HW Week7 Logistic Regression & Model Assessment

짱짱조

[1] Dataset

본 실습에서는 Kaggle의 있는 The Ultimate Halloween Candy Power Ranking 데이터를 이용하였다. (https://www.kaggle.com/fivethirtyeight/the-ultimate-halloween-candy-power-ranking)

1. Data

이름을 나타내는 열인 'competitorname '을 제외하고 Data의 각 변수들의 의미는 다음과 같다.

chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts, peanut butter or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice, wafers, or a cookie component?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bag or box?
sugarpercent	The percentile of sugar it falls under within the data set
pricepercent	The unit price percentile compared to the rest of the set
winpercent	The overall win percentage according to 269,000 matchups

위 변수는 크게 세 가지로 분류할 수 있다.

1) chocolate ~ crispedricewafer

- Taste 속성에 관하여 해당 속성이 있으면 1, 없으면 0으로 표시 되어있고(binary response), 한 제품이 여러 특성을 가지고 있을 수 있다. (범주형)

2) hard ~ bar

- Type 속성에 관하여 one-hot-encoding이 되어 있고, 한 제품당 한 type만을 나타낸다. (범주형)

pluribus

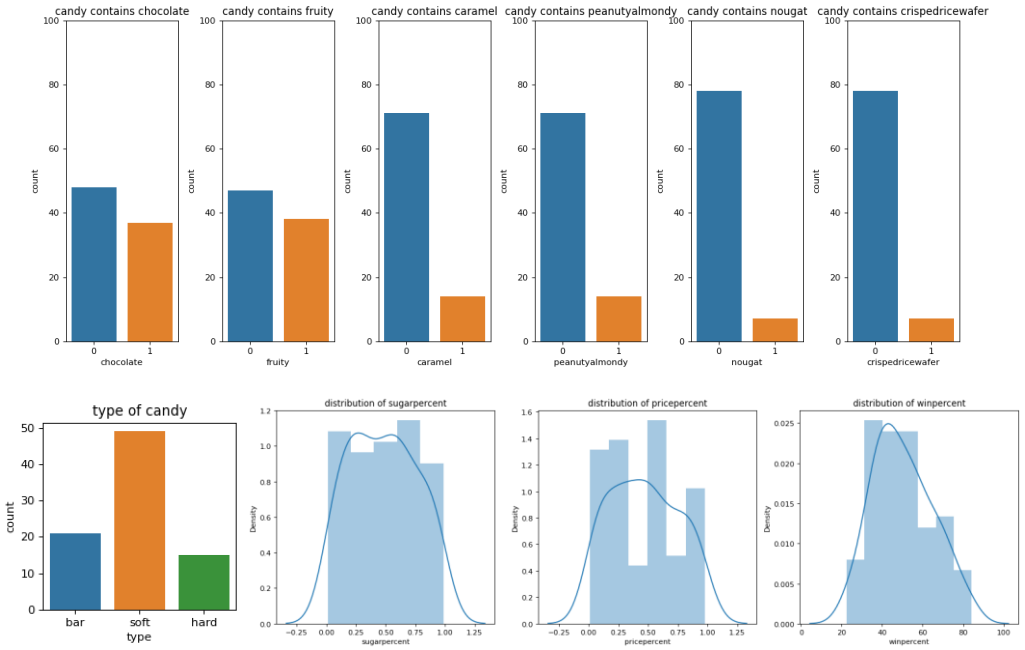
- Type 속성에 관하여 binary encoding 되어 있다. (범주형)

3) Sugarpercent ~ winpercent

- 제품에 관한 여러 정보가 비율로 나타나 있다. (연속형)

2. Data Visualization

각 변수별로 분포를 한 눈에 파악하기 위하여 시각화를 진행하였다.



fruity하거나 chocolate이 함유된 candy가 가장 많은 것을 확인할 수 있고 또, one-hot encoding되었던 bar와 hard 변수의 경우 bar와 hard에 해당되지 않는 경우 soft type으로 볼 수 있다. 따라서 이를 그래프로 표현해보면 soft type이 가장 많은 것을 확인할 수 있다. 본 실습에서는 사용할 Target변수는 'winpercent'로 이를 범주화해 logistic regression을 수행할 것이다.

