2024 - 1

다변량분석

# Logistic Regression &

# Dimensionality Reduction


담당 교수: 강필성

강의명: 다변량분석

제출자: 정동은

전공: 경영학과

학번: 2020120120


제출 날짜: 2024-04-22

**[Q1] 본인이 스스로 Logistic Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 선정 이유를 설 명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결 정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오**

### 1.1. Data name

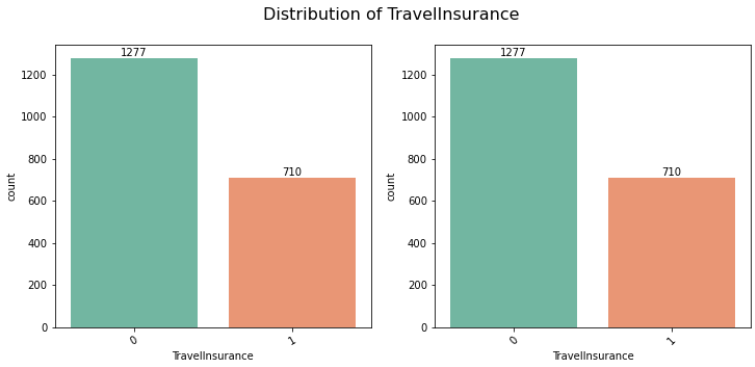
- Travel Insurance Prediction Data

### 1.2. Data source

- https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data

### 1.3. Reason Choosing this data

해당 데이터셋은 인도 내 2019년 당시, 여행자 보험 가입을 제시한 사람들 중 실제로 가입한 사람들을 담아놓았다. 총 1987 개의 개체와 binary와 종속 변수를 포함하여 총 9개의 변수들로 구성되어 있다. 고객들의 특성을 담아 놓은 변수와 그에 따른 여행자 보험 가입 여부를 담아 놓은 데이터셋은 분류 모델 그 중에서도 로지스틱 회귀 분석과 어울린다고 판단하여 선정하게 되었다.




또한 (종속변수)여행자 보험 가입 여부는 1277:710으로, 보험에 가입하지 않은 승객이 가입한 승객보다 많아 약간의 class Imbalance를 확인할 수 있지만, 9:1 정도의 심각한 불균형이 아니기에 분석에 있어 문제가 되는 정도는 아니라고 판단하였다.

**[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 세 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.**

## 2.1. 변수설정

**종속변수**

1. (명목형)TravelInsurance: 여행자보험 가입 여부

**설명변수**

1. (연속형) Age:승객 나이
2. (명목형) Employment Type: 승객이 종사하는 분야, 정부 or 개인
3. (명목형) GraduateOrNot: 승객의 대학 졸업
4. (연속형) AnnualIncome: 승객의 연 소득
5. (연속형) FamilyMembers: 승객의 가족 구성원 수
6. (명목형) ChronicDisease: 승객의 만성 질병 여부
7. (명목형) FrequentFlyer: 승객의 비행 횟수에 따른 분류
8. (명목형) EverTravelledAbroad: 승객의 해외여행 경험 여부

## 2.2. 설명변수들 중 높은 상관관계가 있을 것으로 예상되는 변수

| | 설명변수1 | 설명변수2 |
|---|---|---|
| 양의 상관관계 | GraduateOrNot | AnnualIncome |
| | AnnualIncome | FrequentFlyer |
| | Age | AnnualIncome |
| | FrequentFlyer | EverTravelledAbroad |
| 음의 상관관계 | ChronicDisease | EverTravelledAbroad |
| | ChronicDisease | FrequentFlyer |

1. **GraduateOrNot – AnnualIncome**

   해당 데이터는 인도에서 2019년에 수집된 데이터로, 인도의 2018년 대학진학률

통계는 26.3%이다. 그리고 2018년 당시 OECD 국가의 평균 대학 진학률은 44%로 상대적으로 인도는 낮은 수치이다.

이는 대학 졸업의 유무가 추후 사회의 위치를 결정짓는데 유의미한 영향을 미칠 것으로 보이며 이에 따라 연 소득 또한 유의미하게 양의 영향을 미칠 것으로 보인다. 또한 대학 졸업을 한다는 것은 상대적으로 자본의 여유가 있는 집안이 더 많을 것으로 보이기에, 이미 높은 소득을 올릴 기반이 마련되었다고 볼 수 있을 것 같다.

**2. AnnualIncome – FrequentFlyer**

비행기의 경우, 차 및 버스에 비해 이동수단으로써 높은 가격에 형성된다. 그렇기에 일정수준 이상의 소득이 뒷받침되어야만 이동수단으로써 향유할 수 있다고 생각한다. 그 이유는 해당 데이터셋에서 연소득의 min-max는 30만루피(500만원) ~ 180만루피(3000만원)이며, 인도 내 비행기 편도 가격은 평균적으로 약 30만원임을 비추어 볼 때, 비행기의 이용은 min의 경우 0.75*월급, 그리고 max의 경우 약 0.12*월급을 사용하여야 한다. 그렇기에 연소득이 높을수록 상대적으로 비행기를 더 쉽게 이용할 수 있기에 FrequentFlyer일 가능성이 높을 것으로 보인다

**3. Age – AnnualIncome**

대학 이전의 나이대에서는 연소득이 거의 존재하지 않을 것이고, 이후의 나이대에서는 시간이 흐를수록, 일반적으로 연봉은 증가한다. 그렇기에 나이의 증가는 AnnualIncome의 증가로 이어진다.

**4. FrequentFlyer – EverTravelledAbroad**

해외여행 경험 여부가 True라면, 비행횟수가 최소 1번 이상일 것이다. 해외경험 여부가 얼마나 자주 비행기를 탔는지와 직접적으로 연관되어 있지는 않겠지만, 해외여행 경험 여부가 없는 경우, FrequentFlyer가 아닌 특정 고객들과 상관관계가 있을 것으로 보인다.

**5. ChronicDisease – EverTravelledAborad, FrequentFlyer**

만성 질환이 있는 사람들은 증상 발현에 대한 조치가 필요하다. 그렇기에 비행기를 이용할 경우에 두 가짐 위험사항이 존재한다. 비행기 내에서 증상이 발현되는 경우와 이동한 지역에서 증상이 발현되는 경우이다. 위의 경우에는 긴급 조치가 상대적으로 어려울 수 있기에 만성질환을 가지고 있는 사람들은 해외여행을 기피 그리고 비행기 이용을 자제할 것으로 보인다.

## 2.3. 종속변수 예측에 필요하지 않을 것으로 예상되는 변수

### 1. Employment Type: 승객이 종사하는 분야, 정부 or 개인

여행자 보험은 국내나 해외여행 중 발생 가능한 상해, 질병, 배상책임손해을 보장하는 상품이다. 그리고 해당 가격대는 1만 내외이다. 여행자 보험 가입 여부는 종사하고 있는 분야보다는, 승객의 개인적 위험에 대한 인식에 의해 결정되는 경향이 크다고 판단되기 때문이다.

물론, 정부에 종사하고 있는 사람들이 리스크를 감수하려 하지 않은 경향이 있다고 생각할 수 있지만, Private Sector/Self-employed도 여행 중 발생할 수 있는 불가피한 사고나 질병에 대비하고 싶어할 수 있다. 그렇기에 종사 분야와는 별개로 개인적인 위험에 대한 인식이 여행자 보험 가입에 큰 영향을 미치는 것으로 보여, 종사하는 분야보다는 승객 개인의 필요와 선호에 주목하는 것이 더 중요할 것으로 판단된다

### 2. GraduateOrNot: 승객의 대학 졸업

보험 가입 여부는 앞서 말한 것처럼 개인의 여행 관련 리스크에 더 많은 영향을 받을 것으로 보인다. 그리고 고객의 대학 졸업 유무는 개인의 리스크 성향에 유의미한 영향을 미칠 것으로 보이지 않는다. 그렇기에 대학 졸업 유무와 보험 가입 여부 사이에는 상관관계가 없다고 판단하였다.

