예측애널리틱스

과제 #2

산업경영공학부

2015170852 이무원

0. 각종 분석을 통해 데이터 파악하시오. (예, 변수 별 통계치, 변수간의 상관관계, 변수들의 분포, 등등)

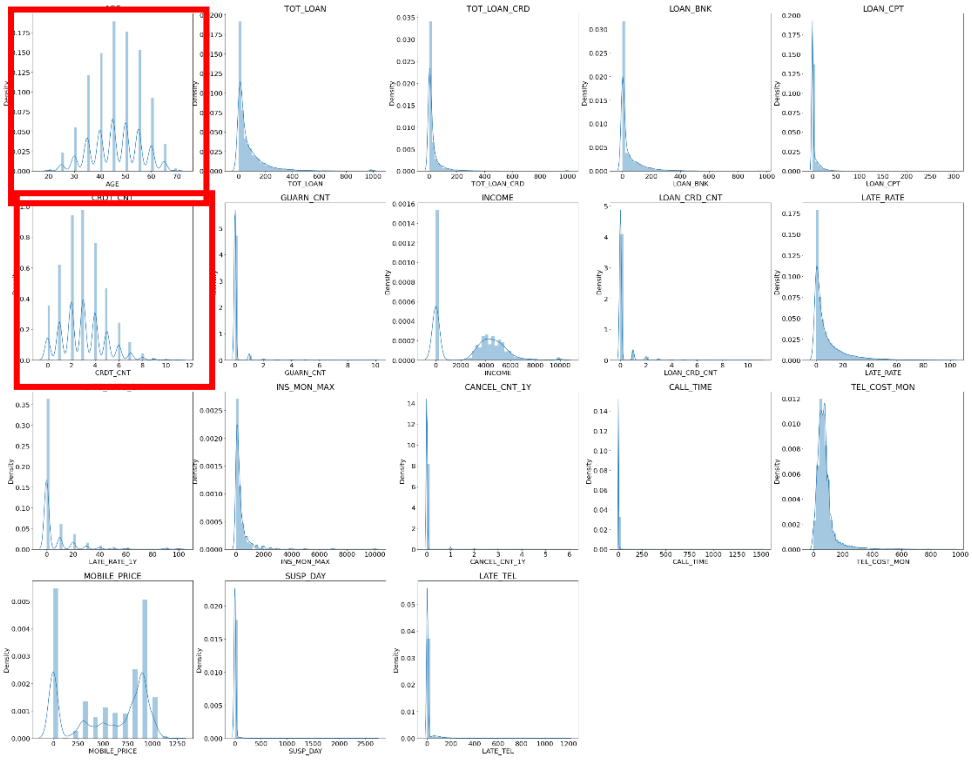
- 변수 별 통계치

	AGE	TOT_LOAN	TOT_LOAN_CRD	LOAN_BNK	LOAN_CPT	CRDT_CNT	GUARN_CNT	INCOME	LOAN_CRD_CNT
count	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000
mean	46.250980	82.019407	32.829622	49.324897	4.288641	3.004264	0.098695	2778.629051	0.163855
std	9.693741	126.702976	83.419760	92.443944	12.660968	1.842478	0.529664	2470.097227	0.617522
min	20.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	40.000000	12.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000
50%	45.000000	36.000000	9.000000	9.000000	0.000000	3.000000	0.000000	3600.000000	0.000000
75%	55.000000	102.000000	27.000000	60.000000	3.000000	4.000000	0.000000	4700.000000	0.000000
max	70.000000	994.000000	994.000000	944.000000	301.000000	11.000000	10.000000	10000.000000	11.000000

LATE_RATE	LATE_RATE_1Y	INS_MON_MAX	CANCEL_CNT_1Y	CALL_TIME	TEL_COST_MON	MOBILE_PRICE	SUSP_DAY	LATE_TEL
43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000	43386.000000
8.216406	6.389619	373.254506	0.024662	2.098970	75.477804	534.423547	18.433320	13.757664
12.120840	14.556618	690.067030	0.206476	15.364253	62.311464	382.237230	133.523351	53.272289
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	70.000000	0.000000	0.430000	40.000000	0.000000	0.000000	0.000000
3.000000	0.000000	190.000000	0.000000	0.990000	60.000000	700.000000	0.000000	0.000000
11.000000	10.000000	390.000000	0.000000	1.887500	80.000000	900.000000	0.000000	0.000000
100.000000	100.000000	10000.000000	6.000000	1520.000000	950.000000	1200.000000	2700.000000	1200.000000

숫자형 변수들의 통계량입니다. GUARN_CNT, LOAN_CRD_CNT, CANCEL_CNT_1Y, SUSP_DAY, LATE_TEL 5개의 변수를 보면 75%까지 모두 0이나 max 값은 꽤 큰 차이를 보이고 있습니다. 다른 변수들에 비해 분포가 많이 치우쳐져 있을 것이라고 생각되었습니다.

