원본 진행

전처리

: 원본 진행

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

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

3. **다중공선성**

   : VIF 값 10 이상 다중공선성 문제 판단

Feature	VIF
CustomerId	1.000260
CreditScore	1.000737
Gender	1.008577
Age	1.044343

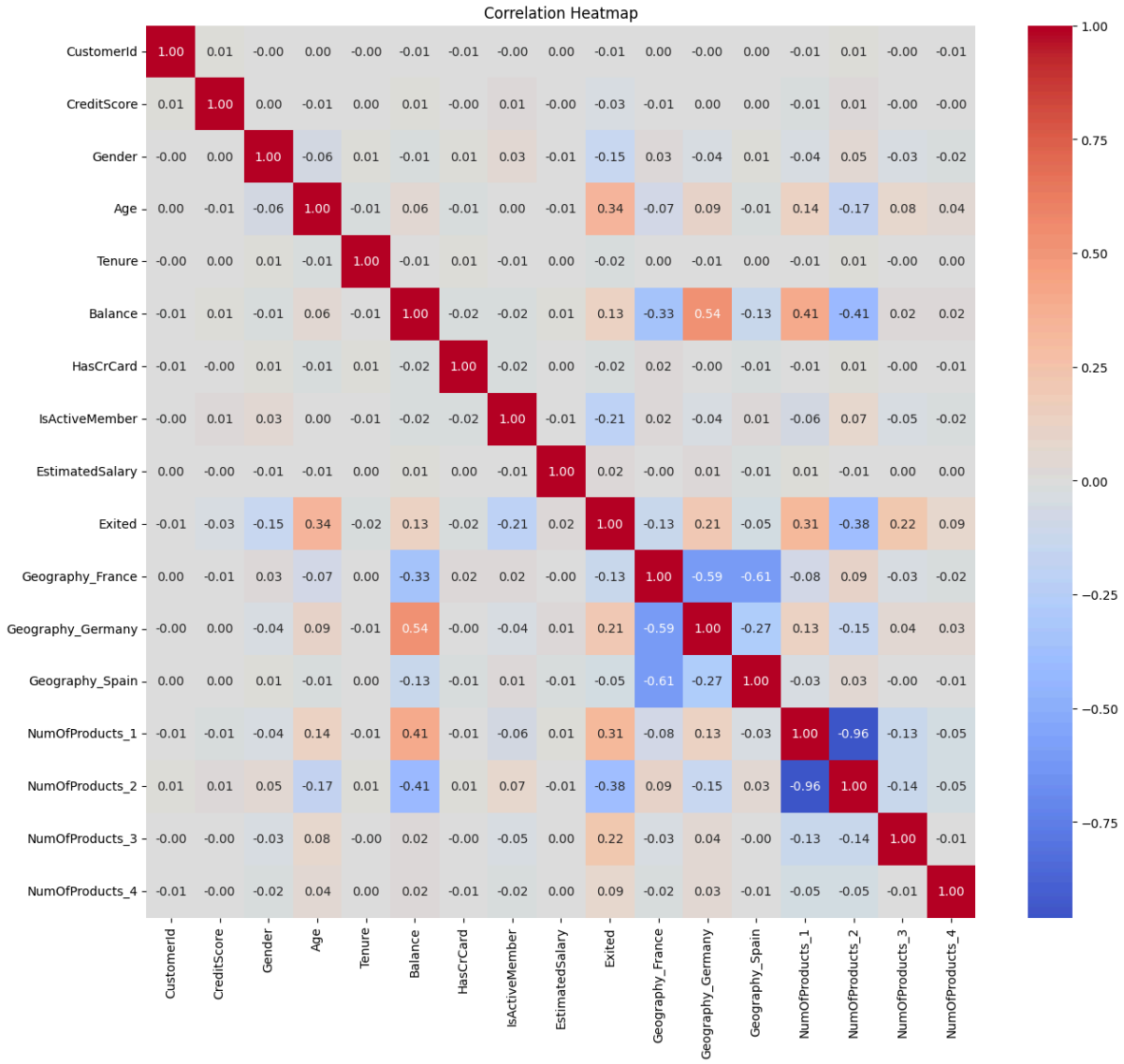
Feature	VIF
Tenure	1.000365
Balance	1.696069
HasCrCard	1.001405
IsActiveMember	1.010977
EstimatedSalary	1.000438
Geography_France	inf
Geography_Germany	inf
Geography_Spain	inf
NumOfProducts_1	inf
NumOfProducts_2	inf
NumOfProducts_3	inf
NumOfProducts_4	inf

