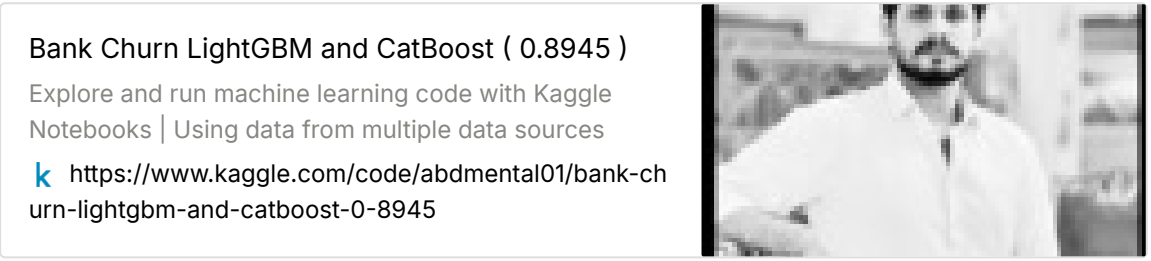
변수 전처리 + 파생변수(성능 중심)

전처리

- **변수 전처리 진행 + 파생변수(성능중심)**
- 전처리 Kaggle 참조



1. **이상치**
    - **의미**
        - 의미상 존재할 수 없는 이상치 존재하지 않음
    - **시각화**
        - BoxPlot을 통한 이상치 Age, CreditScore 존재
            - 의미상 제거하기엔 무리가 있으므로 데이터 보존

2. **로그변환 & 스케일링**
    - **연속형 변수 대상으로 진행**
        - Age, Balance, CreditScore, EstimatedSalary
    - **로그 변환**
        - Age, Balance
    - **스케일링**
        - **Age, CreditScore**
            - Robust Scailing

    - **Balance**
        - MinMax Scaling
    - **EstimatedSalary**
        - Standard Scaling

3. **파생변수**
    - **TF-IDF, SVD를 통한 파생변수**
        - **TF-IDF**
            - 상위 1000개 주요 단어 벡터화 사용
            - TF-IDF 계산 및 변환
        - **SVD**
            - TF-IDF 행렬 3개 주요 성분으로 압축
            - SVD 학습 및 변환
        - **대상**
            - Surname
            - Sur_Geo_Gend_Sal
                - CustomerId + Surname + Geography + Gender + EstimatedSalary

    - **단일 범주 파생변수**
        - Senior : 고령층
            - Age ≥ 60 : 1
            - Age < 60 : 0
        - AgeCat : 연령대
            - 연령 20살 단위로 분해
                - 0-9 : 0
                - 10-29 : 1

        - 30-49 : 2
        - 50-69 : 3
        - 70-89 : 4
        - 90-109 : 5

- **변수 조합**
    - **Active_By_CreditCard** : 신용카드 보유 여부와 활성화 멤버 관계
        - HasCrCard * IsActiveMember
    - **Products_Per_Tenure** : 이용 상품 대비 은행 이용 기간
        - Tenure / NumOfProducts

4. **인코딩**
    - **OneHot Encoding**
        - AgeCat
        - Geography
        - Gender
        - NumOfProducts

다중공선성

: VIF 값 10 이상 다중공선성 문제 판단

Feature	VIF	Feature	VIF
Tenure	9.336355	Products_Per_Tenure	12.44187
HasCrCard	2.043815	Geography_France	inf
IsActiveMember	4.08547	Geography_Germany	inf
CreditScore_Scaled	1.001171	Geography_Spain	inf
EstimatedSalary_Scaled	1.001988	Gender_Female	inf
Age_Scaled	3.398987	Gender_Male	inf

Feature	VIF	Feature	VIF
Balance_Scaled	1.793245	NumOfProducts_1	inf
Surname_tfidf_0	1.014683	NumOfProducts_2	inf
Surname_tfidf_1	1.001599	NumOfProducts_3	inf
Surname_tfidf_2	1.019276	NumOfProducts_4	inf
Sur_Geo_Gend_Sal_tfidf_0	1.011713	AgeCat_1	inf
Sur_Geo_Gend_Sal_tfidf_1	1.004023	AgeCat_2	inf
Sur_Geo_Gend_Sal_tfidf_2	1.007042	AgeCat_3	inf
Senior	1.675373	AgeCat_4	inf
Active_by_CreditCard	5.035984	AgeCat_5	inf

