항공 서비스에 대한
만족을 높이는 요인

[참여자]

신현식(PPT)
정승기(보고서 작성)
남궁다비(보고서 작성)
신석현(발표)

[요약]

'위드 코로나'의 시행으로 해외여행을 다니는 여행객들의 기대감이 높아져 항공 산업이 다시 눈을 뜨고 있다. 항공 산업이 주목을 받고 있는 만큼 우리나라의 많은 항공사에서는 고객 수요를 더 늘리기 위해 고객들이 무엇을 원하는지 찾아 볼 것이다.
그렇기 때문에 우리는 '범주형 자료분석'방법을 활용하여 고객의 만족도를 높일 수 있는 요인들을 찾아보고자 한다.
먼저 인적사항과 12가지의 서비스 만족도를 기초 분석하였고, 카이제곱 독립성 검정을 통하여 각 변수들 사이에 연관성이 있는지 알아보았다.
그 다음 전체 변수에 대해 로지스틱회귀모형을 만들고 Stepwise Model Selection, Backward Elimination 방법을 활용하여 유의하지 않은 변수들을 제거하며 최종 모형을 만들었다.
마지막으로, Classification Table, ROC-Curve를 활용하여 모델의 정분류를 파악하였고 민감도, 특이도를 계산하여 ROC-Curve만으로 부족했던 평가 방법을 뒷받침하였다.

[Introduction]

'위드 코로나' 기대감...해외여행 '기지개'

김진이 기자 | 승인 2021.10.19 09:52 | 댓글 0

G마켓, 국제선 항공권 매출 '쑥'...장거리 여행 수요 꿈틀
티웨이항공, 일상준비 중...중대형 항공기 본격 훈련 시작
에어서울, 660일만 국제선 재개...12월말 인천-괌 운항

[이지경제=김진이 기자] 해외 항공권 수요가 활기를 찾고 있다.

19일 국내 대표 온라인쇼핑몰 G마켓과 옥션이 9월 국제선 항공권 매출을 분석한 결과, 해외 항공권 매출이 전년 동기 대비 69% 증가했다.

전월(8월)과 비교해도 29% 늘었다.

코로나19 이후 1년 7개월만에 해외여행 되살아나

서미영 기자 pepero99@chosun.com

기사입력 2021.11.12 11:01

2021년 10월, 정부가 위드 코로나를 발표하면서 코로나의 직접적인 타격을 받은 산업 중 하나인 항공 산업이 다시 주목을 받고 있다. 이에 따라 해외 여행객들의 수요가 급증하고 있으며 항공사의 서비스에 대한 기대감 또한 커지고 있다. 제한이 풀려가고 있는 시점에서 조금이라도 수요를 더욱 더 활성화하고 고객이 항공사 서비스에 대해 만족도를 높이는 요인들을 찾아 니즈를 파악하고 이에 따른 인사이트를 도출함에 본 프로젝트의 목적이 있겠다.

[Data Description]

-Data Preprocessing
본 프로젝트를 수행하기 위해 Kaggle 플랫폼의 'Airline Passenger Satisfaction'의 test데이터를 이용하였다. 이 데이터는 미국의 어느 항공사가 고객을 상대로 만족도를 조사한 데이터로, 기본 인적사항과 12개의 서비스 등으로 구성되어 있다. 국내 일반 항공사의 데이터가 아닌 미국 항공사의 데이터를 이용하는 이유는 국내 항공사의 데이터 수집에 한계가 있었고 미국 항공사와 국내 일반 항공사의 가장 큰 차이점인 수화물 규정(중량 초과 시 추가요금 등)을 제외한 다른 점에서 큰 차이가 없다고 판단하였기 때문에 일반화하기에 적합한 데이터라고 생각하였다. 총 13만개의 데이터가 있지만 이 중 랜덤으로 추출한 20%를 test데이터로 지정하여 사용하였다.
test데이터는 총 25개의 변수와 25893개의 데이터로 이루어져 있으며 변수의 설명은 다음과 같다.

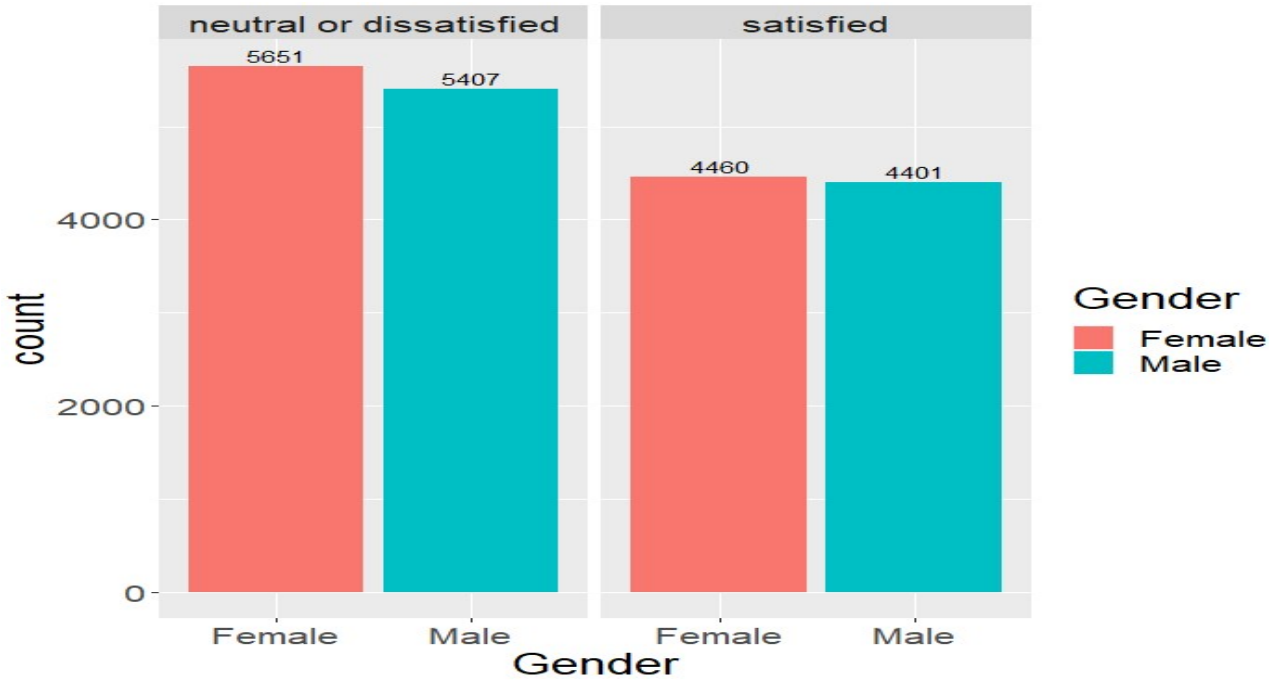
변수	내용	데이터 형태	결측치 수
X	데이터 번호	int	0
id	고객 번호	int	0
Gender	성별(male,female)	chr	0
Customer.Type	고객 유형(Loyal Cust,disloyal cust)	chr	0
Age	고객의 나이	int	0
Type.of.Travel	승객의 비행 목적(Personal,Business)	chr	0
Class	승객의 좌석 클래스(Business,Eco,Eco Plus)	chr	0
Flight.Distance	승객이 비행한 거리	int	0
Inflight.wifi.service	무선 인터넷 서비스 만족도(1~5, 0:미해당)	int	0
Departure.Arrival.time.convenient	출발/도착 시간 편의 만족도(1~5, 0:미해당)	int	0
Ease.of.Online.booking	온라인 예약 만족도(1~5, 0:미해당)	int	0
Gate.location	게이트의 위치 만족도(1~5)	int	0
Food.and.drink	기내식 만족도(1~5, 0:미해당)	int	0
Online.boarding	온라인 탑승권 만족도(1~5, 0:미해당)	int	0
Seat.comfort	좌석의 편의 만족도(1~5)	int	0
Inflight.entertainment	기내 오락거리 만족도(1~5,0: 미해당)	int	0
On.board.service	탑승 서비스 만족도(1~5, 0: 미해당)	int	0
Leg.room.service	발 뻗을 수 있는 공간의 만족도(1~5, 0:미해당)	int	0
Baggage.handling	수하물 시스템 만족도(1~5)	int	0
Checkin.service	체크인 서비스 만족도(1~5)	int	0
Inflight.service	기내 서비스 만족도(1~5, 0:미해당)	int	0
Cleanliness	기내 청결 만족도(1~5, 0:미해당)	int	0
Departure.Delay.in.Minutes	예정된 출발시간보다 늦게 출발한 정도	int	0
Arrival.Delay.in.Minutes	예정된 도착시간보다 늦게 도착한 정도	num	83
satisfaction	만족 여부(neutral or dissatisfied,satisfied)	chr	0

