변수 전처리 + 파생변수(의미 중심)

1️⃣ 전처리

: 변수 처리 진행 + 파생변수(의미중심)

1. **변수 제거**
   - ID : 순번
   - Surname : 성

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

3. **중복치**
   - 모든 열 동일한 경우는 발생할 수 없는 수치 판단 후 제거
     - 30개

4. **인코딩**
   - **Object**
     - Geography : OneHot Encoding
     - Gender : Label Encoding
   - **int64**, **float64**
     - Tenure : Label Encoding
     - NumOfProducts : OneHot Encoding

- HasCrCard : Label Encoding
- IsActiveMember : Label Encoding

5. **로그변환 & 스케일링**
   - **연속형 변수 대상으로 진행**
     - Age, Balance, CreditScore, EstimatedSalary
   - **로그 변환**
     - Age, Balance
   - **스케일링**
     - **Age, CreditScore**
       - Robust Scaling
     - **Balance**
       - MinMax Scailing
     - **EstimatedSalary**
       - Standard Scaling

6. **파생 변수**
   - **의미 중심으로 생성**
     - **변수 파악**
       - **이탈**
         - **낮은 고객 서비스**
           - Tenure, NumOfProducts, HasCrCard, IsActiveMember
         - **시장 이해 부족**
           - Age, Geography, Gender
         - **가격 최적화 부족**
           - Balance, EstimatedSalary
         - **상품 시장 적합성 부족**
           - Tenure, Balance, EstimatedSalary, NumOfProducts, CreditScore

- **참여도 낮은 상품**
    - IsActiveMember, NumOfProducts
- 유입
    - **고객 니즈 파악**
        - Age, NumOrProducts, Geogrphy
    - **선제적 지원 제공**
        - CreditScore, IsActiveMember, Balance
    - **타겟 고객 파악**
        - EstimatedSalary, Geography, NumOfProducts
    - **적극적인 소통**
        - IsActiveMember, Tenure
    - **고객 이탈 경고 징후 인지**
        - Balance, CreditScore, NumOfProducts
    - **고객 긍정적 경험 이해**
        - Age, HasCrCard, EstimatedSalary
    - **고객 여정 지도 구축**
        - Tenure, NumOfProducts, Balance

- **단일 변수 범주화**
    - **Age_group**
        - Age 이상치 기반
            - 10대, 20대, 30대, 40대, 50대, 57세 이상
    - **Balance_group**
        - 0과 나머지
    - **NumOfProducts_group**
        - 0과 1, 나머지
- **도메인 지식 기반 변수 조합**
    - **고객 충성도(Tenure, NumOfPrudcts)**
        - Age + Tenure : 연령에 따른 은행과의 관계 지속성
        - Age + NumOfProducts : 연령별 이용 상품 수

- Balance + NumOfProducts : 잔고 수준과 상품 이용 패턴 분석
- EstimatedSalary + NumOfProducts : 상품 이용 개수 대비 예상 소득

- **경제적 상황(Balance, EstimatedSalary, HasCrCard)**
  - Balance + EstimatedSalary : 예상 소득 대비 잔고 수준
  - Balance + Tenure : 은행 이용 기간 대비 잔고 변화
  - Balance + CreditScore : 신용 점수 대비 잔고 수준
  - Age + EstimatedSalary : 연령별 예상 소득
- **고객의 활동성(CreditScore, NumOfProducts, Tenure, IsActiveMember, HasCrCard)**
  - IsActiveMember + NumOfProducts : 활성회원 여부와 상품 이용 개수의 관계
  - Tenure + NumOfPRoducts : 은행 이용 기간 대비 상품 이용 개수
  - IsActiveMember + CreditScore : 활동성과 신용점수의 연관성
  - Tenure + IsActivemember : 이용 기간과 고객 활동성의 상관관계
- **지역적 특성(Geography)**
  - Geography + Balance : 국가별 잔고 수준 차이 분석
  - Geography + IsActiveMember : 국가별 고객 활동성
  - Geogrphy + EstimatedSalary : 국가별 예상 소득
- **리스크 관리(CreditScore, HasCrCard)**
  - CreditScore + Balance : 신용점수 대비 잔고 수준
  - CreditScore + Age : 연령별 신용점수 분포
  - CreditScore + Tenure : 은행 이용 기간 대비 신용점수
  - Balance + HasCrCard : 신용카드 보유 여부와 잔고의 관계