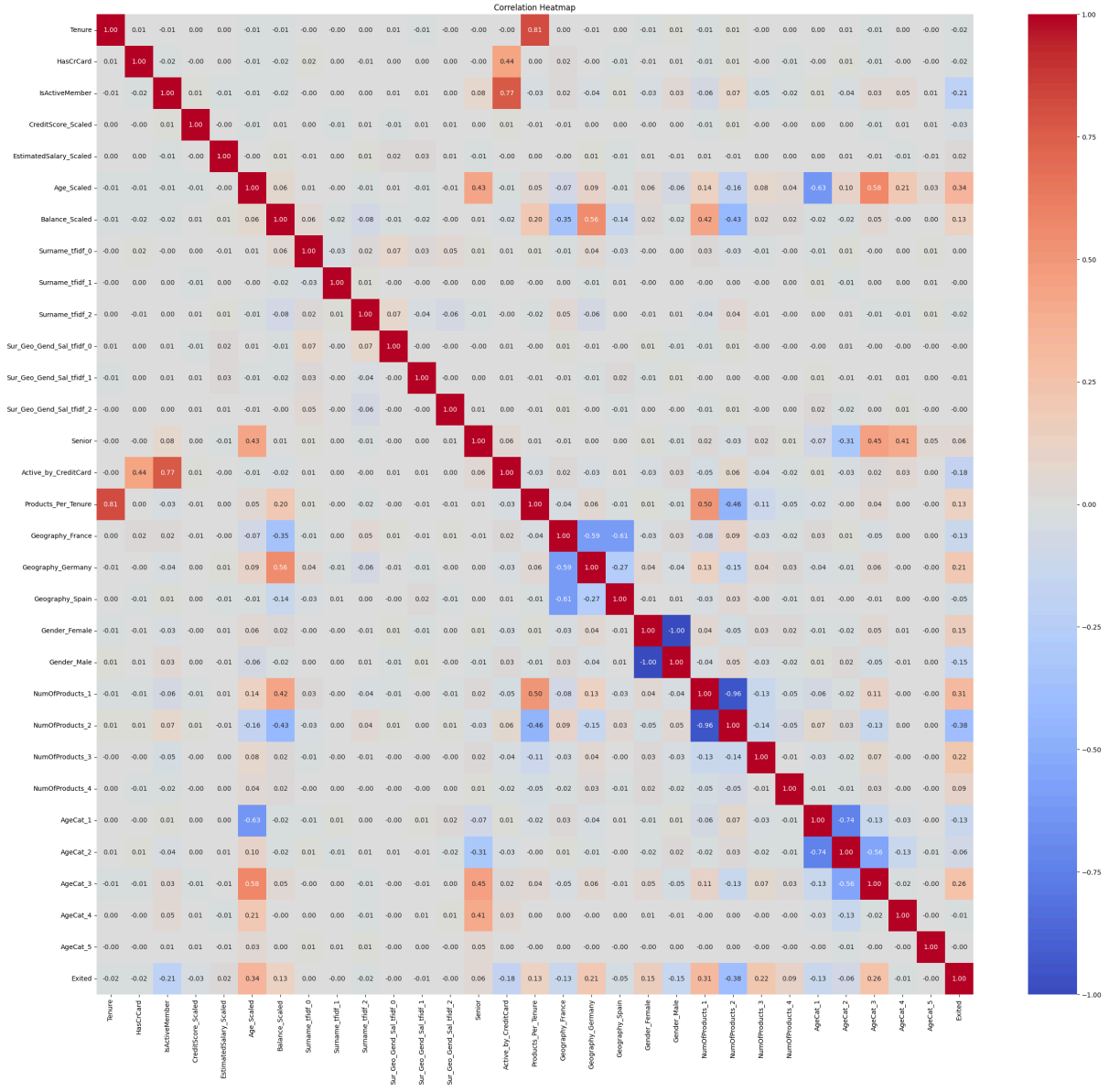
- **5 - 10**
  - Tenure : 9.336355
  - Active_by_CreditCard : 5.035984
- > 10
  - Products_Per_Tenure : 12.44187

- 의미적으로 유의할 수 있는 가능성이 있으므로 유지하고 분석 진행
  - 최종 모델에서의 feature_importance, permutation_importance를 통해 제거 여부 재차 확인