하지만 결측값과 이상값 존재 등의 이유로 위의 데이터를 그대로 분석하기에는 무의미한 결과를 초래할 수 있기 때문에 아래와 같이 전처리를 실시하였다.
⑴ X,id와 같은 고유의 값 변수 제거
⑵ 각 변수별로 결측값 존재 여부를 확인한 결과 Arrival.Delay.in.Minutes변수에서 83개의 결측값을 찾을 수 있었고 결측값을 제거하였음
⑶ 만족도 척도(1~5)에 해당하지 않는 값을 결측치로 판단하여 제거하였음.
⑷ Departure Delay in Minutes와 Arrival Delay in Minutes에 이상치가 존재하는 데이터들 제거(나이와 비행거리는 이상치가 존재하지만 의미가 있을 것 같아 제거하지 않았음)
⑸ 출발,도착 지연 시간Arrival Delay in Minutes와 Departure Delay in Minutes 차이를 계

산하여 파생변수(ontime)생성(Arrival.Delay.in.Minutes-Departure.Delay.in.Minutes)
전처리를 통해 남은 데이터는 19919개의 데이터와 24개의 변수로 유의미한 분석 결과를 도출할 수 있는 형태로 가공되었다.

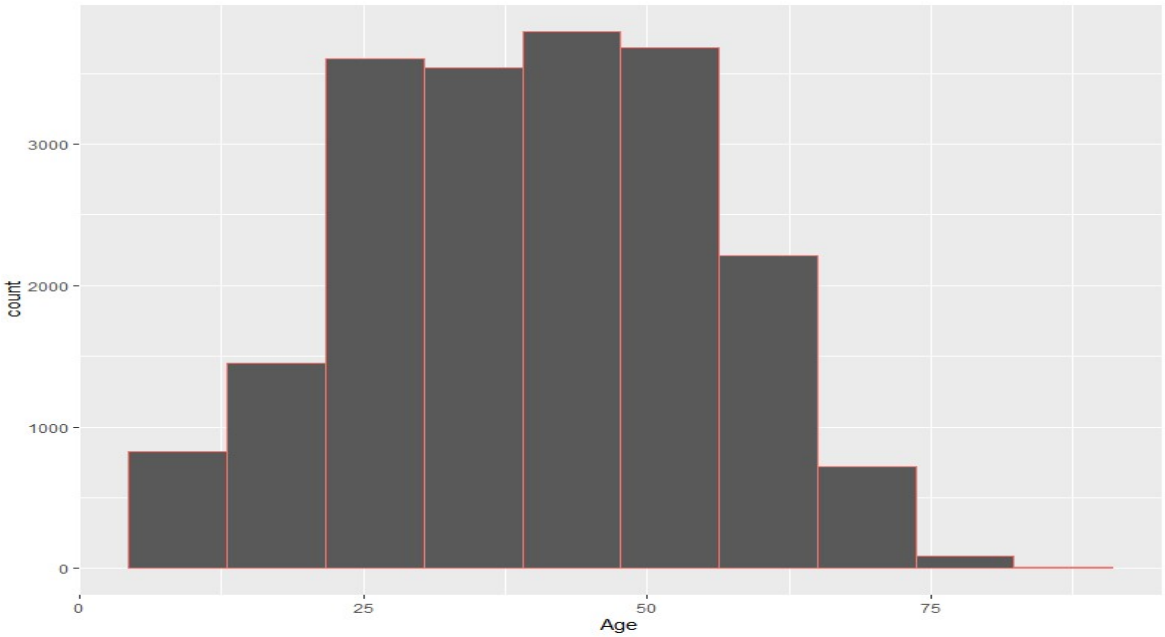
-Exploratory Data Analysis

성별



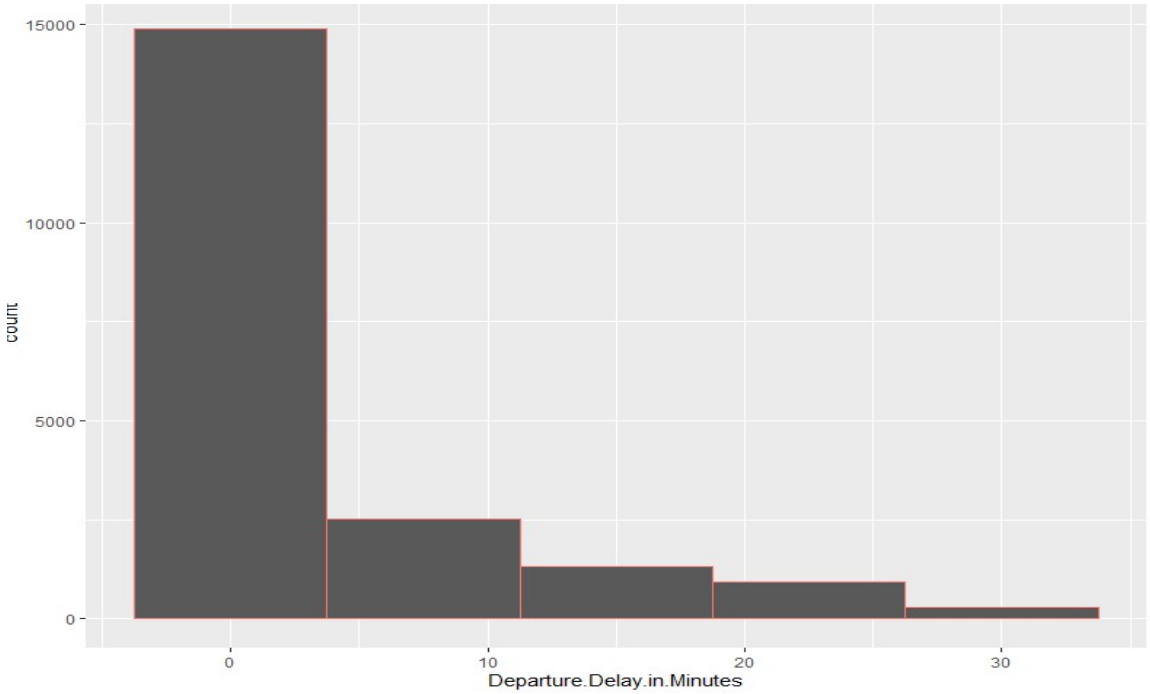
두 범주 모두 여성 응답자가 남성보다 많았지만 남성과 여성의 만족 여부의 비율은 비슷하게 나타났다.

