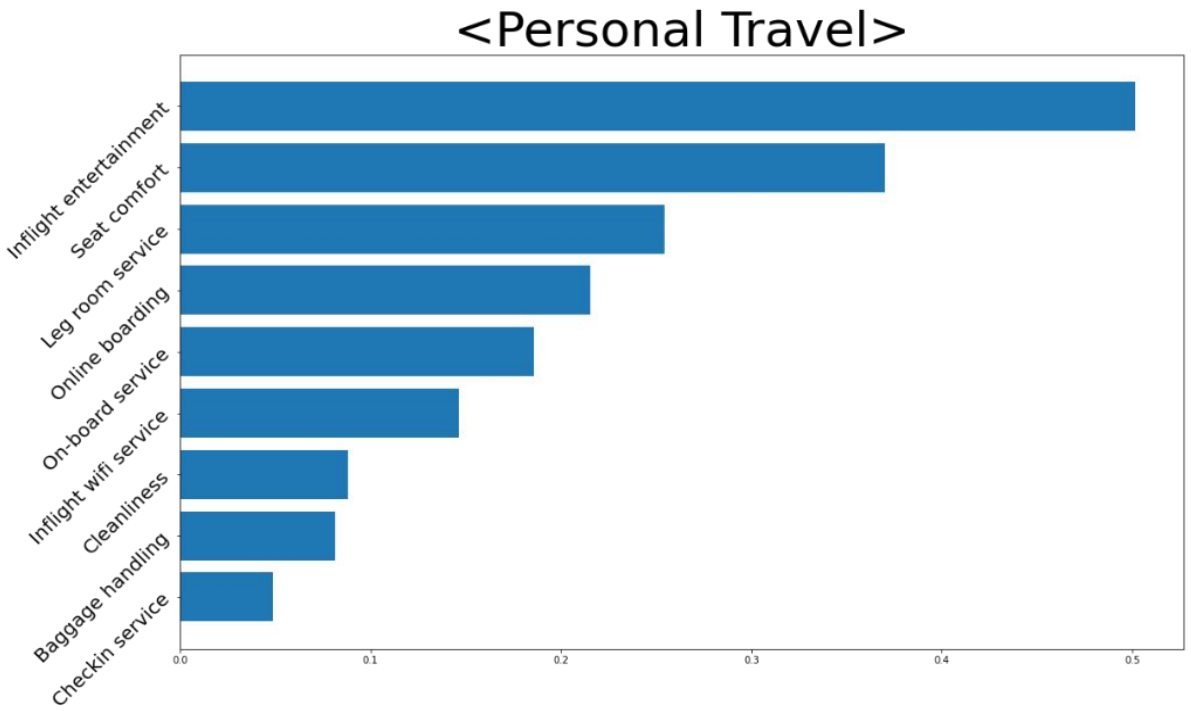
<해석 및 추론>

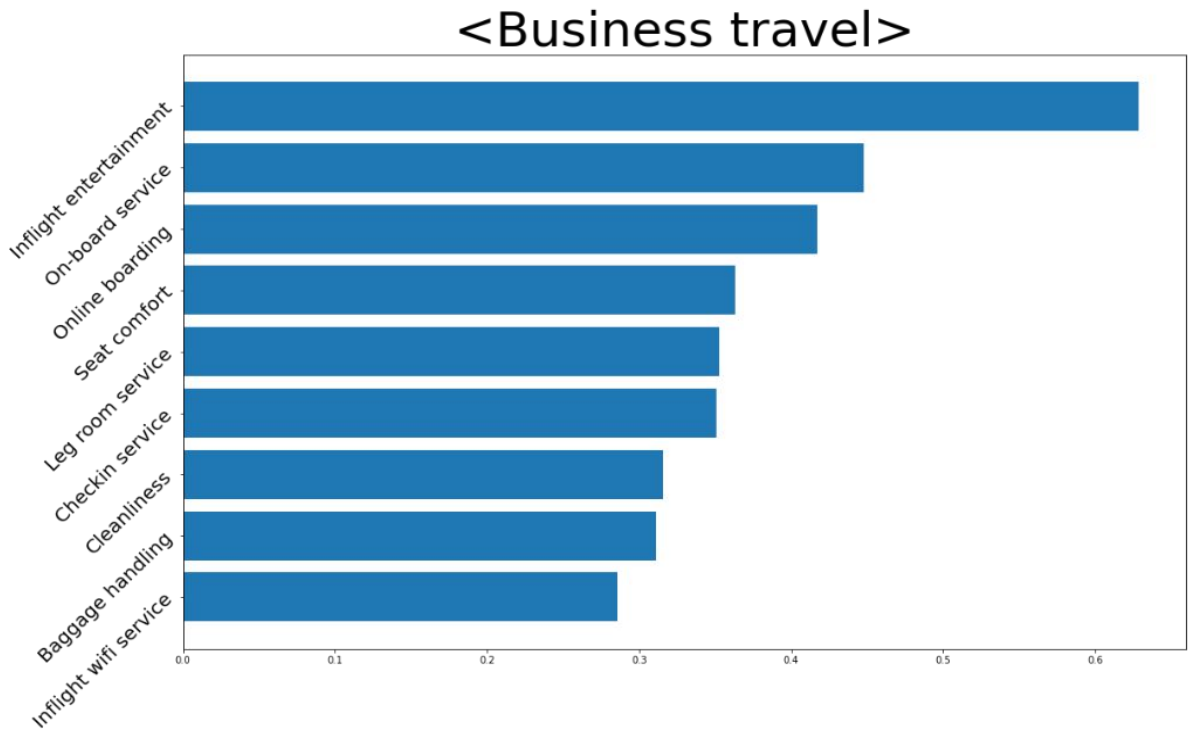
대표적으로 **Male**그룹은 **Seat comfort**의 상관도가 낮은반면 **Female** 그룹은 상관도가 높게 나왔다.

이는 일반적으로 남성의 체격이 더 크기때문에 **Seat comfort**에 민감한 것으로 추론 할 수 있다.

한편, **Cleanliness** 또한 남/여 그룹에서 차이를 보이는데, 이를 통해 남성은 상대적으로 청결도에 크게 민감하지 않다는 점을 추론 할 수 있다.

5. Object Feature 값들에 대한 해석 및 처리





<해석 및 추론>

출장 목적의 고객의 경우 전반적인 항목에 대해 고르게 평가하는 경향성 나타남

5. Object Feature 값들에 대한 해석 및 처리

<기타 Feature 들 중 (4 항목) Group2를 채택한 이유>

Group2에 속해있는 **seat comfort**의 경우 **Business/Eco**로 나눠 분석할 경우 상관도 차이를 크게 보였으며 특히 **Business**의 경우 타 항목보다 높은 상관도를 보인다. 따라서 해당 **Feature**를 버리기 보다 활용하는 것이 타당할 것이다.

<Object Feature의 활용**>**

실제 필드에서 활용하기 위해서는 **5**의 분석처럼 고객 **Group**을 나누어 분석하고 이를 활용하는 것이 보다 정확한 타겟팅 효과를 기대할 수 있을 것.

6. 결론 및 고찰

<데이터 전처리 전후의 성능 값이 크게 증가하지는 않았다.>

이는 3000*24개 데이터 중 값을 변환 해 준 데이터가 약 180개(0.25%) 밖에 되지 않았다.
따라서 성능 향상이 크게 이루어 지지 않았을 것이다.
그러나, 전처리 과정을 통해 약 10*3000개의 데이터(전체의 약 41%)를 제거한 후 분석을 진행하였으므로 분석 비용 및 시간 측면에서 효율성이 증대되었을 것이다.

Object Feature의 결과를 보면, 고객의 소속 그룹에 따라 중요하게 평가하는 척도가 다르게 나타남을 확인 할 수 있었다.
실제 항공사의 경우 이러한 그룹별 중요 feature 들에 맞춘 서비스를 강화하고 약점을 보완하거나, 마케팅의 키워드로 사용함으로써 고객 만족도 향상 및 신규 고객 유치를 기대할 수 있을 것으로 생각된다.

감사합니다 !