**[Q3] 모든 연속형 숫자 형태를 갖는(명목형 변수 제외) 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?**

### 3.1. 연속형 변수 단변량 통계량

| | 평균 | 표준편차 | 첨도 | 왜도 | Shapiro 통계량 | P-Value |
|---|---|---|---|---|---|---|
| Age | 29.65 | 2.91 | -1.10 | 0.24 | 0.93 | 0.00 |
| Annual Income | 932762.96 | 376855.68 | 1.01 | 0.08 | 0.97 | 0.00 |
| Faimily Members | 4.75 | 1.61 | -0.09 | 0.56 | 0.94 | 0.00 |

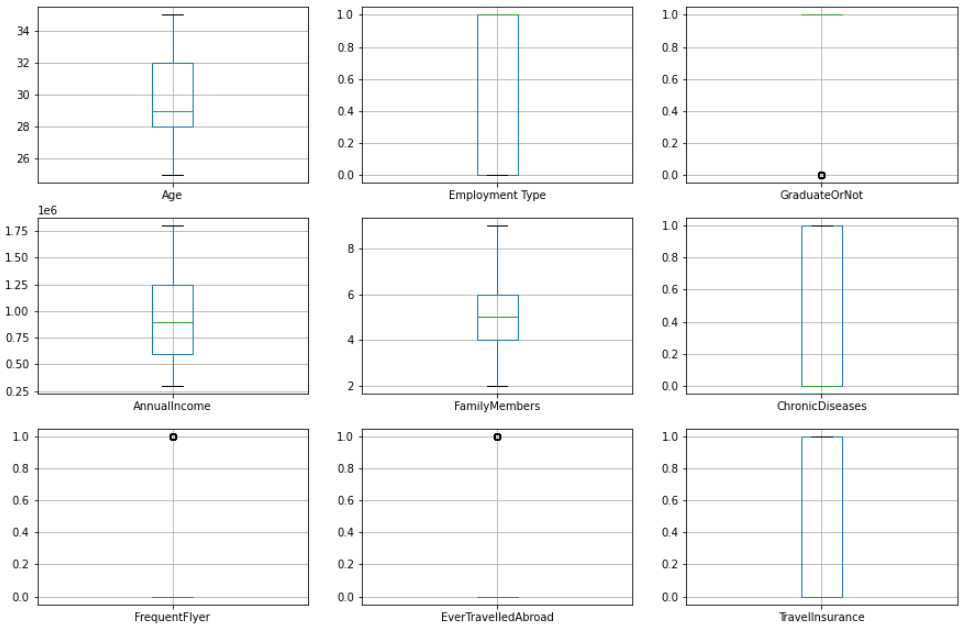
정규분포 확인의 경우 Shapiro 검정을 이용한다.

H0: 데이터가 정규분포를 따른다

H1: 데이터가 정규분포를 따르지 않는다

검정 결과, 3가지 변수의 P-Value 모두 유의수준(0.05)보다 작기에 정규분포를 따르지 않는다는 결론을 도출하였다.

### 3.2. Box Plot



**[Q4] [Q3]의 Box plot을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.**

### 4.1. 이상치 제거 수식 및 Histogram 이용 Outlier 정의

Box Plot에서 벗어나는 경우는 이진 분류에 해당하는 'FrequentFlyer', 'EverTraveledAborad', 'GraduateOrNot'이다. 이를 일반적으로 사용되는 이상치 제거 수식(IQR의 1.5배를 초과하는 관측치를 이상치로 간주 및 제거)을 바탕으로 재확인해보았다. 결과는 다음과 같다.

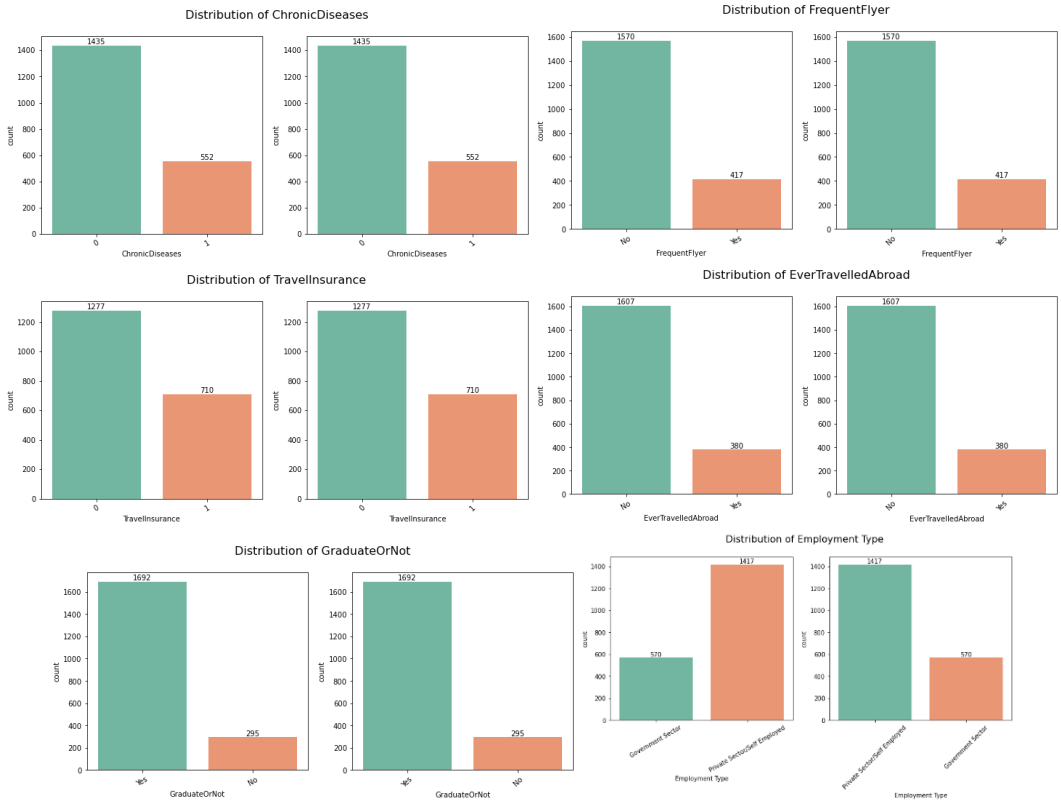
```
GraduateOrNot
Upper bound보다 큰 값의 개수:     0
Lower bound보다 작은 값의 개수:   295

FrequentFlyer
Upper bound보다 큰 값의 개수:     417
Lower bound보다 작은 값의 개수:   0

EverTravelledAbroad
Upper bound보다 큰 값의 개수:     380
Lower bound보다 작은 값의 개수:   0
```

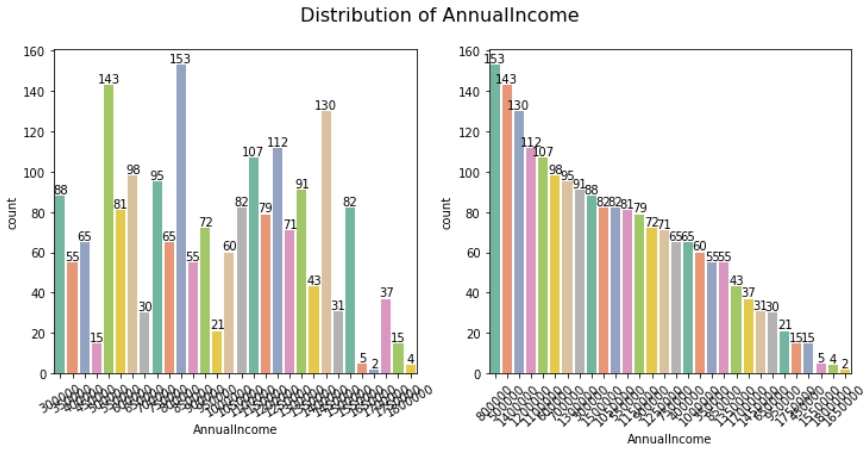
하지만 해당 변수들은 명목형 변수로 해당 변수의 값들의 이진적인 특성을 나타내기에 해당 변수들의 이상치들은 이상치로 정의할 만한 근거는 없다고 판단하였다.

또한 아래의 명목형 변수 분포와 같이, 확연한 클래스 불균형 문제가 식별되지 않았기에, 이상치를 정의할 변수가 없었다




## 4.2. 빈도수 이용 Outlier 정의

AnnualIncome의 분포의 경우, 인근 구간 대비 빈도수가 급락하는 구간이 있다. 정렬된 분포표를 보면 우측에 해당하는 3개 구간이 이에 해당한다. 인근 소득 구간 대비 급락 및 적은 수의 빈도를 지닌 1650000(2개),1800000(4개),1550000(5개)의 경우 유독 적은 수의 빈도를 지닌 9개의 행을 이상치로 정의하여 삭제하였다.




최종적으로 [1976 x 9]의 크기로 데이터 전처리를 마무리하였다.