추가 시각화

: 더미변수끼리의 상관계수 비교 제외

[Correlation Heatmap]



- **Exited**
  - 양
    - **Age_Scaled(0.34)**
      - 연령이 높을수록 이탈 가능성이 다소 증가
    - NumOfProducts_1(0.31)
      - 이용중인 상품이 1개일 경우 이탈 가능성 다소 증가
    - AgeCat_2(0.26)
      - 30세에서 49세 사이의 고객의 이탈 가능성 다소 증가
  - 음

        - **NumOfProducts_2(-0.38)**
            - 이용중인 상품이 2개일 경우 이탈 가능성이 다소 감소
        - IsActiveMember(-0.21)
            - 활성 멤버일 수록 이탈 가능성이 다소 감소
        - Active_By_CreditCard(-0.18)
            - 신용카드를 보유한 고객 중 활성 고객은 이탈 가능성이 다소 낮음
- **전체 상관계수**
    - 다수의 파생변수 생성으로 변수간 상관관계가 복잡해지고 중복성이 높아짐
    - 수치의 신뢰성과 해석력 저하
    - 전체 상관계수 분석은 참고용으로 활용

분석 모델링

1. **데이터 분리**
    - train = 70%
    - test = 30%

2. **사용 모델 결정**
    - **AutoML Top 5(AUC 기준)**
        - CatBoost : CatBoost Classifier
        - LightGBM : Light Gradient Boosting Machine
        - GBC: Gradient Boosting Classifier
        - XGBoost: Extreme Gradient Boosting
        - AdaBoost: Ada Boost Classifier

3. **하이퍼 파라미터 최적화**

- **Optuna + StratifiedKFold : AWS에서 제공하는 모델 별 하이퍼 파라미터 목록 사용**
  - **CatBoost Hyper Parameters**

```
Best AUC: 0.8929926718960395
Best hyperparameters:
  iterations: 567
  learning_rate: 0.02900858676615259
  depth: 7
  l2_leaf_reg: 1.2647515608981543
  random_strength: 0.23498389010132387
  bagging_temperature: 0.024432250592590334
  grow_policy: Depthwise
  border_count: 174
  od_wait: 50
```

  - **LightGBM Hyper Parameters**

```
Best AUC: 0.8932907005145936
Best hyperparameters:
  num_boost_round: 633
  learning_rate: 0.03342462805497592
  num_leaves: 56
  max_depth: 14
  min_data_in_leaf: 52
  feature_fraction: 0.6364882192462488
  bagging_fraction: 0.8584003437311537
  bagging_freq: 2
  min_gain_to_split: 0.14446805858185238
  lambda_l1: 0.07945719422821337
  lambda_l2: 0.09228860158082569
  tree_learner: feature
  max_bin: 277
  early_stopping_rounds: 15
  num_threads: 8
  scale_pos_weight: 4.955235602871113
```

  - **GBC Hyper Parameters**

```
Best AUC: 0.8872064342320748
Best hyperparameters:
  n_estimators: 473
  learning_rate: 0.021945351322294467
  max_depth: 10
  min_samples_split: 2
  min_samples_leaf: 3
  subsample: 0.9750612643600708
  max_features: sqrt
  loss: exponential
  ccp_alpha: 0.00015702247402449507
  validation_fraction: 0.2984753709846782
  n_iter_no_change: 15
  tol: 0.005547267099418254
  min_impurity_decrease: 0.09411712873217182
  max_leaf_nodes: 86
```

- **XGBoost Hyper Parameters**

```
Best AUC: 0.8862153263809549
Best hyperparameters:
  num_round: 360
  alpha: 0.7558611471242505
  base_score: 0.5566055700158926
  booster: gbtree
  colsample_bylevel: 0.652980718854562
  colsample_bynode: 0.6207456906259616
  colsample_bytree: 0.7084669537044522
  eta: 0.2931701770417505
  eval_metric: auc
  gamma: 0.38038567264515777
  grow_policy: lossguide
  lambda: 7.6102845156108465
  max_bin: 142
  max_delta_step: 7
  max_depth: 11
  max_leaves: 33
  min_child_weight: 3.6279534873415384
```

```
  objective: binary:logistic
  scale_pos_weight: 4.199769049496863
  seed: 675
  subsample: 0.7974634701882836
  verbosity: 1
  early_stopping_rounds: 100
```

