Part 2 : Modeling part

202035349 심승민

Part 1 수정 :

분석 target feature 생성 :

- Revenue_YN : threshold 이상 구매 시 1 else 0

  Total_Purchases(총 구매 횟수)의 중간값을 threshold로 설정

  threshold 이상 구매 시 매출 발생

- 분석에 필요한 feature 생성
  - Age : current_year - Year_Birth
  - membership_Years : current_year - df["Dt_Customer"].dt.year
  - Total_Mnt : Mnt_*의 합
  - Total_Purchases : *Purchases의 합

Modeling 시 불필요한 feature 추가 제거:

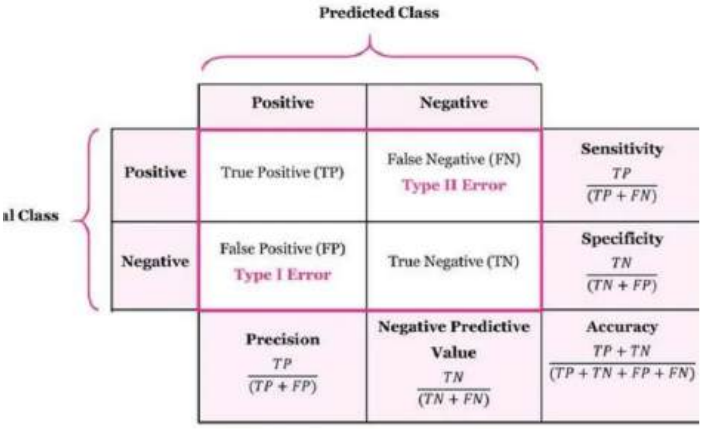
- ID, Dt_Customer, membership_Years, Year_Birth
- scale 되지 않은 raw data : Income, Recency, NumWebVisitsMonth, Total_Purchases, Age, membership_Years

---

Part 2 :

Classification 평가 지표 :

Confusion Matrix와 ROC 그래프



- **Accuracy**

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{21 + 26}{60} = 0.78$$

- **Precision**
  - "Out of predicted 1s, how many were actually 1?"

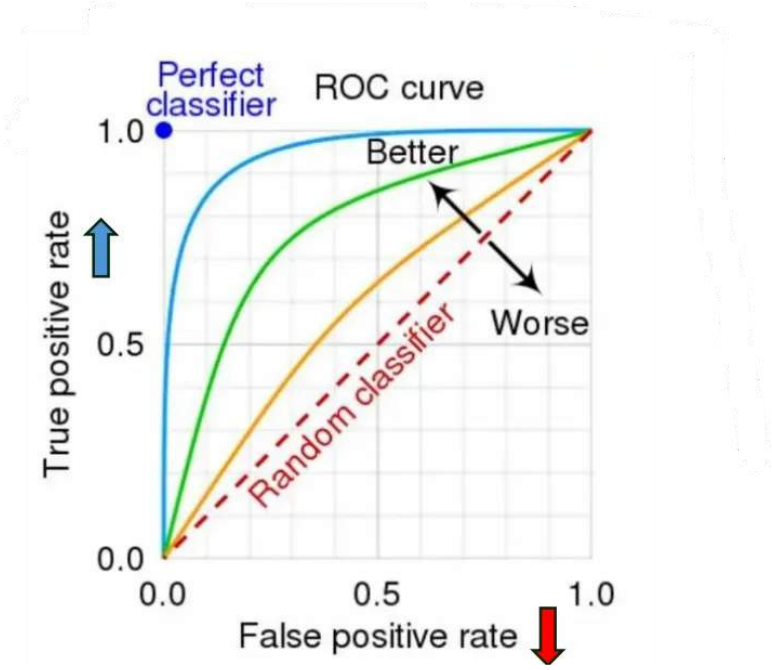
$$\text{Precision} = \frac{TP}{TP + FP} = \frac{21}{21 + 6} \approx 0.78$$

- **Recall (Sensitivity)**
  - "Out of actual 1s, how many did we catch?"

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{21}{21 + 7} \approx 0.75$$

- **F1-score**
  - Balance of precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \approx 0.765$$



part2_classification.py

logistic을 활용하여 매출 발생(Revenue_YN) classification

각 feature의 coefficient 참고하여 독립변수(X) 설정

	Feature	Coefficient
10	Income_scaled	1.574212
15	Total_Mnt_scaled	1.231605
4	AcceptedCmp5	-0.923689
12	NumWebVisitsMonth_scaled	0.840472
1	Teenhome	0.544220
5	AcceptedCmp1	-0.480640
18	Education_Master	-0.468882
0	Kidhome	-0.449450
25	Marital_Status_Widow	-0.417266
17	Education_Graduation	-0.373574
23	Marital_Status_Single	-0.336684
8	Response	0.313453
29	Income_eq_freq_Medium	0.311689
14	membership_Years_scaled	0.286158
19	Education_PhD	-0.255410
21	Marital_Status_Divorced	-0.217420
27	Income_eq_width_Medium	0.197898
28	Income_eq_width_High	-0.197721
16	Education_Basic	-0.153085
13	Age_scaled	-0.137051
2	AcceptedCmp3	-0.128277
24	Marital_Status_Together	-0.120107
22	Marital_Status_Married	-0.113168
11	Recency_scaled	-0.099813
6	AcceptedCmp2	0.058950
7	Complain	0.050145
30	Income_eq_freq_High	-0.039149
26	Marital_Status_YOLO	0.027916
3	AcceptedCmp4	0.012197
9	Total_Mnt	0.002990
20	Marital_Status_Alone	0.002155

선택된 독립변수(x)의 coefficient (상위 10개)

```
Top 10 Logistic Regression Coefficients (by absolute value):
                   Feature  Coefficient
         Total_Mnt_scaled     2.638558
             AcceptedCmp5    -1.821571
            Income_scaled     1.516900
                 Teenhome     0.790940
             AcceptedCmp2     0.754407
NumWebVisitsMonth_scaled     0.723787
                 Kidhome    -0.576977
    Marital_Status_Widow    -0.422543
 membership_Years_scaled     0.342254
   Marital_Status_Single    -0.329987
```