다음 각 물음에 대해서는 [Q4]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

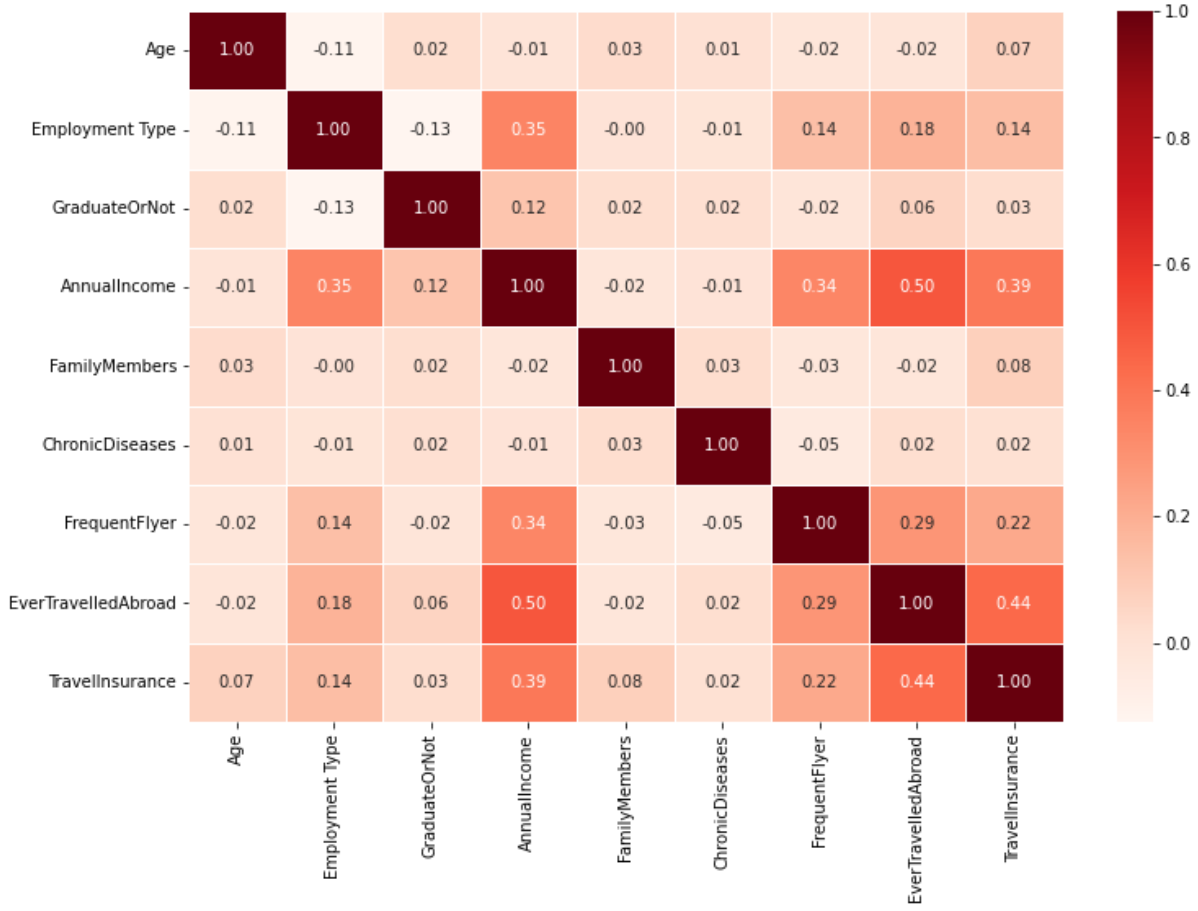
[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot)를 도시하고 적절한 정량적 지표를 사용하여 상관관계를 판단해 보시오.

1. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가?

2. 강한 상관관계가 존재하는 변수 조합들 중에 대표 변수를 하나씩만 선택해서 전체 변수의 개수를감소시켜 보시오 ([Q7]에서 사용함)

### 5.1. Scatter Plot

### 5.2. Correlation Plot




### 5.3. 결과 해석

전반적으로 데이터셋의 상관관계가 낮기 때문에, 데이터의 특성을 고려하여, 기준을 다음과 같이 정의하였다

| 상관계수 | 상관관계 |
|---|---|
| ±) 0.4 =< X | 높은 상관관계 |
| ±) 0.2 =< X < 0.4 | 중간 상관관계 |
| ±) 0.1 =< X < 0.2 | 약한 상관관계 |
| ±) X < 0.1 | 상관관계 없음 |

| 상관관계 | 부호 | 설명변수1 | 설명변수2 |
|---|---|---|---|
| 높은 상관관계 | (+) | EverTravelledAbroad | AnnualIncome |
| | | TravelInsurance | EverTravelledAbroad |
| | (-) | - | |

| | | | |
|---|---|---|---|
| 중간 상관관계 | (+_ | AnnualIncome | Employment Type |
| | | AnnualIncome | FrequentFlyer |
| | | EverTravelledAbroad | FrequentFlyer |
| | | TravelInsurance | AnnualIncome |
| | | TravelInsurance | FrequentFlyer |
| | (-) | - | |
| 약한 상관관계 | (+) | AnnualIncome | GraduateOrNot |
| | | FrequentFlyer | Employment Type |
| | (-) | Employment Type | Age |
| | | Employment Type | GraduateOrNot |

### 5.3.1. 강한 상관관계 해석

- **EverTravelledAbroad - AnnualIncome**

EverTravelledAbroad는 해외여행 경험 여부로 여행 경험이 있으면 1, 경험이 없으면 0의 값을 가진다. AnnualIncome은 승객의 연소득액을 보여주는 변수이다.

이는 타당해 보인다. 해외여행의 경우, 이동 경비 및 체류비 등의 비용이 소모된다. 이를 위해서는 소득이 보장되어야 한다. 그렇지 않으면 해외여행을 경험하기는 어렵다. 따라서 두 변수 간 높은 상관관계를 보이는 것은 타당해 보인다.

- **TravelInsurance – EverTravelledAbroad**

TravelInsurance는 보험 가입 여부로 보험에 가입하였으면 1, 가입하지 않았으면 0의 값을 가진다 EverTravelledAbroad는 해외여행 경험 여부로 여행 경험이 있으면 1, 경험이 없으면 0의 값을 가진다.

예상치 못한 상관관계이지만 그 의미를 해석해보자면, 해외 여행을 경험해본 사람들이 여행 중의 리스크(질병, 상해, 배상책임)를 줄이려는 것으로 보인다. 하지만 국내의 사례 조사 시, 해외 여행 시 여행자 보험의 가입 비율은 76%로 타 국가의 사례에서 비춰 볼 수 있듯 이 상관관계는 데이터셋의 잘못된 선정으로 비롯된 것이 아닌, 실제 상관관계가 있음으로 받아들일 수 있다.

### 5.3.2. 강한 상관관계 변수 제거

TravelInsurance는 종속변수이기에, 다음과 같은 조합이 가능하다.

1. 독립변수에서 EverTravelledAbroad 제거

2. 독립변수에서 AnnualIncome 제거

조합 1, 2를 바탕으로 Q7을 수행하도록 하겠다.

**[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 Logistic Regression 모델을 학습해 보시오. 이 때 70:30으로 구분하는 random seed를 저장하시오.**

### 6.1. 유효 변수 결과 해석

sm.Logit() 함수의 method='newton'으로 진행하였다.

유효 변수의 수는 총 5개로 결과는 다음과 같다.

```
Optimization terminated successfully.
         Current function value: 0.565893
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:        TravelInsurance   No. Observations:                 1383
Model:                          Logit   Df Residuals:                     1375
Method:                           MLE   Df Model:                            7
Date:                Sat, 20 Apr 2024   Pseudo R-squ.:                 0.1312
Time:                        17:58:18   Log-Likelihood:                -782.63
converged:                       True   LL-Null:                       -900.83
Covariance Type:            nonrobust   LLR p-value:                 2.161e-47
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1911      0.062      3.063      0.002       0.069       0.313
x2             0.0053      0.066      0.080      0.936      -0.124       0.134
x3            -0.0630      0.062     -1.011      0.312      -0.185       0.059
x4             0.5316      0.076      6.970      0.000       0.382       0.681
x5             0.2761      0.062      4.480      0.000       0.155       0.397
x6             0.0218      0.062      0.352      0.725      -0.100       0.143
x7             0.1650      0.067      2.456      0.014       0.033       0.297
x8             0.8037      0.084      9.582      0.000       0.639       0.968
==============================================================================
```

| 독립변수명(P-Value) | 유효성 판단 |
|---|---|
| X1: Age(0.002) | 보험은 리스크에 대한 경향성과 관련이 있다. 나이가 들수록 본인의 리스크를 감소시키려는 경향이 있다. 이에 비추어 보았을 때 나이가 많을수록, 여행자 보험을 가입하려는 것은 상식적으로 타당해 보인다. |
| X4: AnnualIncome(0.000) | 소득이 높을수록 리스크를 덜 지려고 한다고 단정지을 수 없다. 하지만 소득이 높을 수록, 소득 대비 리스크 관 |

| | |
|---|---|
| | 리를 위한 투자 비용은 낮아진다. 그렇기에 리스크 감수를 위한 상대적 비용이 감소하므로 더 쉽게 보험을 구매하게 된다고 판단할 수 있다. |
| X5: FamilyMembers(0.000) | 가족 구성원의 수는 리스크 성향과 어느정도 밀접한 관계가 있다. 가족 구성원 수가 많을수록, 자신이 부담해야 되는 리스크 또한 증가한다. 그렇기에 여행자 보험 가입을 통해 해당 리스크를 감소시키고자 한다. 그렇기에 해당 결과는 타당하다고 볼 수 있다. |
| X7: FrequentFlyer(0.014) | 비행기를 자주 탈수록 여행자 보험 가입을 할 가능성이 높다. 이 또한 타당해 보인다. 여행을 한 경험이 많을 수록, 여행 중에 발생할 수 있는 질병/상해 등을 목격했을 가능성이 높다. 그리고 이에 대한 조치가 어려움을 목격했을 가능성이 높다. 그렇기에 이에 대비하고자 여행자 보험을 가입하고자 하게 된다고 볼 수 있다. 그렇기에 해당 결과는 상식 선에서 타당하다고 볼 수 있다. |
| X8: EverTraelledAborad(0.000) | 앞서 말한 것처럼 예상치 못한 관계이다. 하지만 국내의 사례 조사 시, 해외 여행 시 여행자 보험의 가입 비율은 76%로 타 국가의 사례에서 비춰 볼 수 있듯 이 상관관계는 데이터셋의 잘못된 선정으로 비롯된 것이 아닌, 실제 상관관계가 있음으로 받아들일 수 있다. 그렇기에 상식 선에서 납득할 만한 이유는 형성하기 어렵지만, 국내(대한민국)의 사례 또한 유사한 결과가 나왔으므로, 관계를 받아들이도록 한다. |

### 6.2. 유효X 변수 확인

EmploymentType, GraduateOrNot, ChronicDisease가 유효하지 않은 변수로 선정함.

