예측애널리틱스 Homework #4

고려대학교
KOREA UNIVERSITY

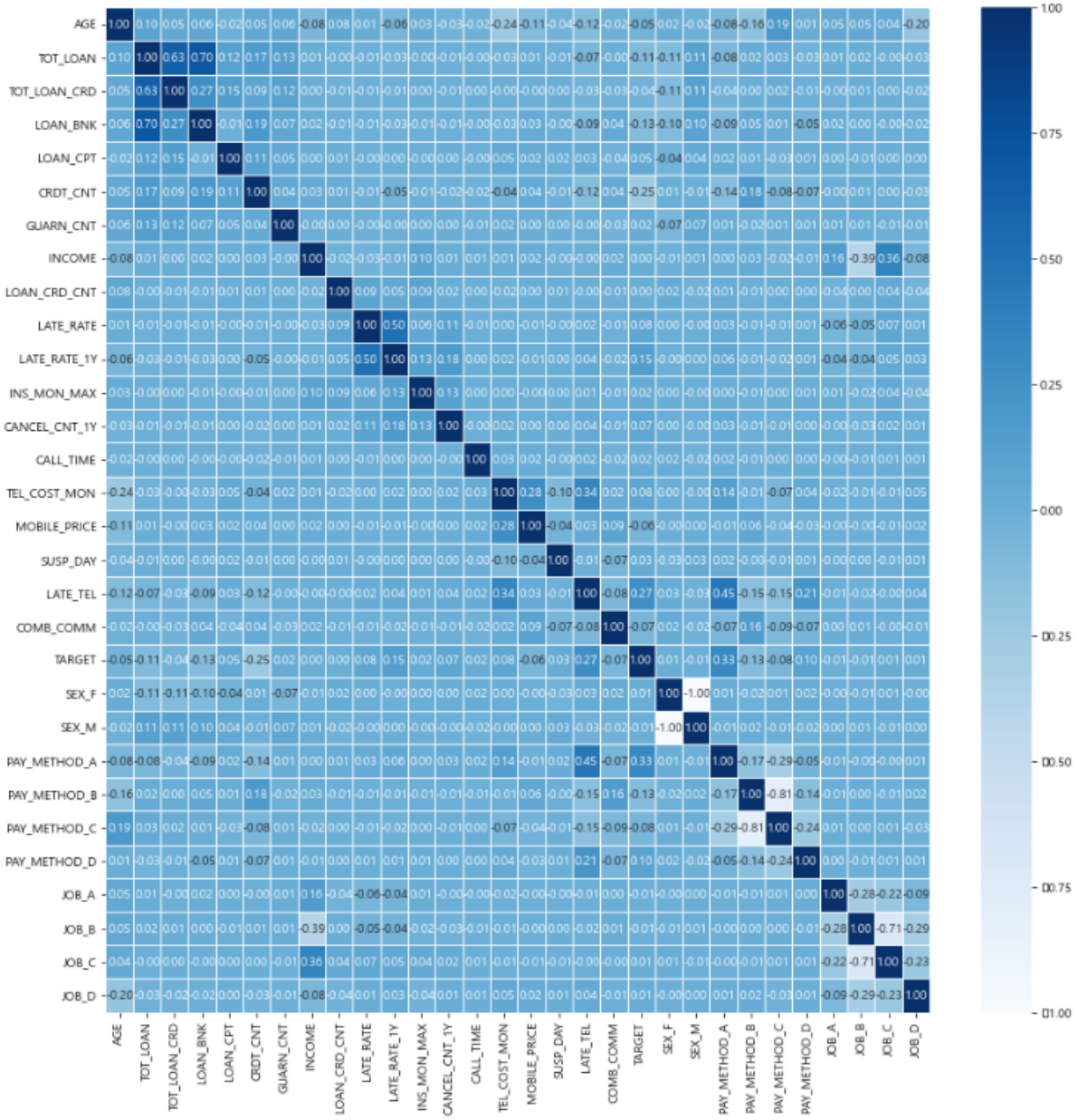
대학	고려대학교 공과대학
학과	산업경영공학부
학번	2017170819
이름	박상민

로지스틱 회귀모델 구축 및 해석

Loan_Data 는 22 개의 독립변수와 1 개의 종속변수, 총 43,386 개의 instance 로 구성된 데이터이다. 파이썬 코드 내에서 데이터 명칭은 "data"라 하였다. 변수에 대해서 먼저 알아보았다.