나이



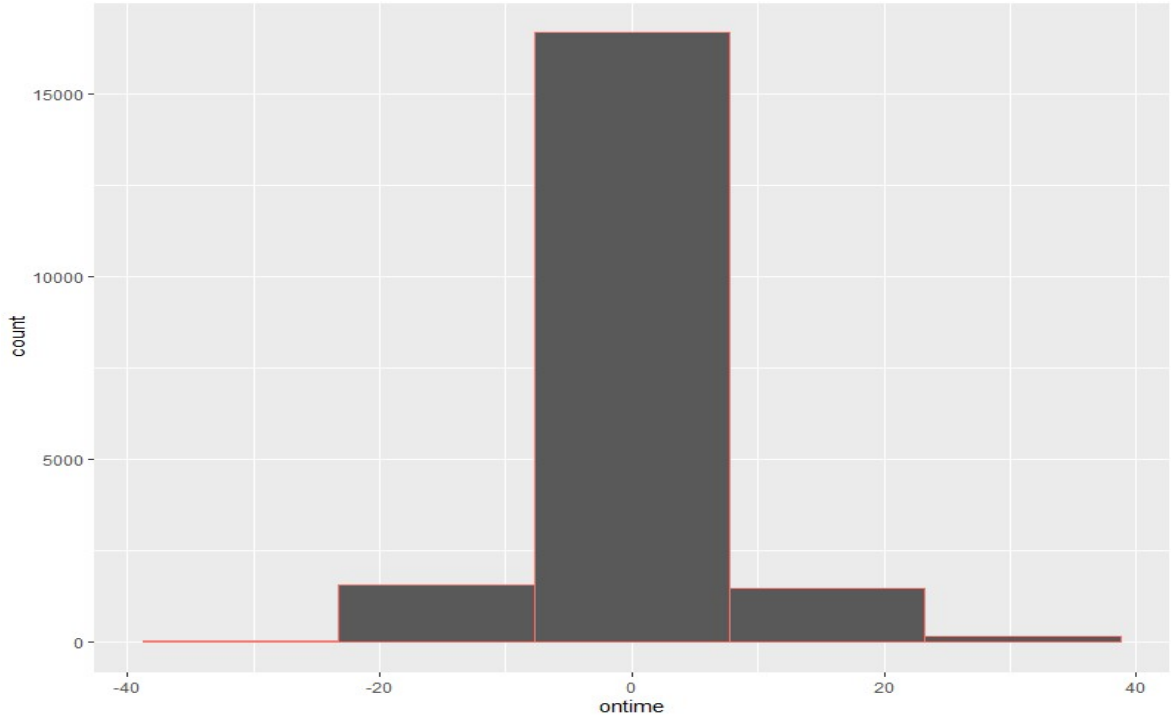
25~50세의 청,중장년층쪽에서 빈도가 높다는 것을 알 수 있었다.

출발 딜레이 시간



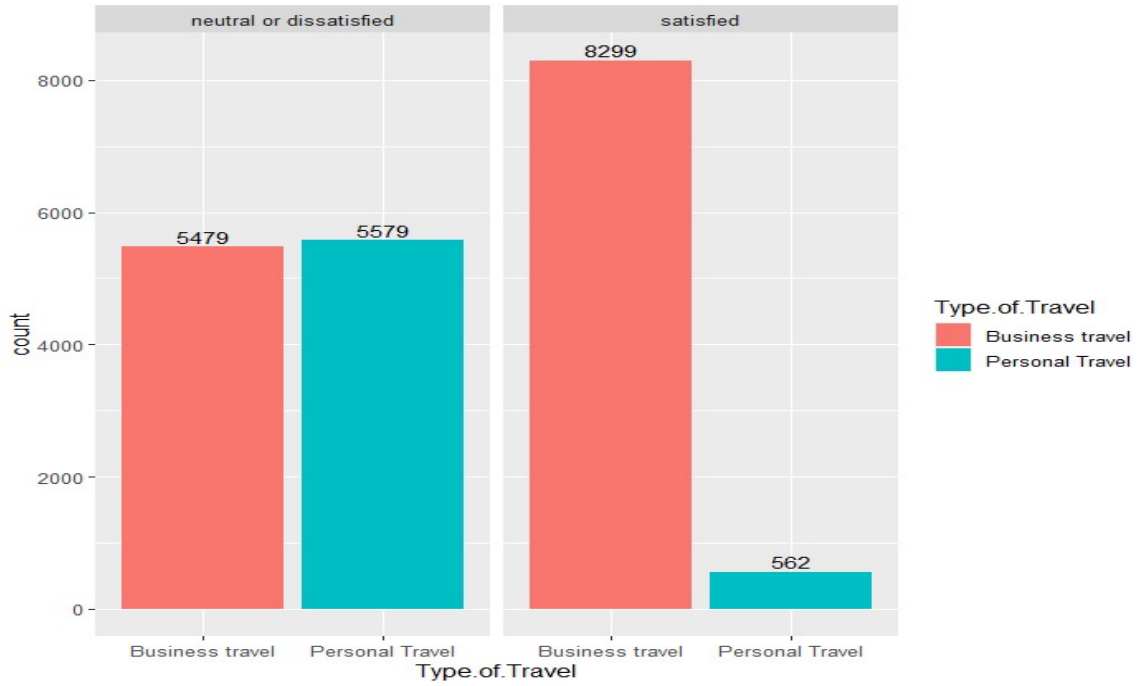
상당수가 0분으로 대체로 제 시간에 출발하였으나 대략 3000건정도가 출발시간을 잘 지키지 않은 것으로 나타났다.

ontime



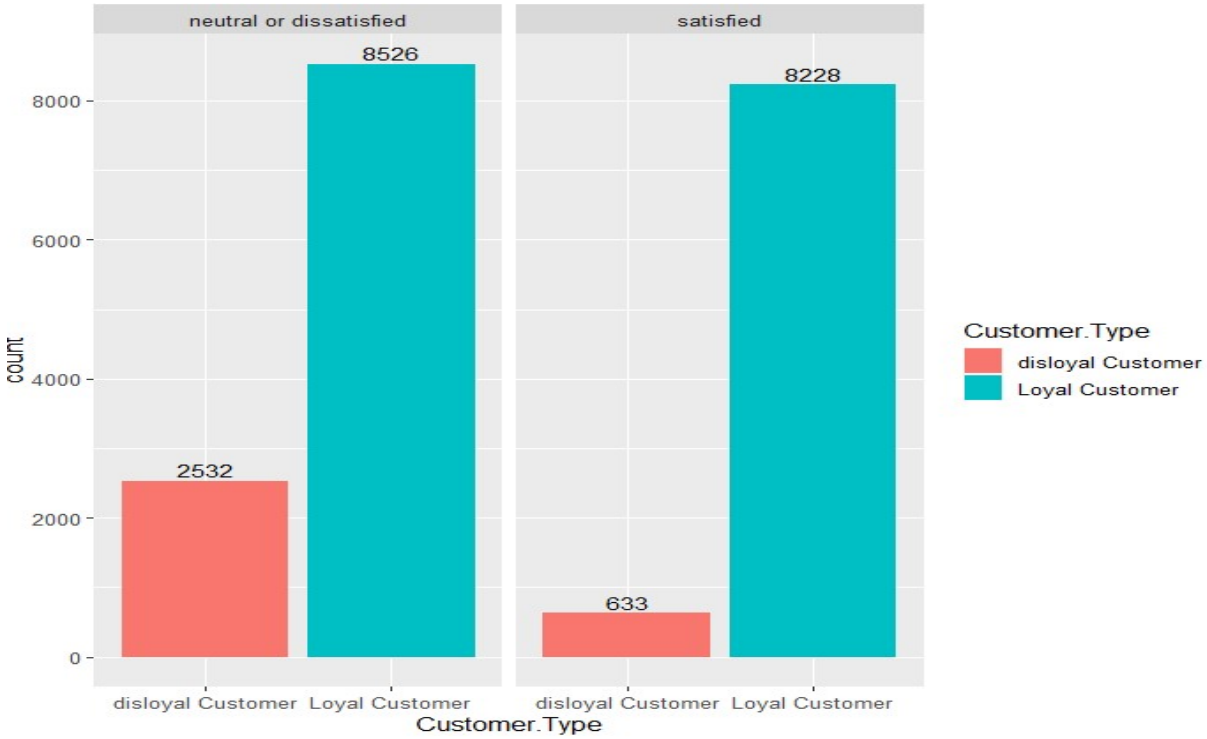
상당수가 도착 예정시간에 맞춰 딜레이 오차가 없었지만(예를 들어 출발시간이 5분 딜레이가 된다면 도착시간도 5분 딜레이가 되므로 ontime=0) 예정보다 빨리 도착하거나 늦게 도착한 경우도 있던 것으로 나타났다.

여행 유형



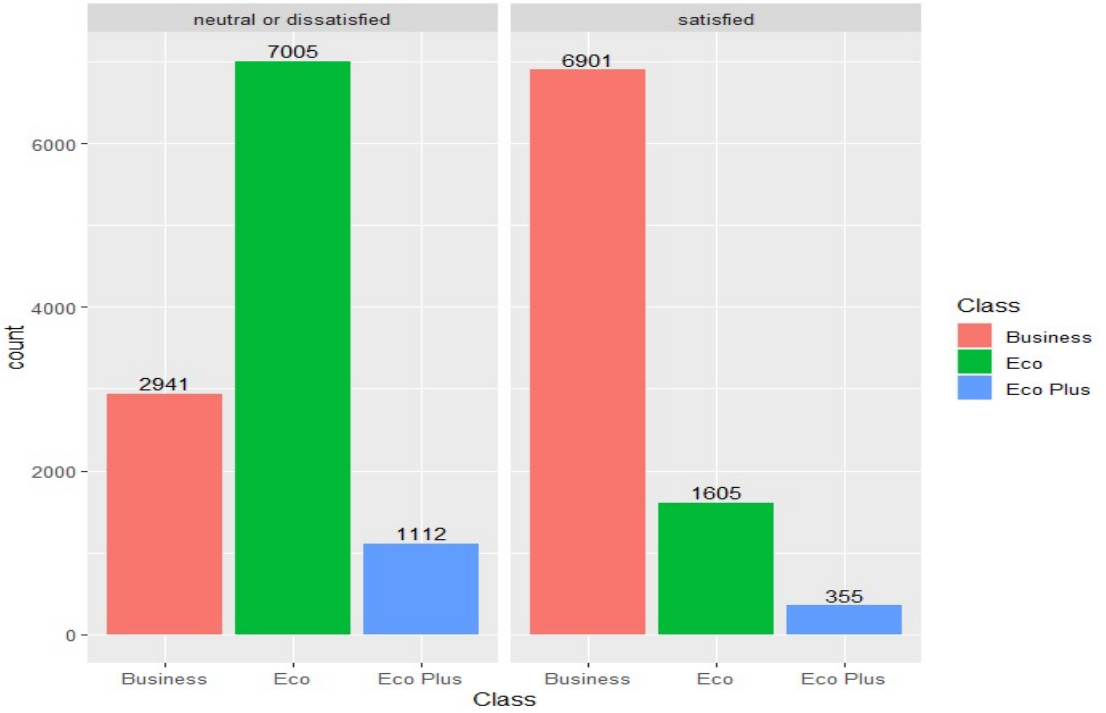
만족하는 집단의 경우 출장을 이유로 항공서비스를 이용하는 고객이 여행 고객보다 훨씬 많았지만 만족하지 않은 집단의 경우 두 유형의 수가 비슷했다.

소비자 유형



만족, 불만족 집단의 Loyar Customer 수는 비슷했지만 불만족 집단의 경우 disloyal Customer의 수가 2532건으로 많다는 것을 알 수 있다.

클래스



만족하는 집단에서는 비즈니스클래스가, 그렇지 않은 집단에서는 에코클래스가 가장 많았다.

서비스 만족도

	satisfaction	Inflight.wifi.servi~	Departure.Arrival.time.con~	Ease.of.Online.boo~	Gate.location	Food.and.drink	Online.boarding	Seat.comfort
*	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	2.41	3.29	2.63	2.99	2.97	2.72	3.05
2	1	3.35	3.10	3.22	2.97	3.55	4.17	4.01

	satisfaction	Inflight.entertainment	On.board.service	Leg.room.service	Baggage.handling	Checkin.service	Inflight.service	Cleanliness
*	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	2.87	3.01	2.99	3.37	3.05	3.41	2.92
2	1	4.05	3.91	3.89	4.01	3.67	4.03	3.79

만족 여부 집단별로 12개의 서비스 만족도의 평균을 구한 결과이다. 단순 평균으로만 볼 때 출발,도착시간의 만족도는 두 집단이 비슷하였으나 나머지 모든 서비스만족도에서 만족하는 집단의 평점 평균이 높다는 것을 알 수 있다.

[Methodology]

1.분할표 및 독립성검정
만족 여부 satisfaction 변수를 종속 변수로 두어 각 변수들간의 독립성 관계를 알아보고자 각 변수값을 범주화하여 이원분할표를 그려보았고, 카이제곱 독립성검정을 실시하였다.

2.일반화 선형 모형
binary형태인 satisfaction변수를 반응변수로 하여 먼저 전체 변수에 대한 로지스틱회귀모형과 변수가 없는 회귀모형을 Likelihood Ratio검정을 통해 다중공선성을 확인하고 Stepwise Model Selection과 Backward방법을 통해 유의하지 않은 변수들을 하나씩 제거하여 최종 모형을 만들었다.

3.최종 모형 평가
표본비율에 대한 반응변수 satisfaction 에 대한 분산을 구하고 설명된 분산을 cut-off하여 분산보다 작으면 0, 크면 1로 예측하였고 satisfaction의 예측값과 실제값을 비교하기 위해 Classfication Table과 ROC-Curve를 이용하여 분류가 잘 되었는지 평가하였다. 또한 민감도와 특이도를 구하여 정분류율만 표시해주는 ROC-Curve의 설명을 뒷받침하였다.

[Results]

1.분할표 및 독립성 검정

```
> str(data)
'data.frame':   19919 obs. of  24 variables:
 $ Gender                           : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 1 1 1 1 1 ...
 $ Customer.Type                    : Factor w/ 2 levels "disloyal Customer",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Age                              : int  36 49 16 77 47 46 47 33 46 52 ...
 $ Type.of.Travel                   : Factor w/ 2 levels "Business travel",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Class                            : Factor w/ 3 levels "Business","Eco",..: 1 2 2 1 2 1 2 1 1 1 ...
 $ Flight.Distance                  : int  2863 1182 311 3987 556 1744 1235 325 1009 925 ...
 $ Inflight.wifi.service            : int  1 2 3 5 5 2 4 2 5 2 ...
 $ Departure.Arrival.time.convenient: int  1 3 3 5 2 2 1 5 5 2 ...
 $ Ease.of.Online.booking           : int  3 4 3 5 2 2 1 5 5 2 ...
 $ Gate.location                    : int  1 3 3 5 2 2 1 5 5 2 ...
 $ Food.and.drink                   : int  5 4 5 3 5 3 5 1 4 5 ...
 $ Online.boarding                  : int  4 1 5 5 5 4 1 3 5 5 ...
 $ Seat.comfort                     : int  5 2 3 5 5 4 5 4 5 4 ...
 $ Inflight.entertainment           : int  4 2 5 5 5 4 3 2 5 4 ...
 $ On.board.service                 : int  4 2 4 5 2 4 3 2 5 4 ...
 $ Leg.room.service                 : int  4 2 3 5 2 4 4 2 5 4 ...
 $ Baggage.handling                 : int  4 2 1 5 5 4 3 2 5 4 ...
 $ Checkin.service                  : int  3 4 1 4 3 5 1 3 5 3 ...
 $ Inflight.service                 : int  4 2 2 5 3 4 3 2 5 4 ...
 $ Cleanliness                      : int  5 4 5 3 5 4 4 4 3 5 ...
 $ Departure.Delay.in.Minutes       : int  0 0 0 0 1 28 29 18 0 10 ...
 $ Arrival.Delay.in.Minutes         : int  0 20 0 0 0 14 19 7 0 0 ...
 $ satisfaction                     : int  1 1 1 1 1 1 1 0 1 1 ...
 $ ontime                           : int  0 20 0 0 -1 -14 -10 -11 0 -10 ...
```

```
data <- transform(data, Age = ifelse(Age < 10, "0",
  ifelse(Age >= 10 & Age < 20, "1",
  ifelse(Age >= 20 & Age < 30, "2",
  ifelse(Age >= 30 & Age < 40, "3",
  ifelse(Age >= 40 & Age < 50, "4",
  ifelse(Age >= 50 & Age < 60, "5",
  ifelse(Age >= 60 & Age < 70, "6",
  ifelse(Age >= 70 & Age < 80, "7", "8")))))))))
```

```
data$Age<-as.factor(data$Age)
data$Inflight.wifi.service<-as.factor(data$Inflight.wifi.service)
```

이원분할표와 독립성 검정에 앞서 데이터 타입이 int인 범주형 변수를 factor로 변환하였고 데이터 타입이 int인 연속형 변수는 범위를 나누어 factor로 변환하였다.

Gender와 satisfaction 변수간의 독립성 검정

```
        satisfaction
Gender       0    1
  Female  5651 4460
  Male    5407 4401
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  Gender_table
X-squared = 1.1374, df = 1, p-value = 0.2862
```

=> p-value>0.05이므로 귀무가설 H0을 기각하지 못하기 때문에 Gender는 만족도와 독립이라고 볼 수 있다.

Customer Type와 satisfaction 변수간의 독립성 검정

```
                  satisfaction
Customer.Type          0    1
  disloyal Customer 2532  633
  Loyal Customer    8526 8228

        Pearson's Chi-squared test with Yates' continuity correction

data:  CustomerType_table
X-squared = 912.31, df = 1, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Customer Type은 만족도와 독립이라고 볼 수 없다.

Age와 satisfaction 변수간의 독립성 검정

```
   satisfaction
Age    0    1
  0  320   41
  1 1115  310
  2 2447 1256
  3 2228 1699
  4 1939 2766
  5 1603 2271
  6 1190  448
  7  200   53
  8   16   17

        Pearson's Chi-squared test

data:  Age_table
X-squared = 1583.8, df = 8, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Age는 만족도와 독립이라고 볼 수 없다.

Type of Travel와 satisfaction 변수간의 독립성 검정

```
                 satisfaction
Type.of.Travel       0    1
  Business travel 5479 8299
  Personal Travel 5579  562

        Pearson's Chi-squared test with Yates' continuity correction

data:  TypeTravel_table
X-squared = 4486.1, df = 1, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Type of Travel은 만족도와 독립이라고 볼 수 없다.

Class와 satisfaction 변수간의 독립성 검정

```
           satisfaction
Class          0    1
  Business  2941 6901
  Eco       7005 1605
  Eco Plus  1112  355

        Pearson's Chi-squared test

data:  Class_table
X-squared = 5191.6, df = 2, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Class는 만족도와 독립이라고 볼 수 없다.

Flight Distance와 satisfaction 변수간의 독립성 검정

```
               satisfaction
Flight.Distance    0    1
              1 3701 1958
              2 3732 1752
              3 1549  861
              4  805 1184
              5  925 1824
              6  165  657
              7  173  613
              8    8   12

         Pearson's Chi-squared test

data:  FlightDist_table
X-squared = 2139.7, df = 7, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Flight Distance는 만족도와 독립이라고 볼 수 없다.

Inflight wifi service와 satisfaction 변수간의 독립성 검정

```
                     satisfaction
Inflight.wifi.service    0    1
                    1 2158 1325
                    2 3758 1350
                    3 3623 1382
                    4 1489 2507
                    5   30 2297

          Pearson's Chi-squared test

data:  InflightWifi_table
X-squared = 4619.6, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Inflight wifi service는 만족도와

독립이라고 볼 수 없다.

Departure / Arrival time convenient와 satisfaction 변수간의 독립성 검정

```
                                satisfaction
Departure.Arrival.time.convenient    0    1
                                1 1544 1673
                                2 1917 1627
                                3 1964 1629
                                4 3079 2026
                                5 2554 1906
```

```
        Pearson's Chi-squared test

data:  DATimeConvenient_table
X-squared = 130.76, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Departure / Arrival time convenient만족도와 독립이라고 볼 수 없다.

Ease of Online booking와 satisfaction 변수간의 독립성 검정

```
                     satisfaction
Ease.of.Online.booking    0    1
                     1 2065 1438
                     2 3288 1561
                     3 3179 1596
                     4 1754 2168
                     5  772 2098
```

```
        Pearson's Chi-squared test

data:  EaseOnlineBook_table
X-squared = 1686.6, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Ease of Online booking은 만족도와 독립이라고 볼 수 없다.

Gate location와 satisfaction 변수간의 독립성 검정

```
             satisfaction
Gate.location    0    1
            1 1652 1754
            2 1930 1766
            3 3572 1882
            4 2697 1883
            5 1207 1576
```

```
        Pearson's Chi-squared test

data:  GateLocation_table
X-squared = 491.25, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Gate location은 만족도와 독립이

라고 볼 수 없다.

Food and drink와 satisfaction 변수간의 독립성 검정

```
                satisfaction
Food.and.drink    0    1
             1 1985  438
             2 2482 1652
             3 2412 1754
             4 2220 2594
             5 1959 2423

        Pearson's Chi-squared test

data:  FoodDrink_table
X-squared = 1107.6, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Food and drink는 만족도와 독립이라고 볼 수 없다.

Online boarding와 satisfaction 변수간의 독립성 검정

```
                satisfaction
Online.boarding    0    1
              1 1686  255
              2 3025  391
              3 3534  626
              4 2276 3886
              5  537 3703

        Pearson's Chi-squared test

data:  OnlineBoard_table
X-squared = 7755.6, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Online boarding은 만족도와 독립이라고 볼 수 없다.

Seat comfort와 satisfaction 변수간의 독립성 검정

```
            satisfaction
Seat.comfort    0    1
           1 1791  460
           2 2095  583
           3 2711  781
           4 2702 3578
           5 1759 3459

        Pearson's Chi-squared test

data:  SeatComfort_table
X-squared = 3179.8, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에 Seat comfort는 만족도와 독립이라고 볼 수 없다.

Inflight entertainment와 satisfaction 변수간의 독립성 검정

```
                       satisfaction
Inflight.entertainment    0    1
                     1 2104  261
                     2 2653  595
                     3 2518  973
                     4 2133 3636
                     5 1650 3396

        Pearson's Chi-squared test

data:  InflightEntertain_table
X-squared = 4228.8, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Inflight entertainment은 만족도와 독립이라고 볼 수 없다.

On board service와 satisfaction 변수간의 독립성 검정

```
                satisfaction
On.board.service    0    1
               1 1761  425
               2 2078  667
               3 2919 1349
               4 2845 3268
               5 1455 3152

         Pearson's Chi-squared test

data:  OnBoardService_table
X-squared = 2562.5, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, On board service는 만족도와 독립이라고 볼 수 없다.

Leg room service와 satisfaction 변수간의 독립성 검정

```
                satisfaction
Leg.room.service    0    1
               1 1554  336
               2 2833 1016
               3 2697 1051
               4 2158 3335
               5 1816 3123

        Pearson's Chi-squared test

data:  LegRoomService_table
X-squared = 2754.8, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Leg room service는 만족도와 독립이라고 볼 수 없다.

Baggage handling와 satisfaction 변수간의 독립성 검정

```
                satisfaction
Baggage.handling    0    1
               1  924  377
               2 1555  584
               3 3005  971
               4 3703 3561
               5 1871 3368

        Pearson's Chi-squared test

data:  BaggageHandling_table
X-squared = 1922.9, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Baggage handling은 만족도와 독립이라고 볼 수 없다.

Check in service와 satisfaction 변수간의 독립성 검정

```
                satisfaction
Checkin.service    0    1
              1 1881  568
              2 1765  595
              3 2901 2566
              4 2987 2611
              5 1524 2521

        Pearson's Chi-squared test

data:  CheckinService_table
X-squared = 1349.6, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Check in service는 만족도와 독립이라고 볼 수 없다.

Inflight service와 satisfaction 변수간의 독립성 검정

```
                satisfaction
Inflight.service    0    1
               1  887  358
               2 1524  608
               3 2894  902
               4 3700 3556
               5 2053 3437

        Pearson's Chi-squared test

data:  InflightService_table
X-squared = 1794.9, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Inflight service는 만족도와 독립이라고 볼 수 없다.

Cleanliness와 satisfaction 변수간의 독립성 검정

```
           satisfaction
Cleanliness    0    1
          1 2089  478
          2 2399  577
          3 2556 2095
          4 2382 2858
          5 1632 2853

        Pearson's Chi-squared test

data:  Cleanliness_table
X-squared = 2333.9, df = 4, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Cleanliness는 만족도와 독립이라고 볼 수 없다.

Departure delay와 satisfaction 변수간의 독립성 검정

```
                          satisfaction
Departure.Delay.in.Minutes    0    1
                         1 8319 6971
                         2  926  692
                         3  644  461
                         4  499  329
                         5  633  389
                         6   37   19

         Pearson's Chi-squared test

data:  DepartDelay_table
X-squared = 40.101, df = 5, p-value = 1.425e-07
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Departure delay는 만족도와 독립이라고 볼 수 없다.

Arrival delay와 satisfaction 변수간의 독립성 검정

```
                        satisfaction
Arrival.Delay.in.Minutes    0    1
                       1 8014 7140
                       2 1036  602
                       3  815  422
                       4  529  312
                       5  576  337
                       6   88   48

        Pearson's Chi-squared test

data:  ArrivDelay_table
X-squared = 180.45, df = 5, p-value < 2.2e-16
```

=> p-value<0.05이므로 귀무가설 H0을 기각하기 때문에, Arrival delay는 만족도와 독립이라고 볼 수 없다.

ontime와 satisfaction 변수간의 독립성 검정

```
      satisfaction
ontime    0    1
     1   33   36
     2  484  351
     3 2185 1697
     4 7543 6324
     5  617  338
     6  179  104
     7   17   11

        Pearson's Chi-squared test

data:  ontime_table
X-squared = 50.786, df = 6, p-value = 3.271e-09
```

=> p-value<0.050이므로 귀무가설 H0을 기각하기 때문에, ontime은 만족도와 독립이라고 볼 수 없다.

