DACA

NGO 머신러닝 : 1주차

2020.12.16 ~ 2020.12.31

빅데이터 응용학과 석사 1기 양선욱

목차

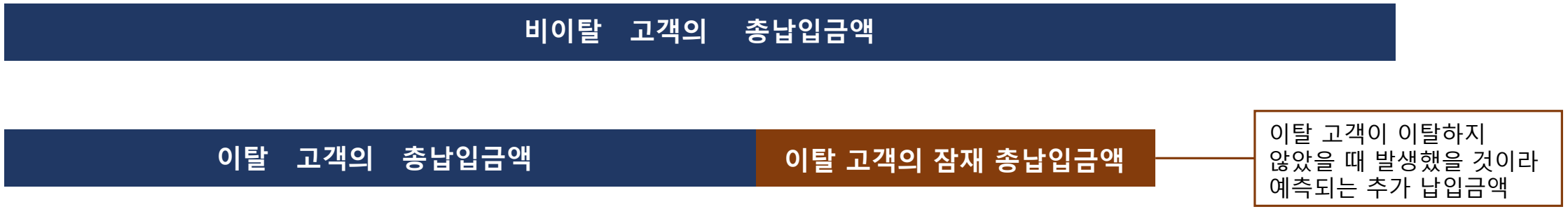
분석의 목적

* 분석 목적

- 이탈 고객을 복귀시키기 위한 프로젝트가 구상 중이다.
프로젝트의 유효성 여부를 결정하기 위해 이탈 고객이 복귀 했을 때 얻을 실리적 이익과 프로젝트에 들어가는 비용을 비교할 필요가 있다.
따라서, 현재 고객 데이터를 바탕으로 총납입금액을 추론하는 학습 모델을 개발한다.
이 모델에 이탈 고객의 데이터를 삽입시켜 이탈 고객이 이탈하지 않았을 때 발생할 잠재 총납입금액을 도출한다.

* 분석에 앞서

구분		총납입금액(원)	비율(%)
비이탈 고객 701명 80.9%	총액	225,935,000	84%
	평균	322,303	
이탈 고객 165명 19.1%	총액	42,751,100	16%
	평균	259,097	
총합 866명	총액	268,686,100	100%