- OneHot Encoding으로 생성된 더미 변수를 제외한 모든 변수의 VIF 수치가 10 미만으로 확인
    - 더미 변수의 경우, 인코딩 특성상 다중공선성이 발생하여 VIF 값이 높게 나타나는 것은 자연스러운 결과로 확인

추가 시각화

: 더미변수끼리의 상관계수 비교 제외

[Correlation Heatmap]



- **Exited**
  - 양
    - **Age(0.34)**
      - 연령이 높을수록 이탈 가능성이 다소 증가
    - NumOfProducts_1(0.31)
      - 은행 상품을 1개 보유한 고객의 이탈 가능성이 다소 증가
    - Geography_Germany(0.21)
      - 독일에 거주하는 고객의 이탈 가능성이 다소 증가
    - Balacne(0.13)
      - 잔고가 높은 고객의 이탈 가능성이 약간 증가

- 음
  - **NumOfProducts_2(-0.38)**
    - 은행 상품을 2개 보유한 고객의 이탈 가능성이 다소 감소
  - IsActiveMember(-0.21)
    - 활동적인 고객일수록 이탈 가능성이 다소 감소
  - Gender(-0.15)
    - 남성 고객의 이탈 가능성이 약간 낮음
  - Geography_France(-0.13)
    - 프랑스에 거주하는 고객의 이탈 가능성이 다소 감소

- **전체**
  - 양
    - **Geography_Germany ↔ Balance(0.54)**
      - 독일에 거주하는 고객일수록 잔고가 높은 경향
    - NumOfProducts_1 ↔ Balance(0.41)
      - 은행 상품을 1개 보유한 고객일수록 잔고가 높은 경향
    - Age ↔ NumOfProducts_1(0.14)
      - 연령이 높을수록 은행 상품을 1개 보유할 가능성
    - NumOfProducts_1 ↔ Geography_Germany(0.13)
      - 은행 상품을 1개 보유한 고객이 독일에 거주할 가능성
  - 음
    - **NumOfProducts_2 ↔ Balance (-0.41)**
      - 은행 상품을 2개 보유한 고객일수록 잔고가 낮은 경향
    - Balance ↔ Geography_France (-0.33)
      - 프랑스에 거주하는 고객일수록 잔고가 낮은 경향
    - NumOfProducts_2 ↔ Age (-0.17)
      - 은행 상품을 2개 보유한 고객은 연령이 낮을 가능성
    - NumOfProducts_2 ↔ Geography_Germany(-0.15)

- 은행 상품을 2개 보유한 고객은 독일에 거주할 가능성이 낮음

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: 0.8879903418757644
Best hyperparameters:
  n_estimators: 441
  learning_rate: 0.045696033561884446
  max_depth: 7
  min_samples_split: 2
  min_samples_leaf: 2
  subsample: 0.9753636860597981
  max_features: log2
```

```
  loss: exponential
  ccp_alpha: 2.8658360294722587e-05
  validation_fraction: 0.27267380097000604
  n_iter_no_change: 17
  tol: 0.005721221703263734
  min_impurity_decrease: 0.04492762681125898
  max_leaf_nodes: 65
```

- **LightGBM Hyper Parameters**

```
Best AUC: 0.8893180320851476
Best hyperparameters:
  num_boost_round: 424
  learning_rate: 0.04631415040823912
  num_leaves: 38
  max_depth: 9
  min_data_in_leaf: 47
  feature_fraction: 0.9024601968835392
  bagging_fraction: 0.7342793249627353
  bagging_freq: 2
  min_gain_to_split: 0.3017007245252093
  lambda_l1: 0.07886266683723922
  lambda_l2: 0.0859776680854032
  tree_learner: serial
  max_bin: 310
  early_stopping_rounds: 38
  num_threads: 3
  scale_pos_weight: 3.8048327134060944
```

- **CatBoost Hyper Parameters**

```
Best AUC: 0.8894975131185816
Best hyperparameters:
  iterations: 622
  learning_rate: 0.15620856452750326
  depth: 3
  l2_leaf_reg: 0.023092783762797234
  random_strength: 0.013271842224732251
```

```
  bagging_temperature: 6.81082830068841
  grow_policy: SymmetricTree
  border_count: 117
  od_wait: 15
```

- **XGBoost Hyper Parameters**