[2] Data Preprocessing

1. 결측치 제거 및 winpercent 범주화

	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent	winpercent_cat
0	1	0	1	0	0	1	0	1	0	0.732	0.860	66.971725	1.0
1	1	0	0	0	1	0	0	1	0	0.604	0.511	67.602936	1.0
4	0	1	0	0	0	0	0	0	0	0.906	0.511	52.341465	1.0
5	1	0	0	1	0	0	0	1	0	0.465	0.767	50.347546	1.0
6	1	0	1	1	1	0	0	1	0	0.604	0.767	56.914547	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
80	0	1	0	0	0	0	0	0	0	0.220	0.116	45.466282	1.0
81	0	1	0	0	0	0	1	0	0	0.093	0.116	39.011898	0.0
82	0	1	0	0	0	0	0	0	1	0.313	0.313	44.375519	1.0
83	0	0	1	0	0	0	1	0	0	0.186	0.267	41.904308	1.0
84	1	0	0	0	0	1	0	0	1	0.872	0.848	49.524113	1.0

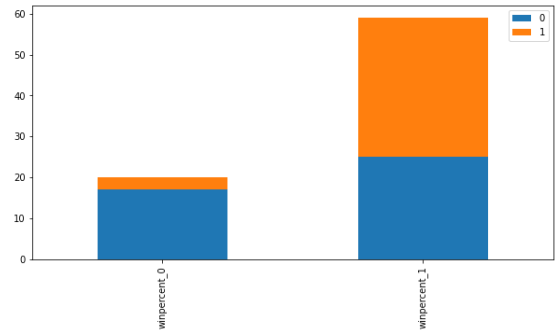
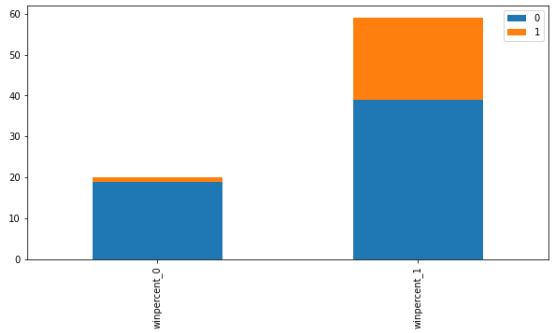
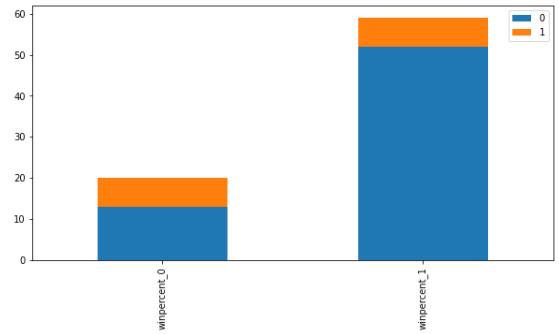
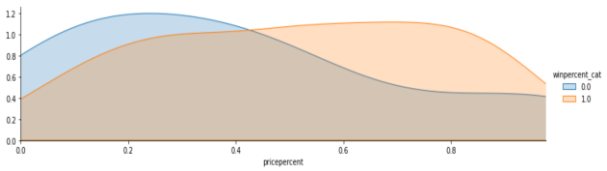
79 rows × 13 columns

taste를 나타내는 속성들 중 아무 속성도 가지고 있지 않은 row의 경우 결측치로 판단하고 해당 열을 제거하였다. 또, 연속형 변수인 winpercent를 범주형으로 변형시켜주기 위해 winpercent의 값이 40% 이상일 경우 1로 40% 미만일 경우 0으로 범주화했다. 위의 table은 기존 candy dataset에서 85개의 row 중 결측치 6개를 제거하고 winpercent를 범주형으로 바꾼 값을 winpercent_cat에 저장한 결과이다.