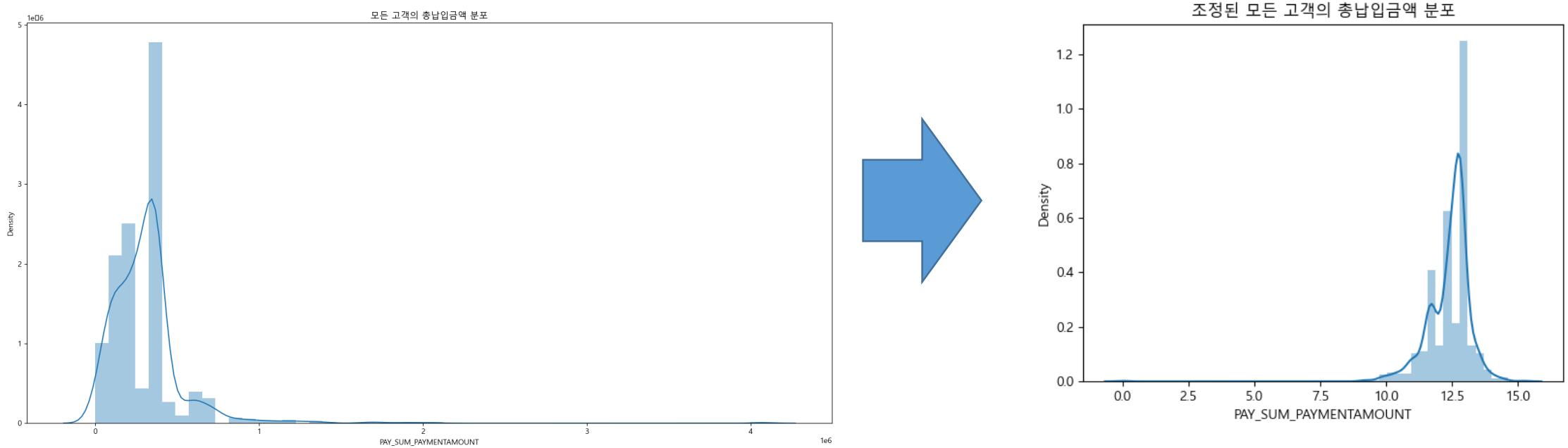
분석 과정

* 사용 데이터

- 독립변수 : 나이, 가입나이, 성별, 가입일수, 납부횟수, 미납율
- 종속 변수 : 총납입금액

* 데이터 전처리

- 종속 변수인 총납입금액의 분포가 한쪽으로 하게 치우쳐져 있어서 로그 변환을 사용
- 독립변수 표준화와 명목형에 대한 원핫인코딩




```
df1['PAY_SUM_PAYMENTAMOUNT'] = np.log1p(df1['PAY_SUM_PAYMENTAMOUNT'])
```

분석 과정

* 분석 과정

① 데이터 분할

비이탈 고객 701명 80.9%

이탈 고객 165명 19.1%

② 모델 학습

③ 모델 평가

④ 이탈 고객 데이터 삽입

완성된 모델

⑤ 이탈 고객의 잠재 총납입금액 도출

⑥ 프로젝트 경비와 비교

프로젝트 수행 비용

이탈 고객의 잠재 총납입금액

1. 표준 선형 회귀분석

* 분석 결과

학습용 데이터 세트 결정계수 : 0.42

평가용 데이터 세트 결정계수 : 0.3

RMSE : 0.5451

절편	12.468
나이	2.046
가입나이	-2.06
가입 일수	-0.639
미납율	-0.112
납부횟수	0.396
남성	0.021
여성	-0.021

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	44,381,774	+3.8%
	평균	268,980	
	잠재	+1,630,674	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	270,316,774	+0.6%

* 실무적 인사이트

- 현재까지의 모든 이탈 고객의 잠재 총납입금액의 총액은 163만원 가량으로 기존 이탈고객의 총납입금액의 3.8%에 해당하는 수치이다.

- 이탈->비이탈 총납입금액 총액의 상승량이 있다고 해도 이탈 고객의 수가 전체의 19.8%이기 때문에 전체 고객의 총납입금액에 대한 상승폭은 0.6%에 그친다.

- 이탈한 고객의 복귀에 대한 효과는 미비한 것으로 사료되며 해당 프로젝트에 대한 결정을 재차 확인할 필요가 있다.

- 학습용 데이터 세트의 결정계수가 평가용에 비해 높게 나타났다. 과잉적합이 의심되며 다른 선형 회귀분석을 통한 확인이 필요하다.

2&3. 릿지&라쏘 선형 회귀분석

* 릿지 분석 결과

학습용 데이터 세트 결정계수 : 0.42

평가용 데이터 세트 결정계수 : 0.3

alpha = 1

RMSE : 0.5439

절편	12.467
나이	0.18
가입나이	-0.197
가입 일수	-0.006
미납율	-0.112
납부횟수	0.396
남성	0.019
여성	-0.019

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	44,381,774	+3.8%
	평균	268,980	
	잠재	+1,630,674	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	270,316,774	+0.6%

* 라쏘 분석 결과

학습용 데이터 세트 결정계수 : 0.42

평가용 데이터 세트 결정계수 : 0.3

alpha = 0.001

max_iter=10000

RMSE : 0.5439

절편	12.451
나이	0
가입나이	-0.017
가입 일수	0.054
미납율	-0.111
납부횟수	0.396
남성	0.033
여성	0

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	44,381,774	+3.8%
	평균	268,980	
	잠재	+1,630,674	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	270,316,774	+0.6%

선형 회귀분석 모델 비교

* 회귀분석 모델 비교

회귀분석 모델	RMSE
표준	0.5451
릿지 a=1	0.5439
라쏘 a=0.001	0.5439

- 매우 근소하지만 표준 모델보다 릿지, 라쏘 모델이 우수하다.
- 대신 최적의 alpha값을 찾는 것에 노력과 시간이 소모된다.

