Logistic Regression

Banking Dataset - Marketing Targets 분석

데이터마이닝이론및응용
에이쁠원하조

2020147024 김우영
2020251009 김혜리
2020195053 노가원
2020147018 조윤영

1. data preprocessing & target 변수 설명

Detailed Column Descriptions

bank client data:

1 - age (numeric)
2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary: "yes","no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes","no")
8 - loan: has personal loan? (binary: "yes","no")
related with the last contact of the current campaign:
9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", …, "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)
other attributes:
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):
17 - y - has the client subscribed a term deposit? (binary: "yes","no")

Missing Attribute Values: None

변수는 크게 고객의 특성(age, job,marital,education), 고객의 finance와 관한 특성(default,balance,housing,loan) 그리고 캠페인과 관련한 고객에 대한 변수 (contact,day,month,duration,campaign,pdays,previous,poutcome)으로 나눌 수 있다. 이때 해당 실습에서는 고객의 특성을 제외한 고객의 finance에 관한 변수와 캠페인과 관련한 고객에 대한 변수를 독립변수로 사용한다
binary인 타켓변수 y는 캠페인을 통해 고객이 정기예금에 가입했는지 여부에 대한 변수다.

```
data.describe()
```

	age	duration	campaign	pdays	previous
count	32950.000000	32950.000000	32950.000000	32950.000000	32950.000000
mean	40.014112	258.127466	2.560607	962.052413	0.174719
std	10.403636	258.975917	2.752326	187.951096	0.499025
min	17.000000	0.000000	1.000000	0.000000	0.000000
25%	32.000000	103.000000	1.000000	999.000000	0.000000
50%	38.000000	180.000000	2.000000	999.000000	0.000000
75%	47.000000	319.000000	3.000000	999.000000	0.000000
max	98.000000	4918.000000	56.000000	999.000000	7.000000

연속형 변수로 이뤄진 age,duration,campaign,pdays,previous의 경우 변수마다 데이터의 범위가 많이 차이나고 평균도 많이 차이나는 것을 알 수 있다. age의 경우 조사한 고객들의 평균연령은 3-40대쯤임을 알 수 있다. duration의 경우 평균시간이 4분임을 알 수 있다. campaign의 경우 평균적으로 2번 캠페인 기간동안 고객에게 연락을 취했음을 알 수 있다
pdays는 대부분의 연락이 마지막 연락했을때보다 꽤 시간이 지난후 취해졌음을 알 수 있다.
previous의 경우 평균적으로 전에 연락을 취한적 없는 고객을 대상으로 했음을 알 수 있다다

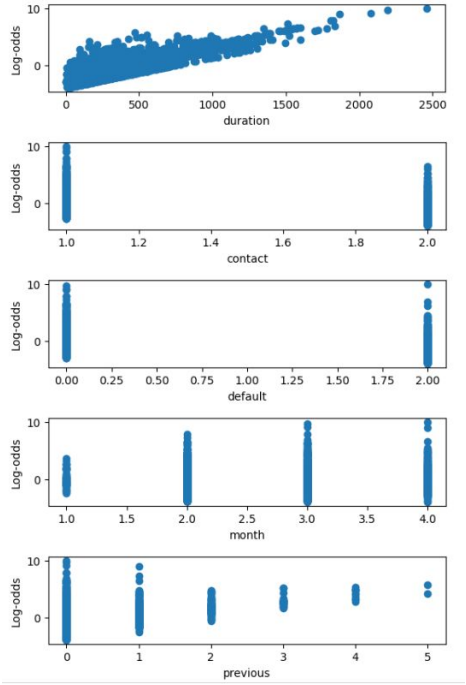
```
data.drop(['job', 'marital', 'education'], axis = 1, inplace = True)

data.loc[data['y'] == 'yes' , 'y'] = 1
data.loc[data['y'] == 'no' , 'y'] = 0
```

