변수 전처리 진행

전처리

: 변수 처리 진행

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 수치라 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**
        - CreditScore, Age, BAlance, EstimatedSalary
    - **로그 변환**
        - Balance, Age
    - **스케일링**
        - **CreditScore, Age**
            - Robust Scaling
        - **Balance**
            - MinMax Scailing
        - **EstimatedSalary**
            - Standard Scaling

6. **다중공선성**

    : VIF 값 10 이상 다중공선성 문제 판단

Feature	VIF
CustomerId	1.000239
CreditScore	1.000778
Gender	1.008338
Age	1.043246
Tenure	1.000369
Balance	1.784245
HasCrCard	1.001502
IsActiveMember	1.010725

Feature	VIF
EstimatedSalary	1.000418
Geography_France	inf
Geography_Germany	inf
Geography_Spain	inf
NumOfProducts_1	inf
NumOfProducts_2	inf
NumOfProducts_3	inf
NumOfProducts_4	inf

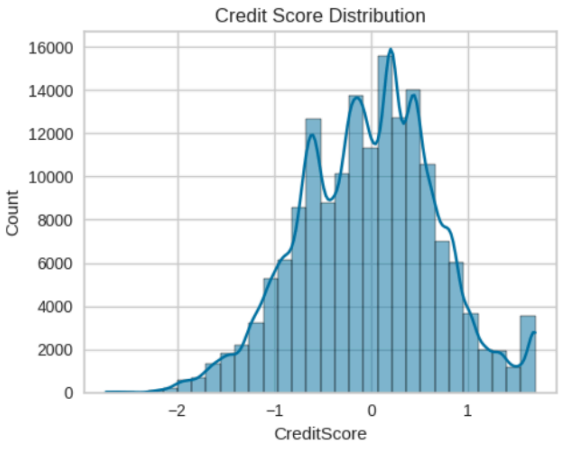
- OneHot Encoding으로 생성된 더미 변수를 제외한 모든 변수의 VIF 수치가 10 미만으로 확인
    - 더미 변수의 경우, 인코딩 특성상 다중공선성이 발생하여 VIF 값이 높게 나타나는 것은 자연스러운 결과로 확인

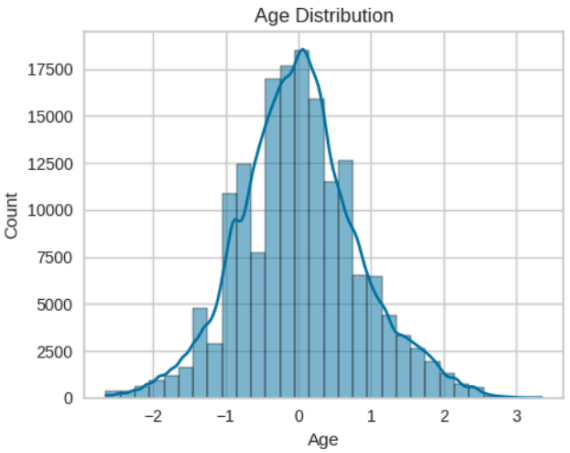
추가 시각화

1. **스케일링 변수 EDA 시각화**

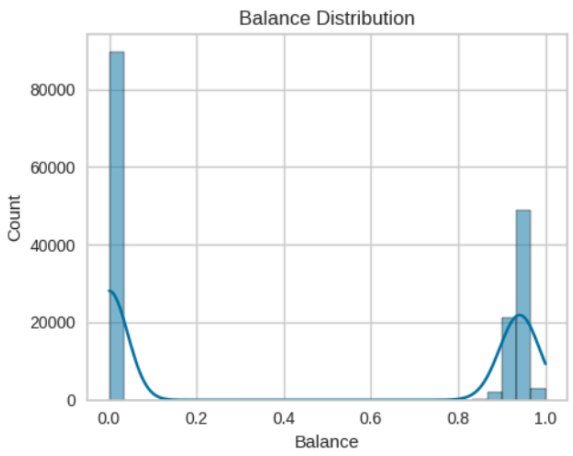
   : train, test 분포 동일

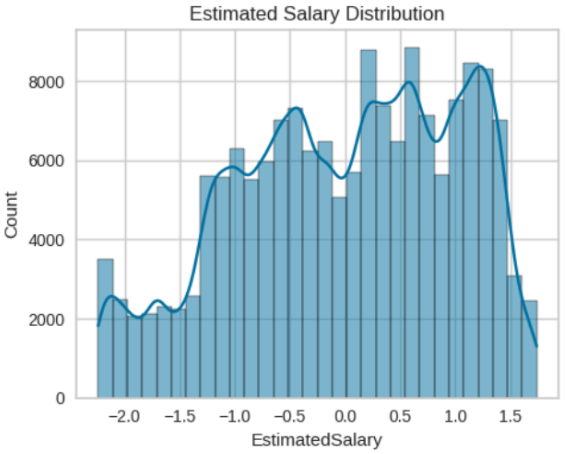
[Train : Histogram,BarPlot]