* 회귀분석 가중치 비교

표준		릿지 a=1		라쏘 a=0.001	
절편	12.468	절편	12.467	절편	12.451
나이	2.046	나이	0.18	나이	0
가입나이	-2.06	가입나이	-0.197	가입나이	-0.017
가입 일수	-0.639	가입 일수	-0.006	가입 일수	0.054
미납율	-0.112	미납율	-0.112	미납율	-0.111
납부횟수	0.396	납부횟수	0.396	납부횟수	0.396
남성	0.021	남성	0.019	남성	0.033
여성	-0.021	여성	-0.019	여성	0

- 릿지 모형과 라쏘 모형의 가중치를 보아 총납부금액을 예측함에 있어서 나이는 중요한 변수가 아니라고 판단된다.
- 릿지 모형의 경우 alpha값이 작아 독립변수의 영향력을 대부분 유지시켰다.
- 반대로 라쏘 모형의 경우 alpha값이 작아 독립 변수의 영향력을 대부분 유지시켰다.
- 릿지 모형과 라쏘 모형의 alpha값에 의하여 데이터에 과잉적합은 없다고 판단된다.

* 전처리 추가 사항

- k-nn의 특성을 고려해 독립변수를 정규화

* 최적의 k 구하기

학습용 데이터의 결정계수 최대

거리	유클리드
K개수	3
학습용 결정계수	0.7
평가용 결정계수	0.24
RMSE	0.5662

- 결정계수 차이가 큼
- 평가용 결정계수는 0.2를 겨우 넘는 수준이다.

학습용, 평가용 데이터의 결정계수 차 최소

거리	맨해튼
K개수	13
학습용 결정계수	0.46
평가용 결정계수	0.33
RMSE	0.5316

- 결정계수 간의 차이는 작으나 결정계수 값이 타 모델 보다 낮다.
- RMSE가 특별히 낮은 것은 아니다.

RMSE 최소

거리	유클리드
K개수	10
학습용 결정계수	0.54
평가용 결정계수	0.37
RMSE	0.5180

모델 채택

* 비교된 모델

거리	k	학습용 결정계수	평가용 결정계수	RMSE
유클리드	2	0.78	0.18	0.58998217
	3	0.7	0.24	0.56622018
	4	0.67	0.29	0.54850826
	5	0.65	0.3	0.54608442
	6	0.62	0.31	0.54170081
	7	0.59	0.33	0.53131338
	8	0.57	0.32	0.53532843
	9	0.55	0.35	0.52270401
	10	0.54	0.37	0.51800690
	11	0.52	0.36	0.52191192
	12	0.5	0.36	0.52226683
	13	0.48	0.34	0.52777000
	14	0.47	0.34	0.52865997

거리	k	학습용 결정계수	평가용 결정계수	RMSE
맨해튼	2	0.79	0.13	0.60703542
	3	0.7	0.27	0.55751011
	4	0.67	0.29	0.54721227
	5	0.64	0.28	0.55089895
	6	0.6	0.33	0.53160273
	7	0.59	0.31	0.52727019
	8	0.56	0.34	0.52736761
	9	0.55	0.36	0.52135391
	10	0.52	0.35	0.52327505
	11	0.5	0.35	0.52404317
	12	0.49	0.34	0.52785328
	13	0.46	0.33	0.53161142
	14	0.45	0.32	0.53636331

4. K-NN

* 분석 결과

K = 10

거리 : 유클리드

학습용 데이터 세트 결정계수 : 0.54

평가용 데이터 세트 결정계수 : 0.37

RMSE : 0.5180

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	44,011,400	+2.9%
	평균	266,735	
	잠재	+1,260,300	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	269,946,400	+0.4%

* 실무적 인사이트

- 현재까지의 모든 이탈 고객의 잠재 총납입금액의 총액은 126만원 가량으로 기존 이탈고객의 총납입금액의 2.9%에 해당하는 수치이다.

- 이탈->비이탈 총납입금액 총액의 상승량이 있다고 해도 이탈 고객의 수가 전체의 19.8%이기 때문에 전체 고객의 총납입금액에 대한 상승폭은 0.4%에 그친다.

- 이탈한 고객의 복귀에 대한 효과는 없는 것으로 사료되며 해당 프로젝트의 취소를 고려할 것을 추천한다.