해당 실습에서 필요한 변수들만 남기고 나머지는 drop했다. 또 타겟변수가 yes,no로 돼있어서 각각 1,0으로 바꿔줬다

우리가 구하고자하는 타겟변수의 1이 나올 확률은 0.5로 했다

2. Logistic Regression 기본 가정 확인

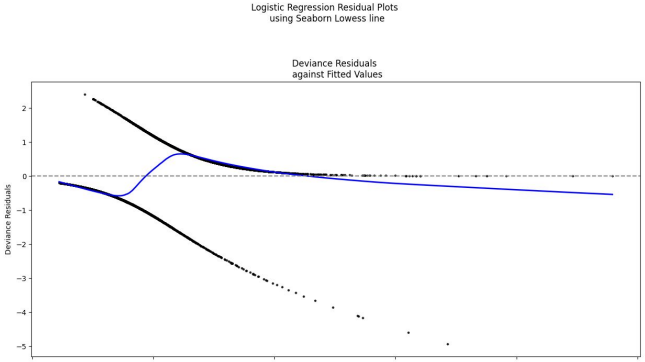
1. Outcome Type : Y의 값은 0,1로 binary 형태를 만족한다.
2. 관측치간 독립성 : 개별 고객에 대해 관측한한 데이터이므로 충족한다.
3. 독립변수와 log odds 간의 linearity : 첨부한 그래프를 통해 “duration” 이외의 변수는 log odds와 선형 관계를 보이지 않고 있음을 확인할 수 있다.
“duration”은 log odds와 선형관계를 보이기 때문에 이는 모델 성능에 영향을 미칠 수 있다.

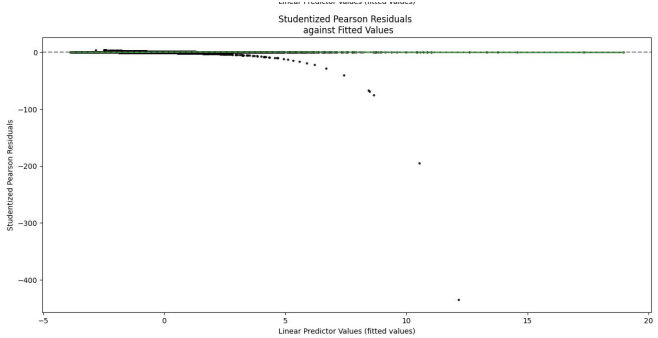


2. Logistic Regression 기본 가정 확인

4. 이상치 존재 확인: studentized residuals의 값을 확인하여 2~3을 넘어가는 데이터를 이상치로서 고려해볼 수 있기 때문에 첨부한 그래프를 이용해 이상치를 확인했다. 첨부한 그래프에서 추이선(실선)은 y=0에 근사하지만 이상치로 판단될 수 있는 몇개의 데이터가 보인다. 이는 최종 logistic regression 모델의 성능을 저하시킬 가능성이 있다.
5. 다중공선성 확인: “contact”/”month”가 5에 가까운 VIF를 가지고 있는 것을 확인할 수 있는데, 두 변수 사이의 유의미한 연관성은 찾기 힘들고, VIF의 값도 극단적으로 크지 않기 때문에 모델의 성능을 떨어뜨릴 수 있다는 가능성만 인지하고 변수를 제거하지는 않았다.

	variables	VIF
0	duration	2.341763
1	default	1.301769
2	contact	4.613792
3	month	4.764993
4	previous	1.201810
5	y	1.444089





6. 잔차의 등분산성 확인: 첨부한 deviance residuals를 보면 실선이 굴곡을 가지는 것을 볼 수 있는데 이를 통해 잔차가 완전히 무작위로 분포하지는 않는다는 것을 확인할 수 있다. 따라서 완벽한 등분산성을 만족하지는 못한다. 이는 모델의 성능에 영향을 미치는데, ROC 그래프를 확인하면 분류 성능이 나쁘지 않음을 알 수 있다. 하지만 이상치로 인해 모델 성능과 R-squared 값에 영향을 미쳤을 가능성은 남아 있다고 판단된다.

3. 결과해석 - 변수 및 계수 [model.summary() and model.summary2()] + odds ratio (승산비) 해석

```
model_sm.summary()
```

Logit Regression Results

Dep. Variable:	y	No. Observations:	4800
Model:	Logit	Df Residuals:	4794
Method:	MLE	Df Model:	5
Date:	Fri, 21 Apr 2023	Pseudo R-squ.:	0.3285
Time:	23:55:33	Log-Likelihood:	-2234.2
converged:	True	LL-Null:	-3327.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0889	0.196	-0.453	0.650	-0.473	0.296
duration	0.0052	0.000	28.837	0.000	0.005	0.006
default	-0.5411	0.058	-9.360	0.000	-0.654	-0.428
contact	-1.2452	0.095	-13.167	0.000	-1.431	-1.060
month	-0.1014	0.048	-2.116	0.034	-0.195	-0.007
previous	1.0011	0.069	14.479	0.000	0.866	1.137

```
model_sm.summary2()
```