결과

```
tic Regression Results>
acy-
: [0.9157239819004525, 0.9089366515837104, 0.9055429864253394, 0.9072398190045249, 0.91515837104
 [0.8981900452488688, 0.8981900452488688, 0.9230769230769231, 0.9140271493212669, 0.895927601809
sion-
: [0.9011764705882352, 0.8943661971830986, 0.8950471698113207, 0.8957345971563981, 0.90845886442
 [0.8976744186046511, 0.8967136150234741, 0.9112149532710281, 0.8933333333333333, 0.868686868686
l-
: [0.9217809867629362, 0.9147659063625451, 0.9068100358422939, 0.9086538461538461, 0.91695906432
 [0.8935185185185185, 0.8925233644859814, 0.9285714285714286, 0.9348837209302325, 0.895833333333

: [0.9113622843545509, 0.9044510385756677, 0.9008902077151335, 0.9021479713603818, 0.91268917345
 [0.8955916473317865, 0.8946135831381733, 0.9198113207547169, 0.9136363636363637, 0.882051282051
```

1. **Accuracy**

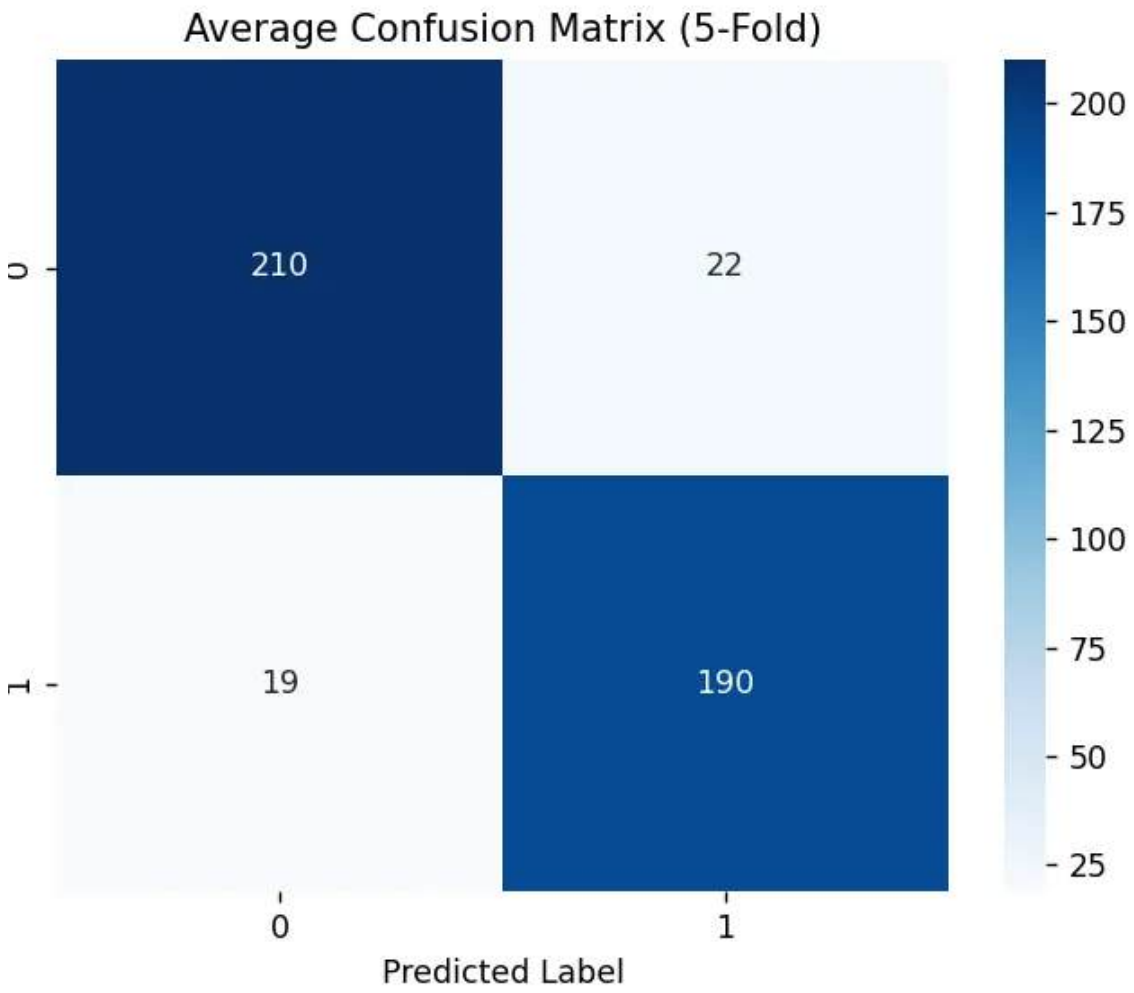
   높은 정확성을 보여주며 train 정확도와 test 정확도가 비슷한 수준을 보임

   →overfitting 없이 잘 학습됨

2. **Precision : 예측 positive 중 실제 positive 비율** - TP / (TP + FP)
3. **Recall : 실제 positive 중 예측 성공한 positive 비율** - TP / (TP + FN)
4. **F1-score : Precision과 Recall 비율**

지표	Train - avg	Test - avg
Accuracy	0.911	0.906
Precision	0.899	0.894
Recall	0.914	0.909
F1 Score	0.906	0.901