- 숫자형 변수 분포

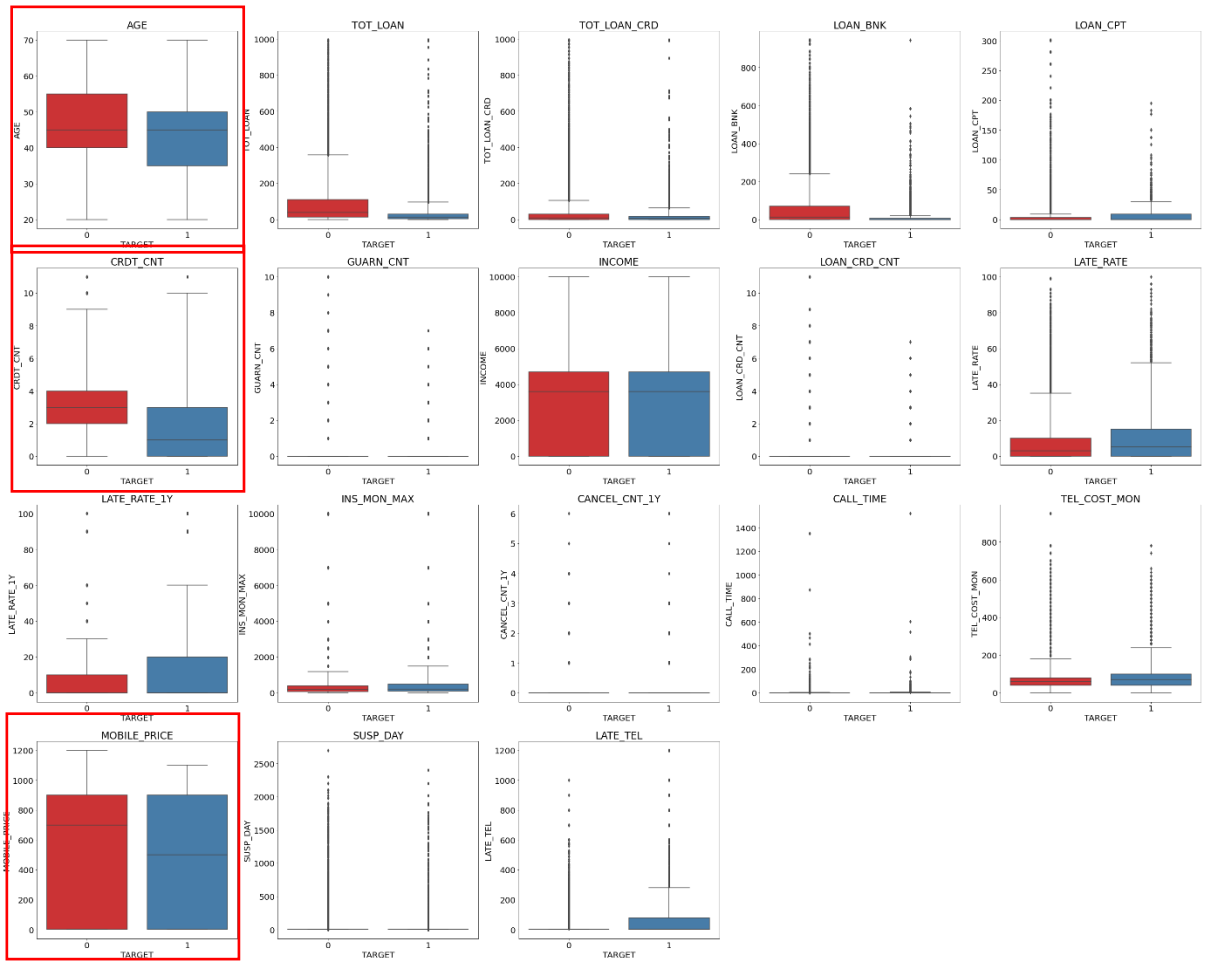


숫자형 변수들의 히스토그램입니다. 변수들이 한쪽에 치우쳐져 있었으며, 이 외에도 분포가 한곳에 치우쳐져 있는 경우가 많았습니다. 변수들의 분포가 고른 경우엔 Standard Scaler를 사용하고, 그렇지 않은 경우엔 Robust Scaler를 사용하여 이를 해결하려 했습니다. 빨간색 박스 부분이 분포가 고르다고 판단한 경우이며, AGE와 CRDT_CNT에 해당합니다.

