


# RecSys-06 Wrap Up Report

## 비트코인 상승/하락 시계열 분류 예측

### 1. 프로젝트 개요



- 프로젝트 기간 : 2024/09/10 ~ 2024/09/26
- 데이터 정보 : [Cryptoquant Catalog](#) 참고
- 프로젝트 목표

upstage의 비트코인 상승/하락 시계열 분류 예측 대회 참가를 위한 프로젝트.  
비트코인의 시간별 Market Data와 Network Data를 이용해 다음 시점의 등하락을 분류한다.  
2023년의 시간 별 데이터들을 학습해 2024년 1~4월 동안의 다음 시간 비트코인 등락률을 아래의 클래스에 맞게 분류한다.

▼ 타겟 값


클래스	설명	등락률
0	하락	-0.5% 미만
1	소폭 하락	-0.5% ~ 0%
2	소폭 상승	0% ~ 0.5%
3	상승	0.5% 이상

### 2. 프로젝트 팀 구성 및 역할

이름	역할
<a href="#">김건울</a>	도메인 스터디 진행, EDA, 피쳐 엔지니어링, 문서화
<a href="#">백우성</a>	데이터 분석, 시각화,EDA
<a href="#">유대선</a>	프로젝트 구조 설계, 모델 파이프라인 설계 ,EDA, 데이터 전처리, 모델링
<a href="#">이제준</a>	Feature Engineering, 모델링
<a href="#">황태결</a>	특징 분석, TSCV 구축, EDA

### 3. 프로젝트 수행 절차 및 방법

#### 프로젝트 진행 과정



- 비트코인 도메인 지식을 위한 스터디 진행 후, 주어진 데이터에 대해서 토론.
- 프로젝트를 위한 기본 구조 설립 및 코드 작성. (유대선)
- 팀원 별로 EDA와 feature engineering을 진행.
- 각자 도출한 결과에 대해 공유하고 토론을 진행해 feature와 model 선택
- 모델 훈련 및 하이퍼파라미터 튜닝
- 최종 제출 선택

## 협업 방식

- Slack : 팀 간 실시간 커뮤니케이션, 이슈 공유, 질의 응답을 위한 소통 채널
- Zoom : 정기적인 회의와 토론을 위해 사용
- GitHub : 버전 관리와 코드 협업을 위해 사용. 각 팀원은 EDA를 제외하면 기능 단위로 이슈와 브랜치를 만들어 작업했고, Pull Request를 통해 코드 리뷰 후 병합하는 방식으로 진행

## 4. 프로젝트 수행 결과

### EDA & Feature Engineering 과정



우선 도메인 지식 스터디를 진행한 뒤, 팀원 전체가 각자 가설을 세우고, 그에 따른 EDA를 따로 진행한 뒤, 토론을 통해 새로 Feature Engineering을 진행.  
(각자 EDA 한 내용은 github의 EDA-개인별 폴더에 정리)

#### 1. 주어진 Market 데이터 셋에서 각 거래소 별 데이터는 All Exchange에 포함되어 있기 때문에 사용하지 않았다.



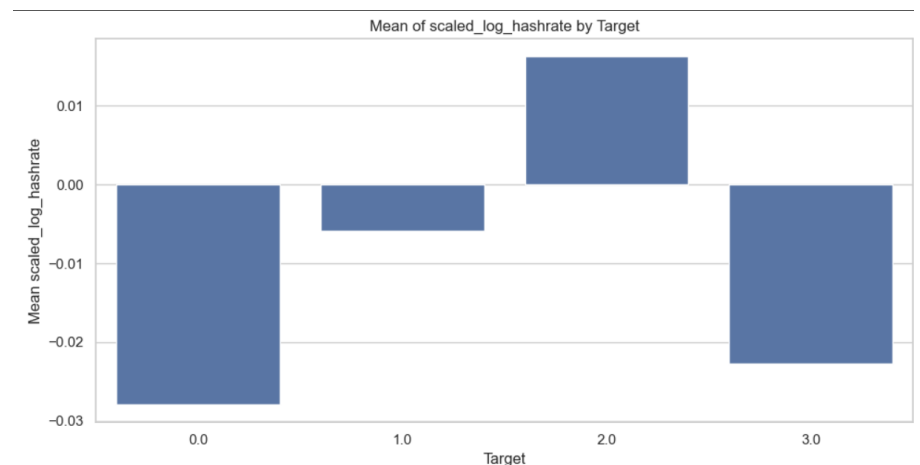
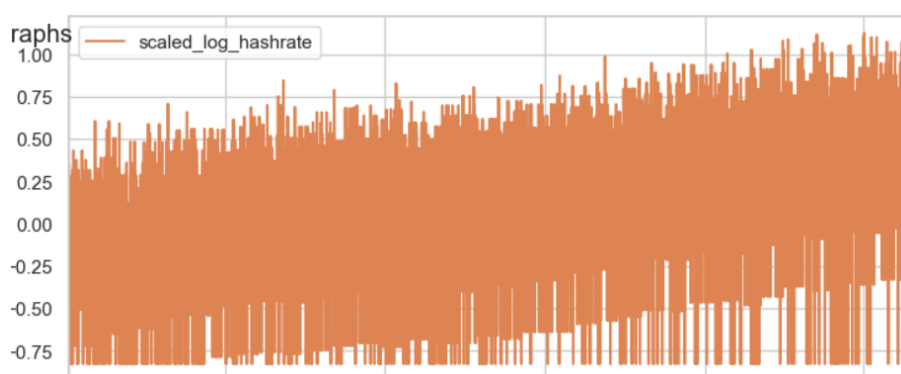
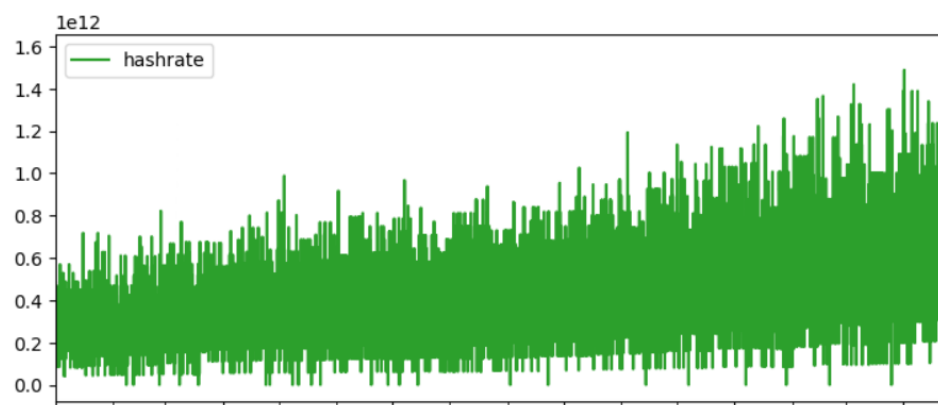
Funding-Rates, Liquidations, Open-Interest, Buy-Sell-Stats 데이터들은 각 거래소 데이터 셋도 포함되어 있지만, 이는 전체 비트코인 데이터를 대표하지 못하기 때문에 All Exchange 데이터만 사용한다.