변수 이름	변수 설명	변수 형태
AGE	연령	연속형
TOT_LOAN	대출 총액	연속형
TOT_LOAN_CRD	신용대출 총액	연속형
LOAN_BNK	은행권에서 발생한 대출 총액	연속형
LOAN_CPT	카드사/캐피탈에서 발생한 대출 총액	연속형
CRDT_CNT	신용카드 발급 수	연속형
GUARN_CNT	보증 건수	연속형
INCOME	소득	연속형
LOAN_CRD_CNT	신용대출 건수	연속형
LATE_RATE	보험료 연체율	연속형
LATE_RATE_1Y	최근 1년 보험료 연체율	연속형
INS_MON_MAX	월납입보험료 (최대값)	연속형
CANCEL_CNT_1Y	최근1년 실효해지건수	연속형
CALL_TIME	월별 통화시간	연속형
TEL_COST_MON	서비스 납부요금	연속형
MOBILE_PRICE	사용중인 핸드폰단말기 가격	연속형
SUSP_DAY	회선의 사용정지일수	연속형
LATE_TEL	핸드폰 요금 연체금액	연속형
COMB_COMM	결합상품가입 여부	이진형
SEX	성별: 남자(1), 여자(F)	명목형
PAY_METHOD	핸드폰 요금 납부방법	명목형
JOB	직업군	명목형
TARGET	대출연체여부: 미발생(0), 발생(1)	이진 (타겟변수)

22개의 독립변수 중 연속형 변수는 18개, 이진형 변수는 1개, 명목형 변수는 3개이며 종속변수이자 타겟변수는 'TARGET'으로 이진형 변수임을 확인할 수 있었다. 먼저 명목형 변수를 binary 변수로 바꾸었다. SEX는 female을 기준으로 하여 male을 1, female을 0으로 하였다. PAY_METHOD는 D를 기준으로 하여 A는 (1, 0, 0), B는 (0, 1, 0), C는 (0, 0, 1), D는 (0, 0, 0)으로 정했다. JOB는 D를 기준으로 하여 A는 (1, 0, 0), B는 (0, 1, 0), C는 (0, 0, 1), D는 (0, 0, 0)으로 정했다. 따라서 26개의 X 변수와 1개의 Y 변수로 정해졌다.



변수들 사이의 상관관계를 파악하기 위해 시각화 툴을 이용하여 상관관계 matrix 를 그려보았다. 대출 총액을 나타내는 'TOT_LOAN' 변수와 신용대출 총액을 나타내는 'TOT_LOAN_CRD', 은행권에서 발생한 대출 총액을 나타내는 'LOAN_BNK' 변수 사이에서 유의미하다고 볼 수 있는 상관관계가 발견되었다. 모두 대출 총액에 관련된 변수이므로 상관관계가 있음을 예상할 수 있었다. 나머지 변수들 사이에서는 유의미한 상관관계를 찾을 수는 없었다.

데이터 전처리

데이터를 Train set과 test set으로 나눴다. 비율은 7:3이다. 그리고 독립변수들의 단위가 서로 다르기 때문에 데이터 칼럼 단위 정규화를 진행했다. 이후 로지스틱 회귀모델을 구현해주는 프로그램을 사용해 로지스틱 회귀모델을 구축해 학습시켰다. 이후 각 데이터 칼럼에 따른 파라미터 β와 e^{β}를 도출했다. β는 X 값이 1단위 증가할 때 log(Odds) 값의 변화량이고, e^{β}는 X 값이 1단위 증가할 때 Odds 값의 변화 비율이다. 즉 Odds Ratio를 의미한다. Odds ratio는 나머지 입력변수는 모두 고정시킨 상태에서 한 변수를 1단의 증가시켰을 때 변화하는 Odds의 비율이다. 회귀계수인 β가 양수이면 Odds가 1보다 크다는 뜻이고, β가 음수이면 Odds가 0 이상 1 미만이라는 뜻이다. 즉 Odds ratio가 1보다 크면 대출연체 발생(arrear), 0보다 작으면 대출연체 미발생(no)으로 분류했다. AGE를 살펴보면 β 값이 음수이기 때문에 나이를 1살 더 먹을수록 대출연체가 발생할 확률이 감소한다. Odds ratio가 0.98이므로 대출 연체될 확률(Odds)이 0.98배가 된다. LOAN_CPT는 카드사/캐피탈에서 발생한 대출 총액이며, β 값이 양수이기 때문에