```
Best AUC: 0.8855466532778464
Best hyperparameters:
  num_round: 173
  alpha: 0.8296136024447837
  base_score: 0.549188827783536
  booster: gbtree
  colsample_bylevel: 0.6341674144934313
  colsample_bynode: 0.7675482440337507
  colsample_bytree: 0.8777999787302608
  eta: 0.28259482675888387
  eval_metric: auc
  gamma: 0.2819080182819892
  grow_policy: depthwise
  lambda: 3.426745960231184
  max_bin: 340
  max_delta_step: 10
  max_depth: 9
  max_leaves: 33
  min_child_weight: 4.375849619589128
  objective: binary:logistic
  scale_pos_weight: 7.596583814614301
  seed: 55
  subsample: 0.9365805881615846
  verbosity: 2
  early_stopping_rounds: 54
```

- **AdaBoost Hyper Parameters**

```
Best AUC: 0.887546021751948
Best hyperparameters:
  n_estimators: 194
```

```
learning_rate: 0.07457707279949315
algorithm: SAMME.R
random_state: 641
max_depth: 4
min_samples_split: 9
min_samples_leaf: 2
max_features: None
max_leaf_nodes: 52
min_impurity_decrease: 0.0003868616444805657
```

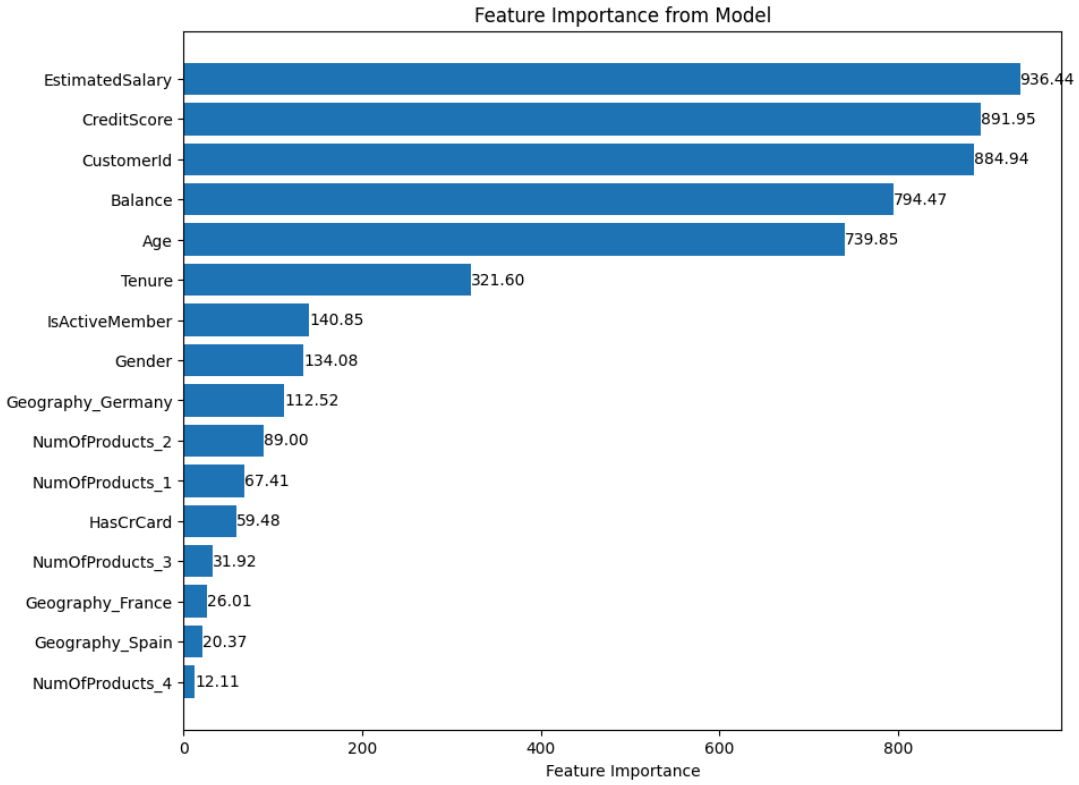
4. **보팅**
    - **4_1. 조합 생성**
        - **사용 모델**
            - CatBoost, LightGBM, GBC, AdaBoost, XGBoost
        - **단일 모델부터 최대 5개 모델의 조합까지, 모든 경우의 수를 생성**
            - 총 31개의 조합에 대해 소프트 보팅 방식으로 평가 진행
    - **4_2. 가중치 생성**
        - 각 모델의 ROC AUC 점수 기반 가중치 생성
            - 가중치 계산