7. **다중공선성**

: VIF 값 10 이상 다중공선성 문제 판단

Feature	VIF	Feature	VIF
CustomerId	0.000159	EstimatedSalary_Age	1.001223
Gender	1.000821	IsActiveMember_NumOfProducts	1.009847

Feature	VIF	Feature	VIF
Tenure	1.109649	NumOfProducts_Tenure	8.294561
HasCrCard	1.000419	IsActiveMember_CreditScore	1.00526
IsActiveMember	1.001078	Tenure_IsActiveMember	1.109744
Age_group	1.010235	Balance_Geography	2.129831
Balance_group	7.060322	Geography_IsActiveMember	1.228856
CreditScore_Scaled	1.005243	EstimatedSalary_Geography	1.000013
EstimatedSalary_Scaled	1.000079	CreditScore_Age	1.006829
Age_Scaled	1.010357	CreditScore_Tenure	10.72824
Balance_Scaled	7.03589	Balance_HasCrCard	4.92636
NumOfProducts_group	1.024231	Geography_France	1.113864
Tenure_Age	1.090783	Geography_Germany	1.367807
NumOfProducts_Age	1.411414	Geography_Spain	1.016694
NumOfProducts_Balance	3.844244	NumOfProducts_1	1.68119
EstimatedSalary_NumOfProducts	1.168519	NumOfProducts_2	1.603802
Balance_EstimatedSalary	1.000541	NumOfProducts_3	1.007192
Balance_Tenure	2.138256	NumOfProducts_4	1.001844
Balance_CreditScore	4.43513		

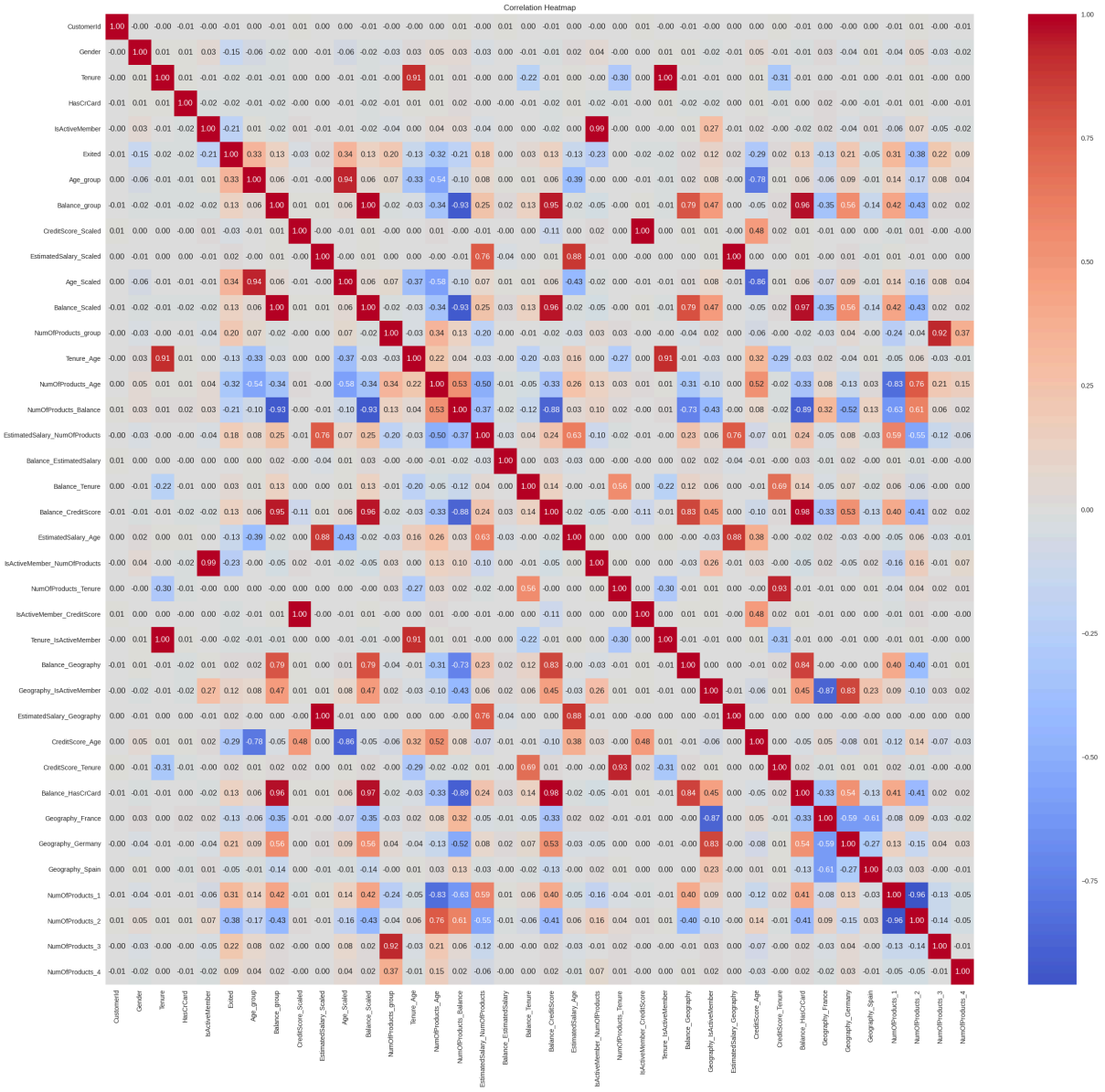
- **5 - 10**
  - Balance_group : 7.060322
  - Balance_Scaled : 7.03589
  - NumOfProducts_Tenure : 8.294561
- **> 10**
  - CreditScore_Tenure : 10.72824