- **AdaBoost Hyper Parameters**

```
Best AUC: 0.8885549835886202
Best hyperparameters:
  n_estimators: 339
  learning_rate: 0.028444996574740207
  algorithm: SAMME.R
  random_state: 582
  max_depth: 5
  min_samples_split: 8
  min_samples_leaf: 3
  max_features: sqrt
  max_leaf_nodes: 12
  min_impurity_decrease: 0.00011591970091207504
```

4. **보팅**
   - **4_1 : 조합 생성**
     - **사용 모델**
       - CatBoost, LightGBM, GBC, XGBoost, AdaBoost
     - 단일 모델부터 최대 5개 모델의 조합까지, 모든 경우의 수를 생성
       - 총 31개의 조합에 대해 소프트 보팅 방식으로 평가 진행
   - **4_2 : 가중치 생성**
     - 각 모델의 ROC AUC 점수 기반 가중치 생성
       - 가중치 계산
         - 전체 모델 ROC AUC 점수 합산하여 전체 점수 계산
         - 각 모델의 가중치

        - 해당 모델 ROC AUC score / 전체 모델 ROC AUC 점수 합산

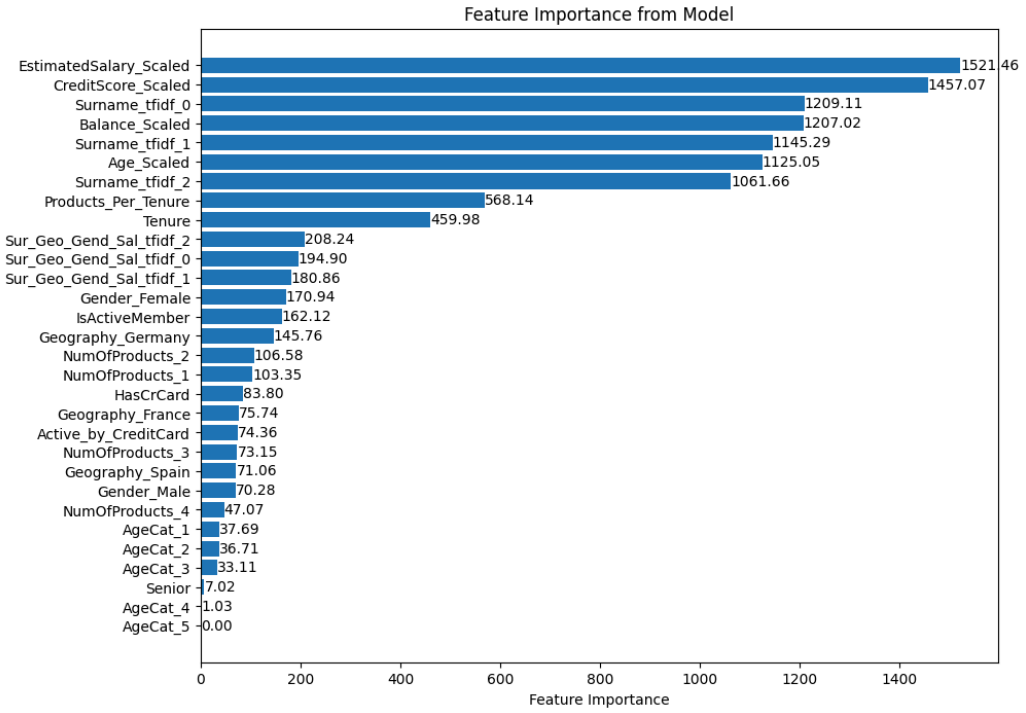
    - **단일 모델 점수**
        - **LightGBM: 0.8932907005145936**
        - CatBoost: 0.8929926718960395
        - AdaBoost: 0.8885549835886202
        - GBC: 0.8872064342320748
        - XGBoost: 0.8862153263809549

    - **생성된 가중치**
        - LightGBM: 0.20081800009368792
        - CatBoost: 0.20075100117484296
        - AdaBoost: 0.1997533777915236
        - GBC: 0.19945021445997535
        - XGBoost: 0.1992274064799702

5. **보팅 최적 모델**
    - **모델**
        - LightGBM + CatBoost + AdaBoost
    - **성능**
        - **Best AUC : 0.8938**
    - **Fold AUCs**
        - 0.8934978365688402
        - 0.8938326442565854
        - 0.8955069275443579
        - 0.894669072207884
        - 0.8917147260538993
    - **가중치**