2. 일반화 선형 모형

먼저, 변수 전체에 대하여 로지스틱 모형을 적합시켜보았다.

```
Call:
glm(formula = satisfaction ~ ., family = binomial, data = data3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3935  -0.2698  -0.0414   0.2889   4.0489

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.180e+01  2.474e-01 -47.705  < 2e-16 ***
GenderMale                       1.203e-01  5.268e-02   2.283  0.02242 *
Customer.TypeLoyal Customer      2.873e+00  8.860e-02  32.428  < 2e-16 ***
Age                             -4.454e-03  1.921e-03  -2.319  0.02038 *
Type.of.TravelPersonal Travel   -3.486e+00  8.994e-02 -38.759  < 2e-16 ***
ClassEco                        -6.819e-01  7.199e-02  -9.471  < 2e-16 ***
ClassEco Plus                   -8.886e-01  1.129e-01  -7.872 3.50e-15 ***
Flight.Distance                  3.537e-05  2.980e-05   1.187  0.23525
Inflight.wifi.service            8.595e-01  3.288e-02  26.143  < 2e-16 ***
Departure.Arrival.time.convenient -3.939e-01  2.883e-02 -13.659  < 2e-16 ***
Ease.of.Online.booking           3.427e-01  3.359e-02  10.204  < 2e-16 ***
Gate.location                   -2.809e-01  2.912e-02  -9.647  < 2e-16 ***
Food.and.drink                  -8.455e-02  2.884e-02  -2.931  0.00337 **
Online.boarding                  9.467e-01  3.019e-02  31.358  < 2e-16 ***
Seat.comfort                     6.767e-03  3.098e-02   0.218  0.82710
Inflight.entertainment           3.217e-02  4.020e-02   0.800  0.42352
On.board.service                 3.820e-01  2.865e-02  13.332  < 2e-16 ***
Leg.room.service                 3.298e-01  2.384e-02  13.834  < 2e-16 ***
Baggage.handling                 1.306e-01  3.142e-02   4.155 3.25e-05 ***
Checkin.service                  4.162e-01  2.315e-02  17.983  < 2e-16 ***
Inflight.service                 1.615e-01  3.355e-02   4.812 1.49e-06 ***
Cleanliness                      2.895e-01  3.163e-02   9.150  < 2e-16 ***
Departure.Delay.in.Minutes      -1.341e-02  4.922e-03  -2.724  0.00645 **
Arrival.Delay.in.Minutes        -3.211e-02  4.802e-03  -6.687 2.28e-11 ***
ontime                                  NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27370.8  on 19918  degrees of freedom
Residual deviance:  9648.2  on 19895  degrees of freedom
AIC: 9696.2

Number of Fisher Scoring iterations: 7
```

Flight.Distance, Seat.comfort, Inflight.entertainment에서 귀무가설을 채택하여 로지스틱 모형에 유의하지 않음을 나타내고 있다.

```
Model 2: satisfaction ~ 1
  #Df   LogLik  Df Chisq Pr(>Chisq)
1  24  -4824.1
2   1 -13685.4 -23 17723  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

절편이 없는 모형을 설정하여 위의 모델과의 유의성을 알아보기 위해 Likelyhood Ratio검정을 실시해 본 결과 p-value가 0.001보다 작으므로 귀무가설을 기각하여 독립변수들 사이에 다중공선성이 존재함을 알 수 있었다.

다중공선성에 의해 모형이 과적합되는 문제를 해결하기 위해 stepwise variable selection방법으로 유의성 여부를 확인해보았다.

설명변수	P-value	검정 결과	비고
Gender	0.2798	유의하지 않음	제외
Age	<0.001	유의	-
Type.of.Travel	<0.001	유의	-
Customer.Type	<0.001	유의	-
Flight.Distance	<0.001	유의	-
Class	<0.001	유의	-
Inflight.wifi.service	<0.001	유의	-
Ease.of.Online.booking	<0.001	유의	-
Inflight.service	<0.001	유의	-
Online.boarding	<0.001	유의	-
Inflight.entertainment	<0.001	유의	-
Food.and.drink	<0.001	유의	-
Seat.comfort	<0.001	유의	-
On.board.service	<0.001	유의	-
Leg.room.service	<0.001	유의	-
Departure.Arrival.time.convenient	<0.001	유의	-
Baggage.handling	<0.001	유의	-
Gate.location	0.388	유의하지 않음	제외
Cleanliness	<0.001	유의	-
Checkin.service	<0.001	유의	-
Departure.Delay.in.Minutes	<0.001	유의	-
Arrival.Delay.in.Minutes	<0.001	유의	-
ontime	<0.001	유의	-

단계적 방법을 통해 Gender와 Gate.location변수를 제외시킬 수 있었다.

다음은 Backward Elimination을 사용하여 각 변수들간의 유의성을 확인해보았다.

설명변수	P-value	검정 결과	비고
Age	0.01415	유의	-
Type.of.Travel	<0.001	유의	-
Customer.Type	<0.001	유의	-
Flight.Distance	0.2844	유의하지 않음	제외
Class	<0.001	유의	-
Inflight.wifi.service	<0.001	유의	-
Ease.of.Online.booking	<0.001	유의	-
Inflight.service	<0.001	유의	-

Online.boarding	<0.001	유의	-
Inflight.entertainment	0.626	유의하지 않음	제외
Food.and.drink	0.005369	유의	-
Seat.comfort	<0.001	유의	-
On.board.service	0.8124	유의하지 않음	-
Leg.room.service	<0.001	유의	-
Departure.Arrival.time.convenient	<0.001	유의	-
Baggage.handling	<0.001	유의	-
Gate.location	0.388	유의하지 않음	제외
Cleanliness	<0.001	유의	-
Checkin.service	<0.001	유의	-
Departure.Delay.in.Minutes	1	유의하지 않음	제외
Arrival.Delay.in.Minutes	1	유의하지 않음	제외
ontime	1	유의하지 않음	제외-

Backward Elimination방법에서 유의하지 않은 변수는 Flight.Distance, Inflight.entertainment, Seat.comfort, Departure Delay in Minutes,Arrival Delay in Minutes, ontime임을 알 수 있었다.

이제 맨 처음 제외했었던 gender와 Gate.location변수를 다시 넣어서 유의한지 판단하였다.

설명변수	P-value	검정 결과	비고
Gender	0.01514	유의	-
Gate.location	<0.001	유의	-

LR검정 결과 두 변수가 유의함으로 처음 제외시켰던 모형에 다시 추가할 수 있었다.
아래 캡쳐본은 변수 선택 방법을 통해 만들어진 최종 모형에 적합시킨 결과이다.

```
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -11.927656   0.236284 -50.480  < 2e-16 ***
GenderMale                       0.121108   0.052282   2.316  0.02053 *
Customer.TypeLoyal Customer      2.883136   0.084222  34.232  < 2e-16 ***
Age                             -0.004537   0.001886  -2.406  0.01613 *
Type.of.TravelPersonal Travel   -3.455519   0.087535 -39.476  < 2e-16 ***
ClassEco                        -0.706808   0.067523 -10.468  < 2e-16 ***
ClassEco Plus                   -0.917803   0.108685  -8.445  < 2e-16 ***
Inflight.wifi.service            0.862320   0.032298  26.698  < 2e-16 ***
Departure.Arrival.time.convenient -0.397707  0.028645 -13.884  < 2e-16 ***
Ease.of.Online.booking           0.343551   0.033378  10.293  < 2e-16 ***
Gate.location                   -0.278244   0.028978  -9.602  < 2e-16 ***
Food.and.drink                  -0.071899   0.025666  -2.801  0.00509 **
Online.boarding                  0.943390   0.028864  32.684  < 2e-16 ***
On.board.service                 0.394500   0.027408  14.394  < 2e-16 ***
Leg.room.service                 0.333299   0.023438  14.221  < 2e-16 ***
Baggage.handling                 0.132915   0.030750   4.322 1.54e-05 ***
Checkin.service                  0.405703   0.022564  17.980  < 2e-16 ***
Inflight.service                 0.167545   0.032102   5.219 1.80e-07 ***
Cleanliness                      0.307027   0.026602  11.541  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