5. Confusion Matrix



True Negative (TN): 210

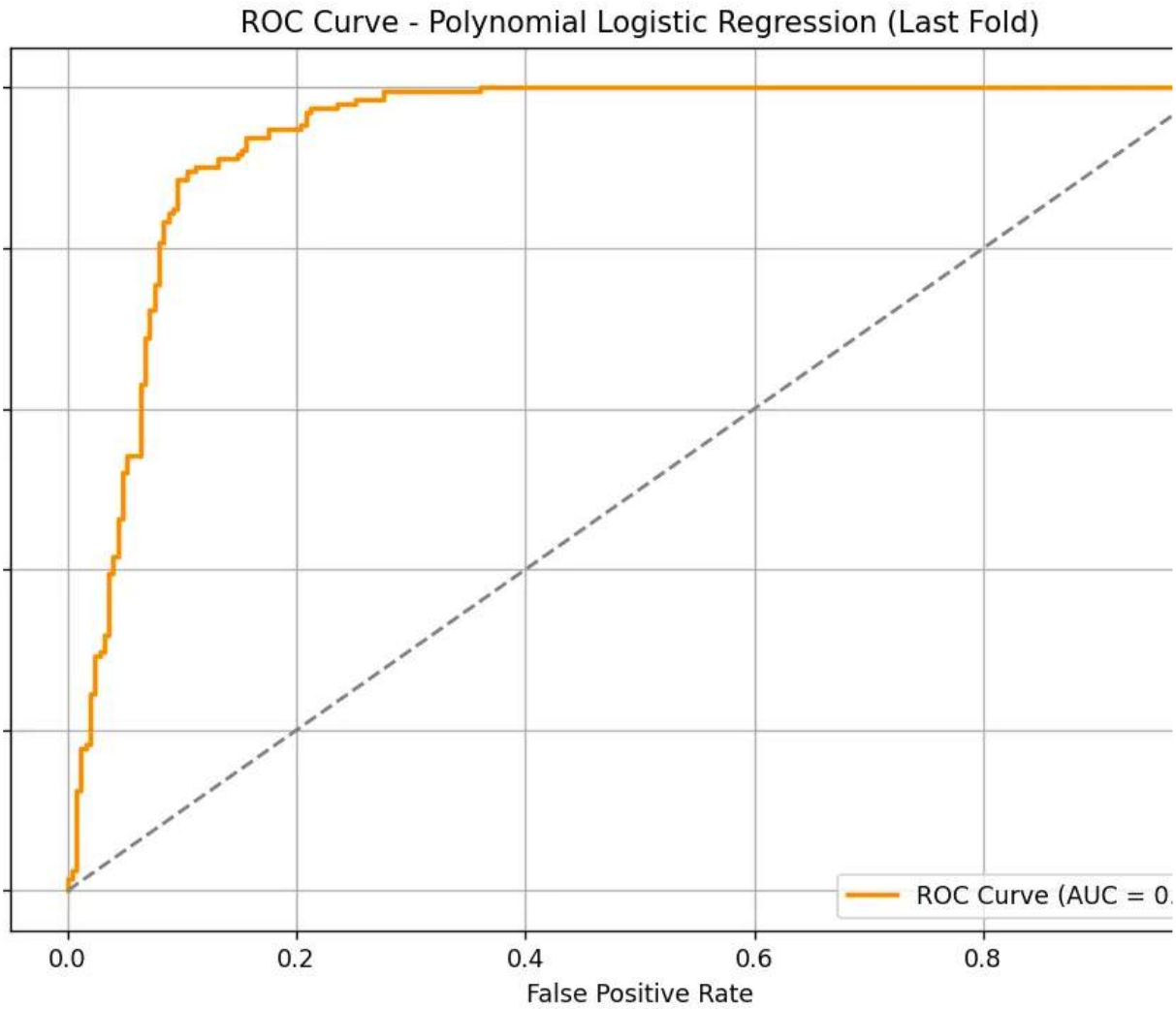
False Positive (FP): 22

False Negative (FN): 19

True Positive (TP): 190

→FP : FN - 오차 비율이 비슷함

6. **ROC Curve:**


ROC Curve - Polynomial Logistic Regression (Last Fold)
ROC Curve (AUC = 0.
0.0
0.2
0.4
0.6
0.8
False Positive Rate

part2_Polynomial_classification.py

polynomial logistic을 활용하여 매출 발생(Revenue_YN) classification

각 feature의 coefficient 참고하여 독립변수(X) 설정

	Feature	Coefficient
10	Income_scaled	1.574212
15	Total_Mnt_scaled	1.231605
4	AcceptedCmp5	-0.923689
12	NumWebVisitsMonth_scaled	0.840472
1	Teenhome	0.544220
5	AcceptedCmp1	-0.480640
18	Education_Master	-0.468882
0	Kidhome	-0.449450
25	Marital_Status_Widow	-0.417266
17	Education_Graduation	-0.373574
23	Marital_Status_Single	-0.336684
8	Response	0.313453
29	Income_eq_freq_Medium	0.311689
14	membership_Years_scaled	0.286158
19	Education_PhD	-0.255410
21	Marital_Status_Divorced	-0.217420
27	Income_eq_width_Medium	0.197898
28	Income_eq_width_High	-0.197721
16	Education_Basic	-0.153085
13	Age_scaled	-0.137051
2	AcceptedCmp3	-0.128277
24	Marital_Status_Together	-0.120107
22	Marital_Status_Married	-0.113168
11	Recency_scaled	-0.099813
6	AcceptedCmp2	0.058950
7	Complain	0.050145
30	Income_eq_freq_High	-0.039149
26	Marital_Status_YOLO	0.027916
3	AcceptedCmp4	0.012197
9	Total_Mnt	0.002990
20	Marital_Status_Alone	0.002155

결과

```
mial Logistic Regression Results>
cy-
 [0.9536199095022625, 0.9440045248868778, 0.9485294117647058, 0.9530542986425339, 0.95475113122171
[0.8846153846153846, 0.9276018099547512, 0.9253393665158371, 0.9095022624434389, 0.889140271493212
ion-
 [0.9299655568312285, 0.9237875288683602, 0.9267734553775744, 0.9309551208285386, 0.93686583990980
[0.8873239436619719, 0.9292452830188679, 0.9116279069767442, 0.8959276018099548, 0.848780487804878
-
 [0.9747292418772563, 0.9603841536614646, 0.967741935483871, 0.9723557692307693, 0.971929824561403
[0.875, 0.9205607476635514, 0.9333333333333333, 0.9209302325581395, 0.90625]

 [0.9518213866039953, 0.9417304296645085, 0.9468147282291058, 0.9512051734273956, 0.95407577497129
[0.8811188811188811, 0.9248826291079812, 0.9223529411764706, 0.908256880733945, 0.8765743073047859
```