2. feature에 따른 종속변수 비교

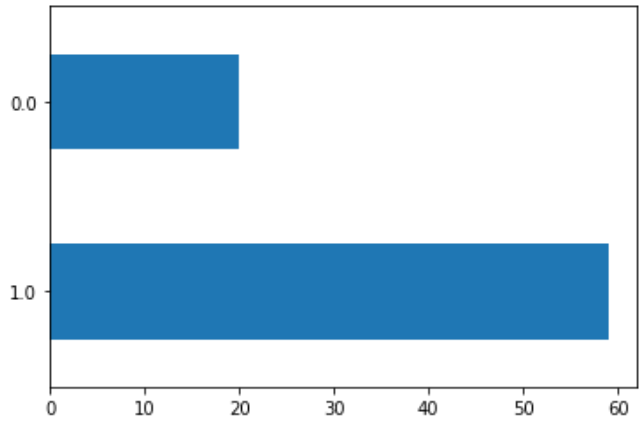
	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent	winpercent_cat
chocolate	1.000000	-0.852828	0.228730	0.361595	0.242939	0.332205	-0.369163	0.583669	-0.341683	0.086282	0.486617	0.614941	0.371451
fruity	-0.852828	1.000000	-0.380447	-0.446794	-0.300181	-0.300181	0.415720	-0.579289	0.342527	-0.067159	-0.519143	-0.479844	-0.255183
caramel	0.228730	-0.380447	1.000000	0.045055	0.321900	0.205248	-0.128571	0.321050	-0.271094	0.224949	0.239644	0.190160	0.041496
peanutyalmondy	0.361595	-0.446794	0.045055	1.000000	0.205248	-0.028055	-0.215385	0.246012	-0.204789	0.080578	0.297140	0.392507	0.193970
nougat	0.242939	-0.300181	0.321900	0.205248	1.000000	-0.097222	-0.144707	0.518188	-0.315777	0.122557	0.141810	0.185669	0.079099
crispedricewafer	0.332205	-0.300181	0.205248	-0.028055	-0.097222	1.000000	-0.144707	0.417358	-0.226683	0.065568	0.324647	0.316653	0.181540
hard	-0.369163	0.415720	-0.128571	-0.215385	-0.144707	-0.144707	1.000000	-0.279256	-0.005875	0.056760	-0.243624	-0.321779	-0.263452
bar	0.583669	-0.579289	0.321050	0.246012	0.518188	0.417358	-0.279256	1.000000	-0.609387	0.089978	0.511321	0.409680	0.284439
pluribus	-0.341683	0.342527	-0.271094	-0.204789	-0.315777	-0.226683	-0.005875	-0.609387	1.000000	-0.006401	-0.208448	-0.233911	-0.050856
sugarpercent	0.086282	-0.067159	0.224949	0.080578	0.122557	0.065568	0.056760	0.089978	-0.006401	1.000000	0.339720	0.237022	0.150739
pricepercent	0.486617	-0.519143	0.239644	0.297140	0.141810	0.324647	-0.243624	0.511321	-0.208448	0.339720	1.000000	0.302980	0.177886
winpercent	0.614941	-0.479844	0.190160	0.392507	0.185669	0.316653	-0.321779	0.409680	-0.233911	0.237022	0.302980	1.000000	0.700585
winpercent_cat	0.371451	-0.255183	0.041496	0.193970	0.079099	0.181540	-0.263452	0.284439	-0.050856	0.150739	0.177886	0.700585	1.000000

'winpercent'와 'winpercent_cat'을 포함한 총 13개의 변수들 간의 상관관계를 분석한 결과이다. 'winpercent_cat'과 관련이 많고 분석에 유의미한 변수들은 'chocolate', 'fruity', 'hard', 'bar', 'pricepercent' 변수임을 알 수 있고, 그 중 'fruity' 변수는 'chocolate' 변수와의 correlation이 높아 chocolate 변수만 logistic regression에 사용했다. 아래의 그림들은 선택한 각 feature들의 값에 따라 'winpercent_cat'이 어떻게 변화하는지 나타낸 그래프들이다.

<table>
<tr><th>chocolate</th><th>bar</th></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><th>hard</th><th>pricepercent</th></tr>
<tr><td> 
</td><td> 
</td></tr>
</table>

먼저 Chocolate flavor를 가진 candy의 경우, 다른 candy들에 비해 선호될 확률이 선호되지 않을 확률보다 확연히 크다. 두번째로 세번째로, bar와 hard의 경우 candy의 형태를 나타내는 변수 bar, hard, soft가 one hot encoding된 변수들로 soft대신 bar와 hard를 dummy 변수로 두어 logistic regression을 진행할 것이다. Bar 변수의 경우 다른 candy들에 비해 선호될 확률이 선호되지 않을 확률보다 확연히 크다. Hard한 candy의 경우 선호되는 candy의 수와 선호되지 않는 수는 비슷하지만, 전체 candy중 선호되지 않는 candy의 절대적인 값이 작은 것을 감안해 선호되지 않을 확률이 다른 candy들에 비해 높은 것을 확인할 수 있다. 마지막으로, pricepercent의 경우 연속형 변수로 kdeplot으로 확인해보았을 때, pricepercent가 낮을수록 선호되지 않고 높을수록 선호되는 것을 확인할 수 있다.