	beta	exp(beta)	interpret
const	-3.17	0.04	no
AGE	-0.02	0.98	no
TOT_LOAN	-0.25	0.78	no
TOT_LOAN_CRD	0.1	1.11	arrear
LOAN_BNK	-0.83	0.44	no
LOAN_CPT	0.21	1.23	arrear
CRDT_CNT	-0.89	0.41	no
GUARN_CNT	0.13	1.14	arrear
INCOME	0.09	1.09	arrear
LOAN_CRD_CNT	-0.05	0.95	no
LATE_RATE	-0.0	1.0	no
LATE_RATE_1Y	0.32	1.38	arrear
INS_MON_MAX	0.03	1.03	arrear
CANCEL_CNT_1Y	0.07	1.07	arrear
CALL_TIME	0.02	1.02	arrear
TEL_COST_MON	-0.02	0.98	no
MOBILE_PRICE	-0.22	0.8	no
SUSP_DAY	0.07	1.07	arrear
LATE_TEL	0.22	1.25	arrear
COMB_COMM	-0.07	0.93	no
SEX_M	0.0	1.0	no
PAY_METHOD_A	0.2	1.22	arrear
PAY_METHOD_B	-0.53	0.59	no
PAY_METHOD_C	-0.4	0.67	no
JOB_A	-0.02	0.98	no
JOB_B	0.05	1.05	arrear
JOB_C	0.02	1.02	arrear

Odds ratio는 1.23으로 1보다 크다. 이는 카드사/캐피탈에서 발생한 대출 총액이 1단위 증가할 경우 대출 연체될 확률(Odds)이 1.23배가 된다는 뜻이다. 이와 같은 방식으로 다른 설명변수들을 분석해보면, constant를 제외하고 가장 β 값이 작은 변수는 CRDT_CNT 이고 가장 β 값이 큰 변수는 LATE_RATE_1Y이다. 이는 CRDT_CNT 값이 커지면 대출 연체될 확률이 가장 많이 감소하고, LATE_RATE_1Y 값이 커지면 대출 연체될 확률이 가장 많이 증가한다는 의미이다.

Test data의 예측정확도 계산

Testing 데이터에 대한 예측정확도를 알아보기 위해서는 정확도, 민감도, 정밀도에 대해 알아야 한다. 로지스틱 회귀모델의 예측정확도 분석은 cut-off 값에 따른 혼동행렬의 관찰을 통해 이루어진다.

True Positive(TP) : 실제 Y=1이고 Y=1로 예측한 경우

True Negative(TN) : 실제 Y=0이고 Y=0으로 예측한 경우

False Negative(FN) : 실제 Y=1이고 Y=0으로 예측한 경우

Fasle Positive(FP) : 실제 Y=0이고 Y=1로 예측한 경우

	예측 Y=1	예측 Y=0
실제 Y=1	True Positive	False Negative
실제 Y=0	False Positive	True Negative

$$\text{정확도} = \frac{TP+TN}{TP+FN+FP+TN}$$

$$\text{민감도} = \frac{TP}{TP+FN}$$

$$\text{정밀도} = \frac{TP}{TP+FP}$$

```
array([[0.95453621, 0.04546379],
       [0.97173605, 0.02826395],
       [0.99204643, 0.00795357],
       ...,
       [0.92722581, 0.07277419],
       [0.98119566, 0.01880434],
       [0.93405779, 0.06594221]])
```

첫 번째 column은 X값이 주어졌을 때 Y=0일 확률이고 두 번째 column은 X 값이 주어졌을 때 Y=1일 확률이다. 이 값들을 토대로 모델의 성능을 평가하기 위해 cut-off 별 정확도, 민감도, 정밀도를 계산해보았다.