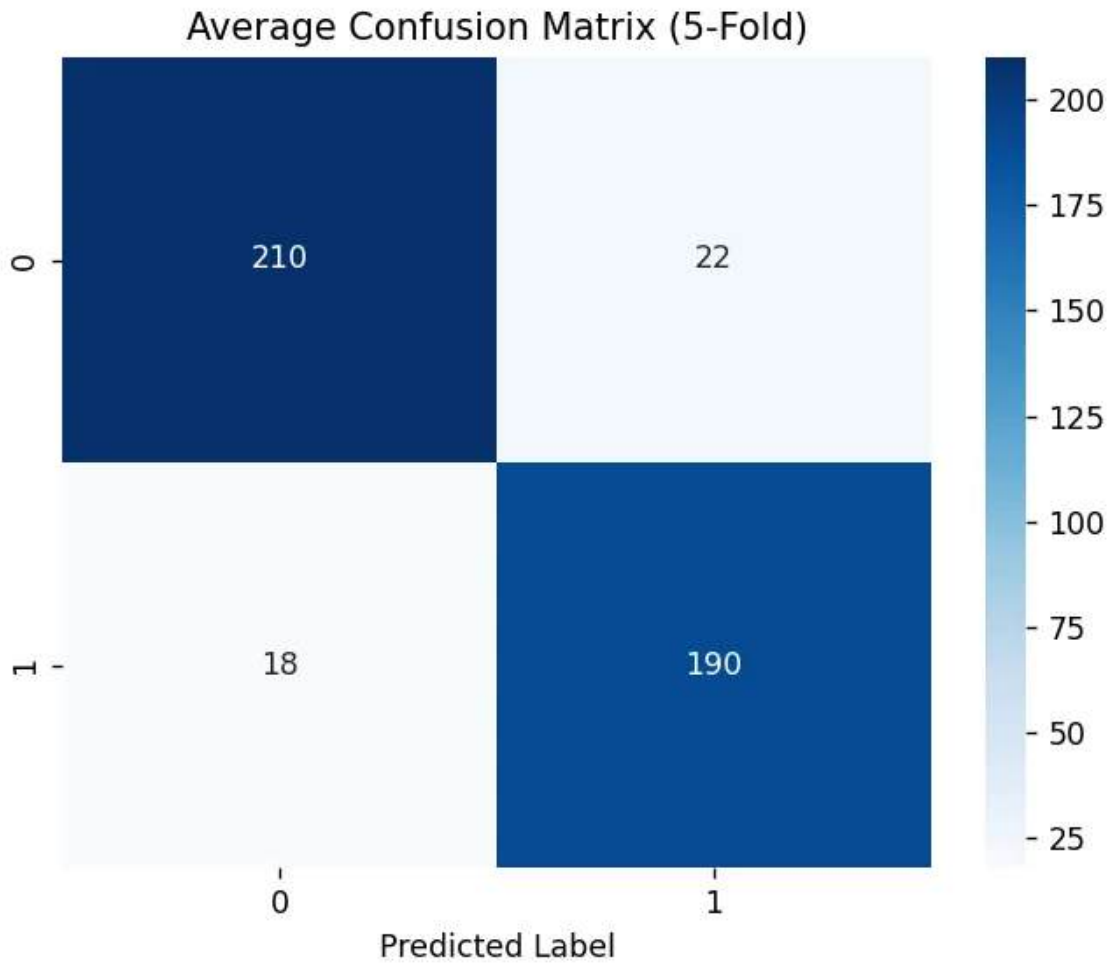
1. **Accuracy :**

   높은 정확성을 보여주지만 overfitting 우려(train 정확도 > test 정확도, train 정확도가 매우 높음)

2. **Precision : 예측 positive 중 실제 positive 비율 - TP / (TP + FP) :**
3. **Recall : 실제 positive 중 예측 성공한 positive 비율 - TP / (TP + FN) :**
4. **F1-score : Precision과 Recall 비율 :**

지표	Train - avg	Test - avg
Accuracy	0.951	0.907
Precision	0.930	0.895
Recall	0.969	0.911
F1 Score	0.949	0.903

5. **Confusion Matrix**



True Negative (TN): 210

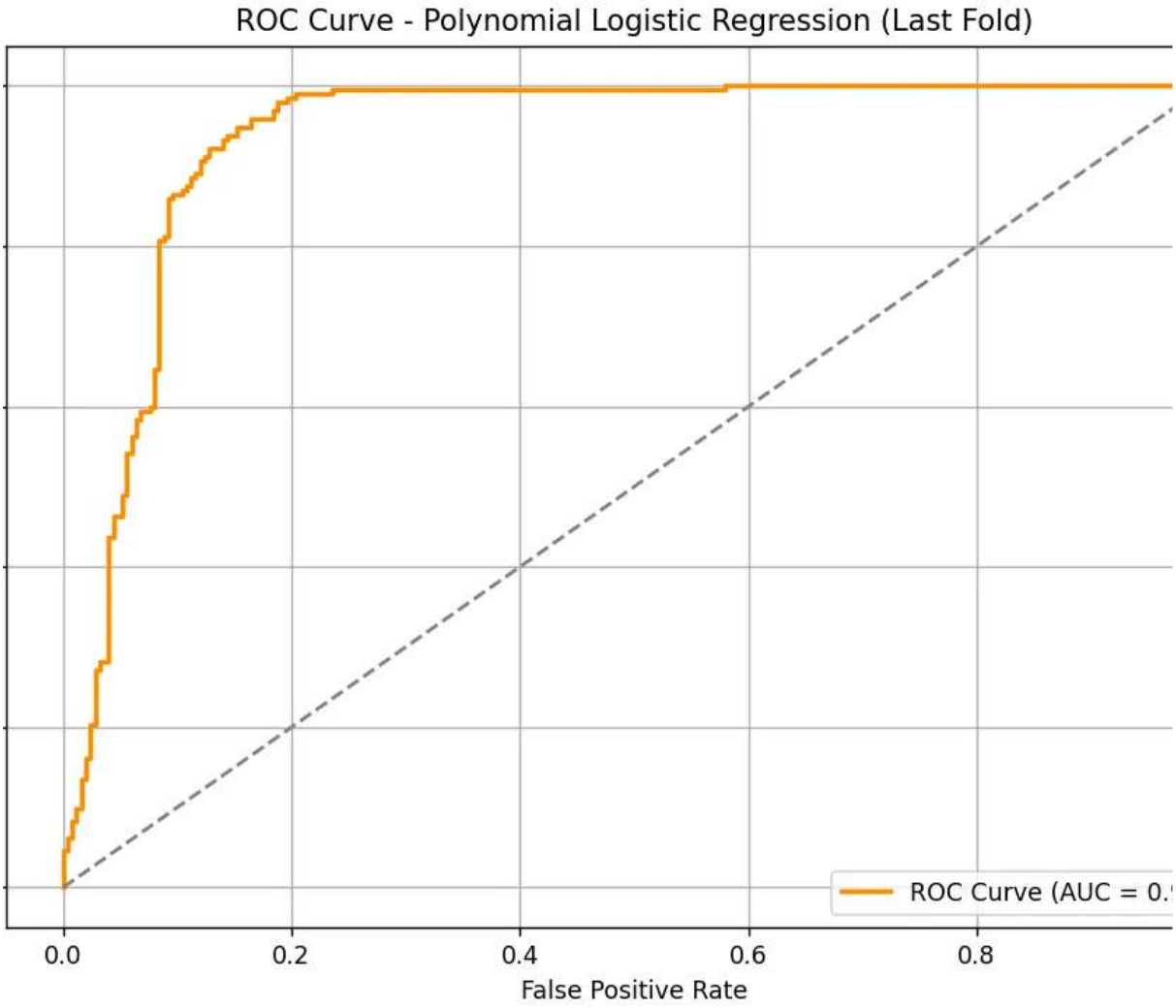
False Positive (FP): 22

False Negative (FN): 18

True Positive (TP): 190

→FP : FN - 오차 비율이 비슷함

6. **ROC :**


ROC Curve - Polynomial Logistic Regression (Last Fold)
ROC Curve (AUC = 0.
0.0
0.2
0.4
0.6
0.8
False Positive Rate