- 의미적으로 유의할 수 있는 가능성이 있으므로 유지하고 분석 진행
  - 최종 모델에서의 Feature importance, Permutation importance를 통해 제거 여부 재차 확인

8. **추가 시각화**

: 더미변수끼리의 상관계수 비교 제외

[Correlation Heatmap]



- **Exited**
  - 양
    - **Age(0.34)**
      - 연령이 높을수록 이탈 가능성이 다소 증가
    - Age_group(0.33)
      - 연령대가 높을수록 이탈 가능성이 다소 증가

        - NumOfProducts_1(0.31)
            - 이용중인 상품이 1개일 경우 이탈 가능성 다소 증가
    - **음**
        - **NumOfProducts_2(-0.38)**
            - 이용중인 상품이 2개일 경우 이탈 가능성이 다소 감소
        - NumOfProducts_Age(-0.32)
            - 연령 대비 이용 상품 수가 높을수록 이탈 가능성 다소 감소
        - CreditScore_Age(-0.29)
            - 연령 대비 신용점수 높을수록 이탈 가능성 다소 감소
- **전체 상관계수**
    - 다수의 파생변수 생성으로 변수간 상관관계가 복잡해지고 중복성이 높아짐
    - 수치의 신뢰성과 해석력 저하
    - 전체 상관계수 분석은 참고용으로 활용

2️⃣ 분석 모델링

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

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

3. **하이퍼 파라미터 최적화**

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

```
Best AUC: 0.8882353624113233
Best hyperparameters:
  n_estimators: 449
  learning_rate: 0.044150878877115926
  max_depth: 6
  min_samples_split: 2
  min_samples_leaf: 4
  subsample: 0.9731842381863204
  max_features: log2
  loss: exponential
  ccp_alpha: 4.9347750239549366e-05
  validation_fraction: 0.2518351573156265
  n_iter_no_change: 18
  tol: 0.004088217530008802
  min_impurity_decrease: 0.03799795246825141
  max_leaf_nodes: 95
```

  - **LightGBM Hyper Parameters**

```
Best AUC: 0.8891460901206308
Best hyperparameters:
  num_boost_round: 532
  learning_rate: 0.04137208387151803
  num_leaves: 46
  max_depth: -1
  min_data_in_leaf: 86
  feature_fraction: 0.8139255851821137
  bagging_fraction: 0.8082068952185747
  bagging_freq: 4
  min_gain_to_split: 0.5814034631622789
  lambda_l1: 0.07098715189710325
  lambda_l2: 0.07241610219010308
  tree_learner: feature
  max_bin: 343
  early_stopping_rounds: 50
```

num_threads: 3
scale_pos_weight: 2.872248166434921

- **XGBoost Hyper Parameters**

Best AUC: 0.8872248463597316
Best hyperparameters:
num_round: 349
alpha: 0.10967146719877353
base_score: 0.8675191316221191
booster: gbtree
colsample_bylevel: 0.6587236837621269
colsample_bynode: 0.4542054877542777
colsample_bytree: 0.8836850766965889
eta: 0.23332571464656857
eval_metric: auc
gamma: 0.2762294374380138
grow_policy: depthwise
lambda: 4.034139666189761
max_bin: 497
max_delta_step: 8
max_depth: 12
max_leaves: 48
min_child_weight: 9.477651521287859
objective: binary:logistic
scale_pos_weight: 8.433168368359874
seed: 434
subsample: 0.9545112785469709
verbosity: 1
early_stopping_rounds: 54