AGE, CRDT_CNT → Standard Scaler

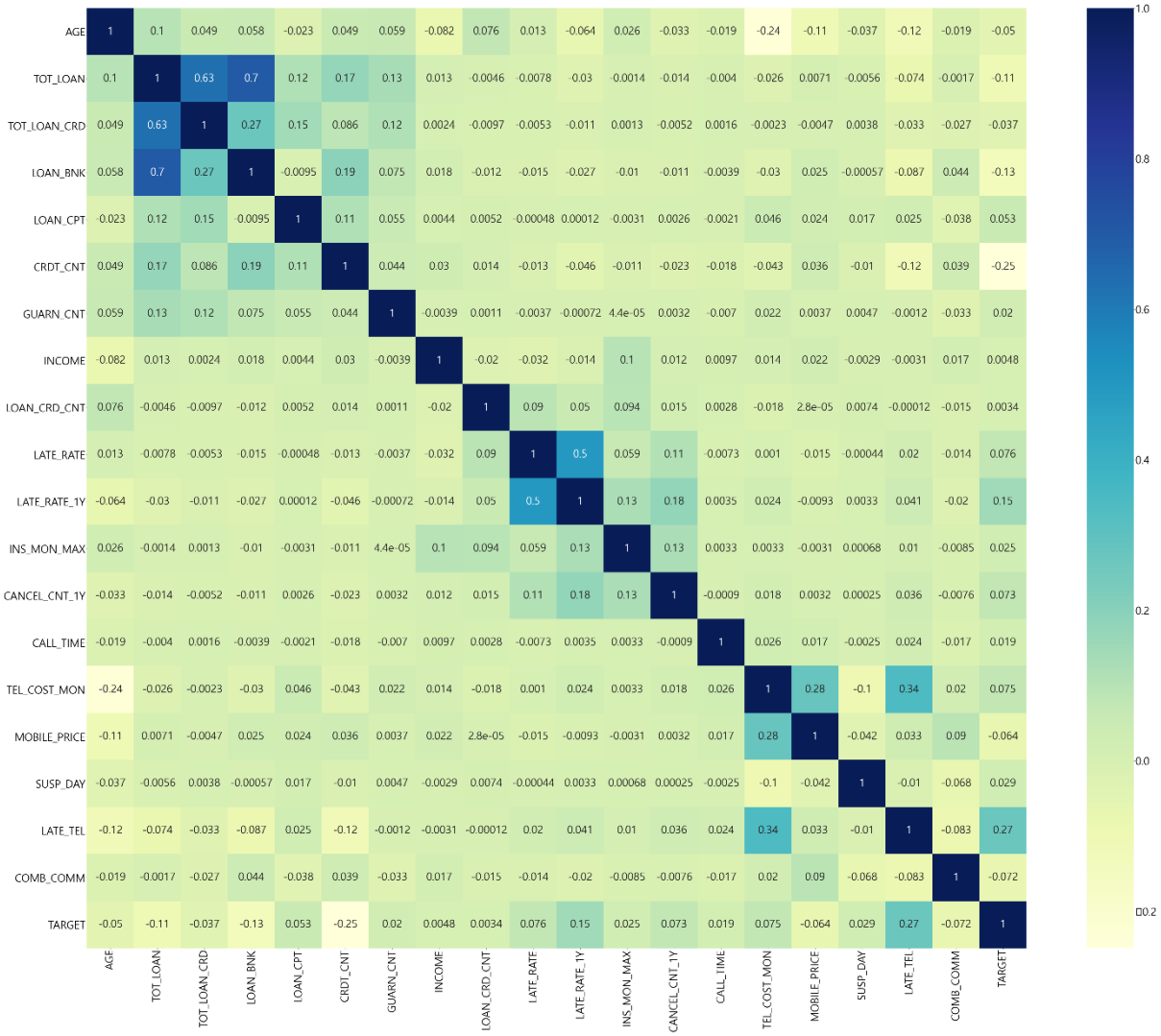
그 외 칼럼 → Robust Scaler

- 숫자형 칼럼들과 타겟 칼럼의 분포



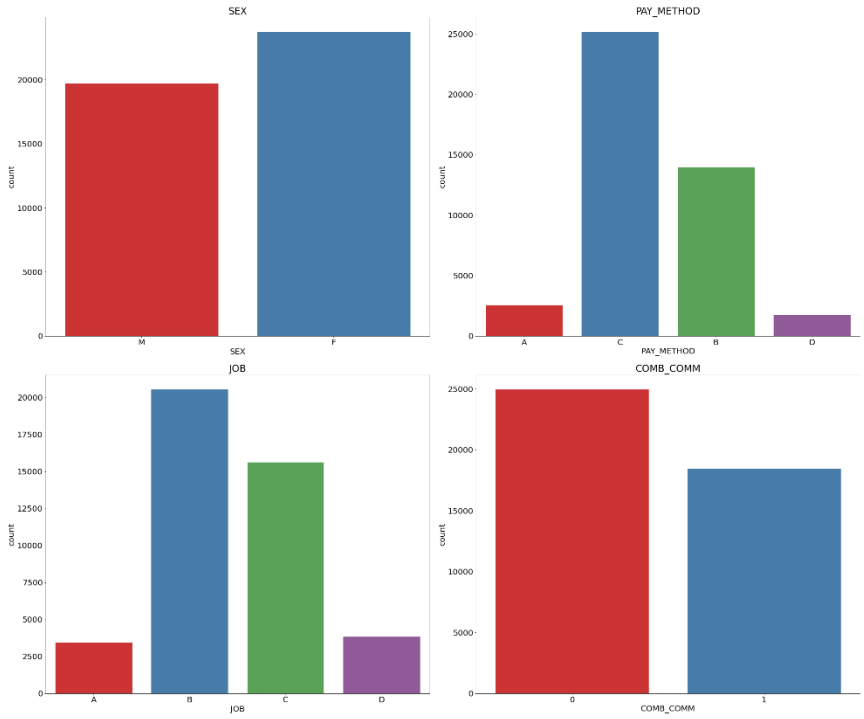
숫자형 칼럼들을 Y축으로, X 축은 TARGET 칼럼으로 하여 그린 box plot입니다. TARGET이 다른 경우에 칼럼들의 통계량 차이가 클수록 도움이 많이 될 것으로 예상하였으며, 그래프 상으로 빨간 박스 부분이 두 평균의 차이가 커 도움이 될 것으로 예상하였습니다. AGE, CRDT_CNT, MOBILE_PRICE가 이에 해당합니다.