Regression 평가 지표 :

Confusion Matrix와 ROC 그래프

Mean Error Squared

$$\text{MSE} = \boxed{\frac{1}{n}\sum_{i=1}^{n}\boxed{(Y_i - \hat{Y}_i)}^2}$$

- MAE (Mean Absolute Error)
  - **Definition:** The average of the absolute differences between the predicted values and the actual values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  - **Interpretation:**
    - **Tells you on average how much the predictions differ from the actual values**
    - Easy to interpret — it's in the **same unit** as the target variable
    - **Robust to outliers** (less sensitive than RMSE)

- RMSE (Root Mean Squared Error)
  - **Definition:** The square root of the average of the squared differences between predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  - **Interpretation:**
    - Penalizes **larger errors more heavily** than MAE due to squaring
    - Useful when **large errors are particularly undesirable**
    - Also in the **same unit** as the target variable

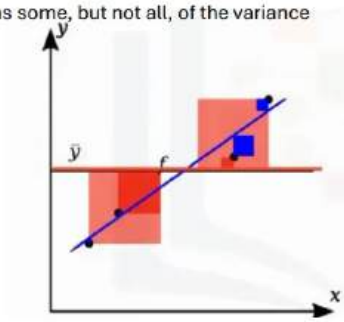
- **R^2** measures how well your regression model explains the variability of the target variable.
  - "How much better is my model than just predicting the mean?"

Where:

- SS_{RES} → Residual sum of squares (model error)
- SS_{TOT} → Total sum of squares (variance of the data)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- Interpretation
  - R^2=1.0: Perfect fit — model explains **100%** of the variance
  - R^2=0.0: Model does no better than just predicting the mean of y
  - R^2<0: Model is **worse than the mean predictor** — it increases error
  - 0<R^2<1 : Model explains some, but not all, of the variance



- The blue line represents the regression line
- The blue squares represents the MSE of the regression line
- The red line represents the average value of the data points
- The red squares represent the MSE of the red line
- We see the area of the blue squares is much smaller then the area of the red squares

part2_regression.py

Lasso regression을 통해 구매 금액(Total_Mnt) regression

각 feature의 coefficient 참고하여 독립변수(X) 설정

	Feature	Coefficient
29	Income_eq_freq_High	226.191601
14	Total_Purchases_scaled	172.735140
9	Income_scaled	97.624770
4	AcceptedCmp5	93.816391
6	AcceptedCmp2	-85.523515
0	Kidhome	-73.244633
30	Revenue_YN	72.799327
18	Education_PhD	68.962232
27	Income_eq_width_High	-65.103221
25	Marital_Status_YOLO	-64.481944
1	Teenhome	-62.104841
17	Education_Master	61.986736
2	AcceptedCmp3	60.051788
3	AcceptedCmp4	58.406770
5	AcceptedCmp1	53.132173
16	Education_Graduation	41.381725
15	Education_Basic	34.017276
13	membership_Years_scaled	29.729376
20	Marital_Status_Divorced	20.419640
28	Income_eq_freq_Medium	20.296025
22	Marital_Status_Single	19.043772
11	NumWebVisitsMonth_scaled	16.009340
23	Marital_Status_Together	15.479270
8	Response	-10.681986
24	Marital_Status_Widow	-10.085876
26	Income_eq_width_Medium	-6.853213
10	Recency_scaled	6.137628
12	Age_scaled	-1.072471
7	Complain	-0.000000
21	Marital_Status_Married	0.000000
19	Marital_Status_Alone	-0.000000

선택된 독립변수(x)의 coefficient (상위 10개)

Top 10 Logistic Regression Coefficients (by absolute value):

	Feature	Coefficient
0	Age_scaled	-0.056283
1	Recency_scaled	-0.056283
2	membership_Years_scaled	-0.056283
3	NumWebVisitsMonth_scaled	-0.056283
4	Kidhome	-0.056283
5	Teenhome	-0.056283
6	Total_Purchases_scaled	-0.056283
7	Income_eq_width_Medium	-0.056283
8	Income_eq_width_High	-0.056283
9	Income_eq_freq_Medium	-0.056283
10	Income_eq_freq_High	-0.056283
11	Education_Basic	-0.056283
12	Education_Graduation	-0.056283
13	Education_Master	-0.056283
14	Education_PhD	-0.056283
15	Marital_Status_Divorced	-0.056283
16	Marital_Status_Widow	-0.056283
17	Marital_Status_YOLO	-0.056283
18	Marital_Status_Single	-0.056283
19	Marital_Status_Together	-0.056283
20	AcceptedCmp1	-0.056283
21	AcceptedCmp2	-0.056283
22	AcceptedCmp3	-0.056283
23	AcceptedCmp4	-0.056283
24	AcceptedCmp5	-0.056283
25	Response	-0.056283

결과

```
egression>

[0.7794768182669775, 0.7746993117701725, 0.7726431164212024, 0.7838731538044432, 0.782386855
0.7672446511989177, 0.7866163578880132, 0.7957751483386026, 0.7486389633923207, 0.754042263

[140.50705758930883, 139.08859153444556, 139.18788727065208, 137.72847969731126, 137.9107274
134.90330838592087, 137.9217171182676, 136.3146291501398, 145.75544512300863, 148.5585935268

[207.93973777718065, 208.02289775424737, 210.34376058281512, 205.88485810724137, 206.8857754
210.43528747332638, 210.31624407336395, 200.5393276212271, 219.01593774984698, 215.275088998

[43238.93454684265, 43273.52599007406, 44244.49761612065, 42388.57479783891, 42801.724068531
44283.01021398152, 44232.9225211268, 40216.021922773856, 47967.980988444855, 46343.363943293
```

1. R^2 - model의 예측 능력
2. MAE - 오차 절댓값의 평균
3. RMSE : MSE의 제곱근
4. MSE : 오차 제곱의 평균

지표	Train - avg	Test - avg
R^2	0.779	0.770
MAE	138.885	140.691
RMSE	207.815	211.116
MSE	43189.451	44608.660