3. Imbalanced Data 처리



위 그래프는 'winpercent_cat' 변수의 값 분포를 나타낸 것으로 각 클래스의 {0,1} 간의 data가 불균형한 것을 확인할 수 있다. 이 데이터로 학습할 경우 1에 치우쳐서 학습될 위험이 있으므로 데이터 불균형을 sampling을 통해 처리해야한다. 이번 실습에서는, random undersampling 과정과 random oversampling과정 모두를 수행해 그 결과를 비교해보았다. (Sampling을 진행하기 전 shuffling을 해 주었다.) Undersampling한 결과 총 40개의 학습데이터를 얻을 수 있었고 oversampling한 결과 총 118개의 데이터를 얻을 수 있었다.

[3] Model 설명

로지스틱 회귀의 경우는 최대우도추정법을 기반으로 모델이 생성된다.
최대우도추정법을 설명하기에 앞서 우도를 최대화한다는 말이 무엇을 의미하는지부터 설명하고자 한다.

우리가 특정한 확률을 구할 때 해당 사건이 임의의 분포를 따른다고 하고 사건을 x, 해당 분포의 파라미터를 θ라고 가정할 때 해당 사건이 발생할 확률은 p(x|θ)라고 기술할 수 있다. 하지만 이러한 확률을 구하고자 하는 상황에서 θ값을 모르는 상황이라면, 베이즈 규칙을 활용해 조건절과 결과절을 뒤집어 해당 식을 θ가 발생할 우도로 바꿔 생각할 수 있다. 이를 L(θ|x)라고 표현하고 이 때 가장 적절한 θ는 해당 우도를 최대화해주는 값이 될 것이다. 즉 관측된 사건(데이터)이 발생할 경우를 가장 높게 만들어주는 파라미터를 찾는 과정이 최대우도추정법이라고 설명할 수 있다.

Logistic Regression은 target variable이 0 or 1로 베르누이 시행을 전제로 하는 모델이다. 그렇기에 P(y=1) = p라고 기술할 경우, 다음 식으로 정리할 수 있고

$$P(Y = y_i) = p^{y_i}(1-p)^{1-y_i} \quad (y_i = 0, 1)$$

이 베르누이 확률변수 Y에 관한 우도함수와 p값은 다음과 같다.

$$L = \prod_i p^{y_i}(1-p)^{1-y_i}$$

$$p_i = \text{logit}^{-1}(\boldsymbol{\beta} \cdot \mathbf{X}_i) = \frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}}$$

이때 p값은 우리가 추정할 logistic function에 x값을 대입해 얻은 결과로, 우리는 모든 데이터에 대해서 위 L값을 최대화하는 logistic function의 coefficient들을 추정하면 되는 것이다.

우리가 사용할 모델에서 yi는 우리가 설정한 target 변수(winpercent 40%를 넘었는지 여부)이며 해당 우도함수에서 p값은 다음과 같이 표현할 수 있다.

해당 식에서 X는 우리가 사용할 독립변수(요거 써줘)이고 β값은 우리가 추정해야 할 파라미터이다. 우리는 이 파라미터를 위에서 설명했던 것과 같이 주어진 데이터의 X,Y에 대해서 L함수를 최대화하는 β값들을 추정하면 될 것이다.

[4] 결과해석

1. Under sampling

1) logistic regression 결과

Under sampling 에서는 y 가 0 인 row 가 20 개 이므로 이에 맞춰 y 가 1 인 row 를 20 개로 줄여 모델링하였다.

다음은 선택된 x 변수 4 개(chocolate, hard, bar, pricepercent)의 y(winpercent)에 대한 odds ratio 값이다.

```
array([1.92352018, 0.63675176, 3.09052628, 1.91075486, 0.59522498])
```

위 값을 통해 다음과 같은 해석을 내릴 수 있다.

1) Chocolate 을 포함한 candy 는 winpercent 가 40% 이상일 확률이 포함하지 않은 경우보다 약 1.9 배 높다.