- 변수들의 상관관계



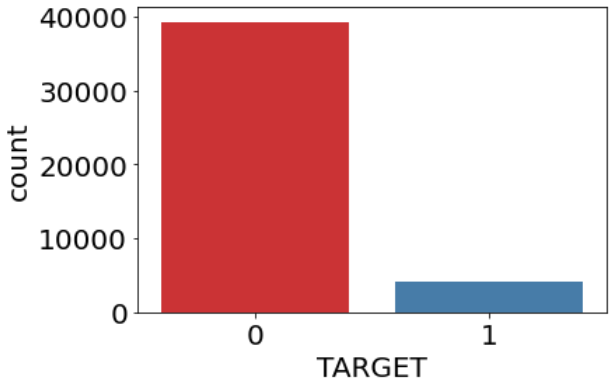
숫자형 변수들의 상관관계를 Heatmap으로 그려보았습니다. TOT_LOAN, TOT_LOAN_CRD, LOAN_BNK 3가지 변수들의 상관관계가 높았습니다.

- Category 형 칼럼의 분포



카테고리형 칼럼들의 분포입니다. 각 칼럼의 class별 개수를 세어 plotting 하였습니다.

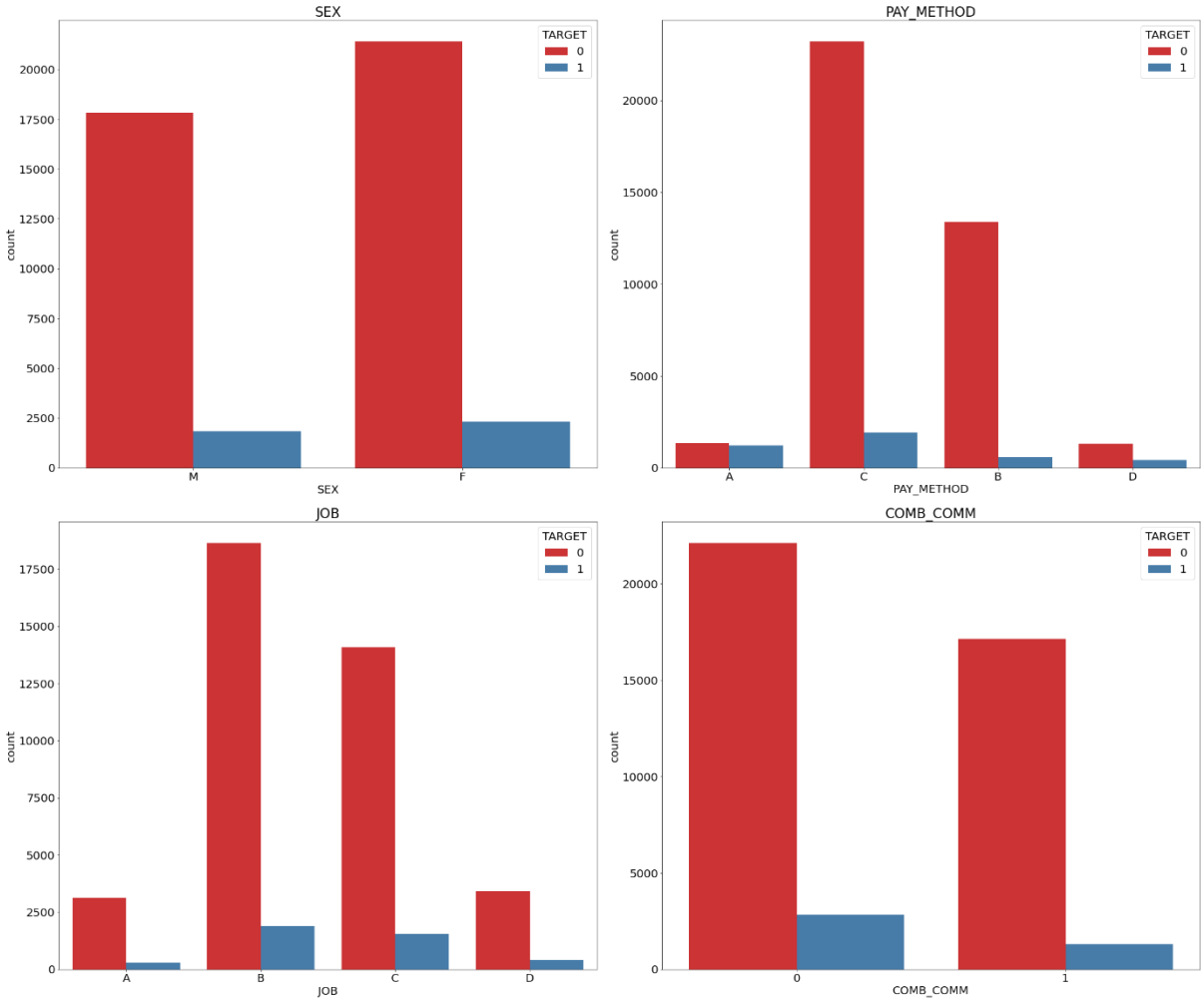
- TARGET 칼럼의 분포



-

TARGET 칼럼의 분포입니다. 대출 연체가 발생(1)한 경우보다 발생하진 않은 경우가 훨씬 많았으며, class가 많이 imbalance한 상황입니다.

- TARGET 칼럼과 category 칼럼들의 분포



TARGET 칼럼의 class 분포와 다른 분포를 가질수록 도움이 많이 될 것이라고 생각합니다. 예를 들면, TARGET의 분포와 비교해보면 PAY_METHOD의 A가 개수는 적지만 도움이 될 것이라고 생각되는데, A일 경우 다른 때보다 훨씬 대출 연체가 발생한 비율이 높기 때문입니다.

1. 로지스틱 회귀모델 구축하고 해석하시오.

- 변수 전처리

Category형 칼럼은 one-hot encoding을 진행하고, 숫자형 칼럼들의 경우 AGE, CRDT_CNT는 standard scaler, 이 외의 칼럼들은 robust scaler를 사용했습니다.

- 모델 해석

	beta	exp(beta)	interpret
const	-1.15	0.32	protective
AGE	0.01	1.01	risky
TOT_LOAN	-0.27	0.76	protective
TOT_LOAN_CRD	0.05	1.05	risky
LOAN_BNK	-0.68	0.51	protective
LOAN_CPT	0.04	1.04	risky
CRDT_CNT	-0.86	0.42	protective