→ 성능이 전체적으로 안정적이고 overfitting의 우려가 적음

→ 정확도는 준수하지만 아쉬운 성능

part2_Polynomial_regression.py

Polynomial Lasso regression을 통해 구매 금액(Total_Mnt) regression

각 feature의 coefficient 참고하여 독립변수(X) 설정

	Feature	Coefficient
29	Income_eq_freq_High	226.191601
14	Total_Purchases_scaled	172.735140
9	Income_scaled	97.624770
4	AcceptedCmp5	93.816391
6	AcceptedCmp2	-85.523515
0	Kidhome	-73.244633
30	Revenue_YN	72.799327
18	Education_PhD	68.962232
27	Income_eq_width_High	-65.103221
25	Marital_Status_YOLO	-64.481944
1	Teenhome	-62.104841
17	Education_Master	61.986736
2	AcceptedCmp3	60.051788
3	AcceptedCmp4	58.406770
5	AcceptedCmp1	53.132173
16	Education_Graduation	41.381725
15	Education_Basic	34.017276
13	membership_Years_scaled	29.729376
20	Marital_Status_Divorced	20.419640
28	Income_eq_freq_Medium	20.296025
22	Marital_Status_Single	19.043772
11	NumWebVisitsMonth_scaled	16.009340
23	Marital_Status_Together	15.479270
8	Response	-10.681986
24	Marital_Status_Widow	-10.085876
26	Income_eq_width_Medium	-6.853213
10	Recency_scaled	6.137628
12	Age_scaled	-1.072471
7	Complain	-0.000000
21	Marital_Status_Married	0.000000
19	Marital_Status_Alone	-0.000000

결과

```
ial Regressionwith Lasso>

[0.856585680712692, 0.8572395480573112, 0.8549326224121041, 0.863048143514004, 0.8622782741556768]
0.7693826101300402, 0.7688864847626062, 0.798017271029245, 0.6915418706236041, 0.7277073947395529]

[110.86646625362185, 108.65341698536861, 110.3379932953364, 107.6953763831981, 107.90475488872761]
124.5183662408545, 137.32057981309342, 127.28766735225426, 146.04722331240865, 146.59014322955275]

[np.float64(167.6897915851556), np.float64(165.5898984021669), np.float64(168.0197988315387), np.float64(163.890
oat64(164.58463825645052)]
np.float64(209.46658812178933), np.float64(218.87941896648223), np.float64(199.43545928051282), np.float64(242.6
.float64(226.50691102470853)]

[28119.866201872926, 27420.01445283996, 28230.652799390733, 26860.124572894834, 27088.10315000667]
43876.25153938333, 47908.20004710486, 39774.50241842909, 58863.99055853586, 51305.380741955225]
```

1. **R^2 - model의 예측 능력**
2. **MAE - 오차 절댓값의 평균**
3. **RMSE : MSE의 제곱근**
4. **MSE : 오차 제곱의 평균**

지표	Train - avg	Test - avg
R^2	0.859	0.751
MAE	109.092	136.353
RMSE	165.955	219.381
MSE	27543.752	48345.665

→train data에서는 안정적이지만 test data에서 fold 별 성능 차이가 큼 = 일부 데이터 분포에 적합하지 않을 수 있음

test 보다 train의 성능이 더 높게 나오지만 overfitting이 의심 될 수준은 아님

part2_clustering_classification.py

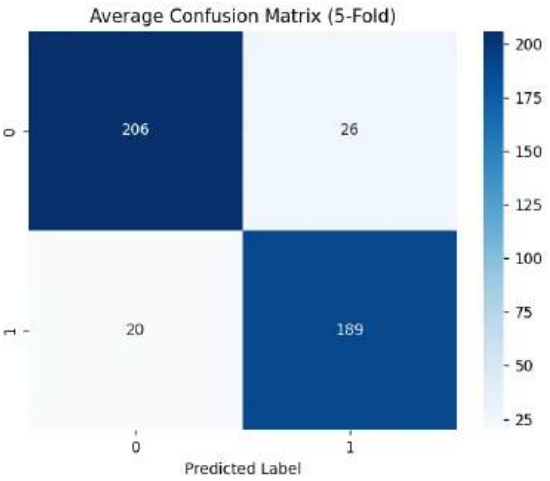
KNN을 통해 Revenue_YN classification

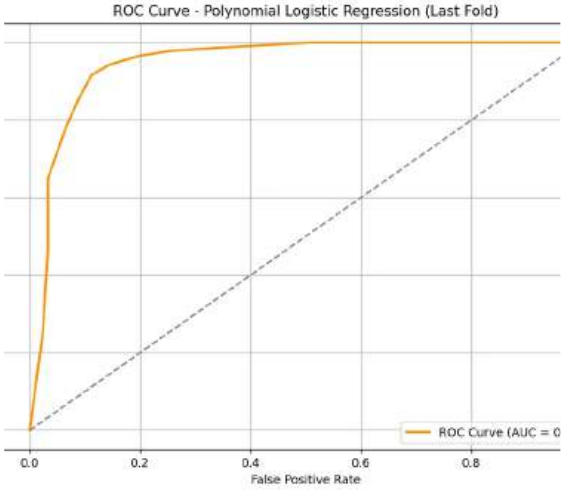
각 feature의 coefficient 참고하여 독립변수(X) 설정

	Feature	Coefficient
10	Income_scaled	1.574212
15	Total_Mnt_scaled	1.231605
4	AcceptedCmp5	-0.923689
12	NumWebVisitsMonth_scaled	0.840472
1	Teenhome	0.544220
5	AcceptedCmp1	-0.480640
18	Education_Master	-0.468882
0	Kidhome	-0.449450
25	Marital_Status_Widow	-0.417266
17	Education_Graduation	-0.373574
23	Marital_Status_Single	-0.336684
8	Response	0.313453
29	Income_eq_freq_Medium	0.311689
14	membership_Years_scaled	0.286158
19	Education_PhD	-0.255410
21	Marital_Status_Divorced	-0.217420
27	Income_eq_width_Medium	0.197898
28	Income_eq_width_High	-0.197721
16	Education_Basic	-0.153085
13	Age_scaled	-0.137051
2	AcceptedCmp3	-0.128277
24	Marital_Status_Together	-0.120107
22	Marital_Status_Married	-0.113168
11	Recency_scaled	-0.099813
6	AcceptedCmp2	0.058950
7	Complain	0.050145
30	Income_eq_freq_High	-0.039149
26	Marital_Status_YOLO	0.027916
3	AcceptedCmp4	0.012197
9	Total_Mnt	0.002990
20	Marital_Status_Alone	0.002155