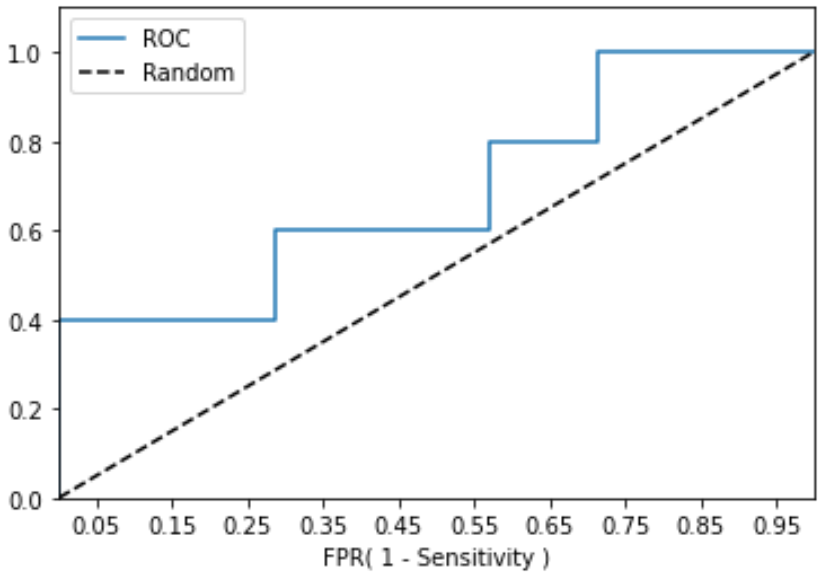
2) Hard 타입의 candy 는 winpercent 가 40% 이상일 확률이 다른 타입인 경우보다 약 0.6 배 낮다.

3) Bar 타입의 candy 는 winpercent 가 40% 이상일 확률이 다른 타입인 경우보다 약 3 배 높다.

4) Pricepercent 의 증가는 winpercent 가 40% 이상일 확률을 감소시킨다.

2) ROC curve

다음은 ROC curve 이다.



위 그래프에서 x 축은 false positive, 즉 실제 값이 negative 인 데이터를 positive 로 예측한 비율이고, y 축은 true positive, 즉 실제 값이 positive 인 데이터를 positive 로 예측한 비율이다. ROC curve 가 random 라인 위로 그려지는 것으로 보아 모델이 어느 정도 유의미한 예측을 하는 것으로 보이나, 표본의 개수가 너무 적어 성능이 낮다.

3) 성능 지표 분석

다음은 cut-off 가 각각 0.5 와 0.7 인 경우에 대한 다양한 성능 지표이다.

[Case 1] cut-off 0.5

```
Accuracy:  0.7500
precision:  1.0000
recall:  0.4000
f1_score:  0.5714
AUC:  0.7000
```

Accuracy 는 모델의 예측이 실제 값과 들어맞는 비율이다. 이 값이 0.75 이므로 반대로 예측이 틀리는 비율인 misclassification rate 은 0.25 라고 할 수 있다.

Precision 은 true positive + false positive 에 대한 true positive 의 비율을 의미한다. 이 값이 1 이므로 모델이 false positive 로 예측한 적이 없음을 알 수 있다.

Recall 은 true positive + false negative 에 대한 true positive 의 비율을 의미한다. 이 값이 1 에 가까울수록 모델의 성능이 좋다고 평가할 수 있는데 다소 부족한 모습을 보인다.

F1 은 2 * (precision * recall) / (precision + recall)의 식으로 계산한다. Precision 과 recall 에 대한 weighted average 라고 할 수 있으며 1 에 가까울수록 모델의 성능이 우수하다고 판단할 수 있다.

AUC 는 ROC curve 를 적분한 값이다. ROC curve 는 sample 라인으로부터 멀어져 위로 올라갈수록 모델의 성능이 좋다고 평가하므로 AUC 값이 클수록 좋다고 말할 수 있다.

[Case 2] cut-off 0.7

```
Accuracy:  0.6667
precision:  1.0000
recall:  0.2000
f1_score:  0.3333
AUC:  0.6000
```

Cuf-off 가 0.5 인 경우보다 전체적으로 성능이 하락했음을 알 수 있다.