GUARN_CNT	0.06	1.06	risky
INCOME	0.15	1.16	risky
LOAN_CRD_CNT	-0.07	0.93	protective
LATE_RATE	0.0	1.0	protective
LATE_RATE_1Y	0.19	1.21	risky
INS_MON_MAX	0.03	1.03	risky
CANCEL_CNT_1Y	0.07	1.07	risky
CALL_TIME	-0.0	1.0	protective
TEL_COST_MON	-0.05	0.95	protective
MOBILE_PRICE	-0.01	0.99	protective
SUSP_DAY	0.0	1.0	protective
LATE_TEL	0.01	1.01	risky
COMB_COMM	-0.46	0.63	protective
SEX_F	-0.64	0.53	protective
SEX_M	-0.5	0.61	protective
PAY_METHOD_A	0.25	1.28	risky
PAY_METHOD_B	-0.66	0.52	protective
PAY_METHOD_C	-0.72	0.49	protective
PAY_METHOD_D	-0.02	0.98	protective
JOB_A	-0.17	0.84	protective
JOB_B	-0.43	0.65	protective
JOB_C	-0.39	0.68	protective
JOB_D	-0.15	0.86	protective

beta값이 0보다 작아질 경우엔 exp(beta)가 0~1사이로 odds값이 1 미만 증가하여 TARGET=1일 확률이 낮습니다. 반대인 beta값이 0보다 클 경우엔 exp(beta)가 1 이상의 값으로 odds값이 1 이상 증가하여 TARGET=1일 확률이 급증하게 됩니다. 칼럼에서 TARGET=1(대출을 연체하는 경우)의 확률이 급증하는 경우엔 risky로, 아닌 경우엔 protective로 구분이 되어 있습니다.