                - 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**
            - **CatBoost: 0.8894975131185816**
            - LightGBM: 0.8893180320851476
            - GBC: 0.8879903418757644
            - AdaBoost: 0.887546021751948
            - XGBoost: 0.8855466532778464

        - **생성된 가중치**
            - CatBoost: 0.20034185481390882
            - LightGBM: 0.20030143023417502
            - GBC: 0.2000023940758461
            - AdaBoost: 0.19990231968955083
            - XGBoost: 0.19945200118651915

5. **보팅 최적 모델**

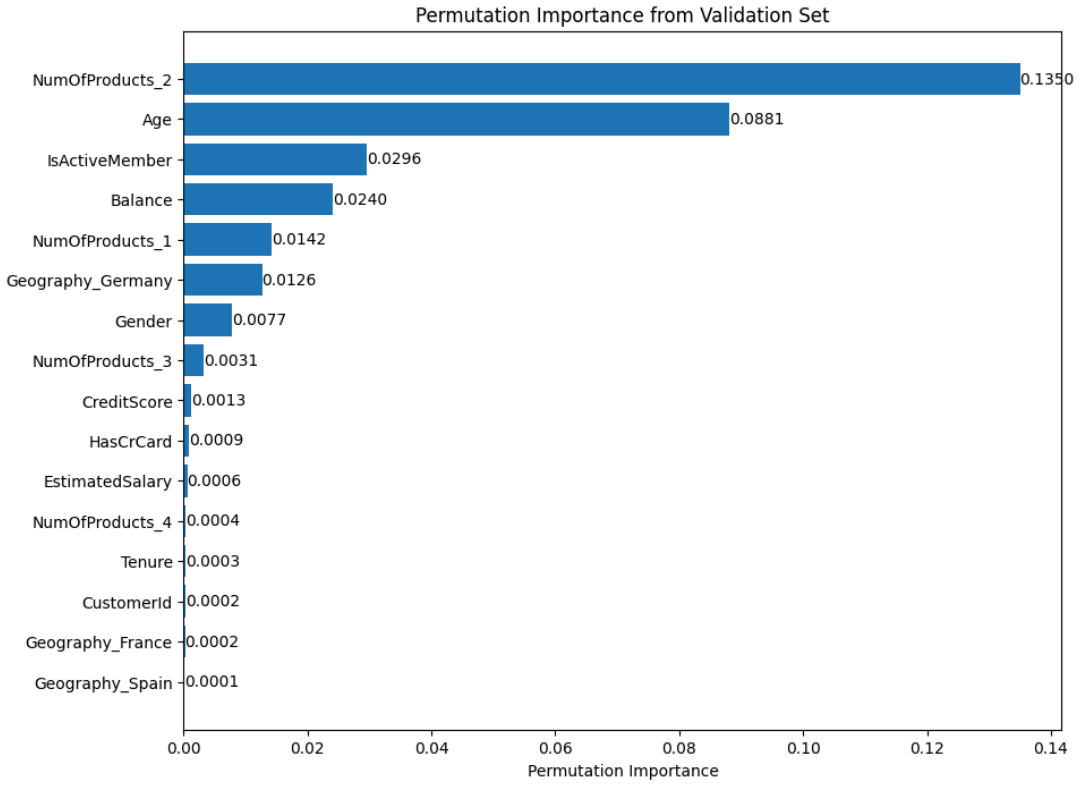
- **모델**
    - CatBoost + LightGBM + GBC
- **성능**
    - **Best AUC : 0.8900**
- **Fold AUCs**
    - 0.8900562631197824
    - 0.8898752206337519
    - 0.8913648217131387
    - 0.8912097065381658
    - 0.8874405758789176
- **가중치**
    - 0.20034185481390882
    - 0.20030143023417502
    - 0.2000023940758461

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



- **Top 5**
  - EstimatedSalary
  - CreditScore
  - **Balance**
  - **Age**
  - Tenure

- **Permutation importance**



- **Top 5**
  - NumOfProducts_2
  - **Age**
  - IsActiveMember
  - **Balance**
  - NumOfProducts_1

- **결론**
  - Age, Balance가 공통적인 중요 변수로 확인

6. 최종 모델 생성

- Colab Final Model ROC Score
  - 0.8902
- Kaggle Final Model ROC Score_Private
  - 0.88890

- Kaggle Final Model ROC Score_Public
  - 0.88641