그리고 [Q2-2]의 정성적으로 선택했던 변수로는 EmploymentType, GraduateOrNot 이 있다. 당시 해당 변수들을 선택했던 이유로는 종사 분야 및 졸업 유무는 개인의 리스크 관리 성향과 관련이 없다고 판단 내렸기 때문이다. 그리고 결과 상으로, 이는 타당해 보인다.

그에 반해 ChronicDisease의 경우, 결과 상으로는 유의미하지 않은 변수로 선정되었다. 또한 ChronicDiseased의 class 비율은 3(없음): 1(있음)로 심각한 불균형이 있지는 않다. 그렇기에 결과적으로, 만성질환의 유무는 보험가입에 유의미한 영향을 미친다

고 보기 어렵다.

*제외해도 되는 변수인지는 아래에서 작성하겠다.

### 6.3. 결과 해석(Confusion Matrix , Accuracy, BCR, F1-Measure)

**Confusion Matrix(모든 변수 사용)**

| Train | Predicted True | Predicted False | Test | Predicted True | Predicted False |
|---|---|---|---|---|---|
| Actual True | 832 | 62 | Actual True | 355 | 28 |
| Actual False | 260 | 229 | Actual False | 122 | 88 |

| Measure | Train | Test |
|---|---|---|
| Simple Accuracy | 0.77 (0.01) | 0.76 (0.03) |
| Balanced Correction Rate | 0.70 (0.02) | 0.70 (0.03) |
| F1-Measure | 0.59 (0.03) | 0.58 (0.06) |

종속변수를 제외한 나머지 변수들을 독립변수로 넣고 모델을 학습시킨 결과이다. Simple Accuracy, Balanced Correction Rate, F1-Measure은 100회의 반복 실험을 바탕으로 결과를 도출하였다. 신뢰구간은 t분포를 사용한 유의수준 0.05(95%)를 바탕으로 결과를 도출하였다. Confusion Matrix는 마지막 실험(100번째)의 값을 바탕으로 작성하였다.

우선 해당 모델은 우수하다고 할 수는 없지만, 비교적 쓸만한 모델이라고 주장할 수 있다. 우선 모델의 성능이 TPR과 TNR이 적절한지를 확인할 수 있는 BCR의 성능과 Recall과 Precision을 종합적으로 고려한 F1-Measure의 성능에서 무작위 예측보다는 준수한 성능에, 그 성능이 훈련데이터와 테스트 데이터에서 일관되게 유지되고 있기 때문이다.

**Confusion Matrix(유효x 변수 제외)**

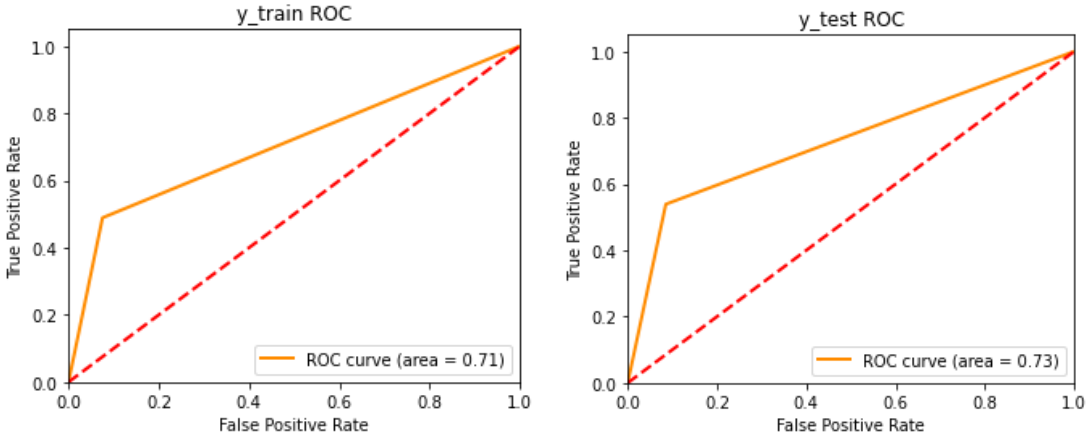
| Train | Predicted True | Predicted False | Test | Predicted True | Predicted False |
|---|---|---|---|---|---|
| Actual True | 833 | 61 | Actual True | 354 | 29 |
| Actual False | 259 | 230 | Actual False | 118 | 92 |

| Measure | Train | Test |
|---|---|---|
| Simple Accuracy | 0.77 (0.01) | 0.77 (0.03) |
| Balanced Correction Rate | 0.70 (0.02) | 0.70 (0.03) |
| F1-Measure | 0.59 (0.03) | 0.59 (0.06) |

'Employment Type', 'GraduateOrNot', 'ChronicDiseases' 변수를 제외하고서 학습하였다. Simple Accuracy와 F1-Measure 측면에서 약 0.01 정도의 수치 감소만 있었다. 유의미한 성능 감소가 보이지 않았기 때문에 해당 변수들은 유효하지 않다고 결론 내릴 수 있다.

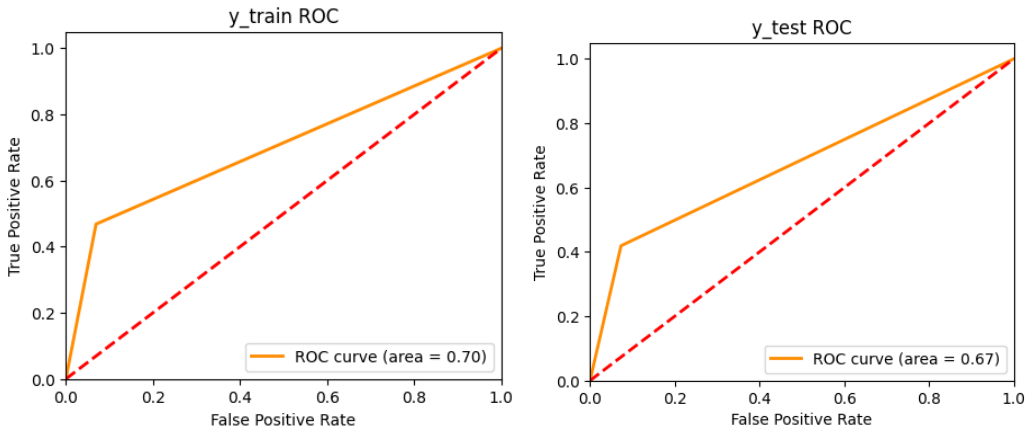
### 6.4. 결과 해석(AUROC)

**ROC Curve(모든 변수 사용)**



Train 데이터 AUROC = 0.71, test 데이터 AUROC = 0.73로 train에서의 성능이 평균적으로 좋다고 보이지만, 신뢰 구간을 고려하지 않았기에 아직 확신할 수는 없다.

**ROC Curve(유효x 변수 제외)**




변수가 감소하였기에 약간의 성능 감소는 있었지만 변수 수가 줄은 것에 비해, 성능 감소는 유의미하지 않았기에 0.05의 유의 수준을 바탕으로 변수의 유효성을 판단한 것은 적절한 선택이었음을 확인 가능하다.

**[Q7] [Q5]에서 변수 간 상관관계를 기준으로 선택한 변수들만을 사용하여 [Q6]에서 사용한 학습/테스트 70:30분할 데이터으로 Logistic Regression 모델을 학습해보시오.**

### 7.1. [Q5] 기반 변수 선택한 모델

"TravelInsurance는 종속변수이기에, 다음과 같은 조합이 가능하다.

1. 독립변수에서 EverTravelledAbroad 제거
2. 독립변수에서 AnnualIncome 제거

조합 1, 2를 바탕으로 Q7을 수행하도록 하겠다."

Q5 당시 위와 같은 결과가 나왔다. 어떤 변수를 제거할 지는 VIF를 기준으로 선정하겠

| | VIF Factor | Features |
|---|---|---|
| 0 | 1.386970 | ChronicDiseases |
| 1 | 1.469365 | FrequentFlyer |
| 2 | 1.847312 | EverTravelledAbroad |
| 3 | 2.039302 | TravelInsurance |
| 4 | 3.951996 | Employment Type |
| 5 | 6.910327 | GraduateOrNot |
| 6 | 9.289720 | FamilyMembers |
| 7 | 11.465874 | AnnualIncome |
| 8 | 18.886750 | Age |

다.

VIF가 10보다 클 때 다중공선성이 존재한다고 말할 수 있다. AnnualIncome의 VIF는 10보다 크기에 AnnualIncome 변수를 제거하고 EverTravelledAbroad 변수를 유지하겠다.