- Protective

```
Index(['const', 'TOT_LOAN', 'LOAN_BNK', 'CRDT_CNT', 'LOAN_CRD_CNT',
       'LATE_RATE', 'CALL_TIME', 'TEL_COST_MON', 'MOBILE_PRICE', 'SUSP_DAY',
       'COMB_COMM', 'SEX_F', 'SEX_M', 'PAY_METHOD_B', 'PAY_METHOD_C',
       'PAY_METHOD_D', 'JOB_A', 'JOB_B', 'JOB_C', 'JOB_D'],
```

- Risky

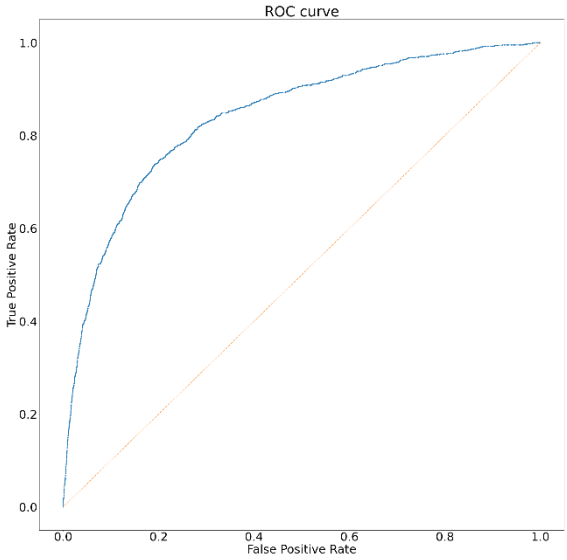
```
Index(['AGE', 'TOT_LOAN_CRD', 'LOAN_CPT', 'GUARN_CNT', 'INCOME',
       'LATE_RATE_1Y', 'INS_MON_MAX', 'CANCEL_CNT_1Y', 'LATE_TEL',
       'PAY_METHOD_A'],
```

각각에 해당하는 칼럼은 위와 같으며 PAY_METHOD_A의 경우 위의 시각화에서 보았듯 TARGET=1인 비율이 TARGET 칼럼의 전체 비율보다 높았으므로 Risky로 나온 것 또한 확인할 수 있었습니다.

2. 데이터를 Training set과 Testing set으로 나누고 Testing 데이터의 예측정확도 계산하시오.

Test 데이터 비율을 30%로 하여 Train set과 Test set을 나누었습니다.

정확도 (Accuracy)	민감도 (Recall)	정밀도 (Precision)
0.91	0.14	0.61



정확도의 경우 0.91로 높은 값을 기록했지만, 민감도(recall)은 0.14로 매우 낮을 값을 기록했고, 정밀도 또한 0.61로 정확도에 비해 많이 낮았습니다. 정확도가 높고 민감도가 낮은 이유는 현재 class imbalance 문제가 커 TARGET=0인 것을 많이 맞추어 Accuracy가 높아진 것으로 보이며 TARGET=1인 경우는 14%밖에 검출을 하지 못한 상황입니다. ROC 커브 또한 그려보았는데, ROC 커브의 면적이 0.5를 넘었으므로 유효한 모델이라고 할 수 있습니다.

3. 분류 Cut off 값 (디폴트=0.5)을 바꾸어 가며 예측정확도 계산하고 적절한 Cut off값을 제시해 보시오.

```
Accuracy:0.09  |  Recall:1.00  |  Precision:0.09  |  cut off:0.00
Accuracy:0.78  |  Recall:0.76  |  Precision:0.26  |  cut off:0.10
Accuracy:0.89  |  Recall:0.52  |  Precision:0.41  |  cut off:0.20
Accuracy:0.91  |  Recall:0.33  |  Precision:0.51  |  cut off:0.30
Accuracy:0.91  |  Recall:0.23  |  Precision:0.57  |  cut off:0.40
Accuracy:0.91  |  Recall:0.14  |  Precision:0.61  |  cut off:0.50
Accuracy:0.91  |  Recall:0.07  |  Precision:0.60  |  cut off:0.60
Accuracy:0.91  |  Recall:0.04  |  Precision:0.65  |  cut off:0.70
Accuracy:0.91  |  Recall:0.02  |  Precision:0.70  |  cut off:0.80
Accuracy:0.91  |  Recall:0.01  |  Precision:0.92  |  cut off:0.90
Accuracy:0.91  |  Recall:0.00  |  Precision:0.00  |  cut off:1.00
```

Cut off가 낮아질수록 class 1로 예측을 많이 하게 되므로 Recall값이 올라가지만, 반대로 Accuracy는 떨어지고, precision 또한 class 1로 예측을 많이 하게 되어 FP가 늘어날 확률이 높아져 떨어지고 있습니다. 기업의 경우 대출이 연체될 경우 타격이 더 클 것이므로 Accuracy보단 Recall값이 조금 더 중요해 보이지만, 그렇다고 Accuracy가 너무 작은 경우를 고를 수도 없는 상황입니다. 따라서 Accuracy가 가장 높은 cut off에 비해 0.02(0.91 → 0.89) 낮지만, Recall값이 그 전 cut off에 비해 0.19(0.33 → 0.52) 높은 0.20이 적당하다고 생각합니다.

4. Odd의 적절한 예시를 기술하시오. (수업시간에 소개한 예 제외하고)

Odd 관련 예시를 찾아보던 중, 상대위험도와 오즈비가 많이 비교되는 것을 보고 오즈비를 사용하는 경우가 어떤 경우인지에 맞춰 기술하겠습니다.