최적의 k를 찾는 것이 중요

k=11



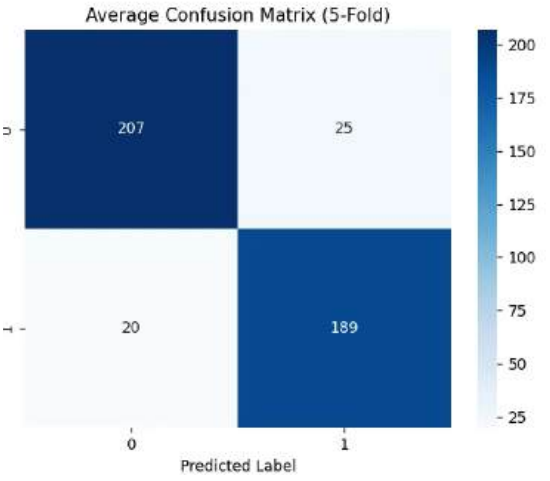


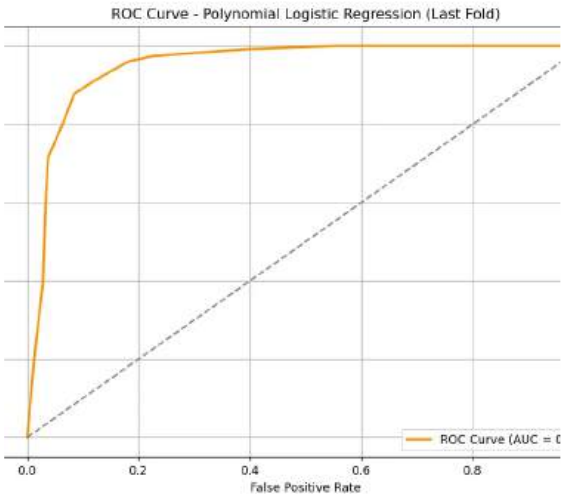

```
st Neighbor Classification(K=11) Results>
y-
[0.9134615384615384, 0.9015837104072398, 0.9055429864253394, 0.9072398190045249, 0.902714932
0.8914027149321267, 0.8981900452488688, 0.9095022624434389, 0.8755656108597285, 0.9027149321
on-
[0.8913525498891353, 0.8873239436619719, 0.8842592592592593, 0.8918296892980437, 0.881036513
0.8465346534653465, 0.8925233644859814, 0.9036697247706422, 0.8483412322274881, 0.8969957081

[0.9359720605355064, 0.9064748201438849, 0.9193742478941035, 0.9171597633136095, 0.913308913
0.9095744680851063, 0.8967136150234741, 0.9120370370370371, 0.8861386138613861, 0.9166666666

[0.9131175468483816, 0.896797153024911, 0.9014749262536873, 0.9043173862310385, 0.8968824940
0.8769230769230769, 0.8946135831381733, 0.9078341013824884, 0.8668280871670703, 0.9067245119
```

k=13



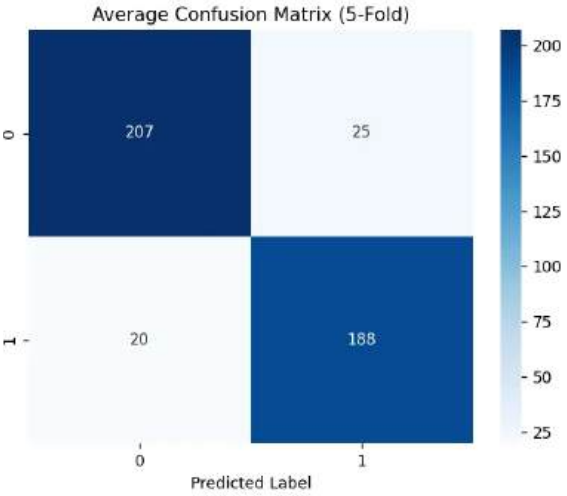


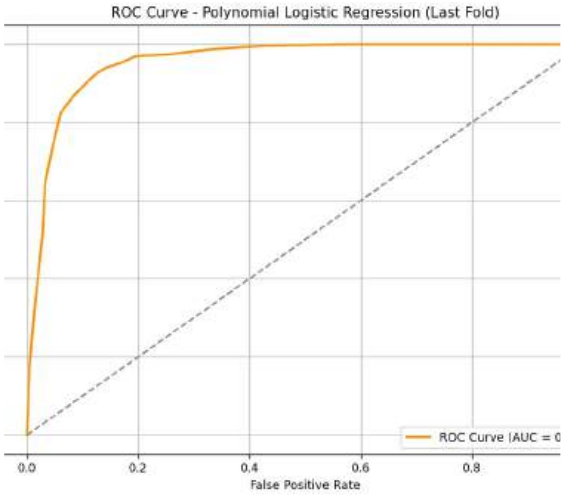
```
est Neighbor Classification(K=13) Results>
y-
[0.9123303167420814, 0.8976244343891403, 0.9038461538461539, 0.9111990950226244, 0.90441176470
0.8959276018099548, 0.9072398190045249, 0.9117647058823529, 0.8733031674208145, 0.895927601809
ion-
[0.8919821826280624, 0.8845700824499411, 0.8865497076023392, 0.9, 0.8805620608899297]
0.8622448979591837, 0.8944954128440367, 0.9116279069767442, 0.8443396226415094, 0.892241379310

[0.9324796274738067, 0.9004796163069544, 0.9121540312876053, 0.9159763313609467, 0.91819291819
0.898936170212766, 0.9154929577464789, 0.9074074074074074, 0.8861386138613861, 0.9078947368421

[0.9117814456459875, 0.8924539512774807, 0.8991696322657177, 0.9079178885630499, 0.89898386132
0.8802083333333334, 0.9048723897911833, 0.9095127610208816, 0.8647342995169082, 0.9]
```

k=15






```
est Neighbor Classification(K=15) Results>
y-
[0.9111990950226244, 0.8987556561085973, 0.9044117647058824, 0.9123303167420814, 0.90101809954
0.9027149321266968, 0.9072398190045249, 0.9027149321266968, 0.8665158371040724, 0.900452488687
ion-
[0.89086859688196, 0.8848413631022327, 0.8884976525821596, 0.9020979020979021, 0.88061465721040
0.8717948717948718, 0.8944954128440367, 0.9061032863849765, 0.8421052631578947, 0.889830508474

[0.9313154831199069, 0.9028776978417267, 0.910950661853189, 0.9159763313609467, 0.909645909645
0.9042553191489362, 0.9154929577464789, 0.8935185185185185, 0.8712871287128713, 0.921052631578

[0.910643141718839, 0.8937685459940653, 0.8995840760546643, 0.908984145625367, 0.89489489489489
0.8877284595300261, 0.9048723897911833, 0.8997668997668997, 0.8564476885644768, 0.905172413793
```