- **AdaBoost Hyper Parameters**

Best AUC: 0.8874794558932251
Best hyperparameters:
n_estimators: 312
learning_rate: 0.06668653847352878
algorithm: SAMME
random_state: 853
max_depth: 8

```
min_samples_split: 3
min_samples_leaf: 6
max_features: log2
max_leaf_nodes: 16
min_impurity_decrease: 0.00026999077869702197
```

- **CatBoost Hyper Parameters**

```
Best AUC: 0.8891316004587454
Best hyperparameters:
  iterations: 193
  learning_rate: 0.05731626408645584
  depth: 5
  l2_leaf_reg: 0.12172037289908884
  random_strength: 0.11159304953837404
  bagging_temperature: 2.0950687664171803
  grow_policy: Lossguide
  border_count: 147
  od_wait: 40
```

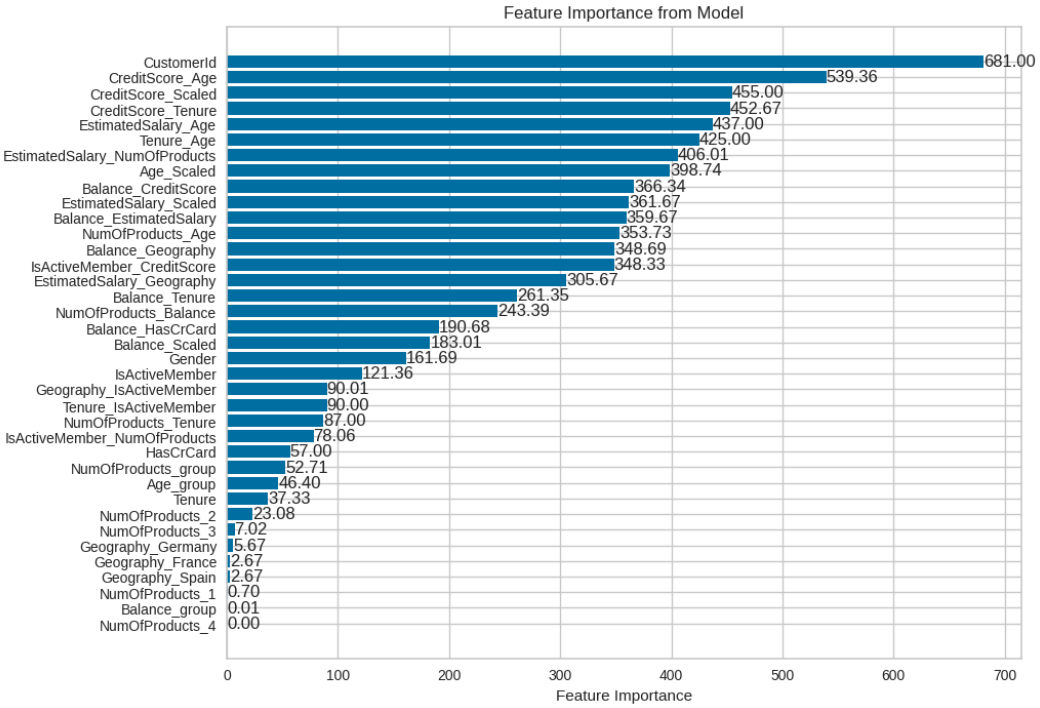
4. **보팅**
   - **4_1 : 조합 생성**
     - **사용 모델**
       - LightGBM, CatBoost, GradientBoosting, AdaBoost, XGBoost
     - 단일 모델부터 최대 5개 모델의 조합까지, 모든 경우의 수를 생성
       - 총 31개의 조합에 대해 소프트 보팅 방식으로 평가 진행
   - **4_2 : 가중치 생성**
     - 각 모델의 ROC AUC 점수 기반 가중치 생성
       - 가중치 계산
         - 전체 모델 ROC AUC 점수 합산하여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

        - **LightGBM: 0.8891460901206308**
        - CatBoost: 0.8891316004587454
        - GBC: 0.8882353624113233
        - AdaBoost: 0.8874794558932251
        - XGBoost: 0.8872248463597316