* 최적의 모델 구하기

C	degree	epsilon	학결	평결	rmse
2	1	0	0.45	0.34	0.529883
2	1	0.1	0.47	0.36	0.518585
2	1	0.2	0.47	0.37	0.517354
2	1	0.3	0.49	0.38	0.511817
2	1	0.4	0.49	0.37	0.515368
2	1	0.5	0.45	0.32	0.536612
2	1	0.6	0.41	0.26	0.557968
2	1	0.7	0.4	0.23	0.571952
2	1	0.8	0.35	0.16	0.597256
2	1	0.9	0.28	0.09	0.620019
2	2	0	0.45	0.34	0.529883
2	2	0.1	0.47	0.36	0.518585
2	2	0.2	0.47	0.37	0.517354
2	2	0.3	0.49	0.38	0.511817
2	2	0.4	0.49	0.37	0.515368
2	2	0.5	0.45	0.32	0.536612
2	2	0.6	0.41	0.26	0.557968
2	2	0.7	0.4	0.23	0.571952
2	2	0.8	0.35	0.16	0.597256
2	2	0.9	0.28	0.09	0.620019
2	3	0	0.45	0.34	0.529883
2	3	0.1	0.47	0.36	0.518585

* 모델 조건

Polynomial kernel로 한정

C : 2~30 / 간격 1

degree : 1~5 / 간격 1

epsilon : 0~0.9 / 간격 0.1

gamma : 0.01

* 이상 징후

- C의 값이 커질 수록 무조건 감소하는 RMSE
- 이론적으로는 당연하지만 최적의 C에 대한 결정 기준 마련 필요
- degree에 영향을 안 받는 RMSE

RMSE 최소

커널	다항식
degree	1~5
gamma	0.01
C	30
epsilon	0.3
학습용 결정계수	0.58
평가용 결정계수	0.45
RMSE	0.48187

모델 채택

* 분석 결과

다항식 커널

C : 30

degree = 2

gamma = 0.01

epsilon = 0.3

학습용 데이터 세트 결정계수 : 0.58

평가용 데이터 세트 결정계수 : 0.45

RMSE : 0.48187

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	54,443,854	+27.3%
	평균	329,962	
	잠재	+11,692,754	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	269,946,400	+4.3%

* 실무적 인사이트

- 현재까지의 모든 이탈 고객의 잠재 총납입금액의 총액은 1169만원 가량으로 기존 이탈고객의 총납입금액의 27.3%에 해당하는 수치이다.

- 전체 고객의 총납입금액 총액이 4.3% 상승할 것으로 전망된다.

- 기존의 분석 결과가 너무나도 상이한 결과였기에 C=10으로도 똑같은 분석을 해보았으나 결과에 변화는 없었다.

- 분석 결과로만 본다면 이탈 고객을 복귀 시키면 유의미한 후원금 상승이 있다고 판단된다. 따라서 이탈 고객 복귀 프로젝트의 진행을 긍정적으로 고려해본다.

* 최적의 모델 구하기

hidden	alpha	학결	평결	rmse
[50, 50]	0	0.5	0.66	0.460205
[50, 50]	0.001	0.51	0.67	0.454114
[50, 50]	0.002	0.51	0.67	0.453172
[50, 50]	0.003	0.52	0.69	0.452166
[50, 50]	0.004	0.51	0.67	0.453673
[50, 50]	0.005	0.51	0.68	0.456935
[50, 50]	0.006	0.51	0.68	0.457456
[50, 50]	0.007	0.49	0.67	0.462557
[50, 50]	0.008	0.5	0.67	0.461519
[50, 50]	0.009	0.5	0.67	0.461172
[50, 50]	0.01	0.49	0.67	0.462513
[50, 50]	0.011	0.49	0.67	0.464385
[50, 50]	0.012	0.49	0.67	0.46543
[50, 50]	0.013	0.49	0.67	0.46367
[50, 50]	0.014	0.49	0.67	0.463226
[50, 50]	0.015	0.49	0.66	0.4651
[50, 50]	0.016	0.48	0.66	0.469182
[50, 50]	0.017	0.48	0.66	0.470839
[50, 50]	0.018	0.48	0.66	0.470976
[50, 50]	0.019	0.48	0.66	0.471122

* 모델 조건

MLP 모형

alpha : 0~1 / 간격 0.001

hidden size = [50,50]

* 한계점

- 모델 탐구에 시간이 오래 걸려서 hidden layer를 [50,50]으로만 제한 했다.