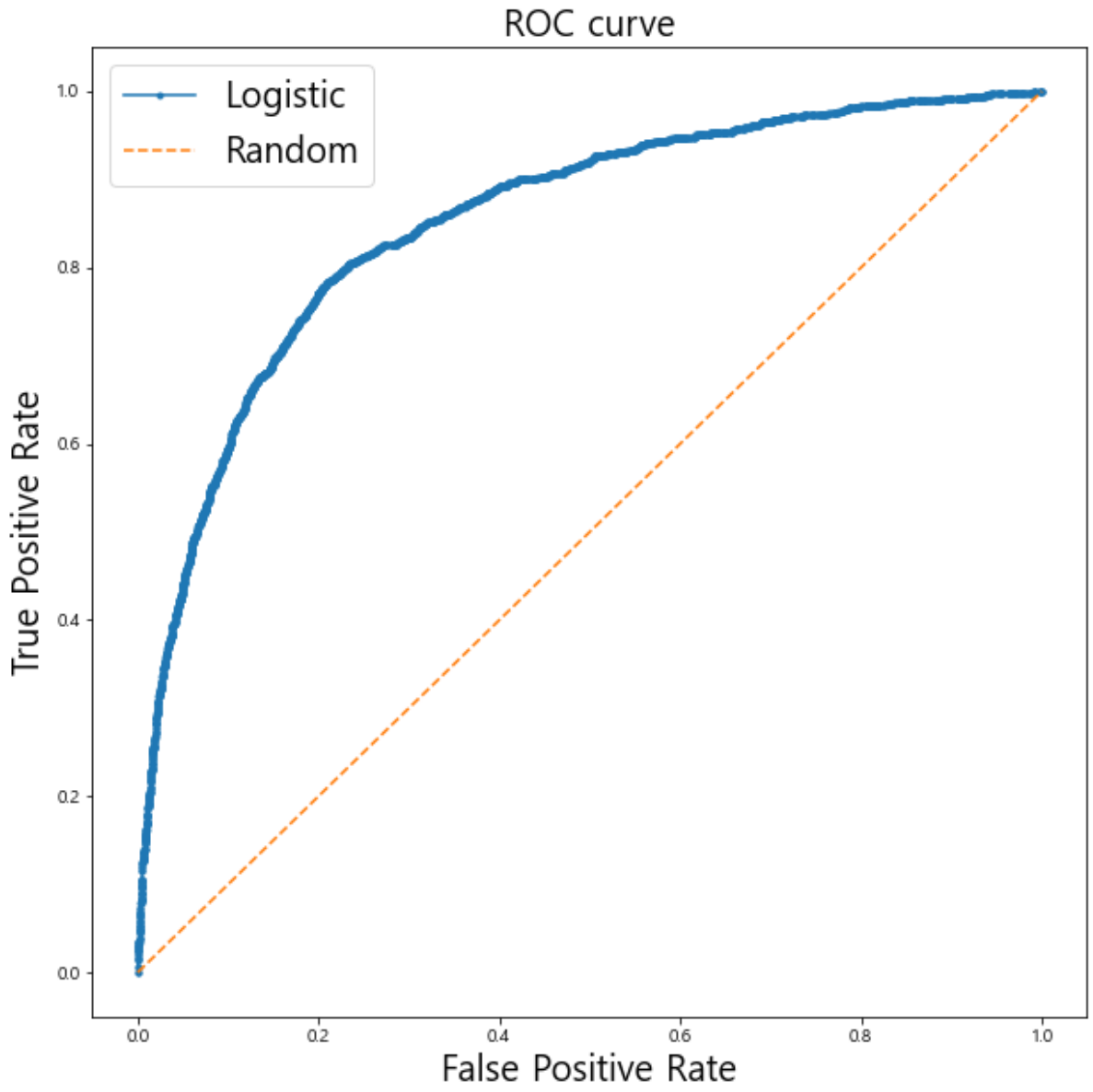
Cut-off 값에 따른 예측정확도 비교

로지스틱 회귀모델은 타겟변수가 0인지 1인지의 확률을 0과 1사이의 확률값으로 제공한다. Cut-off는 사용자가 임의로 정하는 value로써 cut-off보다 Y=1 확률값이 크면 Y=1로 분류하고, cut-off보다 Y=1 확률값이 작으면 Y=0으로 분류한다. 보통 타겟변수가 이진변수일 때 디폴트 cut-off는 0.5이다. Cut-off를 변경해가면서 예측정확도를 계산해보았다.

```
정확도:0.10  |  민감도:1.00  |  정밀도:0.10  |  cut off:0.00
정확도:0.79  |  민감도:0.78  |  정밀도:0.28  |  cut off:0.10
정확도:0.89  |  민감도:0.53  |  정밀도:0.42  |  cut off:0.20
정확도:0.91  |  민감도:0.40  |  정밀도:0.51  |  cut off:0.30
정확도:0.91  |  민감도:0.33  |  정밀도:0.57  |  cut off:0.40
정확도:0.91  |  민감도:0.25  |  정밀도:0.62  |  cut off:0.50
정확도:0.91  |  민감도:0.17  |  정밀도:0.64  |  cut off:0.59
정확도:0.91  |  민감도:0.11  |  정밀도:0.70  |  cut off:0.69
정확도:0.91  |  민감도:0.05  |  정밀도:0.71  |  cut off:0.79
정확도:0.91  |  민감도:0.02  |  정밀도:0.84  |  cut off:0.89
정확도:0.90  |  민감도:0.00  |  정밀도:1.00  |  cut off:0.99
```