    - **생성된 가중치**
        - LightGBM: 0.20020323686045985
        - CatBoost: 0.20019997431761935
        - GBC: 0.19999817423090124
        - AdaBoost: 0.1998279716810968
        - XGBoost: 0.19977064290992266

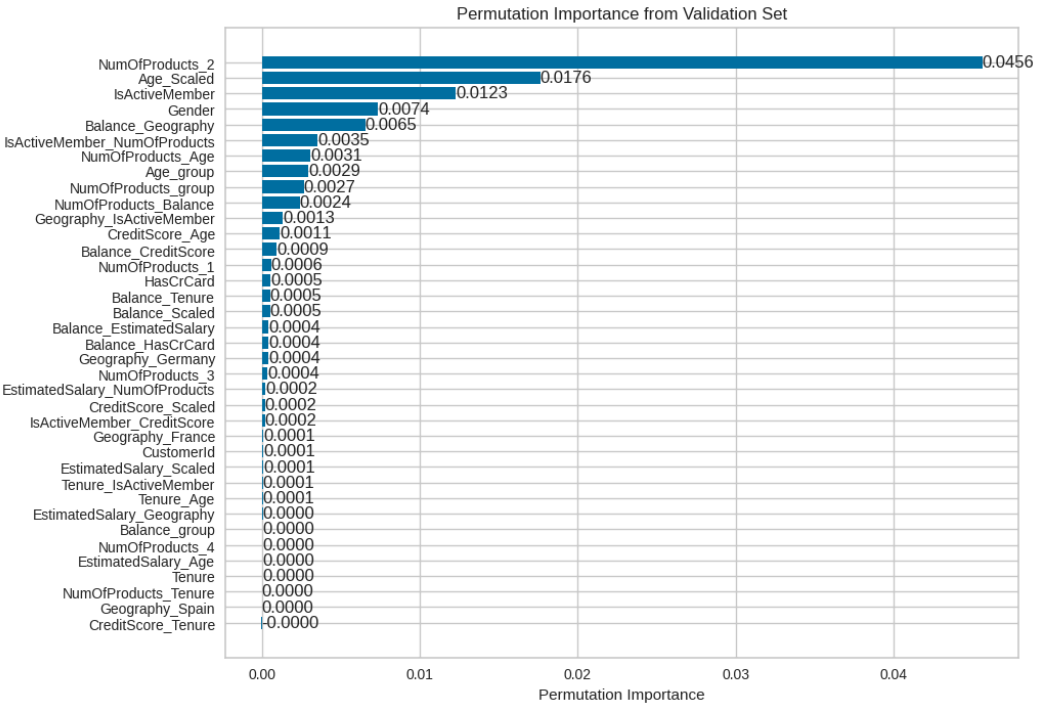
5. **보팅 최적 모델**
    - **모델**
        - CatBoost + LightGBM + Ada
    - **성능**
        - **Best AUC : 0.8892**
    - **Fold AUCs**
        - 0.8875412534269587
        - 0.8897113821342711
        - 0.8917677196721903
        - 0.8890633795677888
        - 0.8880000832908077
    - **가중치**
        - 0.20020323686045985
        - 0.1998279716810968
        - 0.20019997431761935
    - **Feature Importance & Permutation Importance**

- **Feature Importance**



  - **Top 5**
    - CustomerID
    - **CreditScore_Age**
    - CreditScore_Scaled
    - CreditScore_Tenure
    - **EstimatedSalary_Age**

- **Permutation importance**



- **Top 5**
  - NumOfProducts_2
  - **Age_Scaled**
  - IsActiveMember
  - Gender
  - Balance_Geography

- **결론**
  - Age가 포함된 변수가 공통적으로 다수 포함
  - **제거 할 변수**
    - **Balance_group**
      - Feature importance : 0.01
      - Permutation importance : 0.0000
      - VIF : 7.06
    - **CreditScore_Tenure**
      - Feature importance : 681.00
      - Permutation importance : 0.0000
      - VIF : 10.72824