Model:	Logit	Pseudo R-squared:	0.328
Dependent Variable:	y	AIC:	4480.3840
Date:	2023-04-21 23:55	BIC:	4519.2422
No. Observations:	4800	Log-Likelihood:	-2234.2
Df Model:	5	LL-Null:	-3327.1
Df Residuals:	4794	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-0.0889	0.1962	-0.4532	0.6504	-0.4735	0.2956
duration	0.0052	0.0002	28.8367	0.0000	0.0048	0.0056
default	-0.5411	0.0578	-9.3600	0.0000	-0.6544	-0.4278
contact	-1.2452	0.0946	-13.1668	0.0000	-1.4305	-1.0598
month	-0.1014	0.0479	-2.1158	0.0344	-0.1954	-0.0075
previous	1.0011	0.0691	14.4790	0.0000	0.8656	1.1366

Logit(y = 1) = - 0.0889 + 0.0052 * duration - 0.5411 * default - 1.2452 * contact - 0.1014 * month + 1.0011 * previous

- duration: 통화지속시간은 한 단위인 1초에 대한 계수이므로 다른 변수에 비해 작은 계수를 보이지만, 1분 단위로 생각해보면 약 0.3의 계수를 가지므로 영향의 정도가 작은 것은 아니다. 계수가 양의 값을 가지므로 통화지속시간이 증가하면 구독 할 확률 증가한다고 해석할 수 있다.
- default: 대출 상환 불이행이 있는 경우 0.5411만큼의 비율로 구독 확률이 감소한다.
- contact: 연락 수단이 cellular이 아니라 telephone인 경우, 1.2452만큼의 비율로 구독 확률이 감소한다.
- month: 마지막으로 통화한 달이 4분기에 한 분기씩 가까워질 때마다 0.1014의 비율로 구독 확률이 감소한다.
- previous: 이전에 연락했던 횟수가 1번 증가할 때마다 1,0011의 비율로 구독 확률이 증가한다.

LLR p-value: 0.000, 모델의 p-value값이 0으로 수렴하기 때문에 본 모델은 통계적으로 매우 유의미하다고 해석할 수 있다.

개별적인 변수들의 p-value값을 살펴보면, duration, default, contact,, previous의 p-value 값 모두 0으로 수렴하므로 매우 유의미한 변수들이라고 해석할 수 있다. month 변수의 경우 0.034로, 앞의 네 변수보다는 값이 크지만 0.05보다 작기 때문에 영향이 있기는 하다고 해석할 수 있다.

Df Residuals: 4794, Df Residuals는 로지스틱 회귀 모델에서 잔차의 자유도(degree of freedom for residuals)를 나타내는 지표로, 값이 클수록 사용 변수의 수가 많고 모델이 복잡하다는 것을 의미한다. 이 경우 오버피팅이 발생할 가능성이 높아진다. 반면 Df Residuals의 값이 작아지면 모델이 단순해지면 언더피팅 가능성이 높아진다는 것을 의미한다.

Pseudo R-squ: 0.301, Pseudo R-squared는 해당 모델이 y를 설명하는 정도를 나타내는 지표이다. 최대 가능한 설명력은 1이며, 해당 모델이 y를 약 32.85% 설명할 수 있다고 해석할 수 있다.

Log-Likelihood: -2234.2, 주어진 데이터에서 모델의 예측값이 실제 데이터와 얼마나 일치하는지를 계산한 값으로 높을수록 모델 성능이 상대적으로 좋다고 해석할 수 있다.

LL-Null: -3327.1, LL-Null은 모든 예측변수를 빼고, 오직 상수항(constant)만을 가지고 예측한 경우의 Log-Likelihood 값으로 높을수록 모델 성능이 상대적으로 좋다고 해석할 수 있다.

AIC: 4480.3840, AIC(Akaike Information Criterion)는 모델의 잔차 제곱합에 대한 패널티로서, 모델에 사용된 변수의 수를 고려하여 계산된다. 따라서 AIC 값이 낮을수록 모델의 적합도가 높아지며, 동시에 모델의 복잡도가 낮아진다는 것을 의미한다.

BIC: 4519.2422, BIC(Bayesian Information Criterion)는 AIC에 샘플의 수(n)에 대한 페널티를 추가로 적용한다. 따라서 BIC 값이 작을수록 적합한 모델이라고 해석할 수 있다.