항목	K=11	K=13	K=15
Test Accuracy - avg	0.8953	0.8968	0.8951
Test Precision - avg	0.8772	0.8802	0.8821
Test Recall - avg	0.9042	0.8989	0.8935
Test F1 Score - avg	0.8912	0.8899	0.8877
TN	206	207	207
FP	26	25	25
TP	189	189	188
FN	20	20	20
AUC	0.9483	0.9514	0.9556

→k=11, 13, 15 모두 비슷한 성적

전체적으로 우수한 성적, overfitting 안 보임

test 성능이 가장 안정적이고(Accuracy) 균형이 잘 잡힌 k=13 사용 추천

part2_clustering_regression.py

KNN을 통해 Total_Mnt regression

각 feature의 coefficient 참고하여 독립변수(X) 설정

	Feature	Coefficient
29	Income_eq_freq_High	226.191601
14	Total_Purchases_scaled	172.735140
9	Income_scaled	97.624770
4	AcceptedCmp5	93.816391
6	AcceptedCmp2	-85.523515
0	Kidhome	-73.244633
30	Revenue_YN	72.799327
18	Education_PhD	68.962232
27	Income_eq_width_High	-65.103221
25	Marital_Status_YOLO	-64.481944
1	Teenhome	-62.104841
17	Education_Master	61.986736
2	AcceptedCmp3	60.051788
3	AcceptedCmp4	58.406770
5	AcceptedCmp1	53.132173
16	Education_Graduation	41.381725
15	Education_Basic	34.017276
13	membership_Years_scaled	29.729376
20	Marital_Status_Divorced	20.419640
28	Income_eq_freq_Medium	20.296025
22	Marital_Status_Single	19.043772
11	NumWebVisitsMonth_scaled	16.009340
23	Marital_Status_Together	15.479270
8	Response	-10.681986
24	Marital_Status_Widow	-10.085876
26	Income_eq_width_Medium	-6.853213
10	Recency_scaled	6.137628
12	Age_scaled	-1.072471
7	Complain	-0.000000
21	Marital_Status_Married	0.000000
19	Marital_Status_Alone	-0.000000

최적의 k를 찾는 것이 중요

k=25

```
st Neighbor Regression(K=25) Results>

[0.7752449697699408, 0.7721162360375262, 0.764255878485938, 0.7730839770075373, 0.7769846446
0.7387396675867819, 0.7605899602134749, 0.773853023658419, 0.7479312799010938, 0.740990919336

[135.88630090497736, 133.71455882352942, 136.6359841628959, 134.7916628959276, 132.921968325
137.9993212669683, 142.57135746606335, 136.01986425339365, 138.59927601809954, 148.181357466

[209.92544246659665, 209.21199202594977, 214.18843433725863, 210.96122130892647, 209.437985
222.94896989935953, 222.7734522169769, 211.02827677611126, 219.32403049584232, 220.912875687

[44068.691394796384, 43769.65760746607, 45876.68540384615, 44504.63689615385, 43864.269863800
49706.24317918552, 49628.01101266969, 44532.93359909501, 48103.030352941176, 48802.498644343
```

k=27

```
est Neighbor Regression(K=27) Results>

 [0.773165185011249, 0.7701428641683241, 0.7612830182064594, 0.7707249585242337, 0.77544913128
[0.7407976834898917, 0.7575482731599728, 0.7726690458260487, 0.7485482096400342, 0.74073307081

 [136.36631682587566, 134.23451902128372, 137.29358345902463, 135.48366013071893, 133.40223311
[137.94754482989777, 143.79822356292942, 135.62422490363667, 138.9715518686107, 148.3783727166

 [210.89448497438454, 210.11588074703133, 215.53471902085866, 212.054962226506, 210.1577638924
[222.06912012687147, 224.1841458813437, 211.57996806327168, 219.05547196853914, 221.0228098404

 [44476.4837926109, 44148.68334210069, 46455.21510340049, 44967.30700488489, 44166.28572426122
[49314.69411392287, 50258.5312645476, 44766.08288565506, 47985.29979935944, 48851.082469787536
```

k=29

```
est Neighbor Regression(K=29) Results>

: [0.7730688364324136, 0.768665570096093, 0.7606279933619051, 0.7687741753962016, 0.775221612824
 [0.7421937638175767, 0.7562255244312961, 0.7735488798143078, 0.7506949599045221, 0.743511247832

: [136.76743641753785, 134.5935208300827, 137.87265954127008, 136.14516695272275, 133.1499258854
 [138.50748946793573, 143.9269776876268, 135.6135902636917, 138.3872679045093, 147.6012248400686

: [210.939269163917, 210.790007742457, 215.83022351255934, 212.95518457613497, 210.2642045095848
 [221.470273408163, 224.79485700832916, 211.17013448967873, 218.11838300258884, 219.835434528976

: [44495.37527540742, 44432.42736406508, 46582.68538148132, 45349.91063785571, 44211.03569804854
 [49049.082003486474, 50532.72773739515, 44592.825700389, 47575.629003664035, 48327.61827454388]
```

구분	K=25	K=27	K=29
Train R^2	0.7723	0.7686	0.7691
Test R^2	0.7520	0.7521	0.7532
Train MAE	134.79	135.35	135.71
Test MAE	140.67	140.94	140.81
Train RMSE	210.75	211.75	212.16
Test RMSE	219.40	219.58	219.08
Train MSE	44,417.0	44,842.6	45,014.2
Test MSE	48,154.4	48,235.2	48,015.8

→k=25, 27, 29 모두 비슷한 성적

전체적으로 우수한 성적이지만, train에 비해 조금 떨어진 test 성능

Train R^2와 Test R^2의 차이는 k=29가 가장 적지만 k=27과 큰 차이가 없고 나머지 성능을 추가로 고려하여 k=27의 사용을 추