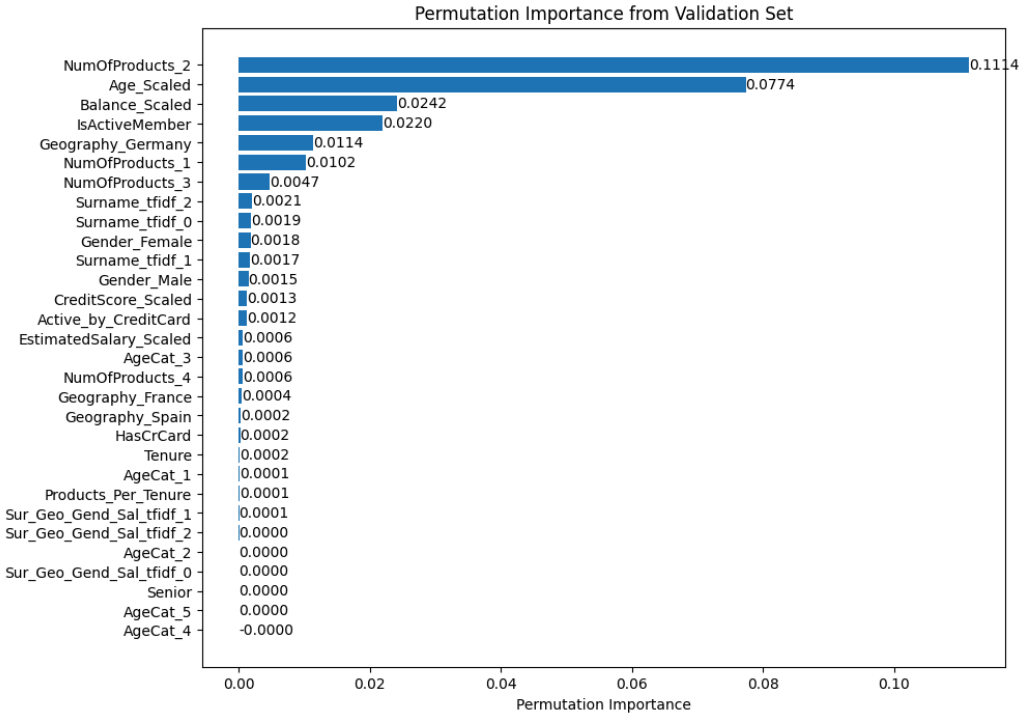
- 0.20081800009368792
- 0.20075100117484296
- 0.1997533777915236

- **Feature Importance & Permutation Importance**
  - **Feature Importance**



    - **Top 5**
      - EstimatedSalary_Scaled
      - CreditScore_Scaled
      - Surname_tfidf_0
      - Balance_Scaled
      - Surname_tfidf_1

  - **Permutation importance**



- **Top 5**
  - NumOfProducts_2
  - Age_Scaled
  - Balance_Scaled
  - IsActiveMember
  - Geography_Germany

- **결론**
  - Balance_Scaled가 공통적으로 포함
  - **제거 할 변수**
    - **Products_Per_Tenure**
      - Feature importance : 568.14
      - Permutation importance : 0.0001
      - VIF : 12.44187

---

6. 중요도 기반 추가 모델링

- **6_1. 사용모델**
  - **AutoML**
    - CatBoost
    - LightGBM
    - GBC
    - XGBoost
    - AdaBoost

- **6_2. 하이퍼 파라미터 최적화**
  - **GBC Hyper Parameters**

```
Best AUC: 0.8873325360161234
Best hyperparameters:
  n_estimators: 473
  learning_rate: 0.021945351322294467
  max_depth: 10
  min_samples_split: 2
  min_samples_leaf: 3
  subsample: 0.9750612643600708
  max_features: sqrt
  loss: exponential
  ccp_alpha: 0.00015702247402449507
  validation_fraction: 0.2984753709846782
  n_iter_no_change: 15
  tol: 0.005547267099418254
  min_impurity_decrease: 0.09411712873217182
  max_leaf_nodes: 86
```