2. Over sampling

1) logistic regression 결과

Over sampling 에서는 y 가 1 인 row 가 59 개이므로 이에 맞춰 y 가 0 인 row 를 59 개로 늘려 모델링하였다.

다음은 odds ratio 값이다.

```
array([2.15226076, 0.43171971, 4.19680893, 1.11707695, 0.6670113 ])
```

1) Chocolate 을 포함한 candy 는 winpercent 가 40% 이상일 확률이 포함하지 않은 경우보다 약 2.1 배 높다.

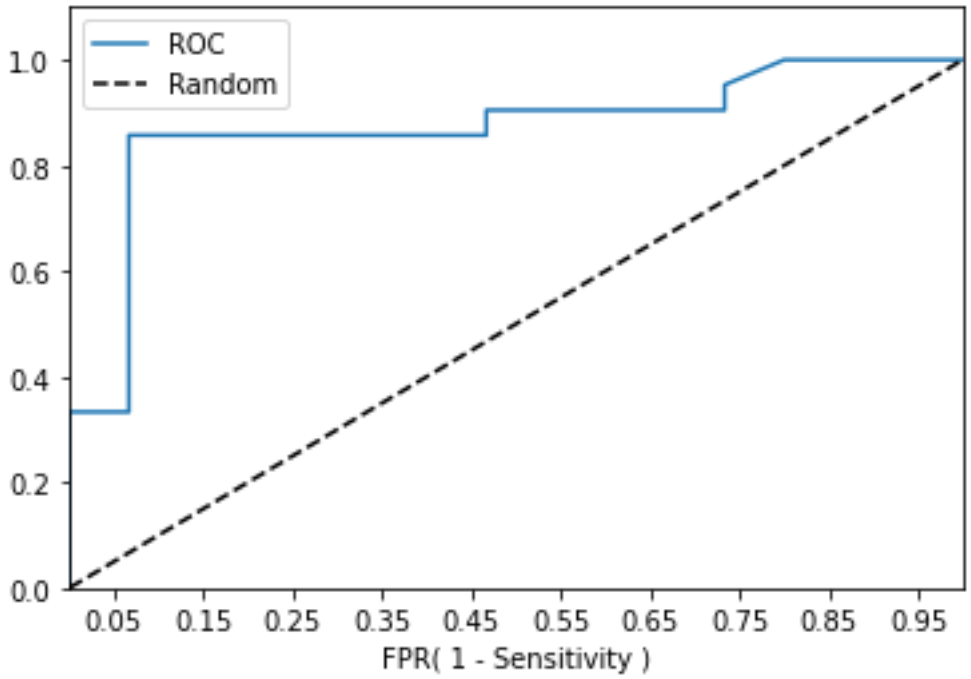
2) Hard 타입의 candy 는 winpercent 가 40% 이상일 확률이 다른 타입인 경우보다 약 0.4 배 낮다.

3) Bar 타입의 candy 는 winpercent 가 40% 이상일 확률이 다른 타입인 경우보다 약 4.1 배 높다.

4) Pricepercent 의 증가는 winpercent 가 40% 이상일 확률을 감소시킨다.

2) ROC curve

다음은 ROC curve 이다.



Under sampling 의 경우보다 ROC curve 가 위로 올라간 것을 볼 수 있다. 이에 대해 표본 개수의 증가가 모델의 성능 향상을 불러왔다고 판단할 수 있다.

3) 성능 지표 분석

다음은 cur-off 가 각각 0.5 와 0.7 인 경우에 대한 다양한 성능 지표이다.

[Case 1] cut-off 0.5

```
Accuracy:  0.8889
precision:  0.9474
recall:  0.8571
f1_score:  0.9000
AUC:  0.8952
```

Under sampling 의 경우보다 전체적으로 향상된 결과를 보인다. 특히 recall 점수가 눈에 띄게 향상되었다. 또한 ROC curve 가 위로 올라간 것의 결과로 AUC 값이 크게 증가하였다.

[Case 2] cut-off 0.7

```
Accuracy:  0.6111
precision:  1.0000
recall:  0.3333
f1_score:  0.5000
AUC:  0.6667
```

Under sampling 에서와 마찬가지로 cut-off 가 0.5 인 경우보다 낮은 성능을 보인다. 다만 모델이 positive 로 예측하는 경향이 강해져 precision 점수는 증가한 것을 확인할 수 있다.