#### 2. Network 데이터는 논리적으로 비트코인 시장 가격과 연관성이 있다고 생각하는 데이터를 추론한 뒤 target과의 연관성을 파악해서 feature engineering을 한다.

##### 2-1. hashrate

비트코인 전체 컴퓨팅 파워를 나타내는 지표.

- target과의 상관관계는 0.001337로 낮지만, ID에 따른 추세가 우상향하는 지표고 분산이 적절해 쓰기 좋다고 판단했다. 분포를 정규 분포에 가깝게 만들기 위해 log 변환 후 StandardScaler()를 사용해 scaled\_log\_hashrate 피처로 사용

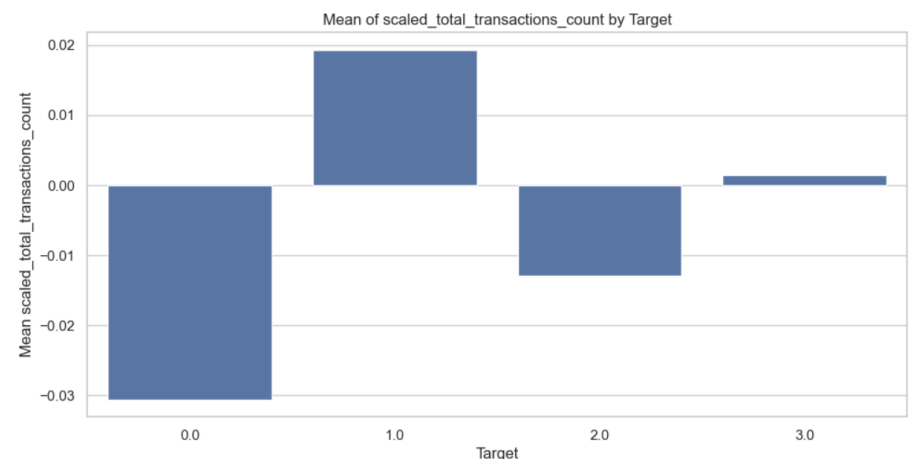
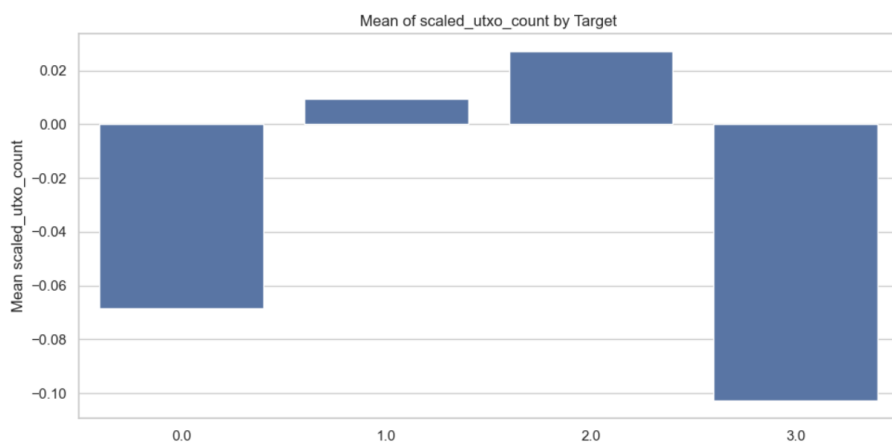


- 네트워크 보안성과 채굴 난이도에 영향을 줘서, 해시레이트 증가가 채굴 비용의 상승을 의미해 비트코인 가격 상승과 상관관계가 있을 수 있다고 추론

## 2-2. utxo\_count & transactions\_count\_total

지정된 시점의 비트코인 네트워크에 총 미사용 트랜잭션을 의미하는 utxo\_count 와 토큰 전송 여부와 무관한 총 트랜잭션의 수를 의미하는 지표인 transactions\_count\_total

- utxo\_count 가 늘어나면서 이동이 적다면 투자자들이 홀딩을 하고 있으니 공급 감소로 이어져 가격 상승을 만들 수 있다고 추론.
- transactions\_count\_total는 비트코인 참여자들의 활동 수준을 반영하기 때문에 단기적인 변동성이나 시장 심리에 영향을 끼칠 수 있다고 추론.

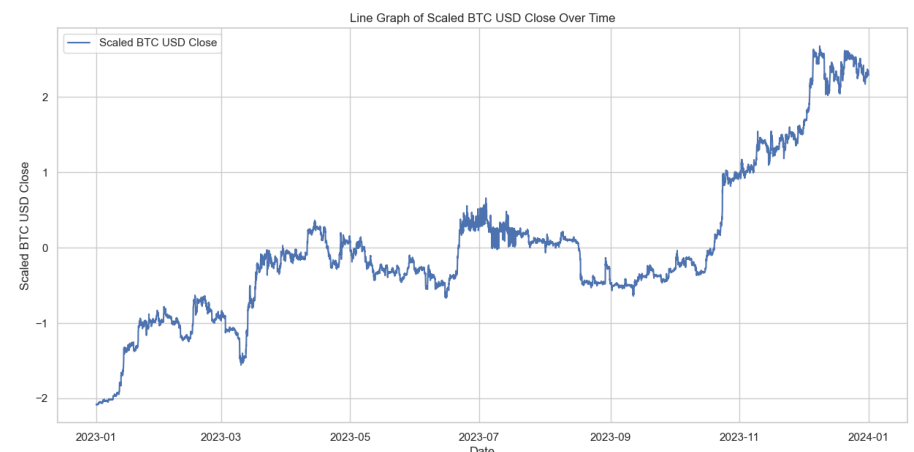
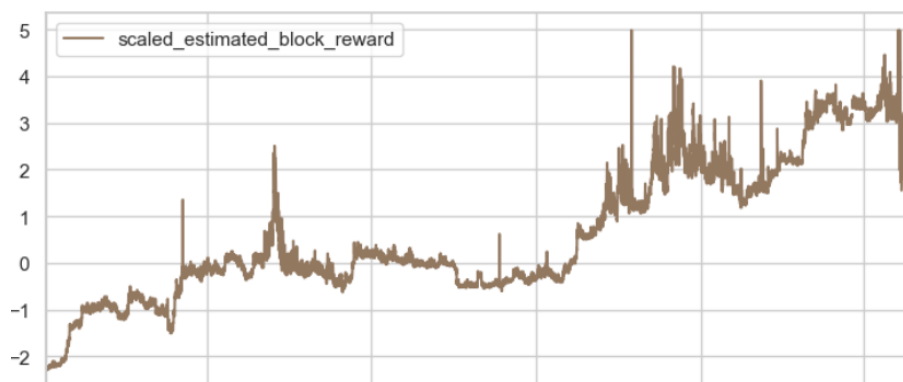


- 두 값 모두 StandardScaler()를 사용해 표준화를 시킨 채로 target에 따른 평균을 확인하니, utxo는 가격 변동성이 커질때 평균이 높아지고, transactions count는 가격 하락시 절대값이 높아지는 패턴을 확인해 피처로 사용.