이 모델은 대출연체 여부에 대한 분류모델이다. 보통 연체되는 경우가 연체되지 않는 경우에 비해 월등히 적기 때문에, 성공 범주의 비중이 낮다고 판단했다. 따라서 cut-off를

낮게 해서 민감하게 분류를 하는 것이 좋다고 판단했다. Cut-off가 0.2에서 0.1이 되면 민감도가 크게 증가하기 때문에, 0.10 정도가 적당한 cut-off라고 생각한다.



ROC curve를 보면 random 값은 대출연체 여부를 random하게 정한 값이므로 True positive rate와 false positive rate가 동일한 비율을 가지고 있다. 로지스틱 회귀모델을 사용하면 True positive 비율이 커지기 때문에 좋은 모델이라고 할 수 있다.

Odds 예시

$$Odds = \frac{p}{1-p} = \frac{\pi(X=x)}{1-\pi(X=x)} = \frac{P(Y=1)}{P(Y=0)}$$

Odds는 승산이라고도 하며 범주 0에 속할 확률 대비 범주 1에 속할 확률이다. 이런 지표를 사용하는 이유는 Odds 가 취할 수 있는 값의 범위 때문이다. Odds는 0부터 양의 무한대까지 값을 가질 수 있다.

	당뇨	정상	전체
비만	10 (40.0%)	10 (13.3%)	20 (20%)
정상 체중	15 (60.0%)	65 (86.7%)	80 (80%)
전체	25 (100%)	75 (100%)	100 (100%)

[출처] 배정민 (2012). 〈〈닥터 배의 술술 보건의학통계〉〉. 한나래출판사. 120p

이 표는 당뇨 환자 25명과 정상 인구 75명에 대한 비만 여부를 정리한 표이다. 이 표를 통해 당뇨가 비만일 확률과 어떤 관계인지 판단해보았다. Y=1이 비만, Y=0이 정상 체중이라고 하면 당뇨 환자들의 비만 Odds는 10/15 = 0.67 이다. 정상 인구의 비만 Odds는 10/65 = 0.15 이다. 당뇨 환자가 비만일 확률은 $\frac{10}{15+10}$=0.4(40%) 이고, 정상인 사람이 비만일 확률은 $\frac{10}{65+10}$=0.13(13%) 이다. 따라서 당뇨가 있는 환자들이 정상 사람보다 비만일 확률이 많이 높다는 것을 알 수 있다.

산업경영공학부

Industrial and Management Engineering

본인이 좋아하는 활동 3가지

1. 테니스

저는 중 고등학교 때 동아리로 테니스부에 있었을 정도로 테니스를 매우 좋아했습니다. 제가 미국에서 살았던 초등학교 때 처음 배웠었는데, 그때까지 배웠던 다른 구기종목과는 또 다른 매력을 느끼게 되었습니다. 그래서 열심히 배우게 되었고, 한국에 와서도 꾸준히 배우면서 실력이 늘게 되니 더 재미있었습니다. 지금은 시간이 부족하고 코로나 이슈 때문에 자주 치지는 못하지만, 여전히 매우 좋아하는 활동입니다.

2. 축구 시청

저는 비록 축구를 잘하지는 못하지만 축구 중계를 보는 것을 매우 좋아합니다. 축구 국가대표 경기나 EPL 경기를 챙겨보는 것이 제 취미 중 하나라고 생각합니다. 특히 2018년부터 토트넘의 챔피언스리그 결승 진출과 2018 러시아월드컵을 계기로 축구를 보는 것이 정말 즐거워졌습니다.

3. 클래식 음악 감상

저는 어릴 때는 대중음악을 주로 들었지만, 조금씩 나이가 들면서 클래식 음악에 관심을 가지기 시작했습니다. 특히 쇼팽의 피아노협주곡을 듣고 나서 피아노곡에 깊은 관심을 가지게 되었고, 피아노곡을 듣고 있으면 마음이 편안해졌습니다. 저는 피아노를 잘 치지

는 못하지만, 언젠가는 꾸준히 연습해서 화려하게 피아노를 치는 것을 버킷리스트로 삼고 있습니다.