- **CreditScore(신용점수)**



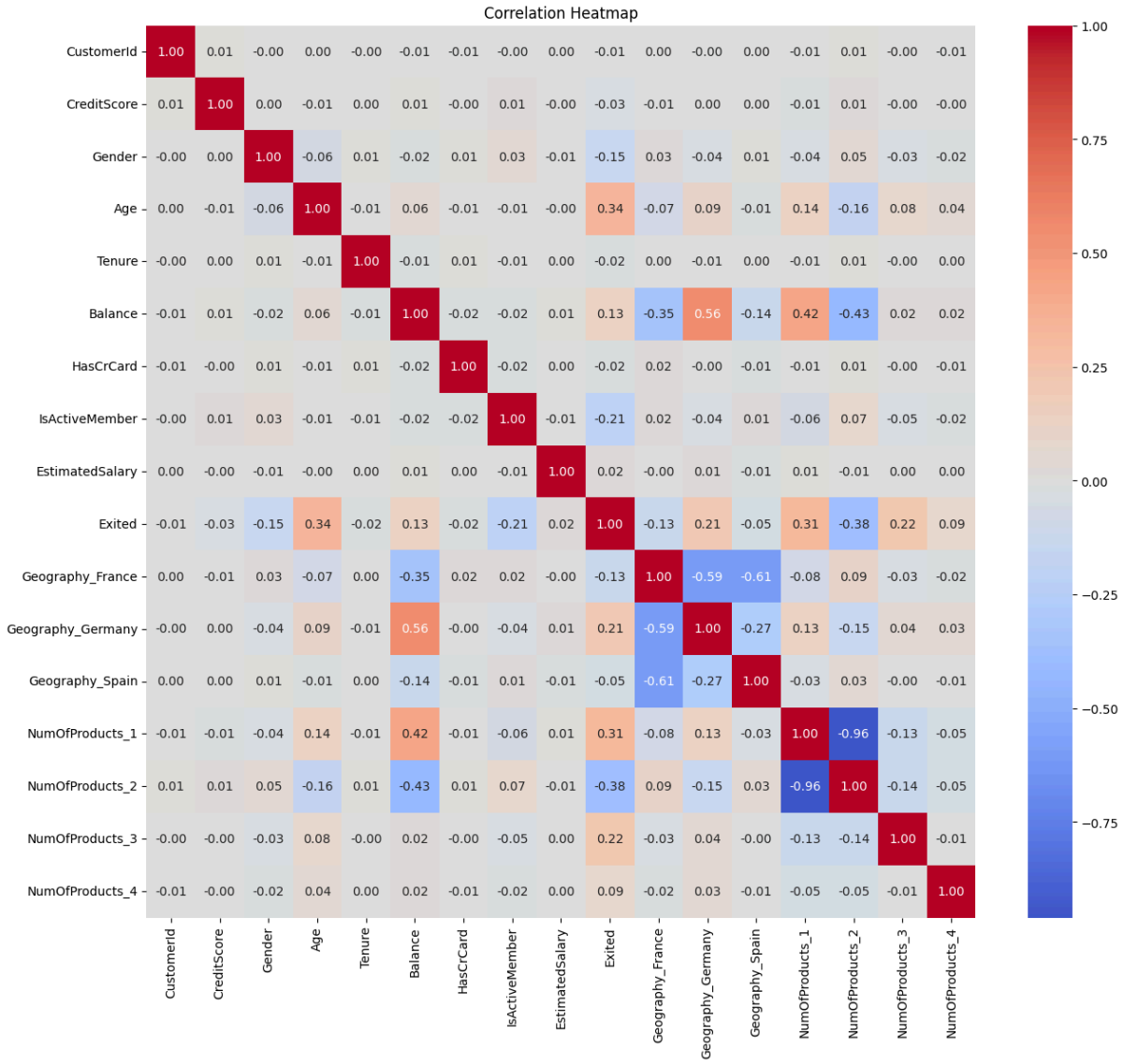


- **Geography(거주 국가)**
- **Gender(성별)**
- **Age(연령)**

2. **상관관계 히트맵**

: 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
    - NumOfProducts_3(0.22)
      - 은행 상품을 3개 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 고객이 여성 고객보다 이탈 가능성이 약간 낮음
        - Geography_France(-0.13)
            - 프랑스에 거주하는 고객의 이탈 가능성이 다소 감소
- **전체**
    - 양
        - **Geography_Germany ↔ Balance(0.56)**
            - 독일에 거주하는 고객일수록 잔고가 높은 경향
        - NumOfProducts_1 ↔ Balance(0.42)
            - 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 (-0.43)**
            - 은행 상품을 2개 보유한 고객일수록 잔고가 낮은 경향
        - Balance ↔ Geography_France (-0.35)
            - 프랑스에 거주하는 고객일수록 잔고가 낮은 경향
        - NumOfProducts_2 ↔ Age (-0.16)

        - 은행 상품을 2개 보유한 고객일수록 연령이 낮은 경향
    - NumOfProducts_2 ↔ Geography_Germany(-0.15)
        - 은행 상품을 2개 보유한 고객은 독일에 거주할 가능성이 낮음

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: 0.8882737168930737
Best hyperparameters:
  n_estimators: 329
  learning_rate: 0.09703126589114959
  max_depth: 6
```

```
  min_samples_split: 2
  min_samples_leaf: 4
  subsample: 0.976775123480814
  max_features: log2
  loss: exponential
  ccp_alpha: 4.9347750239549366e-05
  validation_fraction: 0.22213208226683603
  n_iter_no_change: 18
  tol: 0.007469276485956177
  min_impurity_decrease: 0.09451799070886659
  max_leaf_nodes: 90
```

- **LightGBM Hyper Parameters**

```
Best AUC: 0.8891196891007377
Best hyperparameters:
  num_boost_round: 308
  learning_rate: 0.03385607628427362
  num_leaves: 60
  max_depth: 13
  min_data_in_leaf: 57
  feature_fraction: 0.7527398747336704
  bagging_fraction: 0.9102753061478276
  bagging_freq: 7
  min_gain_to_split: 0.17989632585790152
  lambda_l1: 0.021773719838902653
  lambda_l2: 0.05656663487944709
  tree_learner: data
  max_bin: 261
  early_stopping_rounds: 12
  num_threads: 4
  scale_pos_weight: 1.8556609860594033
```

- **CatBoost Hyper Parameters**

```
Best AUC: 0.8895904379189616
Best hyperparameters:
  iterations: 664
```

learning_rate: 0.15885941694924932
depth: 3
l2_leaf_reg: 0.3127637742132014
random_strength: 0.46541486717008473
bagging_temperature: 0.0028376914309546568
grow_policy: SymmetricTree
border_count: 254
od_wait: 21

- **XGBoost Hyper Parameters**

Best AUC: 0.8861743534751397
Best hyperparameters:
num_round: 230
alpha: 0.6666915824938959
base_score: 0.641507423455919
booster: gbtree
colsample_bylevel: 0.6424294460064837
colsample_bynode: 0.920714093271794
colsample_bytree: 0.9183896232735339
eta: 0.2828202156781443
eval_metric: auc
gamma: 0.08543500110985355
grow_policy: depthwise
lambda: 5.19470705488688
max_bin: 481
max_delta_step: 9
max_depth: 12
max_leaves: 36
min_child_weight: 6.235798272170733
objective: binary:logistic
scale_pos_weight: 7.492561359560108
seed: 149
subsample: 0.7472406764014309
verbosity: 3
early_stopping_rounds: 52

- **AdaBoost Hyper Parameters**