## 2-3. estimated\_block\_reward(fees\_block\_mean\_usd / fees\_reward\_percent)

블록당 평균 수수료(usd)에서 블록 채굴 보상 중 수수료 비율을 나눠서 블록 보상 가치 USD로 추정

- 채굴자들이 받는 보상의 가치가 비트코인 가격에 연동되어있기 때문에 블록 보상 가치도 연관성이 있을 것이라 추론했다.

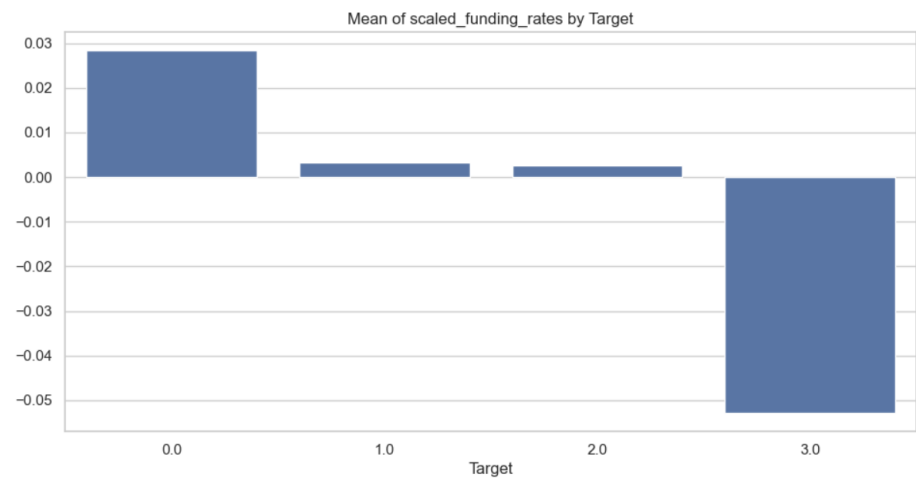
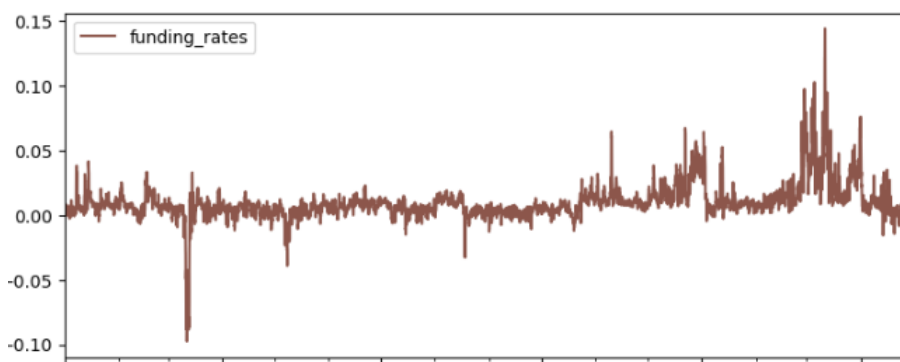


- scaled한 블록 보상 가치 추정 값과 증가를 비교했을때 시계열에 따라 비슷한 추세로 움직이는 것을 확인해 피처로 채택.

## 3. Market 데이터는 1번에서 걸러진 데이터들을 제외하고, 각 영역에서 대표성을 가진 데이터들을 피처 엔지니어링을 통해 최대한 활용한다.

### 3-1. funding\_rates

흔히들 편비라고 하는 이 지표는 시장 참가자들의 투자 심리를 반영해 양수일 때는 강세, 음수일때는 약세를 나타낸다. 단기적으로 비트코인 가격과 밀접하게 움직인다.

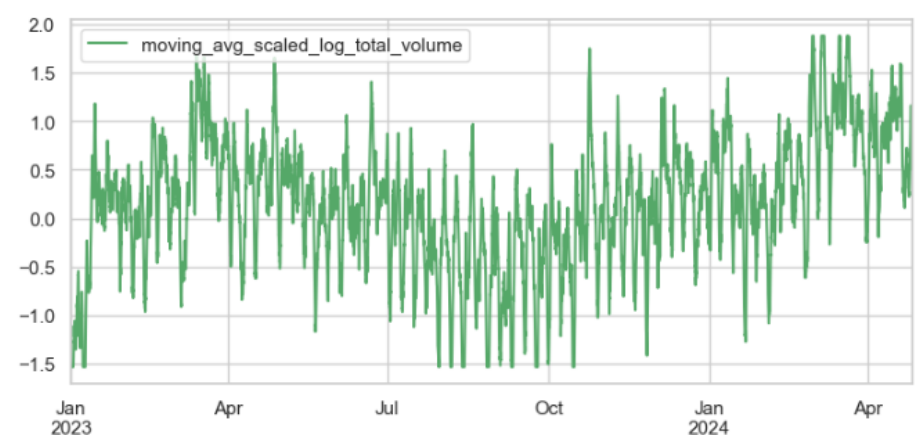
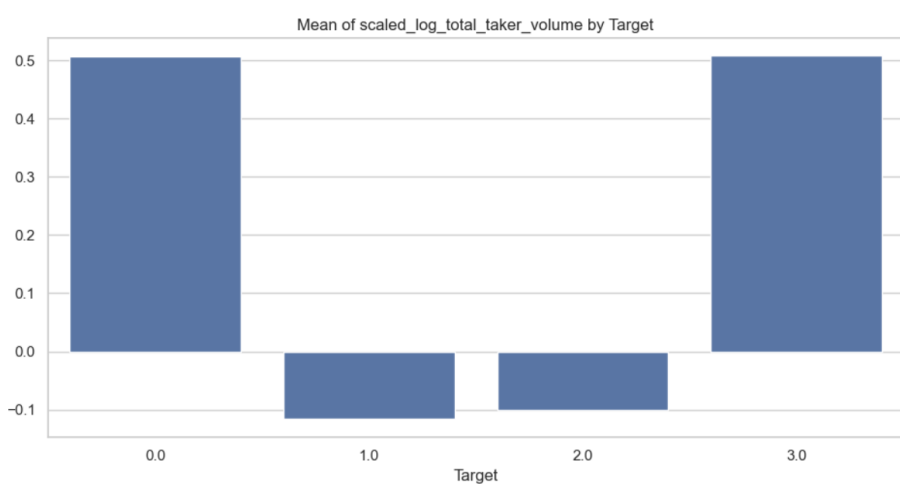


- 정규분포와 유사하지만 첨도가 크고 강세쪽 꼬리가 길어 로그변환과 표준화 처리를 해서 target별 평균을 살펴보니 가격 변동이 클 수록 큰 차이를 보여 피처로 채택했다.

### 3-2. taker\_buy\_sell\_stats

시장가 거래에 관련된 volume과 매수 매도 비율을 모은 데이터.

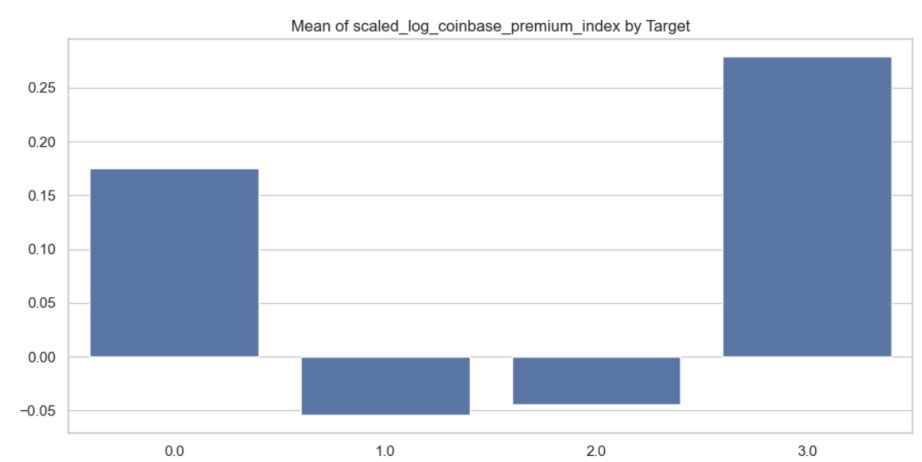
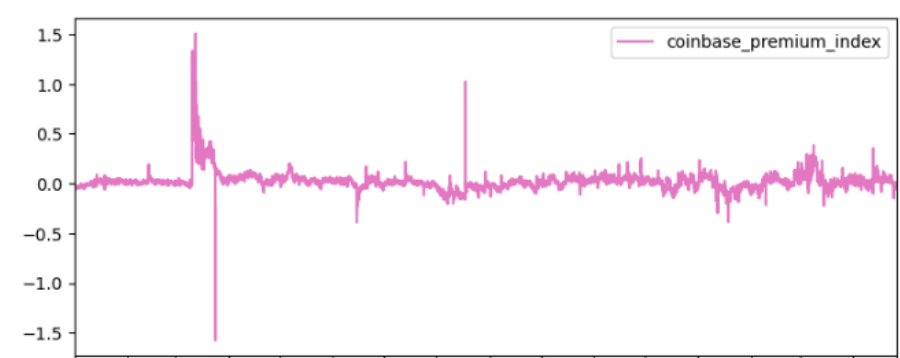
- 우선, 여러 컬럼 중에서 taker\_buy\_ratio와 taker\_sell\_ratio는 taker\_buy\_sell\_ratio 정보 안에 포함되어 있기 때문에 제외했다.
- taker\_buy\_sell\_ratio는 정규분포에 가깝고 매수/매도 심리를 바로 알 수 있기 때문에 전처리를 따로 거치지 않고 바로 피처로 선택했다.
- 그리고 test 데이터 셋에는 전체 거래량(volume) 지표가 존재하지 않기 때문에 taker\_buy\_volume과 taker\_sell\_volume를 더해서 taker\_total\_volume 피처를 생성했다.



- 로그 변환과 표준화를 거쳐서 target 클래스별 평균을 알아보니 변동성이 커질 때(0 or 3) 값이 뚜렷하게 상승함을 보인다.
- 그리고 거래량은 시간별 지표도 중요하지만, 하루동안의 거래량도 중요하기 때문에 위의 지표를 24시간짜리 window를 만들어 MA(Moving Average) 피처를 생성했다.

### 3-3. coinbase\_premium\_index

코인베이스 가격과 바이낸스 가격의 백분율 차이 지표.

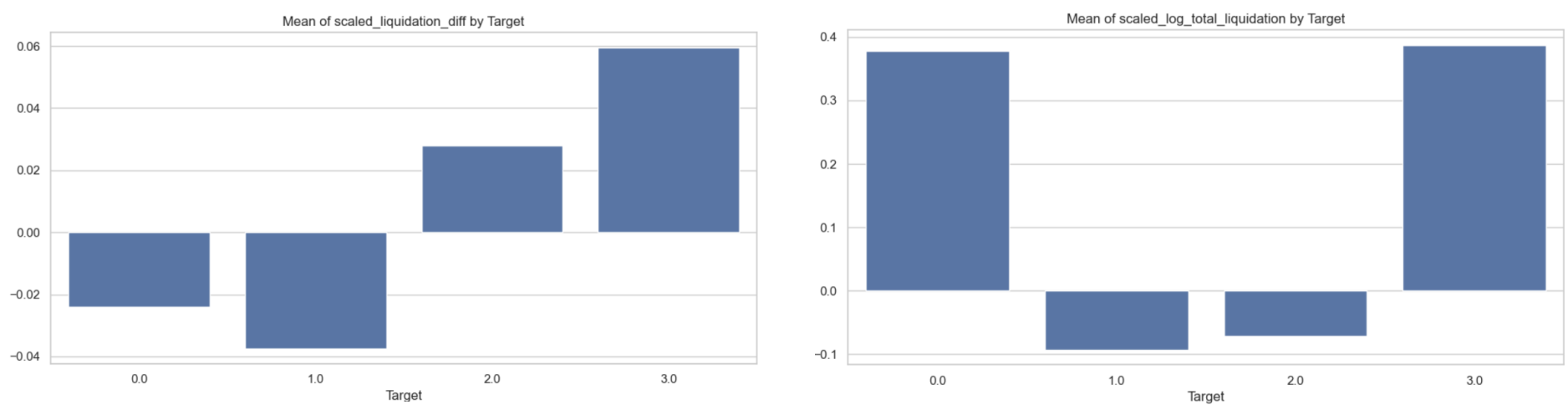


- 비트코인 가격 변동이 심할때 이 지표도 변동이 심해진다. 분포 자체는 정규분포의 모양이나 첨도가 너무 높아 로그 변환과 표준화를 거쳤다.
- 변환 후 target별 평균값을 보면 변동성과 연관이 있는 지표로 보여 피처로 선택했다.

### 3-4. liquidations

일반적으로는 대규모 롱 청산은 강재 매도로 인한 단기적인 가격 하락 압력을 만들고, 숏의 경우는 반대로 작용한다. 과매수/과매도 지표로 활용될 수 있으며 변동성과 연관된다.

- 데이터셋에서는 long/short liquidations 만 나오기 때문에 이를 그대로 활용하는 것보다 두 지표를 더한 total liquidation와 롱에서 숏을 뺀 liquidation diff 피처를 생성했다.
- liquidations diff는 표준화만, total liquidation은 로그변환과 표준화를 거쳤다.

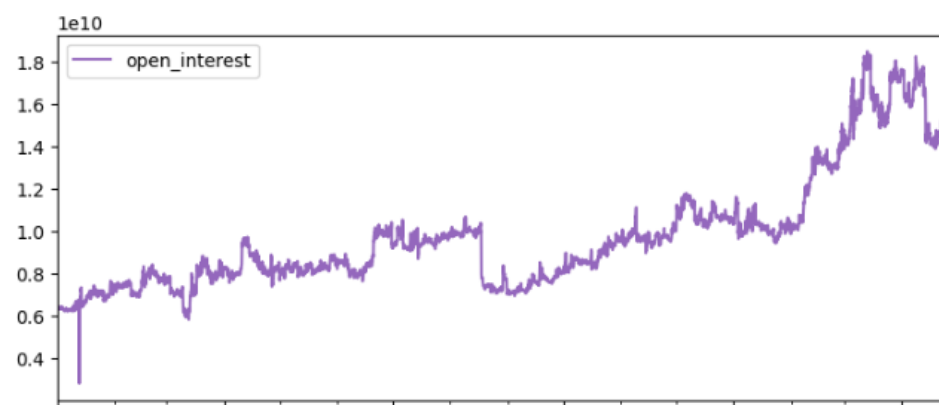


- liquidations diff는 가격이 상승할 때 평균값이 높았으며, total liquidation은 변동성이 클 때 평균값이 높아 두 지표 모두 피처로 채택했다.

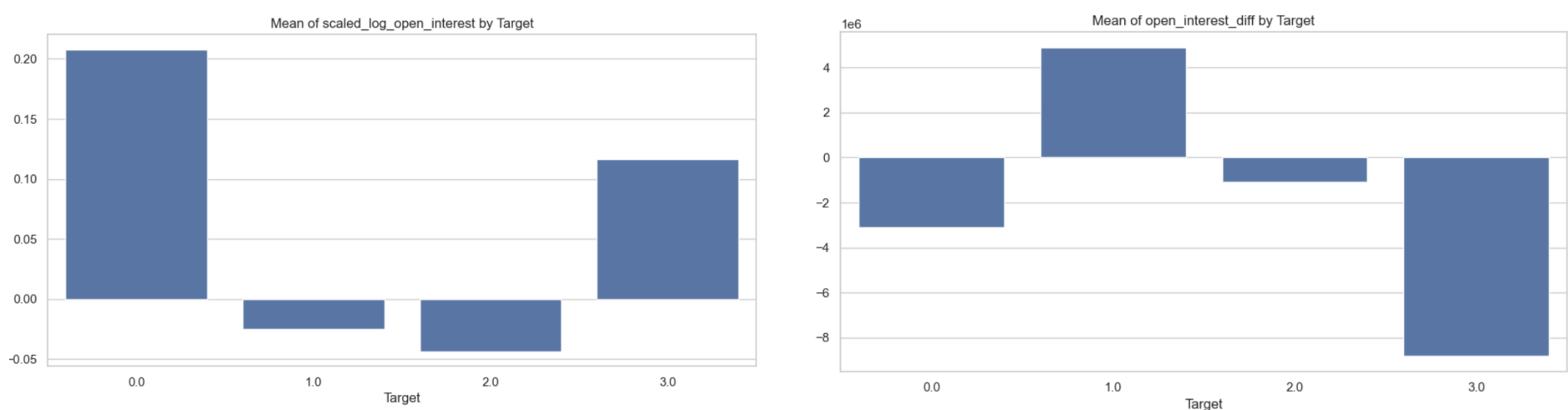
### 3-5. open\_interest

특정 시점에서 청산되지 않은 활성 계약의 총 수로, 시장에 참여하고 있는 자금의 규모를 보여주며 높은 시장 유동성을 나타내는 지표

- open interest가 늘어나면 기존 추세가 강화되고, 줄어들면 추세 반전의 신호로 보기도 한다.



- ID 기준 추세는 종가의 추세와 비슷하게 움직이는 것으로 보인다.



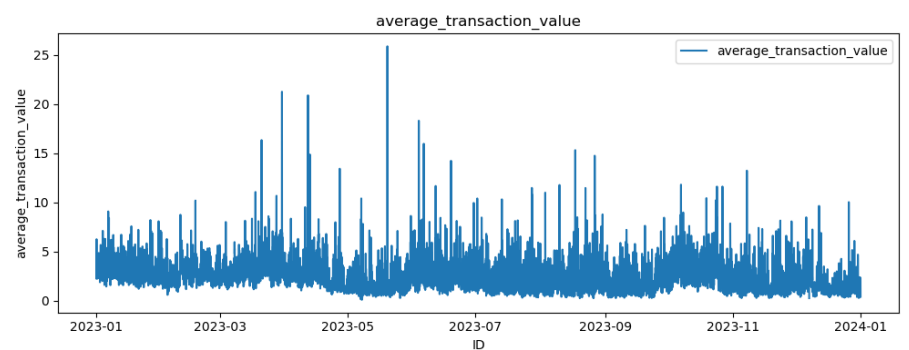
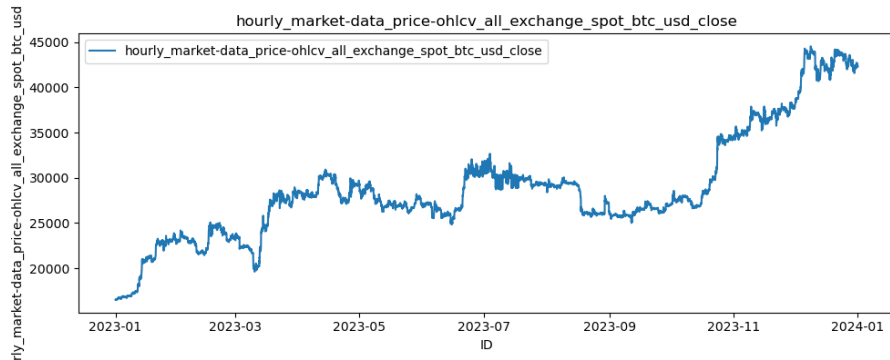
- 로그변환과 표준화를 거친 후 target별 평균값을 보면 변동성과 연관되어 있어 보여 피처로 채택했다.
- 변화량이 중요한 지표기 때문에, 직전값과의 차이를 open\_interset\_diff로 명명해 피처로 만들어 활용했다.

## 4 . 추가적으로 여러 피처들을 조합해서 새로운 피처 생성(최종 제출 기준 XGBoost 모델에서만 추가로 활용)

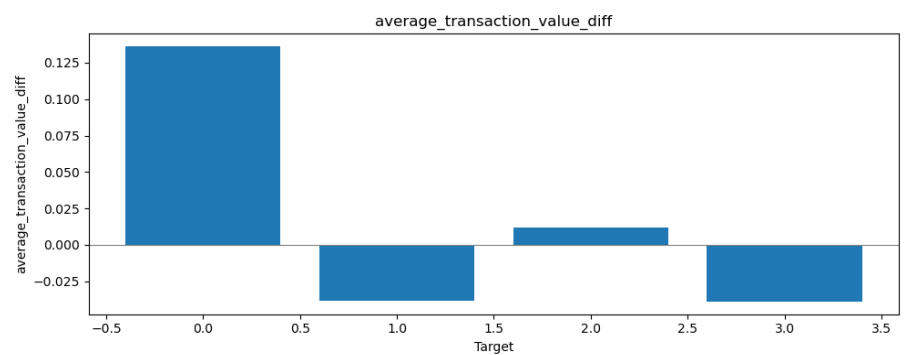
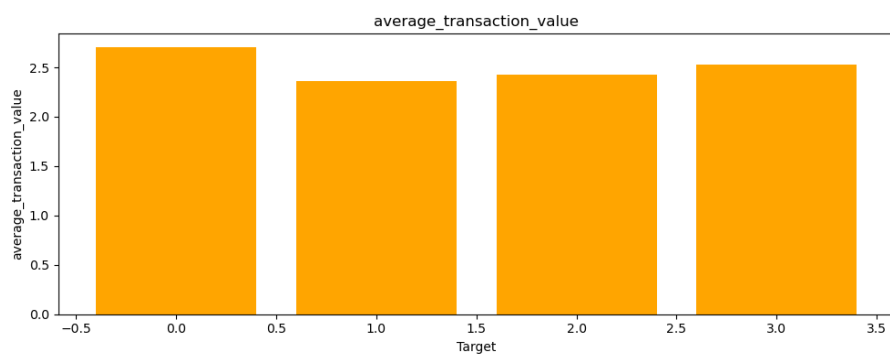
### 4-1. average\_transaction\_value(tokens-transferred\_total / transactions\_count\_total)

평균 트랜잭션의 가치를 나타내는 지표로, 발생한 거래당 비트코인의 수를 나타낸다.

- 수치가 클수록 고액의 거래가 이뤄졌거나 시장의 활동성을 암시할 수 있다.



- 종가의 상승과 하락구간에 같이 움직이는 부분을 확인할 수 있다.

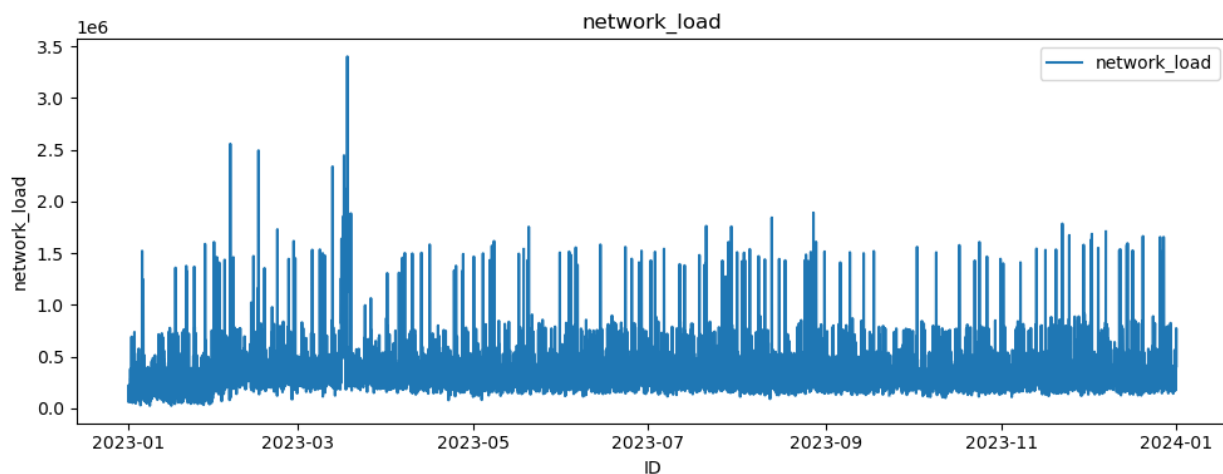


- 분포도에서의 차이가 보이긴하지만 명확한 특징을 추려내기가 어려워 이전 인스턴스와의 차이를 나타내는 추가 피처를 생성해서 분포도를 다시 확인한다.
- 종가 하락률이 클 때 더 많은 고액의 거래가 이뤄지거나(덤핑 예상) 시장이 활발해짐을 확인할 수 있다.

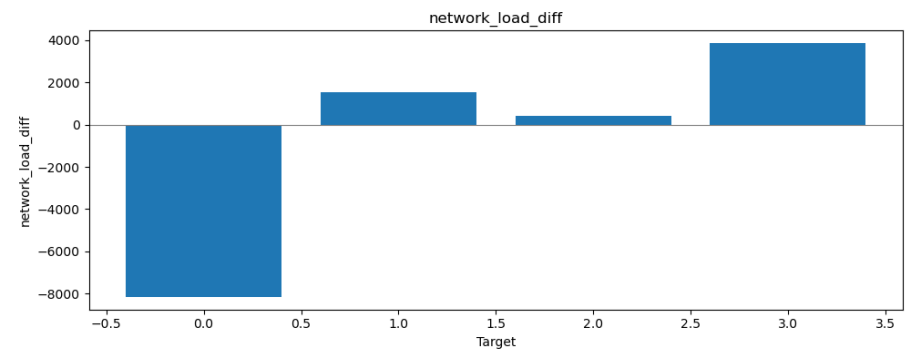
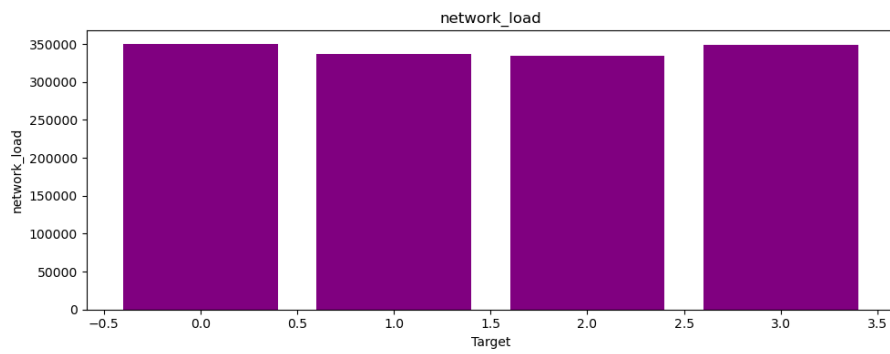
### 4-2. network\_load(block\_bytes / block\_count)

블록에는 비트코인의 거래 정보가 저장므로 새롭게 생성된 블록의 평균 크기를 통해 네트워크 부하를 나타내는 지표다.

- 비트코인 네트워크에서 블록 크기는 1MB로 제한됨에 따라 해당 지표가 높을수록 거래가 활발함을 나타낸다.



- 일정한 특정 구간에서 일반적으로 부하가 확인되는 것으로 보이며, 실질적으로 타겟 분류에 유의미한 지표로 사용할 수 있는지는 확인되지 않음.

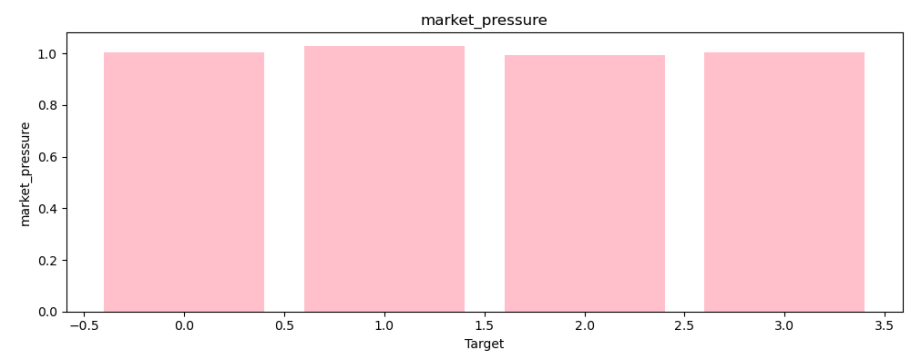
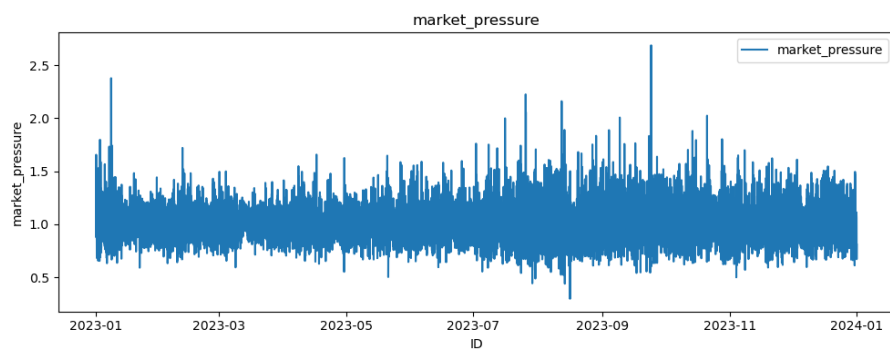


- 분포상으로도 크게 유의미한 차이를 보이고 있지 않으나 시간대별 차이가 극명할 수도 있으므로 이전 인스턴스와의 차이를 구해 분포도 확인.
- Network Load Diff 역시 시장이 활발할 때 부하의 변동이 커짐을 확인할 수 있다.

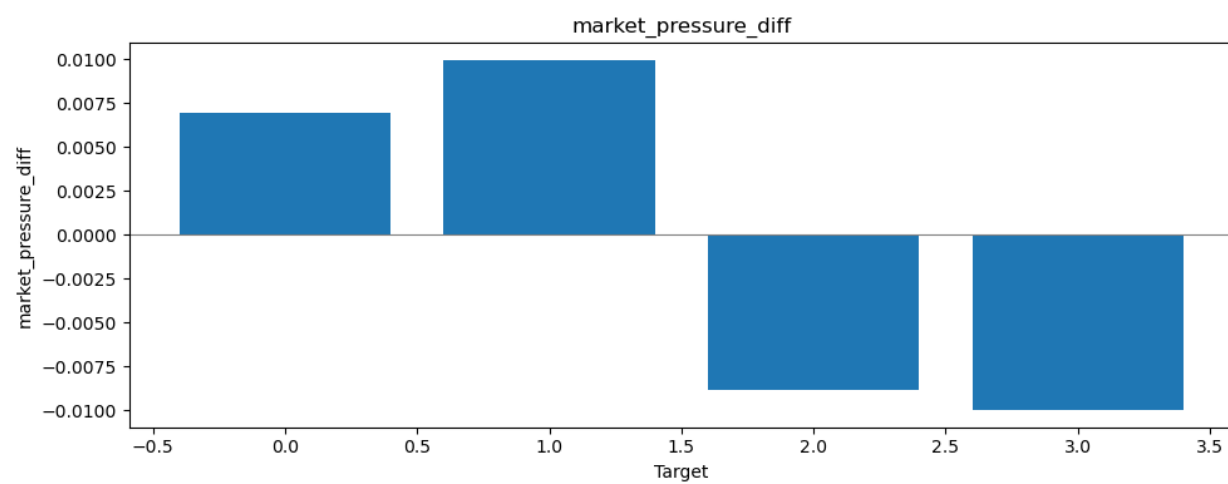
### 4-3. market\_pressure(taker\_buy\_ratio / taker\_sell\_ratio)

시장 압력 지표로, 매수 압력과 매도 압력 간 어떤 세력이 더 우세했는지를 의미한다.

- 주어진 데이터셋이 시간별 데이터임에 따라, 단위 시간에 시장의 단기적인 방향을 파악할 수 있다.



- 피쳐 자체가 타겟 분류에 유의미할 것으로 추정되지만 그래프에서는 확인되지 않음.

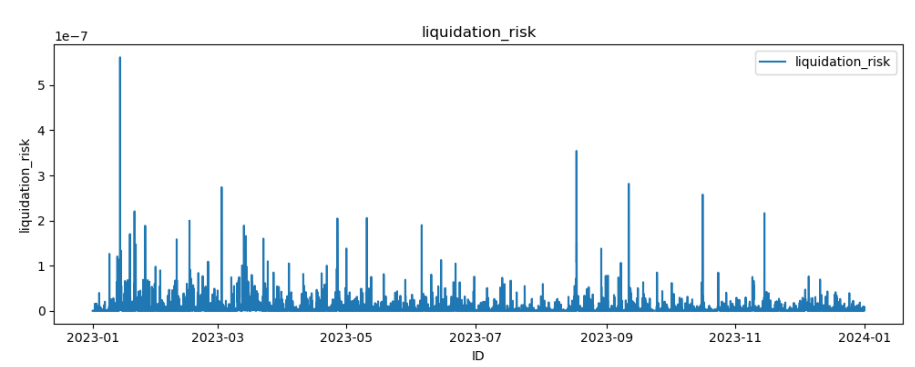
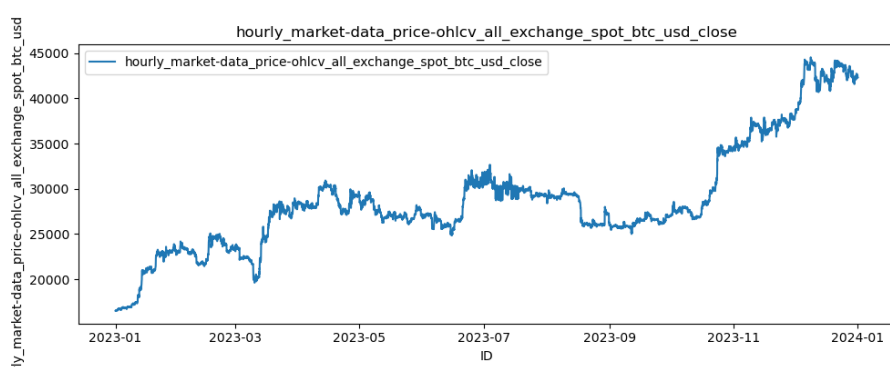


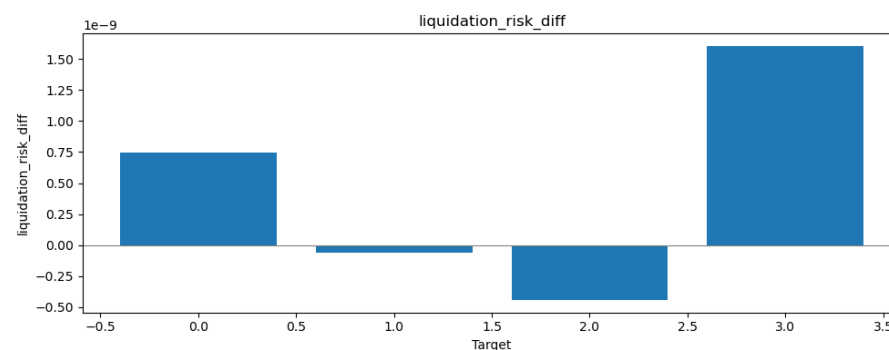
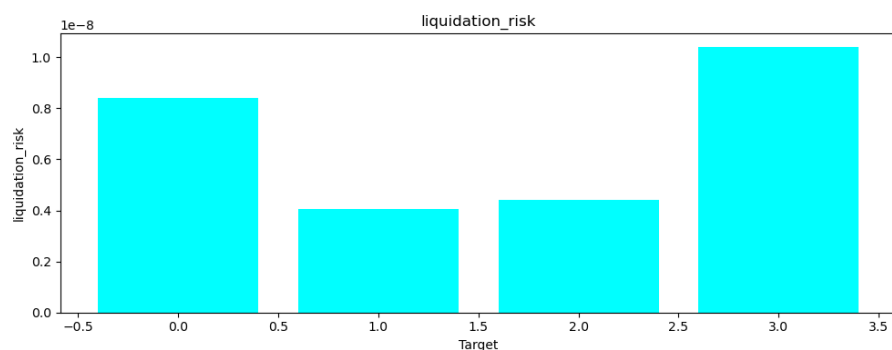
- 이전 인스턴스와의 차이를 계산하고 해당 수치에 따라 명확하게 상승/하강이 확인됨에 따라 모델이 추세를 파악하는데 도움이 될 것으로 추정.

### 4-4. liquidation\_risk((long\_liquidations + short\_liquidations) / open\_interest)

청산 위험 지표로, 청산되지 않고 현재 남아 있는 포지션에 대한 청산된 포지션의 상대적 중요성을 나타낸다.

- 수치가 커질 수록 시장의 변동성이 극도로 커져, 가격이 급격히 상승하거나 하락할 수 있다. 일반적으로 시장의 변동성을 예측하는데 도움이 된다.





- 청산 위험 지표는 시장이 활발할 때 평균적으로 높아지며, 추가로 주가가 하락할 때 높아지는 경향이 있다.
- 이전 인스턴스와의 차이에 따른 분포도를 살펴보면, 극명하게 시장이 활발할 때 청산위험-차이가 커짐을 확인할 수 있다.

## 시도한 모델 목록



1. LightGBM
2. XGBoost (Classifier / Regressor)
3. Random Forest
4. LSTM
5. Prophet

### 1. **LightGBM** - 최종 선택

- 모델 선택 이유 : 빠른 학습 속도와 메모리 효율성으로 대규모 데이터 셋에 적합하고, 시계열 데이터에 강점이 있어 선택.

### 2. **XGBoost Classification** - 최종 선택

- 모델 선택 이유 : 예측 정확도가 높고 결측치를 자동으로 처리를 할 수 있으며, 복잡한 비선형 관계를 잘 포착해 금융 시계열 데이터에 효과적이라 선택.

### 3. XGBoost Regressor

- 사용 이유 : 빠른 학습속도 기반으로 종가(Closed)를 예측 후, 종가 등락률 기반으로 분류 모델로 사용하기 위해 선정.  
해당 모델은 시계열 데이터에도 많이 사용되는 모델로써, 범용성 또한 뛰어나다고 판단
- 사용 결과 : 해당 모델의 예측값이 HyperParameter Tuning에 영향을 많이 받으며, 종가 예측의 전체적인 흐름이 수평이 되어, 부적합 판정.

### 4. LSTM(회귀)

- 사용 이유 : 회귀모델로 종가를 예측후, 제출시에 해당하는 종가 등락률을 클래스로 변경 목적.  
시계열 데이터의 장기 패턴을 효과적으로 포착하여 미래 가격 변동 예측에 유용하다고 판단.
- 사용 결과 : 훈련 시간이 길고, validation 결과 종가가 -log 함수처럼 값이 하락 수렴하는 추세 로 인해 부적합 판정.

### 5. Prophet(시계열)

- 사용 이유 : 비트코인 가격은 직전 가격 혹은 가격에 대한 MA가 가장 큰 영향을 줄 것으로 생각  
해당 흐름으로 진행 할때 가장 적절한 모델은 시계열 모델(prophet)  
기간별 가중치를 별도로 설정하여 사용하면 효율적 모델이 될 것이라고 판단함.
- 사용 결과 : 해당 모델은 종가 예측시, 상승적인 움직임을 보였으나 실제 데이터 흐름과 다르게  
폭발적인 흐름은 나타내지 못하였음.또한 데이터가 일정 시간 후에는 반복적 패턴의 Cos 함수와 같은 움직임을 보였으므로, 부적합 판정



## 최종 결과



- 1. LightGBM : Public Score - 0.3814(Accuracy) / Private Score - 0.4025 (Accuracy)
- 2. XGBoost Classifier : Public Score - 0.4020(Accuracy) / Private Score - 0.3960(Accuracy)

- LGBM HyperParams

```
{
  "boosting_type": "gbdt",
  "objective": "multiclass",
  "metric": "multi_logloss",
  "num_class": 4,
  "num_leaves": 50,
  "learning_rate": 0.05,
  "n_estimators": 30,
  "random_state": 42,
  "verbose": -1,
}
```

- XGBoost HyperParams

```
{
  "objective": "multi:softprob",
  "num_class": 4,
  "eval_metric": "mlogloss",
  "max_depth": 7,
  "learning_rate": 0.01,
  "subsample": 0.8,
  "colsample_bytree": 0.8,
  "seed": 42
}
```