먼저, 담배피는 사람이 피지 않는 사람에 비해 4배 더 발병확률이 높은 질병이 있다고 가정하겠습니다. 담배를 핀 사람과 피지 않은 사람을 각각 100명씩 sampling 한 Table이 밑에 있습니다.

	disease	non-disease	
smoker	20 (a)	80 (b)	100
non-smoker	5 (c)	95 (d)	100
	25	175	200

상대위험도(PR) = (20/100)/(5/100) = 4

오즈비(odds ratio) = (20/80)/(5/95) = 4.8

상대위험도와 오즈비를 계산해보면 비슷한 값이 나왔으나, 4배 더 발병확률이 높은 질병이고 실제로 딱 4배 더 많은 사람이 발병했지만, 오즈비의 경우 4가 넘는 값이 나왔습니다. 이렇게 값이 과대평가 되는 등 왜곡이 일어날 수 있는 것이 오즈비의 단점이라고 할 수 있습니다.

한번 더 sampling하여 계산해보겠습니다. 이번엔 병에 걸린사람 100명과 걸리지 않은 사람 100명을 sampling하여 case-control study를 해보겠습니다. (환자군과 대조군으로 나눈 것을 case-control study라고 합니다.)

	disease	non-disease	
smoker	80 (a)	45 (b)	125
non-smoker	20 (c)	55 (d)	75
	100	100	200

상대위험도(PR) = (80/125)/(20/75) = 2.4

오즈비(odds ratio) = (80/45)/(20/55) = 4.91

이번의 경우엔 상대위험도가 2.4로 나왔습니다. case-control study에서는 환자군과 대조군을 각각 100명씩 뽑았으므로 PR에 왜곡이 일어날 수 밖에 없으며 이경우엔 오즈비가 더 유효함을 알 수 있습니다. 이렇게 case-control study에 주로 쓰이는 것이 오즈비라고 할 수 있습니다.

5. 로지스틱 회귀모델에서 중요한 변수를 선택하는 방법론에 대해 공부하고 요약하시오.

	beta	exp(beta)	interpret
const	-1.15	0.32	protective
AGE	0.01	1.01	risky
TOT_LOAN	-0.27	0.76	protective
TOT_LOAN_CRD	0.05	1.05	risky
LOAN_BNK	-0.68	0.51	protective
LOAN_CPT	0.04	1.04	risky
CRDT_CNT	-0.86	0.42	protective
GUARN_CNT	0.06	1.06	risky
INCOME	0.15	1.16	risky
LOAN_CRD_CNT	-0.07	0.93	protective
LATE_RATE	0.0	1.0	protective
LATE_RATE_1Y	0.19	1.21	risky
INS_MON_MAX	0.03	1.03	risky
CANCEL_CNT_1Y	0.07	1.07	risky
CALL_TIME	-0.0	1.0	protective
TEL_COST_MON	-0.05	0.95	protective
MOBILE_PRICE	-0.01	0.99	protective
SUSP_DAY	0.0	1.0	protective
LATE_TEL	0.01	1.01	risky
COMB_COMM	-0.46	0.63	protective
SEX_F	-0.64	0.53	protective
SEX_M	-0.5	0.61	protective
PAY_METHOD_A	0.25	1.28	risky
PAY_METHOD_B	-0.66	0.52	protective
PAY_METHOD_C	-0.72	0.49	protective
PAY_METHOD_D	-0.02	0.98	protective
JOB_A	-0.17	0.84	protective
JOB_B	-0.43	0.65	protective
JOB_C	-0.39	0.68	protective
JOB_D	-0.15	0.86	protective

변수들의 beta값을 살펴보면 중요도를 알 수 있습니다. Risky한 변수의 경우 odds값이 X가 1단위만큼 증가하였을 때 1이상 증가하여 급증하게 되는데, 이러한 변수들이 TARGET을 나누는 데 큰 도움이 되는 변수라고 생각합니다. 지금 모델을 보면 TARGET이 0인 경우가 90%, TARGET이 1인 경우가 10%인데 여기서 90%를 잘 예측하게 해주는 변수들 보다는 TARGET이 1이 되도록, odds비가 1이상이 되는 경우를 만들어주는 변수들이 더 중요하다고 생각합니다.

더 나아가서 여기서 오즈비를 단위당 1이상 커지게 하는 변수더라도 전체 모델의 성능이 좋지 않다면 중요 변수라고 하기 어려울 것입니다. 어느정도 예측모델의 성능이 나오는 상태에서는 exp(beta)값이 큰 변수들이 중요한 변수가 될 것이고, 그렇지 않은 경우에는 오히려 성능을 떨어뜨리는 변수가 될 수도 있습니다. 이렇게 모델의 성능과 exp(beta)값을 동시에 보면서 중요 변수를 찾아내야 정확한 변수를 찾을 수 있으므로 두 가지를 동시에 기록하며 관찰해야 한다고 생각합니다.