  - **LightGBM Hyper Parameters**

```
Best AUC: 0.8927221334546112
Best hyperparameters:
  num_boost_round: 500
  learning_rate: 0.029888796349948284
  num_leaves: 65
```

```
  max_depth: 9
  min_data_in_leaf: 32
  feature_fraction: 0.6358713698824439
  bagging_fraction: 0.8661769682816917
  bagging_freq: 9
  min_gain_to_split: 0.045680108972356276
  lambda_l1: 0.04081755440857283
  lambda_l2: 0.026788484813059954
  tree_learner: data
  max_bin: 268
  early_stopping_rounds: 23
  num_threads: 4
  scale_pos_weight: 3.4872207884352115
```

- **XGBoost Hyper Parameters**

```
Best AUC: 0.8882229217305595
Best hyperparameters:
  num_round: 106
  alpha: 0.6330437539243525
  base_score: 0.39870903675669067
  booster: gbtree
  colsample_bylevel: 0.9923227855716938
  colsample_bynode: 0.4151277034733893
  colsample_bytree: 0.997498141356141
  eta: 0.27192150582147717
  eval_metric: auc
  gamma: 0.04712099219003624
  grow_policy: lossguide
  lambda: 2.852528179903329
  max_bin: 341
  max_delta_step: 4
  max_depth: 7
  max_leaves: 0
  min_child_weight: 4.542860426764745
  objective: binary:logistic
  scale_pos_weight: 1.6658131783101051
  seed: 141
```

```
subsample: 0.6819570352186918
verbosity: 3
early_stopping_rounds: 24
```

- **AdaBoost Hyper Parameters**

```
Best AUC: 0.8886618978806397
Best hyperparameters:
  n_estimators: 266
  learning_rate: 0.09909004448349125
  algorithm: SAMME.R
  random_state: 519
  max_depth: 5
  min_samples_split: 9
  min_samples_leaf: 6
  max_features: sqrt
  max_leaf_nodes: 41
  min_impurity_decrease: 0.00015191581774246914
```

- **CatBoost Hyper Parameters**

```
Best AUC: 0.8930844927272021
Best hyperparameters:
  iterations: 535
  learning_rate: 0.06964529408502032
  depth: 6
  l2_leaf_reg: 0.027235876696900984
  random_strength: 4.836212370043395
  bagging_temperature: 1.9799984680862954
  grow_policy: Lossguide
  border_count: 166
  od_wait: 28
```

- **6_3. 보팅**
  - **6_3_1. 조합 생성**
    - **사용 모델**
      - CatBoost, LightGBM, GBC, XGBoost, AdaBoost

        - 단일 모델부터 최대 5개 모델의 조합까지, 모든 경우의 수를 생성
            - 총 31개의 조합에 대해 소프트 보팅 방식으로 평가 진행

    - **6_3_2. 가중치 생성**
        - 각 모델의 ROC AUC 점수 기반 가중치 생성
            - 가중치 계산
                - 전체 모델 ROC AUC 점수 합산하여 전체 점수 계산
                - 각 모델의 가중치
                    - 해당 모델 ROC AUC score / 전체 모델 ROC AUC 점수 합산

        - **단일 모델 점수**
            - **CatBoost: 0.8930844927272021**
            - LightGBM: 0.8927221334546112
            - AdaBoost: 0.8886618978806397
            - XGBoost: 0.8882229217305595
            - GBC: 0.8873325360161234

        - **생성된 가중치**
            - CatBoost: 0.2006920628693158
            - LightGBM: 0.20061063425812803
            - AdaBoost: 0.19969822668671516
            - XGBoost: 0.19959958089247345
            - GBC: 0.1993994952933675
- **6_4. 보팅 최적 모델**
    - **모델**
        - LightGBM + AdaBoost + CatBoost
    - **성능**

        - **Best AUC: 0.8937**
    - **Fold AUCs**
        - 0.89347933808576
        - 0.8939094168129269
        - 0.8958242243985506
        - 0.8955682948083079
        - 0.8919839217903821
    - **가중치**
        - 0.20061063425812803
        - 0.19969822668671516
        - 0.2006920628693158

---

7. **최종 모델 선정**
    - Kaggle 점수 기반 판단

	중요도 반영 전 ✔	중요도 반영 후
Colab ROC Score	**0.8936**	0.8939
Kaggle Roc Score_Private	**0.89251**	0.89219
Kaggle Roc Score_Public	**0.88890**	0.88868