```
Best AUC: 0.8876298369756144
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: 0.0003927552574830783
```

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: 0.8895904379189616**
    - LightGBM: 0.8891196891007377
    - GBC: 0.8882737168930737

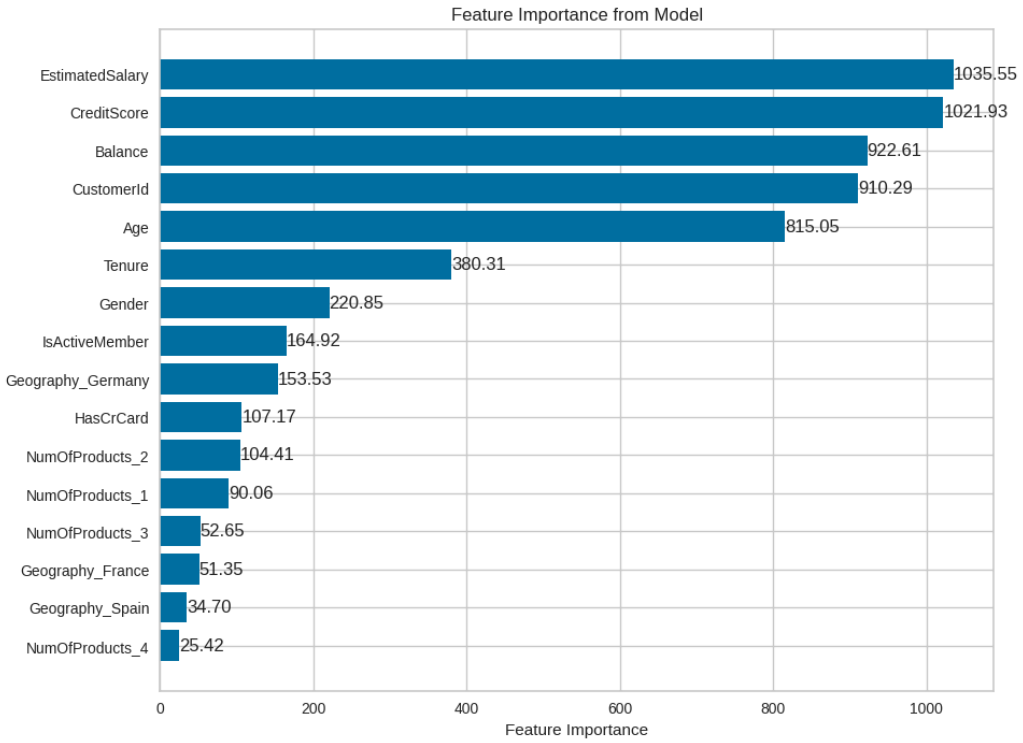
        - AdaBoost: 0.8876298369756144
        - XGBoost: 0.8861743534751397

    - **생성된 가중치**
        - CatBoost: 0.20032265242906638
        - LightGBM: 0.20021664673490103
        - GBC: 0.20002614626491286
        - AdaBoost: 0.1998811539994688
        - XGBoost: 0.19955340057165105

5. **보팅 최적 모델**
    - **모델**
        - CatBoost + LightGBM + AdaBoost
    - **성능**
        - **Best AUC : 0.8900**
    - **Fold AUCs**
        - 0.8880261070649522
        - 0.8908929190970765
        - 0.8924981924905159
        - 0.8899411275730666
        - 0.888671627812549
    - **가중치**
        - 0.20032265242906638
        - 0.20021664673490103
        - 0.1998811539994688

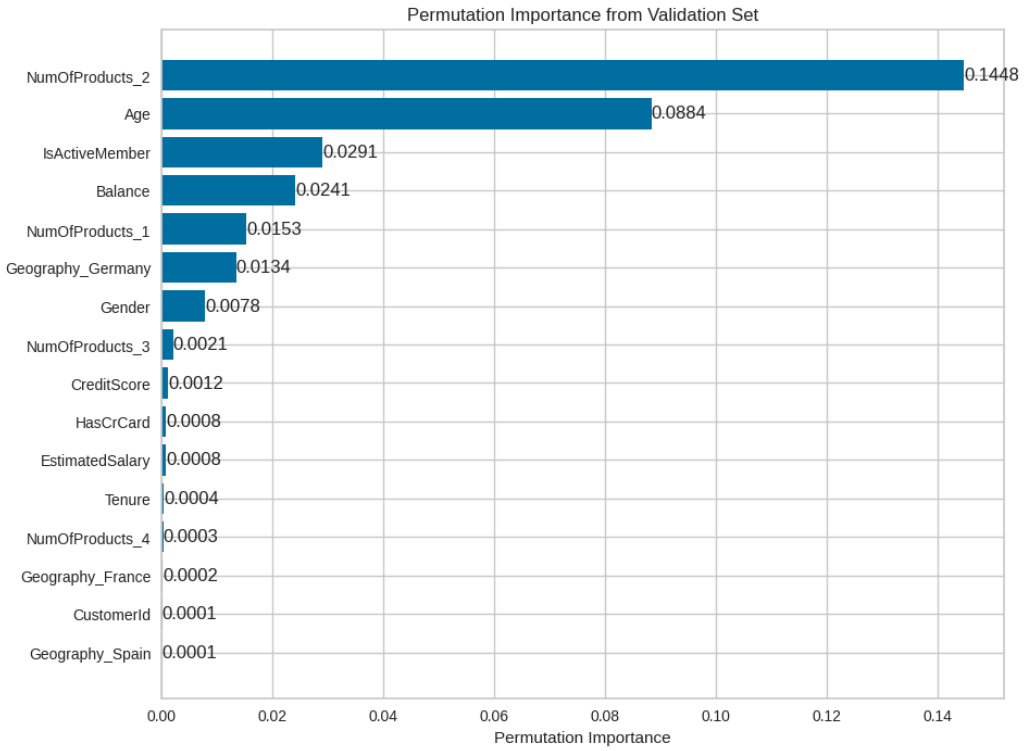
    - **Feature Importance & Permutation Importance**

- **Feature Importance**



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

- **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
  - 0.8911
- Kaggle Final Model ROC Score_Private
  - 0.88891
- Kaggle Final Model ROC Score_Public

- 0.88635