6. 여러분들 스스로 로지스틱 회귀분석 관련 문제 하나를 제출하고 풀어 보시오.

- 모델 구축 전 전처리

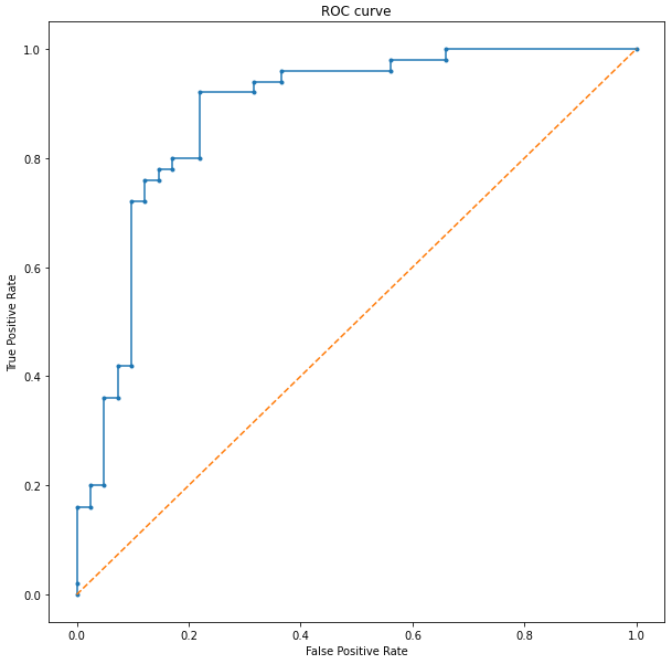
심장 질병에 관련된 데이터로, 발병하지 않은 경우가 0, 발병한 경우가 1인 데이터 입니다.

독립변수로는 age, sex, cp, trestbps, chol, , fbs, restecg, thalach, exang, oldpeak, slope, ca, thal가 있습니다. Sex와 fbs 변수는 바이너리 변수이며, 다른 변수들은 모두 숫자형 변수입니다. 숫자형 데이터셋에 standard scaler를 적용하여 전처리 하였습니다.

Train과 test의 비율은 70%와 30%로 정하였습니다.

- Logistic Regression 모델 구축

	beta	exp(beta)	interpret
const	-0.07	0.93	protective
age	0.03	1.03	risky
sex	-0.65	0.52	protective
cp	0.83	2.29	risky
trestbps	-0.15	0.86	protective
chol	-0.12	0.89	protective
fbs	0.21	1.23	risky
restecg	0.38	1.46	risky
thalach	0.37	1.45	risky
exang	-0.59	0.55	protective
oldpeak	-0.53	0.59	protective
slope	0.57	1.77	risky
ca	-1.1	0.33	protective
thal	-0.76	0.47	protective



Protective

```
Index(['const', 'sex', 'trestbps', 'chol', 'exang', 'oldpeak', 'ca', 'thal'],
```

Risky

```
Index(['age', 'cp', 'fbs', 'restecg', 'thalach', 'slope']
```

Protective 변수와 Risky 변수는 위와 같으며, ROC 커브의 면적이 0.5가 넘었으므로 의미 있는 모델이라 할 수 있습니다.

- Cut off 설정

```
Accuracy:0.55  |  Recall:1.00  |  Precision:0.55  |  cut off:0.00
Accuracy:0.81  |  Recall:0.92  |  Precision:0.78  |  cut off:0.10
Accuracy:0.85  |  Recall:0.92  |  Precision:0.82  |  cut off:0.20
Accuracy:0.85  |  Recall:0.90  |  Precision:0.83  |  cut off:0.30
Accuracy:0.81  |  Recall:0.84  |  Precision:0.82  |  cut off:0.40
Accuracy:0.81  |  Recall:0.84  |  Precision:0.82  |  cut off:0.50
Accuracy:0.80  |  Recall:0.80  |  Precision:0.83  |  cut off:0.60
Accuracy:0.80  |  Recall:0.72  |  Precision:0.90  |  cut off:0.70
Accuracy:0.71  |  Recall:0.56  |  Precision:0.88  |  cut off:0.80
Accuracy:0.62  |  Recall:0.36  |  Precision:0.86  |  cut off:0.90
Accuracy:0.45  |  Recall:0.00  |  Precision:0.00  |  cut off:1.00
```

질병에 관련된 예측이기에 Recall값이 가장 중요하다고 생각합니다. 질병이 있음에도 없다고 진단할 경우의 위험도가 가장 크기 때문에 Recall 값이 가장 큰 cut off 값인 0.10과 0.20 중에 cut off를 설정하는 것이 타당합니다.

이제부터는 Accuracy와 Precision을 생각해야 하는데, cut off의 값이 0.20일 때 Accuracy와 Precision이 모두 높으므로 cut off는 0.20이 타당합니다.