6. **중요도 기반 추가 모델링**
   - **6_1. 사용모델**
     - **AutoML**
       - GBC
       - LightGBM
       - CatBoost
       - AdaBoost
       - XGBoost
   - **6_2. 하이퍼 파라미터 최적화**
     - **GBC Hyper Parameters**

```
Best AUC: 0.8881716017412291
Best hyperparameters:
  n_estimators: 306
  learning_rate: 0.13703881640424945
  max_depth: 4
  min_samples_split: 3
  min_samples_leaf: 3
  subsample: 0.9445270800712589
  max_features: log2
  loss: exponential
  ccp_alpha: 2.8384467094309904e-05
  validation_fraction: 0.28861432470314685
  n_iter_no_change: 18
  tol: 0.003292168493890127
  min_impurity_decrease: 0.06206179531947495
  max_leaf_nodes: 63
```

     - **LightGBM Hyper Parameters**

```
Best AUC: 0.8894890596238593
Best hyperparameters:
  num_boost_round: 787
  learning_rate: 0.01655085455833034
```

```
  num_leaves: 40
  max_depth: 9
  min_data_in_leaf: 56
  feature_fraction: 0.6418202481719244
  bagging_fraction: 0.645916077947058
  bagging_freq: 2
  min_gain_to_split: 0.5064123433302721
  lambda_l1: 0.07012143653987062
  lambda_l2: 0.09764150870955571
  tree_learner: serial
  max_bin: 386
  early_stopping_rounds: 25
  num_threads: 4
  scale_pos_weight: 3.8434903468270747
```

- **CatBoost Hyper Parameters**

```
Best AUC: 0.8894422429062463
Best hyperparameters:
  iterations: 524
  learning_rate: 0.08616825811471435
  depth: 3
  l2_leaf_reg: 0.3995590038558608
  random_strength: 0.32389909847293613
  bagging_temperature: 0.09678886582783942
  grow_policy: SymmetricTree
  border_count: 96
  od_wait: 40
```

- **AdaBoost Hyper Parameters**

```
Best AUC: 0.8874169652866769
Best hyperparameters:
  n_estimators: 146
  learning_rate: 0.08520098897894984
  algorithm: SAMME.R
  random_state: 802
  max_depth: 5
  min_samples_split: 20
  min_samples_leaf: 3
```

```
  max_features: None
  max_leaf_nodes: 15
  min_impurity_decrease: 0.0003927552574830781
```

- **XGBoost Hyper Parameters**

```
Best AUC: 0.8877248497094786
Best hyperparameters:
  num_round: 454
  alpha: 0.40652589353810326
  base_score: 0.8480874400205671
  booster: gbtree
  colsample_bylevel: 0.8469354484548803
  colsample_bynode: 0.9995135598162135
  colsample_bytree: 0.9891525620754378
  eta: 0.27890538121311964
  eval_metric: auc
  gamma: 0.9890045644329901
  grow_policy: depthwise
  lambda: 8.316565190053003
  max_bin: 264
  max_delta_step: 9
  max_depth: 6
  max_leaves: 47
  min_child_weight: 1.3266241122837017
  objective: binary:logistic
  scale_pos_weight: 8.66718236582756
  seed: 775
  subsample: 0.8855554855188931
  verbosity: 3
  early_stopping_rounds: 31
```

- **6_3. 보팅**
  - **6_3_1. 조합 생성**
    - **사용 모델**
      - LightGBM, CatBoost, GBC, AdaBoost, XGBoost
    - 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**
    - **LightGBM: 0.8894890596238593**
    - CatBoost: 0.8894422429062463
    - GBC: 0.8881716017412291
    - AdaBoost: 0.8874794558932251
    - XGBoost: 0.8877248497094786

  - **생성된 가중치**
    - LightGBM: 0.20020323686045985
    - CatBoost: 0.2002236029565053
    - GBC: 0.19993756712432656
    - AdaBoost: 0.1997676898433026
    - XGBoost: 0.19983699814175504

- **6_4. 보팅 최적 모델**
  - **모델**
    - LightGBM + AdaBoost + CatBoost
  - **성능**
    - **Best AUC : 0.8897**
  - **Fold AUCs**
    - 0.887680390844541

        - 0.8901420151662042
        - 0.8925261036359157
        - 0.8896599080597734
        - 0.8885884801161994
    - **가중치**
        - 0.20023414193411043
        - 0.1997676898433026
        - 0.2002236029565053

7. **최종 모델 선정**
    - Kaggle 점수 기반 판단

	중요도 반영 전	중요도 반영 후 ✔
Colab ROC Score	0.8905	**0.8907**
Kaggle Roc Score_Private	0.88829	**0.88871**
Kaggle Roc Score_Public	0.88581	**0.88589**