'Age', 'Employment Type', 'GraduateOrNot', 'FamilyMembers', 'ChronicDiseases', 'FrequentFlyer', 'EverTravelledAbroad'// 'TravelInsurance' 만을 사용하여 모델을 학습시키도록 하겠다. 결과는 다음과 같다.

```
Optimization terminated successfully.
         Current function value: 0.584038
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:        TravelInsurance   No. Observations:                 1383
Model:                          Logit   Df Residuals:                     1376
Method:                           MLE   Df Model:                            6
Date:                Sat, 20 Apr 2024   Pseudo R-squ.:                 0.1034
Time:                        19:51:12   Log-Likelihood:                -807.72
converged:                       True   LL-Null:                       -900.83
Covariance Type:            nonrobust   LLR p-value:                 1.622e-37
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1931      0.062      3.140      0.002       0.073       0.314
x2             0.1582      0.061      2.596      0.009       0.039       0.278
x3             0.0126      0.060      0.211      0.833      -0.105       0.130
x4             0.2539      0.060      4.223      0.000       0.136       0.372
x5             0.0002      0.061      0.004      0.997      -0.119       0.119
x6             0.2651      0.065      4.088      0.000       0.138       0.392
x7             0.9734      0.080     12.172      0.000       0.817       1.130
==============================================================================
```

| 독립변수명(Q6) | P-Value(Q6) | 독립변수명(Q7) | P-Value(Q7) |
|---|---|---|---|
| Employment Type | 0.936 | - | - |
| GraduateOrNot | 0.312 | GraduateOrNot | 0.833 |
| ChronicDiseases | 0.725 | ChronicDiseases | 0.997 |

Q6 에서는 3 개의 변수에서 0.05 보다 큰 P-Value 를 보였지만, Q7 에서는 Q6 에서 확인한 2 개의 변수에서만 0.05 보다 큰 P-Value 를 보였다. 다만 이전보다 P-Value 가 더욱 커진 점을 통해, AnnualIncome 이 제거됨으로써, 기존에 P-value 가 0.05 보다 컸던 2 개의 변수들이 더욱 의미가 없는 변수라는 점을 확인할 수 있다. AnnualIncome 제거 이전에는 AnnualIncome 과의 상관성 때문에 그나마 P-Value 가 작게 나온 것으로 추측된다.

추가로, 이런 변화와 앞서 AGE 에서 높은 VIF 값으로 비롯해 추가적인 다중공선성의 문제가 있을 수로 판단되어 AGE 변수마저 제거하고 결과를 보도록 하겠다.

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | TravelInsurance | No. Observations: | 1383 |
| Model: | Logit | Df Residuals: | 1377 |
| Method: | MLE | Df Model: | 5 |
| Date: | Sat, 20 Apr 2024 | Pseudo R-squ.: | 0.09785 |
| Time: | 20:00:52 | Log-Likelihood: | -812.68 |
| converged: | True | LL-Null: | -900.83 |
| Covariance Type: | nonrobust | LLR p-value: | 3.301e-36 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| x1 | 0.1371 | 0.060 | 2.273 | 0.023 | 0.019 | 0.255 |
| x2 | 0.0151 | 0.059 | 0.254 | 0.799 | -0.101 | 0.132 |
| x3 | 0.2623 | 0.060 | 4.375 | 0.000 | 0.145 | 0.380 |
| x4 | 0.0052 | 0.060 | 0.085 | 0.932 | -0.113 | 0.124 |
| x5 | 0.2536 | 0.064 | 3.943 | 0.000 | 0.128 | 0.380 |
| x6 | 0.9618 | 0.080 | 12.086 | 0.000 | 0.806 | 1.118 |

| 독립변수명(Q6) | P-Value(Q6) | 독립변수명(Q7 + AGE 제거) | P-Value(Q7) |
|---|---|---|---|
| Employment Type | 0.936 | - | - |
| GraduateOrNot | 0.312 | GraduateOrNot | 0.799 |
| ChronicDiseases | 0.725 | ChronicDiseases | 0.932 |

Annual Income만 제거한 경우와 비교하여 회귀 계수 및 p-value 측면에서 큰 차이가 발생하지는 않았다. 하지만 모델 내 변수들이 독립변수에 미치는 영향을 정확하게 측정하기 위해서는 VIF 값이 높게 발생한 AGE 변수는 제거하는 게 좋아 보인다.

### 7.2. 결과 해석(Confusion Matrix , Accuracy, BCR, F1-Measure)

따라서 앞서 높은 p-value 값을 가진 'GraduateOrNo't'와 'ChronicDiseases' 변수를 제거하고, 그리고 VIF 값이 높았던 'Age'와 'AnnualIncome' 변수를 제거하고 모델을 새로 생성해 결과를 비교해보도록 하겠다. 즉 독립변수로 'EmploymentType', 'FamilyMembers', 'FrequentFlyer', 'EverTravelAbroad'를 사용하고 종속변수인 'TravelInsurace'를 예측하도록 하는 모델을 만들도록 하겠다.

| Train | Predicted True | Predicted False | Test | Predicted True | Predicted False |
|---|---|---|---|---|---|
| Actual True | 832 | 62 | Actual True | 353 | 30 |
| Actual False | 257 | 232 | Actual False | 121 | 89 |

| Measure | Train | Test |
|---|---|---|
| Simple Accuracy | 0.77 (0.01) | 0.77 (0.02) |
| Balanced Correction Rate | 0.70 (0.01) | 0.70 (0.03) |
| F1-Measure | 0.59 (0.02) | 0.59 (0.05) |

Simple Accuracy, Balanced Correction Rate, F1-Measure은 100회의 반복 실험을 바탕으로 결과를 도출하였다. 신뢰구간은 t분포를 사용한 유의수준 0.05(95%)를 바탕으로 결과를 도출하였다. Confusion Matrix는 마지막 실험(100번째)의 값을 바탕으로 작성하였다.

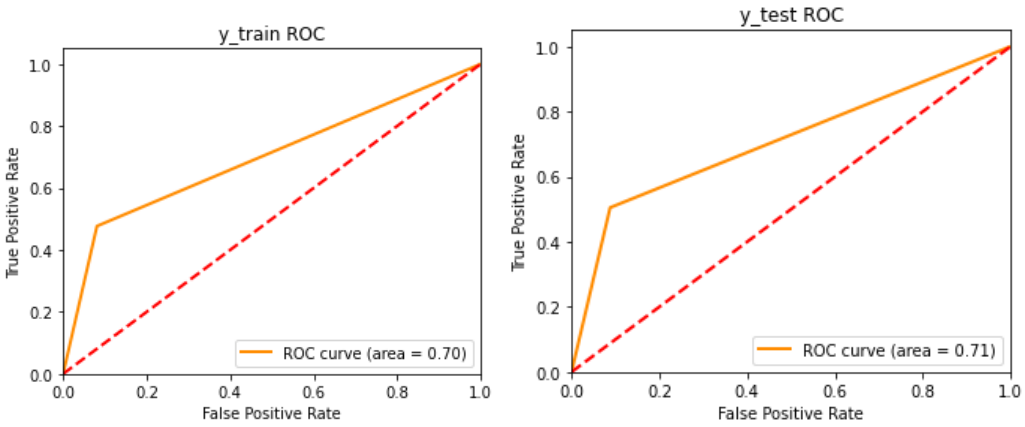
[Q6-3:모든 변수를 사용] 모델 대비, Simple Accuracy와 F-1 Measure에서 평균적으로 약 0.01의 성능 향상이 있었으며, 전반적으로 성능의 분산 또한 감소하게 되었다

[Q6-3:초기 유효X로 판단된 변수 제거] 모델 대비하여, 평균적인 성능은 같으나, 전반적으로 모델 성능의 분산이 감소함을 확인할 수 있다.

이를 통해, VIF가 높은 변수 및 유효X를 제거한 모델이 기존 대비 성능은 유사하나, 모델 성능의 분산이 Simple Accuracy와 F1-Measure 측면에서 감소하였으므로, 보다 일관된 모델이라고 판단할 수 있다. 비록 성능 평가 지표는 우수한 수준은 아니지만, 무작위로 예측하는 경우보다 나은 성능이며 해당 변수들을 수집하는 것은 어렵지 않으므로 해당 모

델은 적어도 보험 예측을 하는 데 있어 쓸만하다고 할 수 있다.

### 7.3. 결과 해석(AUROC)




새로 생선한 모델의 Train 데이터의 AUROC = 0.70, test 데이터의 AUROC = 0.71 로 Test 에서의 성능이 평균적으로 좋다고 보여진다, 이는 다중공선성의 문제를 제거하여 변수 별 정확한 영향력을 유추하였기 때문에 나온 결과로 보인다. 하지만 신뢰 구간을 고려하지 않았기에 확신할 수는 없다.

[Q6-4]의 결과에 비해 감소한 수치는 모든 변수를 사용한 모델 대비 다중공선성의 문제와 유효하지 않은 변수를 제외한 새로운 모델에서 test 데이터에서의 AUROC 가 0.01 감소한 점이다. 다만 수치 차이 또한 작기에 AUROC 의 결과값을 비교해보았을 때 Q6(모든 변수 사용), Q6(유효 X 변수 제외), Q7 에서의 변수선택에 따른 차이가 존재한다고 말할 수 없다.

**[Q8] [Q6]에서 생성한 학습 데이터를 이용하여 Logistic Regression 에 Forward Selection, Backward Elimination, Stepwise Selection 을 적용해보시오. 각 방법론마다 Training dataset 에 대한 AUROC 및 소요 시간, Validation dataset 에 대한 AUROC, Accuracy, BCR, F1-Measure 를 산출하시오.**

### 8.1. 변수 선택

변수 선택은 다음의 두 가지 조건을 만족하여야만 다음 step이 가능하다고 가정하였다.

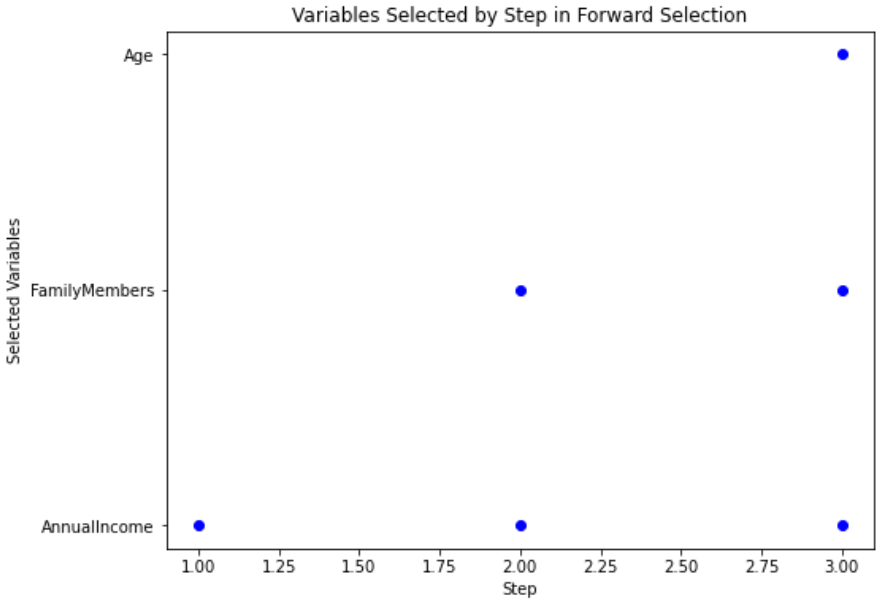
1. P-value < 0.05
2. (현재 step 의 AUROC - 이전 step 의 AUROC) > 0.05

두 가지 조건을 만족한 상태에서, 각 변수 선택 기법별로 선택된 변수는 다음과 같다.

| | Forward Selection | Backward Elimination | Stepwise Selection |
|---|---|---|---|
| 선택된 변수 (변수선택 순서) | AnnualIncome(1), FamilyMembers(2), Age(3) | Age(1), AnnualIncome(2), FamilyMembers(3), FrequentFlyer(4) EverTravelledAbroad(5) | Age(1), AnnualIncome(2), FamilyMembers(3), FrequentFlyer(4) EverTravelledAbroad(5) |

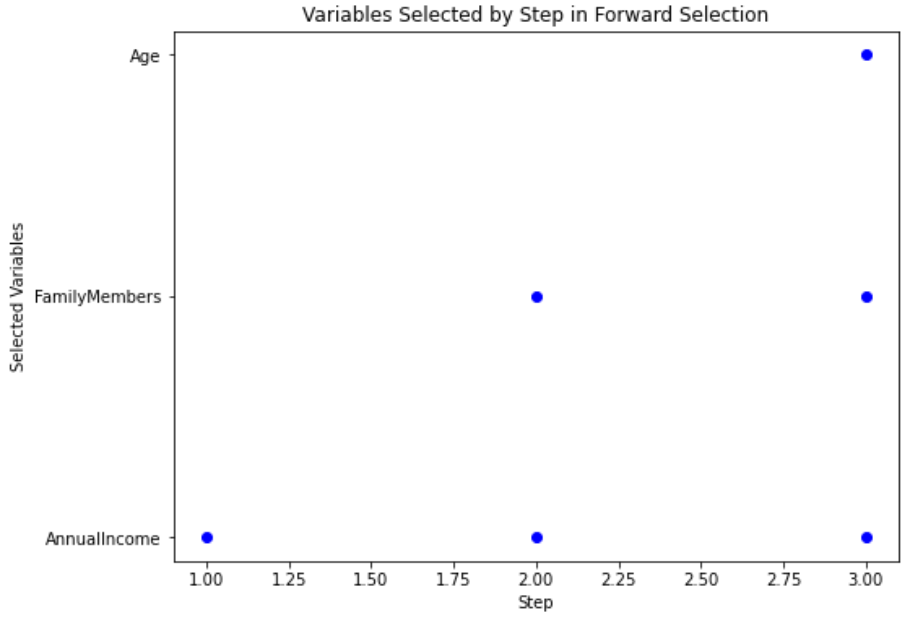
**Forward Selection** 에서는 다음과 같은 순서로 변수가 선택되었다.

1. AnnualIncome
2. AnnualIncome + FamiliyMembers
3. AnnualIncome + FamiliyMembers + Age
4. 변수 추가시 AUROC 를 0.05 이상 높이는 변수가 없기에 종료



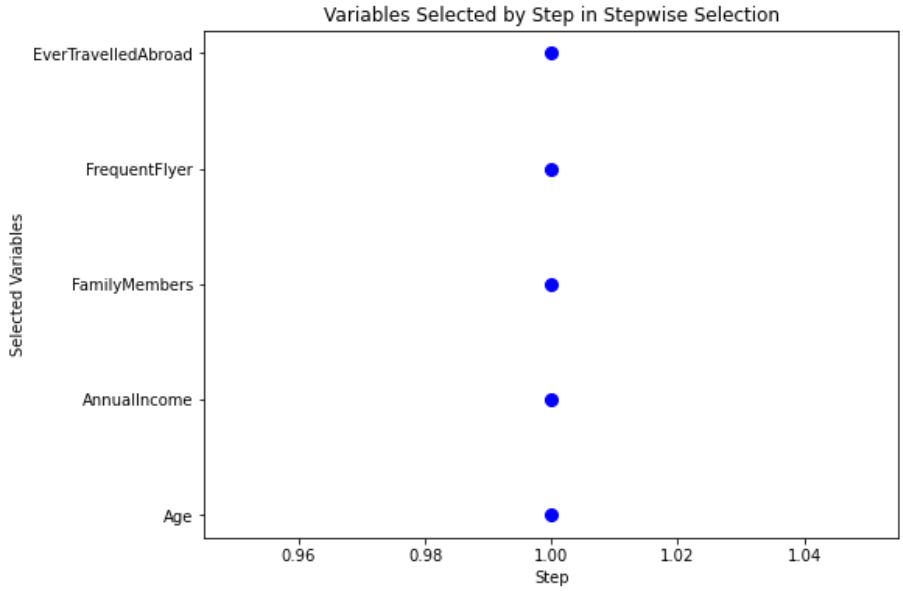

**Backward Elimination**에서는 다음과 같은 순서로 변수가 선택되었다.

1. Age + GraduateOrNot + AnnualIncome + FamilyMembers + ChronicDiseases+FrequentFlyer + EverTravelledAbroad
2. Age + GraduateOrNot + AnnualIncome + FamilyMembers + FrequentFlyer + EverTravelledAbroad
3. Age + AnnualIncome + FamilyMembers + FrequentFlyer + EverTravelledAbroad
4. 변수 제거시 AUROC 를 0.05 이상 높이는 변수가 없기에 종료



**Stepwise Selection** 에서는 다음과 같은 순서로 변수가 선택되었다.

1. Age + AnnualIncome + FamilyMembers + FrequentFlyer + EverTravelledAbroad
2. 변수 선택 및 제거시 AUROC 를 0.05 이상 높이는 변수가 없기에 종료




### 8.2. 소요 시간

| | Forward Selection | Backward Elimination | Stepwise Selection |
|---|---|---|---|
| Train time | 0.146<br>(0.000) | 0.040<br>(0.000) | 0.053<br>(0.000) |

위는 100 번 수행하여 도출한 각 기법 별 소요시간 평균 및 분산이다.

소요 예상시간: Stepwise Selection > Forward Selection = Backward Elimination

실제 소요시간: Forward Selection> Stepwise Selection > Backward Elimination

예상과 정반대로, Forward 와 Bacward 에서의 차이가 큰 결과가 도출되었으며 Forward selection 이 Stepwise Selection 보다 더 큰 시간이 소요되었다. 이는 변수가 적고, 데이터 크기가 작아서 발생한 문제라고 해석된다.

### 8.3. 결과 해석(Accuracy, BCR, F1-Measure)

세 가지 변수 선택에서의 Train, Test 에서의 결과는 다음과 같다.

| | Forward Selection | | Backward Elimination | | Stepwise Selection | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| Simple Accuracy | 0.75 (0.01) | 0.75 (0.03) | 0.77 (0.01) | 0.77 (0.03) | 0.77 (0.01) | 0.77 (0.03) |
| Balanced Correction Rate | 0.70 (0.02) | 0.70 (0.03) | 0.70 (0.02) | 0.70 (0.03) | 0.70 (0.02) | 0.70 (0.03) |
| F1-Measure | 0.60 (0.03) | 0.59 (0.05) | 0.59 (0.03) | 0.59 (0.06) | 0.59 (0.03) | 0.59 (0.06) |

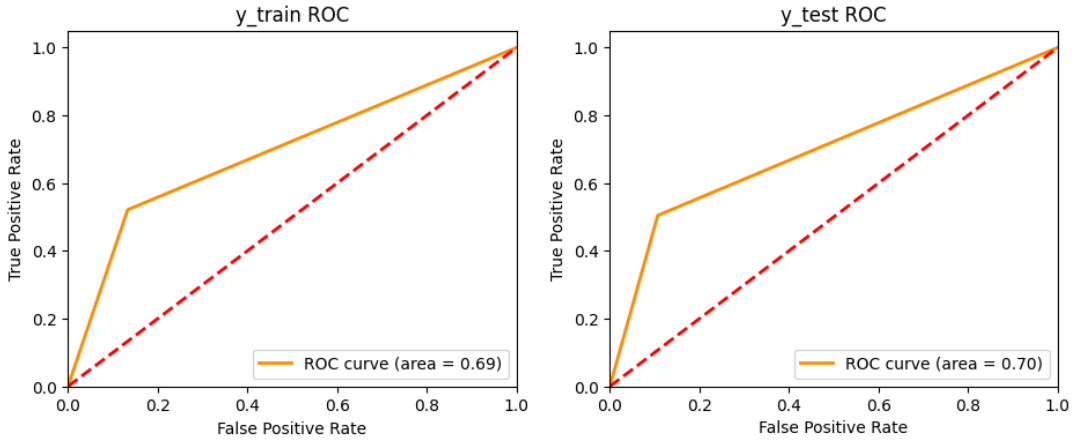
Forward Selection 은 [Q6:유효 X 변수 제거] 모델 대비 약간의 성능 감소가 있었으며, [Q7: 높은 VIF 및 유효 x 변수 제거] 대비 약간의 성능 감소가 있었다.

[Q6:모든 변수 사용], [Q7: 높은 VIF 및 유효 x 변수 제거]에서 학습에 사용한 독립변수는 5 개, 4 개인데 반해, Q8 에서의 Forward Selection 은 독립변수 3 개로 유사한 성능을 발휘하였고, 이를통해 변수 선택 기법은 적절히 수행되었음을 확인할 수 있다.

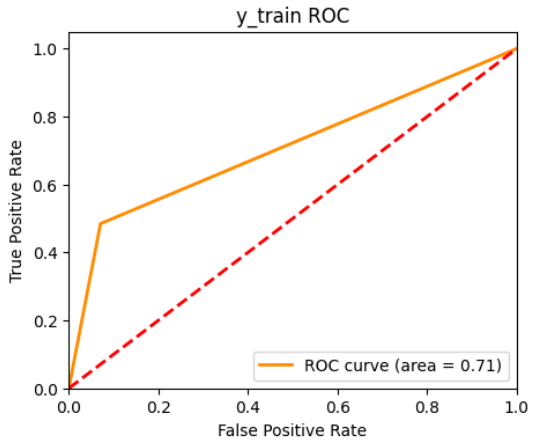
Backward & Stepwise 는 Q6 와 동일한 변수를 선택하였기에, 동일한 성능을 발휘하였다.

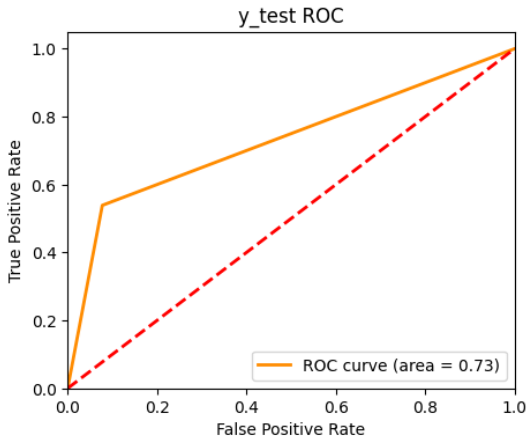
### 8.4. 결과 해석(AUROC)

**Forward Selection**

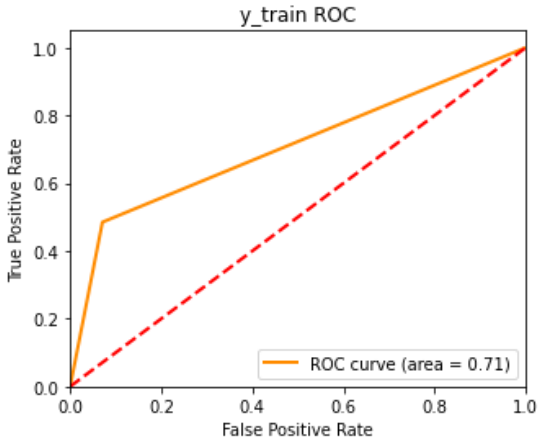


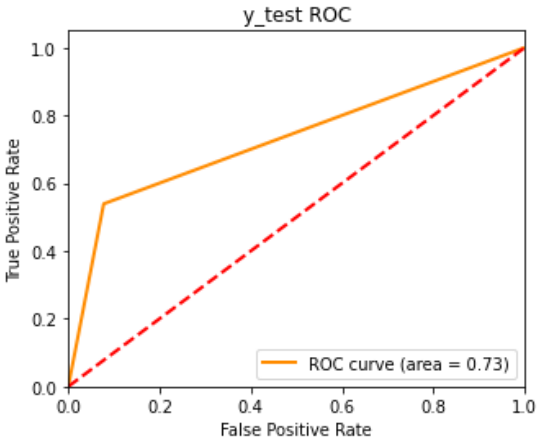

**Backward Elimination**




**Stepwise Seiection**




| | Train AUROC | Test AUROC |
|---|---|---|
| Forward Selection | 0.69 | 0.71 |
| Backward Elimination | 0.71 | 0.73 |
| Stepwise Selection | 0.71 | 0.73 |

Backward & Stepwise 에서의 AUROC 성능이 Forward 대비 Train 에서는 0.02, Test 에서는 0.03 더 우수하지만, 신뢰구간을 고려하지 않았기에 유의미한 차이라고 말할 수는 없다.

**[Q9] AUROC 를 Fitness function 으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 작성해 보시오. 작성한 함수를 이용하여 GA 기반 변수 선택을 수행하고, 선택된 변수를 사용한 Logistic Regression 의 Validation dataset 에 대한 분류 성능(AUROC, Accuracy, BCR, F1-Measure), 변수 감소율, 수행 시간의 세 가지 관점에서 Forward Selection, Backward Elimination, Stepwise Selection 방식을 사용한 Logistic Regression 과 비교해보시오.**

### 9.1. GA 기반의 변수 선택

GA 로 선택된 변수는 다음과 같다.

'Age', 'Employment Type', 'GraduateOrNot', 'AnnualIncome', 'ChronicDiseases', 'EverTravelledAbroad'로 6 가지의 변수 중 3 가지 변수 모두 Q6 에서 P-value 가 0.05 보다 높아 사용하지 않았던 변수들로 AUROC 만을 고려하였기에 이와 같은 결과가 나온 것으로 보인다.

### 9.2. 결과 해석(AUROC, Accuracy, BCR, F1-Measure)

| | Forward Selection | Backward Elimination | Stepwise Selection | Genetic Algorithm |
|---|---|---|---|---|
| AUROC | 0.70 | 0.73 | 0.73 | 0.69 |
| Simple Accuracy | 0.75 (0.03) | 0.77 (0.03) | 0.77 (0.03) | 0.75 (0.02) |
| Balanced Correction Rate | 0.70 (0.03) | 0.70 (0.03) | 0.70 (0.03) | 0.68 (0.03) |
| F1-Measure | 0.59 (0.05) | 0.59 (0.06) | 0.59 (0.06) | 0.55 (0.05) |

일반적으로 메타 휴리스틱 기법은 기존의 변수 선택 기법과 비교하여, 시간은 많이 소모되지만 성능은 더 좋다고 받아들여진다. 그러나 본 모델에서는 예상과 다르게 메타 휴리스틱 기법의 하나인, 유전 알고리즘의 성능이 더 좋지 않았다. 성능이 좋지 않은 이유를 파악하고자 다음과 같은 두 가지 가설을 설정하였다.

1. **적절한 하이퍼파라미터를 설정하지 않아서, 성능이 좋지 않았다.**

**2. p-value 및 변수간의 관계를 고려하지 않아서, 성능이 좋지 않았다.**

1 번은 Q10 에서 다양한 하이퍼파라미터를 비교해보며 확인해볼 예정이며, 만약 1 번을 맞다고할 수 없다면, 잠재적으로 2 번에 따른 결과라고 결론을 내릴 것이다.

### 9.3. 결과 해석(변수 감소율)

| | Forward Selection | Backward Elimination | Stepwise Selection | Genetic Algorithm |
|---|---|---|---|---|
| #of variables | 3 | 5 | 5 | 6 |

기존 8 가지 변수 중 Forward Selection 은 3 가지 변수가 선택되어, 267%의 변수 감소율을 보였다. Backward Elimination 과 Stepwise Selection 은 160%의 감소율, 그리고 GA 에서는 6 가지 변수가 선택되어, 133%의 변수 감소율을 보였다.

일반적으로 Stepwise Selection 은 Forward, Backward 대비 더 많은 변수를 선택할 가능성이 있고, GA 는 Stepwise Selection 보다 더욱 학습에 소요되는 시간과 성능간의 Trade-off 가 존재한다고 알려져있다. 따라서 변수의 수만 놓고 보면 예상에 맞는 결과를 도출하였다. 다만, Backward Elimination 은 변수 수가 많은 것은 예상과 달랐는데 이는 전체 변수 수와 데이터 수가 작아 기인한 문제라고 생각된다.

### 9.4. 결과 해석(수행 시간)

| | Forward Selection | Backward Elimination | Stepwise Selection | Genetic Algorithm |
|---|---|---|---|---|
| Train time | 0.146<br>(0.000) | 0.040<br>(0.000) | 0.053<br>(0.000) | 7.241<br>(8.505) |

Forward Selection, Backward Elimination, Stepwise Selection 은 앞서 살펴본 것처럼, 예상과 약간은 다른 결과를 보이지만, 너무 적은 수행 시간이 소요되었기에 차이가 존재한다고 말하기도 힘들 것으로 보인다. 이들과 비교하여 유전 알고리즘은 학습에 약 100 배 이상의 시간이 더 소요되었다. 유전 알고리즘의 하이퍼 파라미터 설정에 따라 다르겠지만, 기본적으로 유전 알고리즘은 기존의 변수 선택 기법보다 확연히 긴 수행 시간을 필요로 함을 알 수 있다.

**[Q10] Genetic Algorithm 에서 변경 가능한 하이퍼파라미터들(population size, Cross-over rate, Mutation rate 등) 중 세 가지를 선택하고 각각의 하이퍼파라미터마다 최소 세 가지 이상의 후보 값들을 선정(최소 27 가지 이상의 조합)하여 각 조합에 대한 변수 선택 결과에 대해 본인만의 생각을 더해 해석해보시오.**

**10.1. 하이퍼파라미터 조합**

| Index | Population | Generations | Crossover | Mutation | Selection |
|---|---|---|---|---|---|
| Baseline | 50 | 20 | 0.8 | 0.1 | roulette |
| 1 | 100 | 20 | 0.8 | 0.1 | roulette |
| 2 | 50 | 50 | 0.8 | 0.1 | roulette |
| 3 | 50 | 20 | 0.5 | 0.1 | roulette |
| 4 | 50 | 20 | 0.8 | 0.05 | roulette |
| 5 | 50 | 20 | 0.8 | 0.1 | Deterministic |
| 6 | 50 | 20 | 0.8 | 0.1 | tournament |
| 7 | 100 | 50 | 0.8 | 0.1 | roulette |
| 8 | 100 | 20 | 0.5 | 0.1 | roulette |
| 9 | 100 | 20 | 0.8 | 0.05 | roulette |
| 10 | 50 | 50 | 0.5 | 0.1 | roulette |
| 11 | 50 | 50 | 0.8 | 0.05 | roulette |
| 12 | 50 | 20 | 0.5 | 0.05 | roulette |
| 13 | 100 | 50 | 0.5 | 0.1 | roulette |
| 14 | 100 | 50 | 0.8 | 0.05 | roulette |
| 15 | 100 | 50 | 0.5 | 0.05 | roulette |
| 16 | 150 | 20 | 0.8 | 0.1 | roulette |
| 17 | 200 | 20 | 0.8 | 0.1 | roulette |
| 18 | 250 | 20 | 0.8 | 0.1 | roulette |
| 19 | 300 | 20 | 0.8 | 0.1 | roulette |
| 20 | 350 | 20 | 0.8 | 0.1 | roulette |
| 21 | 400 | 20 | 0.8 | 0.1 | roulette |
| 22 | 50 | 100 | 0.8 | 0.1 | roulette |

| | | | | | |
|---|---|---|---|---|---|
| 23 | 50 | 150 | 0.8 | 0.1 | roulette |
| 24 | 50 | 200 | 0.8 | 0.1 | roulette |
| 25 | 50 | 250 | 0.8 | 0.1 | roulette |
| 26 | 100 | 50 | 0.8 | 0.15 | roulette |
| 27 | 100 | 50 | 0.8 | 0.2 | roulette |

### 10.2. 결과 해석 (1 ~ 6 번)

1 ~ 6 번에서는 Baseline 대비, 각 하나의 하이퍼파라미터를 변경했을때의 효과를 확인해본다.

| | #of variables | Time | AUROC | BCR | F1-Measure |
|---|---|---|---|---|---|
| Baseline | 5 | 9.536 | 0.66 | 0.68 (0.03) | 0.55 (0.05) |
| 1 | 6 | 14.285 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 2 | 5 | 21.063 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 3 | 6 | 23.052 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 4 | 7 | 17.562 | 0.73 | 0.69 (0.04) | 0.58 (0.06) |
| 5 | 5 | 15.802 | 0.73 | 0.69 (0.03) | 0.57 (0.06) |
| 6 | 5 | 16.416 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |

가정: 유전 알고리즘은 실행할때마다 값이 바뀌기에 이번 실행이 정답이라고 말할 수 없다. 그럼에도 비교를 위해 해당 결과가 다음 번 실행에도 동일한 결과로 이어질 것이라는 가정하에 비교를 진행한다.

Selection 에서 사용한 기법은 Deterministic, roulette, tournament 이다. 각 기법에서의 성능은 표준편차 내에서 모두 동일하다. 따라서 가장 적은 소요시간을 요구로 하였던 roulette 이 상대적으로 좋다고 가정한다.

Population, Generations, Crossover, Mutation 의 변화에 따른 성능 차이는 표준편차 내에서 차이가 없다.

### 10.3. 결과 해석 (7 ~ 15 번)

7 ~ 15 번에서는 Baseline 대비, 2 개, 3 개, 4 개의 하이퍼파라미터를 변경했을때의 효과를 확인해본다.

| | #of variables | Time | AUROC | BCR | F1-Measure |
|---|---|---|---|---|---|
| Baseline | 5 | 9.536 | 0.66 | 0.68 (0.03) | 0.55 (0.05) |
| 7 | 7 | 29.239 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 8 | 5 | 16.889 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 9 | 6 | 20.142 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 10 | 5 | 25.647 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 11 | 5 | 14.265 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 12 | 6 | 17.854 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 14 | 5 | 21.376 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 15 | 7 | 22.131 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |

Baseline 대비 표준편차 내에서의 성능 차이를 확인할 수 없다. 큰 차이를 주지 않은 문제에서 기인하였다고 생각하여, 이후 변화를 크게 주도록 한다.

### 10.4. 결과 해석(16 ~ 27 번)

16 ~ 27 번에서는 Baseline 대비, 하이퍼파라미터의 변화를 단조적으로 크게 가져가보며 각 하이퍼파라미터의 효과를 확인해본다.

| | #of variables | Time | AUROC | BCR | F1-Measure |
|---|---|---|---|---|---|
| Baseline | 5 | 9.536 | 0.66 | 0.68 (0.03) | 0.55 (0.05) |
| 16 | 6 | 21.244 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 17 | 6 | 30.233 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 18 | 5 | 32.714 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 19 | 6 | 39.725 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 20 | 5 | 46.463 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 21 | 6 | 52.459 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 22 | 3 | 24.346 | 0.73 | 0.70 (0.03) | 0.59 (0.05) |
| 23 | 6 | 30.157 | 0.73 | 0.70 (0.03) | 0.59 (0.05) |
| 24 | 5 | 42.497 | 0.73 | 0.70 (0.03) | 0.59 (0.05) |
| 25 | 6 | 63.708 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 26 | 5 | 30.609 | 0.73 | 0.68 (0.03) | 0.55 (0.05) |
| 27 | 5 | 33.490 | 0.73 | 0.70 | 0.58 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | (0.04) | (0.06) |

**16 ~ 21 번**: Population 을 단조적으로 증가시키며 영향을 확인해보았다. 모델 자체가 작은 문제이겠지만, 시간이 증가한다는 것 빼고는 어떠한 성능 차이도 확인할 수 없었다.

**22 ~ 25 번**: Generations 을 단조적으로 증가시키며 영향을 확인해보았다. 특이하게도 이전과 표준편차 내에서 성능 차이는 없지만, 그럼에도 결과값이 도출되었다. 하지만 이는 Generations 의 영향이라기보다는 유전 알고리즘의 변동성에서 기인한 문제라고 생각된다.

**26~27 번**: Mutation 을 단조적으로 증가시키며 영향을 확인해보았다. Mutation 이 크면, Global 로 가기보다는 Local 조차도 찾지 못할 것이라고 생각하였다. 하지만 해당 데이터는 8 개의 변수만을 지니고 있기에 별다른 성능 차이를 확인하지 못했다.

정리하자면, 하이퍼 파라미터에서 어떠한 변화를 주더라도 Q7에서 나온 결과 이상의 값을 도출하지는 못하였다. 이것은 데이터셋의 한계로 보인다. 보다 구체적으로는 적은 데이터의 수와 조합하기에 부족한 적은 변수들에 의해 기인한 것으로 보인다. 여행자 보험 예측을 위해서는 새로운 변수의 추가에 대한 고민이 필요할 것으로 보인다.