RMSE 최소

MLP	
alpha	0.003
hidden layer	50,50
학습용 결정계수	0.69
평가용 결정계수	0.52
RMSE	0.452166

모델 채택

6. 인공신경망

* 분석 결과

MLP

alpha : 0.003

hidden layer : [50,50]

학습용 데이터 세트 결정계수 : 0.69

평가용 데이터 세트 결정계수 : 0.52

RMSE : 0.452166

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	34,239,802	-19.9%
	평균	207,513	
	잠재	-8,511,297	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	260,174,803	-3.1%

* 실무적 인사이트

- 현재까지의 모든 이탈 고객의 잠재 총납입금액의 총액은 -851만원 가량으로 기존 이탈고객의 총납입금액의 19.9%에 해당하는 수치이다.

- 이탈고객이 비이탈 고객이었다면 전체 고객의 총납입금액 총액이 3.1% 하락할 것으로 전망된다.

- 기존의 분석 결과가 너무나도 상이한 결과이지만 분석 결과로만 본다면 이탈 고객들이 이탈을 하지 않았다면 이탈을 결정한 시기보다 소극적인 후원을 하였을 것이다.

* 최적의 모델 구하기

심층신경망의 특성을 고려해 독립변수를 정규화

* 모델 조건
DNN
input layer : 7
hidden layer : 3~10
output layer : 1
epochs : 5~100 / 간격 1
가중치 업데이트 : MSE
optimizer = 경사하강법(SGD)
activation = relu

* 한계점
- 학습에 lelu모델을 사용하여 마이너스 값을 0으로 도출했다.

MSE 최소

DNN	
epochs	20
Hidden layer	6
학습용 MSE	0.35
평가용 MSE	0.46

모델 채택

7. 심층신경망

* 분석 결과

DNN

epochs : 20

hidden layer : 6

Activation = relu

학습용 데이터 MSE : 0.35

평가용 데이터 MSE : 0.46

구분		총납입금액(원)	증감률(%)
기존 이탈 고객 165명	총액	42,751,100	
	평균	259,097	
이탈 -> 비이탈 고객 165명	총액	44,524,543	+4.1%
	평균	269,719	
	잠재	1,773,443	
기존 모든 고객 866명	총액	268,686,100	
예측된 모든 고객 866명	총액	270,459,543	+0.6%

* 실무적 인사이트

- 현재까지의 모든 이탈 고객의 잠재 총납입금액의 총액은 177만원 가량으로 기존 이탈고객의 총납입금액의 4.1%에 해당하는 수치이다.

- 이탈고객이 비이탈 고객이었다면 전체 고객의 총납입금액 총액이 0.6% 상승할 것으로 전망된다.

- 이탈고객이 복귀 했을 때 얻을 이득이 현저히 낮다. 이것이 이탈고객 복귀 프로젝트 수행 지출보다 이득일지에 대한 의사결정이 재차 요구된다.

모델 비교

분석 방법	학습용 결정계수	평가용 결정계수	RMSE	잠재 총납입금액 총액
표준 선형 회귀분석	0.42	0.3	0.5451	+1,630,674
릿지 선형 회귀분석	0.42	0.3	0.5439	+1,630,674
라쏘 선형 회귀분석	0.42	0.3	0.5439	+1,630,674
K-NN	0.54	0.37	0.5180	+1,260,300
SVM	0.58	0.45	0.4818	+11,692,754
인공신경망	0.69	0.52	0.4521	-8,511,297
DNN	MSE : 0.35	MSE : 0.46	-	+1,773,443

- 회귀 분석 간의 특별한 차이 없음
- SVM과 인공신경망의 잠재 총납입금액 총액의 차이
- RMSE가 가장 낮은 인공신경망이 가장 적합한 것인가
- MSE를 사용한 DNN과 타 모델과의 비교는 어떻게 하는 것이 적합한가