최종 모형을 통해 구할 수 있는 로지스틱 회귀식은

$$\begin{aligned}\log\hat{y}(satisfaction) = &-11.928 + 0.121\,Gender\,Male + 2.883\,Customer.\,Type\,Loyal\,Customer - 0.005\,Age \\ &- 3.456\,Type.of.\,Travel\,Personal\,Travel - 0.707\,Class\,Eco - 0.918\,Class\,Eco\,Plus + 0.862\,Inflight.wifi.service \\ &- 0.398\,Departure.Arrival.time.convenient + 0.344\,Ease.of.Online.booking - 0.278\,Gate.location \\ &- 0.072\,Food.\mathrm{and}.drink + 0.943\,Online.boarding + 0.395\,On.board.service + 0.333\,Leg.room.service \\ &+ 0.133\,Baggage.handling + 0.406\,Checkin.service + 0.168\,Inflight.service + 0.307\,Cleanliness\end{aligned}$$

이다. 모든 변수의 p-value가 유의수준 0.05보다 작아 만족 여부에 유의한 영향을 준다.

이 회귀식에서 satisfaction의 가장 유의한 변수는 Loyar Customer유형으로, 1이 증가할 때마다 satisfaction에 영향을 주는 odds가 exp(2.883)=17.868 증가한다는 것을 알 수 있으며 Online.boarding의 경우 1이 증가할 때 satisfaction에 영향을 주는 odds가 exp(0.943)=2.568 증가한다. 한편 계수가 가장 작은 Personal Travel유형의 경우 odds가 exp(-3.456)=0.031이므로 미미하게 상승하는 것을 알 수 있다.

3. 최종 모형 평가

다음은 최종 모형에 대한 민감도와 특이도에 대한 검사를 진행해보았다.
표본비율에 대한 반응변수 satisfaction 에 대한 분산을 구하고 그보다 설명된 분산이 크면 1로 설정, 작으면 0으로 설정하여 분할표를 만들어보았다.

```
                    predicted
data3$satisfaction    0    1
                 0 9862 1196
                 1  871 7990
```

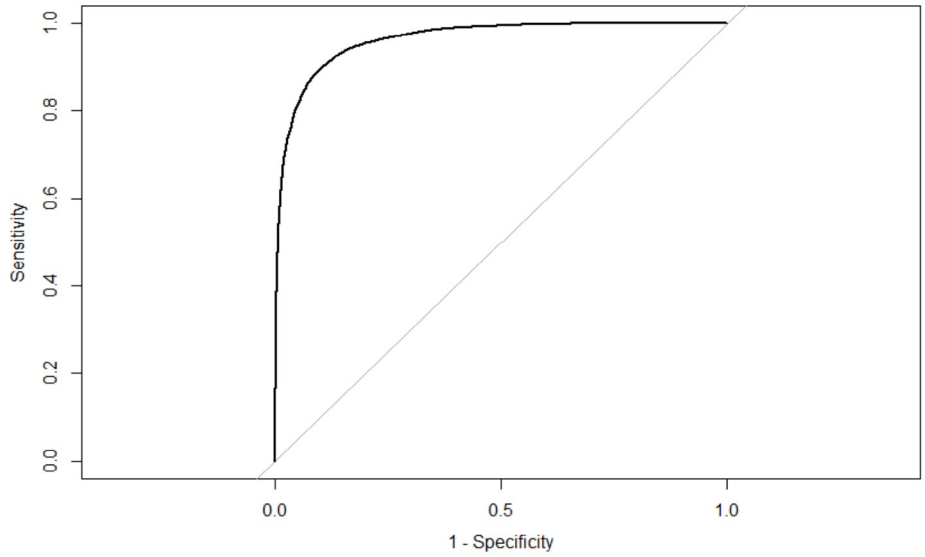
민감도 (실제 결과가 1일 때, 예측결과도 1인 경우) =7990/8861=0.9017
특이도 (실제 결과가 0일 때, 예측결과도 0인 경우) =9862/11058=0.8918
로 양쪽 다 높게 나왔다.

다음은 데이터의 설명정도를 알아보기 위하여 ROC 값을 찾고 그림을 그려보았다.
--AUC의 경우
0.9-1.0 = excellent,
0.8-0.9 = very good,
0.7-0.8 = good,
0.6-0.7 = sufficient,
0.5-0.6 = bad,
<0.5 = not useful)을 기준으로 잡았다

```
Call:
roc.formula(formula = data3$satisfaction ~ fitted(fit.final),     data = data3)

Data: fitted(fit.final) in 11058 controls (data3$satisfaction 0) < 8861 cases (data3$satisfaction 1).
Area under the curve: 0.963
```

```
install.packages("pROC")
library("pROC")
roc(data3$satisfaction~fitted(fit.final),data=data3)
plot.roc(rocplot,legacy.axes=TRUE) # ROC 그림
auc(rocplot) #ROC그림의 넓이(데이터 설명정도)
```



AUROC=0.963으로 이 모형이 데이터를 매우 잘 설명해주고 있다는 것을 알 수 있었다.

[Conclusion]

비행기 이용에서 고객의 만족도에 가장 크게 영향을 미치는 것은 고객의 유형과 좌석 클래스의 차이였다. 하지만 고객의 유형(royal/disroyal)에 따라 좌석의 클래스 선택에는 차이가 있을 것이고, 좌석의 클래스에 따라 제공되는 서비스의 정도도 달라야하니 그 두 요인에 대한 만족도는 당연히 차이가 날 수 밖에 없다. 따라서 이 부분은 회사 측에서 조절할 수 없는 문제이다. 그러므로 그 두 요인을 제외하고 큰 영향을 미치는 온라인 예약 시스템을 관리하고 기내의 와이파이 서비스를 제공하는 것에 집중하는 것이 고객의 만족도를 높이는 중요한 요인이라고 볼 수 있다.

우리나라뿐만 아니라 전 세계적으로 많은 서비스 산업이 중지되고 축소되었다. 특히 그 중의 대표 격으로 큰 타격을 받아 위축되어있는 항공 산업이지만, 조금씩 거리를 좁혀가고 있는 정책에 발맞춰 고객의 만족도를 높이는 서비스를 제공하여 항공 산업 부흥에 조금이라도 도움이 되었으면 좋겠다.

[Reference]

http://www.ezyeconomy.com/news/articleView.html?idxno=113695
- '위드 코로나' 기대감… 해외여행 '기지개'

http://digitalchosun.dizzo.com/site/data/html_dir/2021/11/12/2021111280075.html
- 코로나19 이후 1년 7개월만에 해외여행 되살아나

https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction?select=test.csv
- 분석에 사용된 데이터세트