위의 결과값들을 이용해 모델의 성능을 평가하기 위해서는 비교 대상이 되는 모델이 필요하다. 각각의 요소들에 평가기준이 되는 절대적인 값이 있는 것이 아니므로, 상대적 비교를 통해 성능을 따져보아야 하기 때문이다.

```
# Odds ratio
import numpy as np
np.exp(model_sm.params)

# log odd여서 exp 취해줘야 함
```

```
const        0.914912
duration     1.005216
default      0.582126
contact      0.287896
month        0.903547
previous     2.721197
dtype: float64
```

<odds ratio 해석>

- duration : 통화지속시간이 1단위 증가할 때마다 구독 여부에 대한 odds가 1.005216배 증가
- default : 대출 상환 불이행이 있는 경우 그렇지 않은 경우보다 구독 여부에 대한 odds가 0.582126배 감소한다고 해석할 수 있습니다.
- contact : contact type이 telephone일 때가 cellular일 때보다 구독 여부에 대한 odds가 0.287896배 감소
- month : 마지막으로 통화한 달이 1분기에서 4분기로 한 분기 증가할 때 마다 구독 여부에 대한 odds가 0.903547배 감소
- previous : 이전 연락 횟수가 1단위 증가할 때 마다 구독 여부에 대한 odds가 2.721197배 증가

4. Confusion matrix를 포함한 다양한 성능지표의 해석 (Accuracy, Precision, Specificity, Recall, F1 scor)

cut-off를 0.5로 설정

```
[[3840  832]
 [ 140  460]]
 Accuracy: 0.8156297420333839
Specificity: 0.821917808219178
Precision: 0.3560371517027864
   Recall: 0.7666666666666667
 F1 score: 0.48625792811839325
```

Confusion Matrix:

예측된 Negative(0) 클래스의 실제 Negative(0) 클래스 개수는 3840개, 예측된 Negative(0) 클래스의 실제 Positive(1) 클래스 개수는 832개

예측된 Positive(1) 클래스의 실제 Negative(0) 클래스 개수는 140개, 예측된 Positive(1) 클래스의 실제 Positive(1) 클래스 개수는 460개

정확도(Accuracy): 실제 레이블과 예측 결과가 일치하는 비율, 약 0.8156 (또는 약 81.56%)의 정확도를 보임

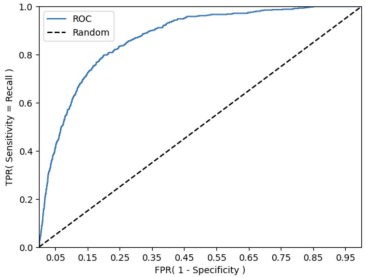
특이도(Specificity): 실제 Negative(0) 클래스 중에서 모델이 Negative(0) 클래스로 예측한 비율로, 약 0.8219 (또는 약 82.19%)의 특이도를 보임

정밀도(Precision): 델이 Positive(1) 클래스로 예측한 샘플 중에서 실제 Positive(1) 클래스인 비율로, 약 0.3560 (또는 약 35.60%)의 정밀도를 보임

재현율(Recall): 실제 Positive(1) 클래스 중에서 모델이 Positive(1) 클래스로 예측한 비율로, 약 0.7667 (또는 약 76.67%)의 재현율을 보임

F1 스코어(F1 Score): 정밀도와 재현율의 균형을 나타내며, 약 0.4863 (또는 약 48.63%)의 F1 스코어를 보임

5. ROC 커브해석 (민감도,



threshold가 높아지건 낮아지건 TPR(=민감도)과 FPR(=특이도)은 어느 정도 비례적으로 함께 커지거나 작아지기 때문에 ROC 커브는 왼쪽과 같은 모양을 가진게 된다.

이때 분류 성능은 ROC 커브가 좌측 상단에 얼마나 붙어있는지로 판단할 수 있는데, TPR과 FPR이 서로의 변화에 얼마나 영향을 받는지를 보여주기 때문이다. True 그룹과 False 그룹을 더 잘 구별할수록 커브는 좌측 상단에 더 붙게 된다.

왼쪽 그림에서는 ROC 커브의 휜 정도가 시각적으로 명확히 보이고, 좌측 상단에 꽤나 가까이 붙어있으므로 꽤 좋은 성능을 가진다고 판단할 수 있다.

6. 한계점

Logistic regression은 독립 변수와 종속 변수 간의 선형 관계를 가정하기 때문에 데이터가 nonlinear한 관계를 갖는 경우 모델의 예측 성능이 저하될 수 있음. 기본 가정 확인에서 명시되어 있듯이 이번 실습에서 사용한 몇몇 데이터는 비선형이기 때문에 예측 성능이 저하될 여지가 있음.

또한 coefficients를 통해 변수의 영향력을 해석할 수 있는 반면, 변수간의 interaction을 표현하기 어려울 수 있음.

Reference

https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets