Wrap-Up Report

팀명	RecSys 02조 추천 닭갈비 팀
구성원	강현구, 서동준, 이도걸, 이수미, 최윤혜

1. Team Wrap Up Report

1) 프로젝트 개요

프로젝트 주제

본 프로젝트는 **암호화폐 시장에서의 비트코인 가격 변동**을 예측하는 인공지능 모델을 개발하는 것을 목표로 진행되었다.

암호화폐의 가격 변동성은 일반 주식시장보다 훨씬 크기 때문에, 이를 예측하는 모델은 투자자들에게 중요한 도구가 될 수 있다.

프로젝트에서 사용한 데이터는 **온체인 데이터**와 **시장 데이터**로, 블록체인 네트워크에서 생성되는 활동 정보 및 거래소에서 수집된 가격, 거래량 등의 정보를 포함하고 있다.

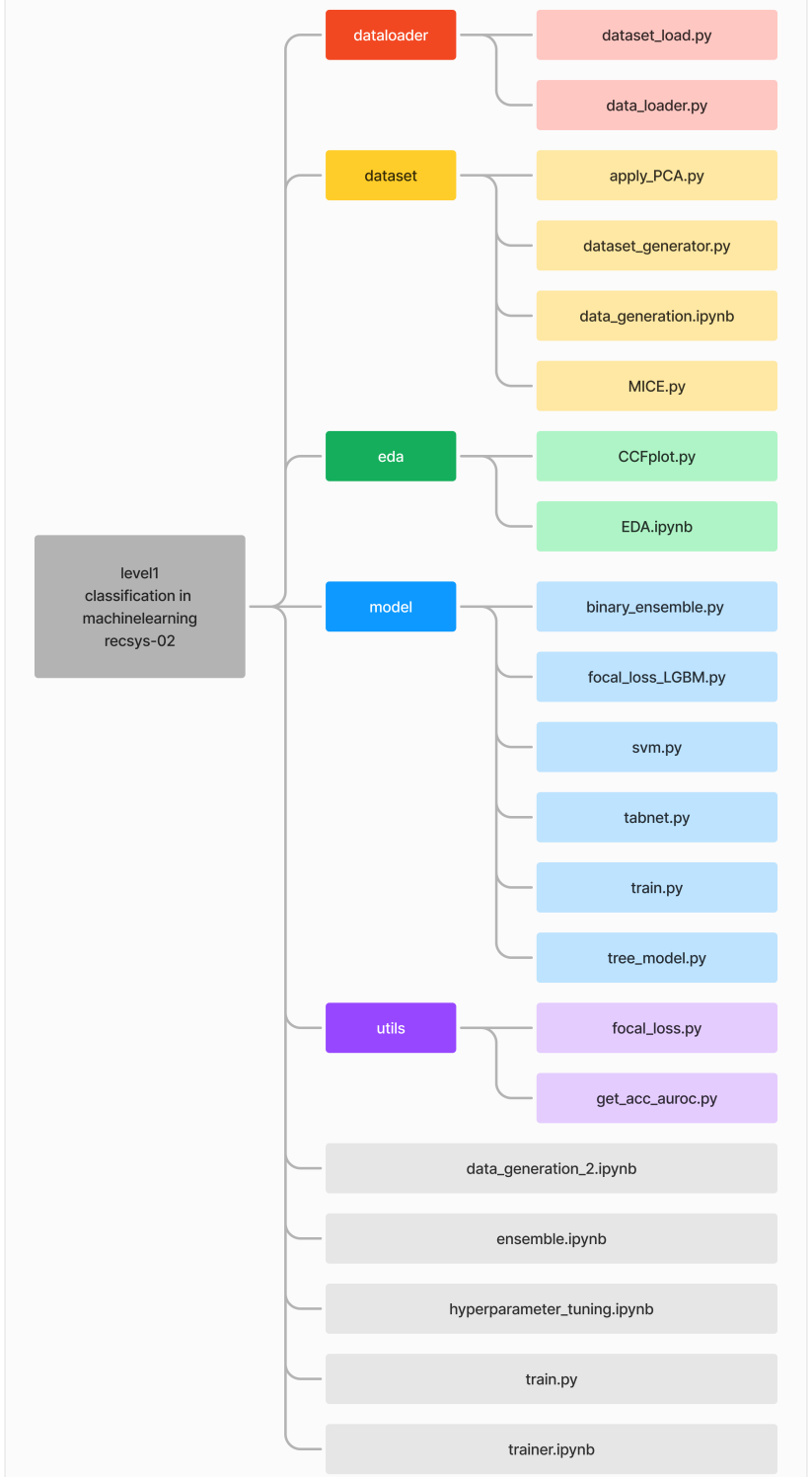
이 데이터를 바탕으로, 비트코인의 가격 등락을 **다중 분류 문제**로 접근하였으며, 성능 평가 지표로 **Accuracy**를 사용했다. 외부 데이터 사용과 미래 데이터 사용은 금지되었으나, **사전 학습된 가중치**를 활용한 모델 개발은 가능하도록 설정됐다.

활용 장비 및 재료

- 서버 스펙 : AI Stage GPU (Tesla V100)
- 협업 툴 : Github / Zoom / Slack / Notion / Google Drive
- 기술 스택 : Python / Scikit-Learn / Scikit-Optimize / Pandas / Numpy

프로젝트 구조도

File Structure

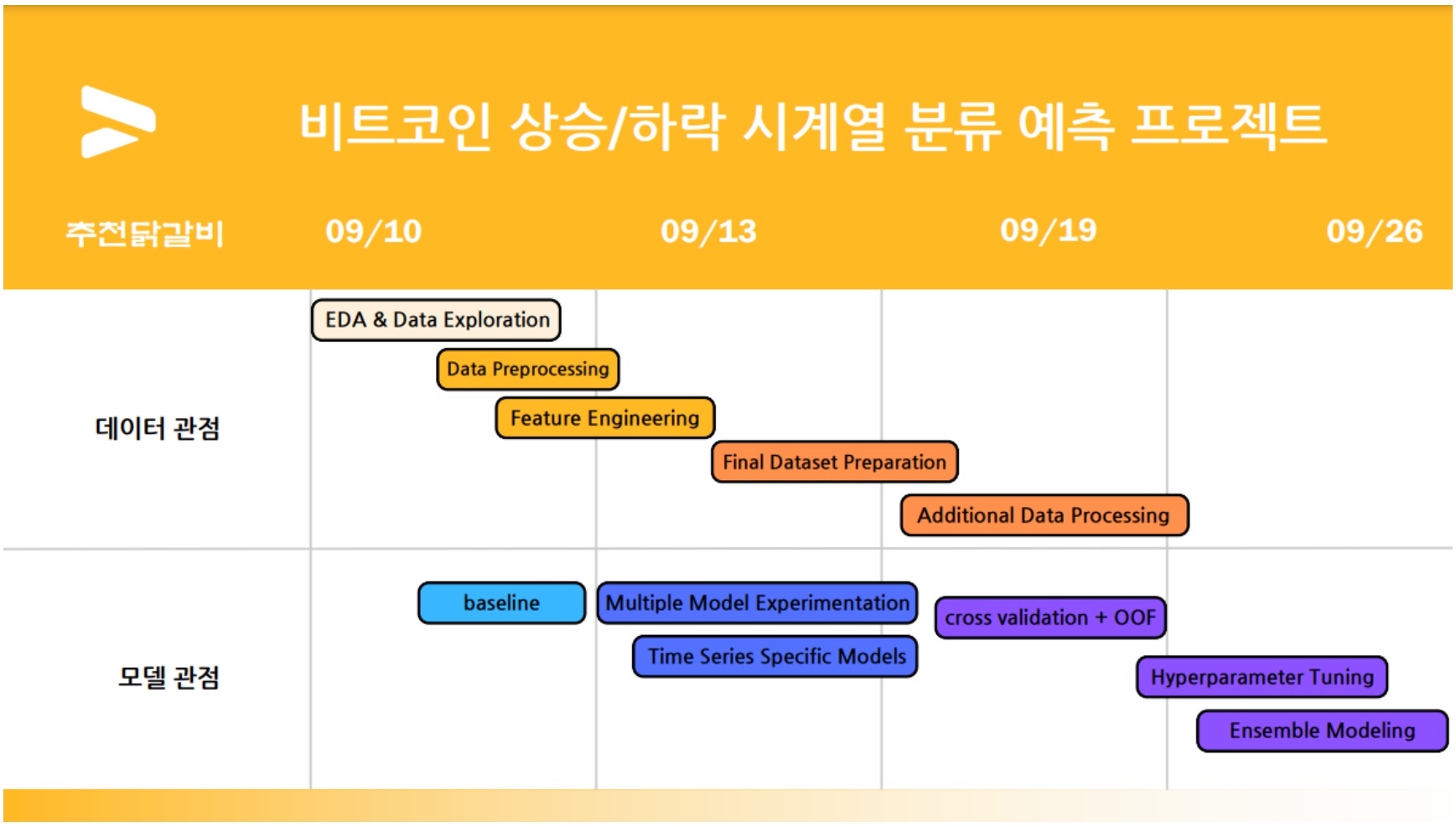


2) 프로젝트 팀 구성 및 역할

이름	역할
강현구	Modeling(XGB, LGBM, Catboost), hyperparameter tunning
서동준	Modeling(focal loss, cross-validation, tabnet), model modularization, Train pipeline refactoring
이도걸	Modeling(Random Forest, SVM), Modularization(train, hyperparameter tunning, ensemble), hyperparameter tunning
이수미	EDA, Feature Engineering, Data Preprocessing, Modeling(Soft Voting)
최윤혜	EDA, Feature Engineering, Data Augmentation, Modeling(direction predctioin)

3) 프로젝트 수행 절차 및 방법

수행 과정



협업 문화

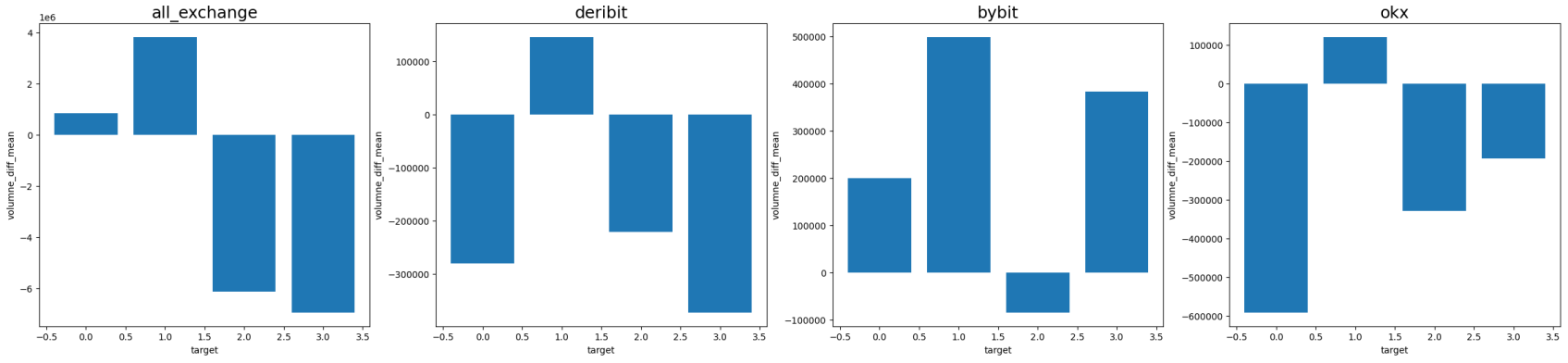
- **GitHub 활용**
  - 체계적이고 효율적인 워크플로우 유지
  - 구조화된 Pull Request(PR) 프로세스 구현
    - 코드 변경 사항에 대한 상세 설명 포함
    - 코드 리뷰 및 승인 절차 체계화
- **Notion을 통한 팀 관리**
  - 팀원 역할 및 업무 명확히 정의하여 중복 작업 방지
  - 실험 결과 및 주요 발견 사항 상세 기록
  - 회의록 작성 및 공유
- **Zoom을 활용한 커뮤니케이션**
  - 정기적인 아이디어 회의 진행
  - 팀원 간 원활한 소통 및 의견 교환 촉진

4) 프로젝트 수행 결과

탐색적 데이터 분석 (EDA)

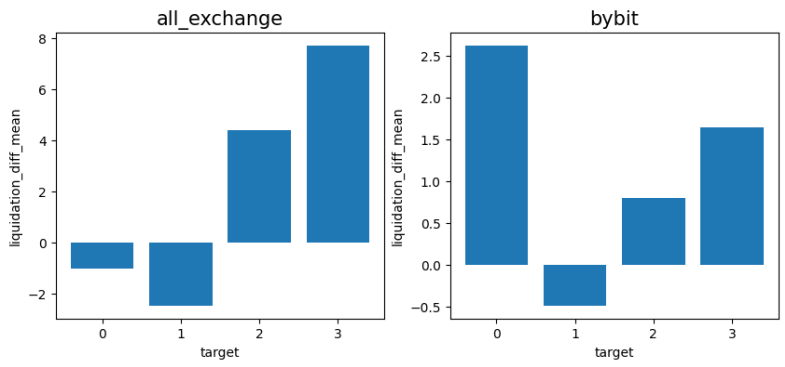
- **데이터 구성**
  - 총 **107개의 데이터**가 합쳐진 형태로, 가상 화폐 거래소(Market Data)와 블록체인 정보(Network Data)가 시간별로 제공된 시계열 데이터이다.
    - 11552 rows × 255 columns
    - 학습 데이터 기간 : 2023.01.01 ~ 2023.12.31
    - 평가 데이터 기간 : 2024.01.01 ~ 2024.04.26
    - 타겟 : 0(-0.5% 미만), 1(-0.5~0%), 2(0~0.5%), 3(0.5% 초과) 로 구성된 불균형 데이터 (0,3이 minor)
  - 가상 화폐 거래소(Market Data)
    - 거래소(`exchange`), 화폐 종류(`symbol`)별 데이터 존재
    - `all_exchange_all_symbol` : 모든 거래소와 화폐 종류에 대해 VWAP (거래량 가중 평균 가격) 방식으로 산출한 대푯값

- **가상 화폐 거래소 변수 선택**
  - `all_exchange_all_symbol`을 대푯값으로 사용
  - 그러나 일부 거래소와 화폐 종류는 대푯값과 다른 양상을 보일 수 있으므로 비트 코인 가격과 연관이 있어 보이는 개별 변수를 추가한다.
    - volume_diff `deribit`, `bybit`, `okx` 거래소 변수 추가



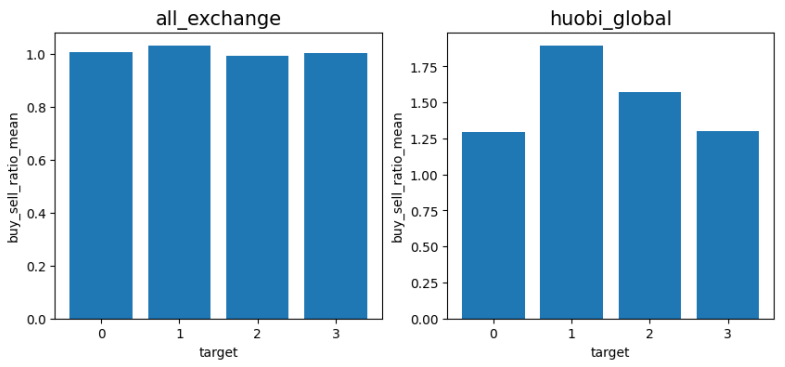
target별 volume_diff 변수 mean 막대그래프

- Liquidations_diff `bybit` 거래소 변수 추가



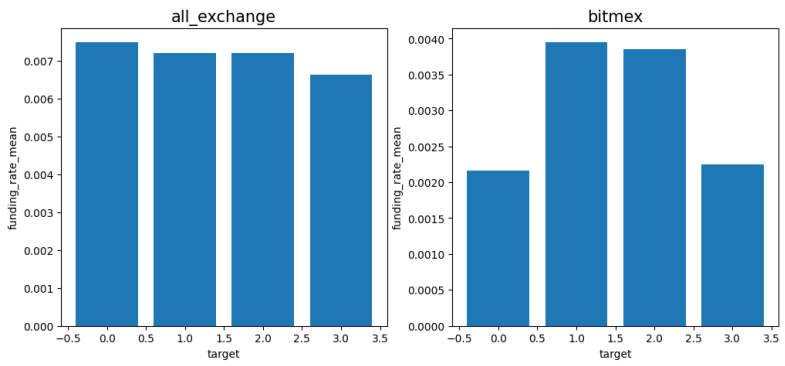
target 별 liquidation_diff 변수 mean 막대그래프

- Buy_sell_ratio `huobi_global` 거래소 변수 추가



target 별 buy_sell_ratio 변수 mean 막대그래프

- funding_rate `bitmex` 거래소 변수 추가



- **Feature Engineering**
  - 원본 변수에 추가적으로 파생 변수를 생성하였으며, 일부는 GPT를 통해 생성함으로써 도메인 지식 부족을 극복하고자 했다.

데이터 구분	의미	파생 변수
Liquidations	청산의 방향성	liquidation_diff = long_liquidations – short_liquidations
	청산의 방향성 비율	liquidation_index = (long_liquidations – short_liquidations) / (long_liquidations + short_liquidations)
Taker-buy-sell-stats	매매의 방향성	volume_diff = buy_volme – sell_volume
	매매의 방향성 비율	volume_index = (buy_volme – sell_volume)/(buy_volme + sell_volume
Address-count	블록체인 발/수신 방향성	addresses_diff = addresses_count_sender - addresses_count_receiver

데이터 구분	의미	파생 변수
상호작용항	시장 압력 지수	market_pressure= (taker_buy_ratio * open_interest) / (taker_sell_ratio * funding_rates)
	네트워크 활성화 효율성	network_active = (addresses_count_active * transactions_count_total) / (block_interval * block_bytes)
	HODLer 신뢰 지수	Hodler = (utxo_count * difficulty) / (velocity_supply_total * supply_new)
	채굴자 수익성 지수	profitability = (blockreward * hashrate) / (difficulty * fees_total)
	기관 투자 유입 지수	investment = coinbase_premium_index * (tokens_transferred_mean / tokens_transferred_median)
	레버리지 위험 지수	leverage = (long_liquidations + short_liquidations) * open_interest / (addresses_count_active * tokens_transferred_total)
	네트워크 성장 대비 수수료 부담 지수	fee_index = (fees_transaction_mean * transactions_count_total) / (addresses_count_active * supply_new)
	시장 건전성 지수	market_health = (hashrate * addresses_count_active) / (long_liquidations + short_liquidations)
	거래소 중앙화 위험 지수	exchange_center = (tokens_transferred_total - tokens_transferred_mean * transactions_count_total) / utxo_count
	청산 비율	(long/short)_liquidation_interest_ratio = (long/short)_liquidation / open-interest
	청산 대비 거래량 비율	(long/short)_liquidation_volume_ratio = (long/short)_liquidation / (taker-buy + taker-sell)
	거래량 대비 오픈 인터레스트 비율	volume_interest_ratio = (taker-buy + taker-sell) / open-interest

- 변수 선택
  - correlation plot 을 통해 상관계수 0.9 이상인 변수 제거
- **Shift**(1~24) 및 **Moving Average**(6, 12, 24, 48, 72) 변수 생성 → 단기 및 중장기 정보를 동시에 제공
  - Shift 변수는 각 변수마다 24개의 새로운 변수를 생성하기 때문에, 비트 코인 종가와 상관 관계가 높은 변수에 대해서만 생성한다.
  - 이를 위해 **CCF(Cross-Correlation Function)** 플롯을 사용하여 비트 코인 종가(`close`)와 각 변수 간의 상관성을 분석 통해 주요 변수 선정
    - 다음의 변수들에 대해 shift 변수 생성

데이터 전처리

- 결측치 처리

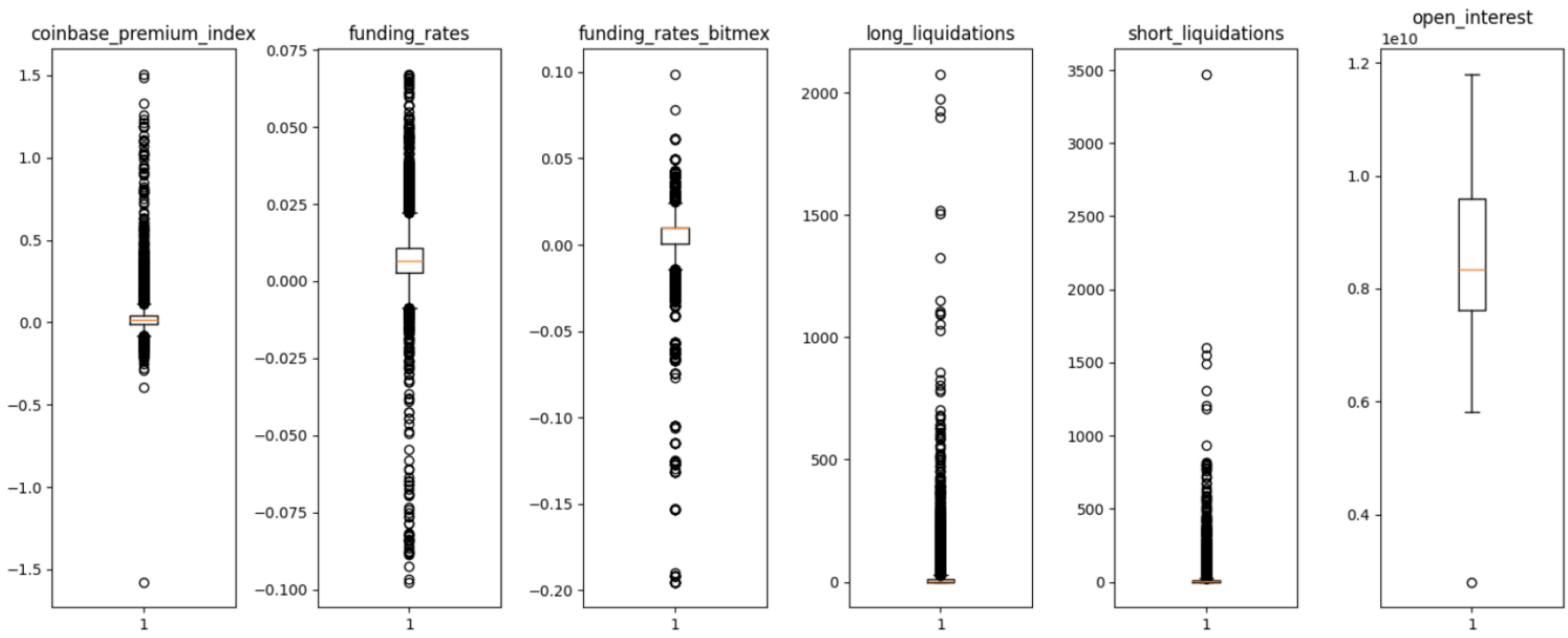
```
fees_transaction_mean_usd          19
fees_transaction_median_usd         0
blockreward_usd                     0
block_interval                     19
tokens_transferred_total            0
tokens_transferred_mean            19
tokens_transferred_median           0
...
funding_taker_ratio                 0
M                                   0
Is_Afternoon_Evening                0
Is_Weekend                          0
dtype: int64
```

EDA를 통해 선택한 변수에서 결측치의 비율은 매우 낮았다.

다른 변수들간의 관계를 고려하여 정확한 추정을 하고 데이터 손실을 최소화하는 **MICE**(Multiple Imputation by Chained Equations) 방법으로 결측치를 대체하였다.

- 이상치 처리, 로그 변환

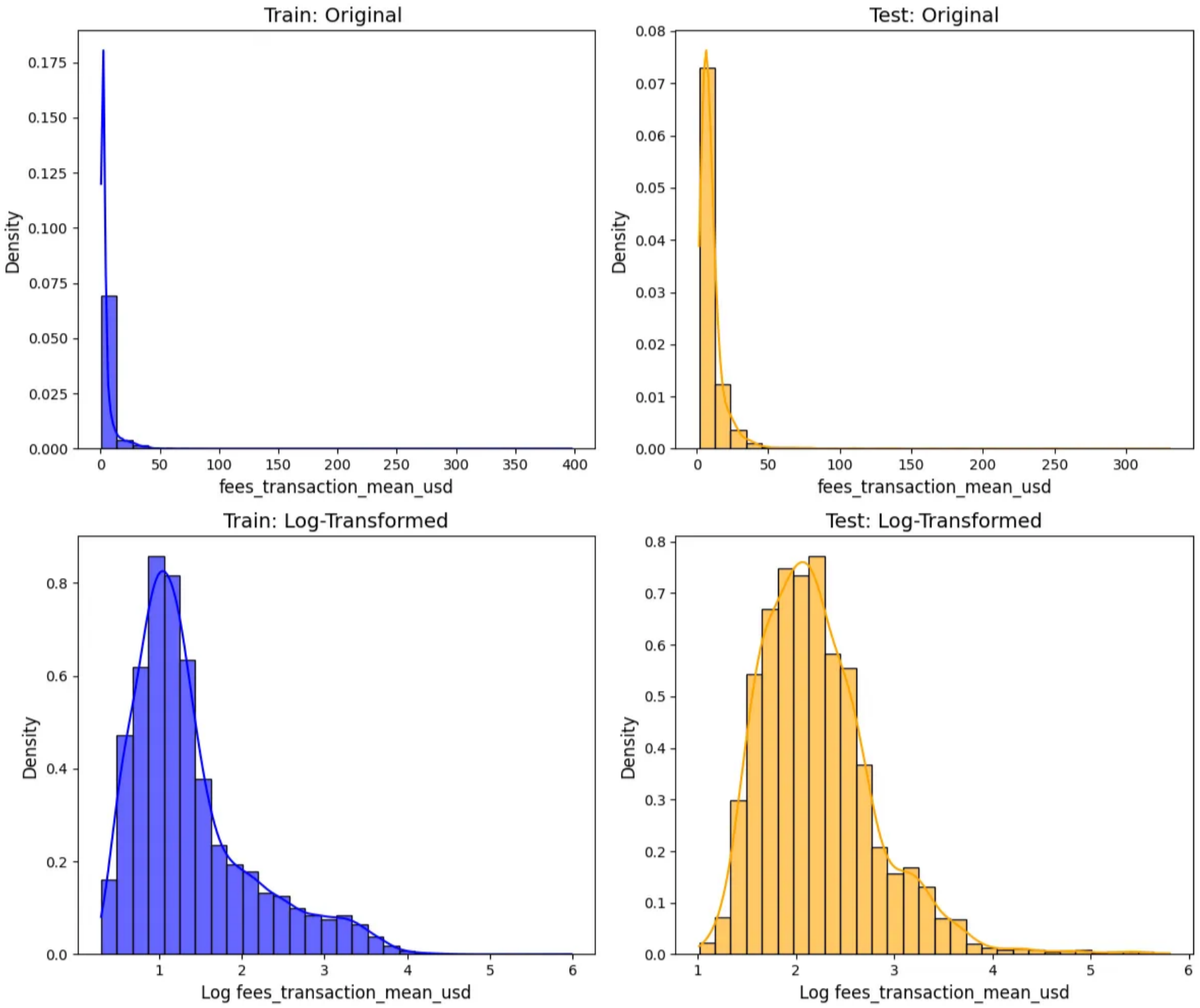
EDA를 통해, 비트코인 데이터는 가격 급등락으로 인해 다수의 **이상치**를 포함하고 있다는 사실을 확인하였다 이러한 이상치는 데이터의 중요한 특징을 담고 있을 가능성이 있기 때문에, 삭제하거나 변환하기 어렵다.



train data의 box-plot 중 일부

이상치를 처리하고자 **seasonal_decompose**를 활용한 잔차 기반 이상치 판별 방법과 **IsolationForest** 알고리즘을 적용해보았지만 처리 기준과 방법을 명확히 정의하는 데에는 한계가 있었다.

따라서 이상치를 변환하거나 삭제하는 대신, **로그 변환**을 통해 **이상치의 영향을 감소**시키고, **데이터 범위를 균등화** 하기로 하였다. 최소값이 0을 초과하며 훈련 데이터와 테스트 데이터 간 분포의 일관성을 높일 필요가 있는 변수들을 선정해 로그 변환을 적용하였다.



로그 변환을 진행한 변수의 예시 → (1) 이상치 영향 감소 (2) 분포의 정규화

- 데이터 정규화 - MinMax Scaling

  처음에는 트리 기반 모델을 사용할 계획이었기 때문에 일반적으로 스케일링이 필요 없다고 판단하였으나 다음 세 가지 이유로 **MinMax Scaling**을 적용하였다.

  (1) **결측치 대체**: 결측치를 -999로 대체하기 위해서는 각 변수의 범위가 999를 초과하지 않아야 한다.

  (2) **데이터 일관성 유지**: 훈련 데이터와 테스트 데이터 간에 변수의 분포 범위가 다를 경우 모델의 성능이 저하될 수 있다.

  (3) **데이터 분포 유지**: Standardization를 사용하면 훈련 데이터와 테스트 데이터의 범위 차이로 인해 값의 분포가 크게 달라질 수 있다. 따라서 MinMax Scaling을 사용해 일관된 스케일링을 제공하고자 하였다.

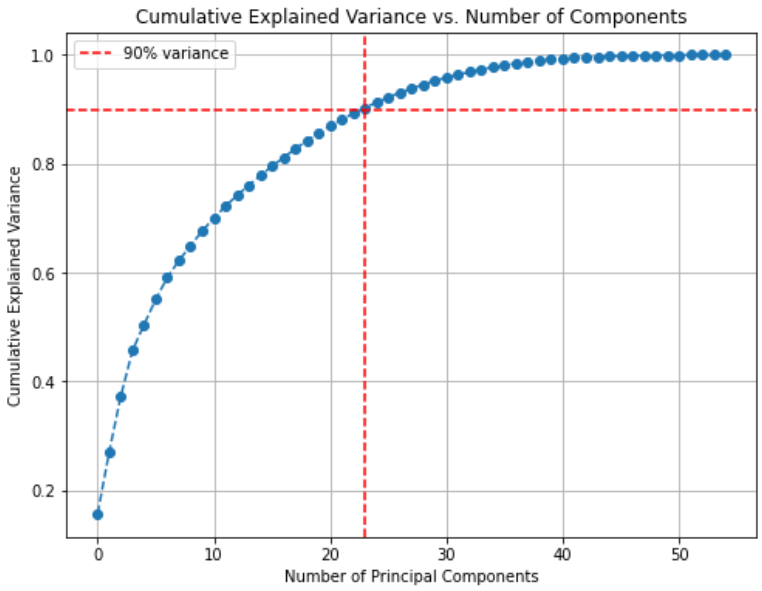
  - PCA를 진행할 때는 MinMax Scaling 대신 Standard Scaling 적용

- PCA

  EDA를 통해 선택된 55개의 변수에 대해 롤링과 이동 평균을 적용하면 **600개 이상**의 변수가 생성되어 **차원이 매우 커지는 문제**를 확인하였다. 따라서, 차원을 줄이기 위해 EDA에서 선정한 변수들에 대표적인 차원 축소 방법인 **PCA**를 적용하였습니다.

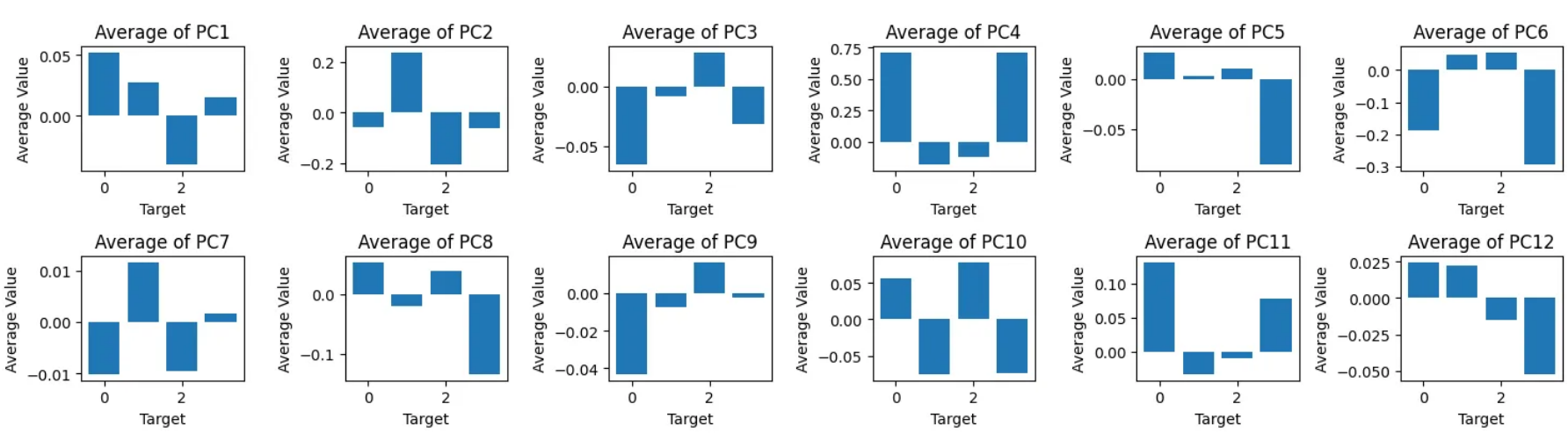
  적용 순서는 다음과 같다.

  1. **누적 설명 분산 비율 계산**: PCA를 적용하여 누적 설명 분산 비율을 계산하고, 90%의 설명력을 가지는 주성분의 개수를 지정하였다.



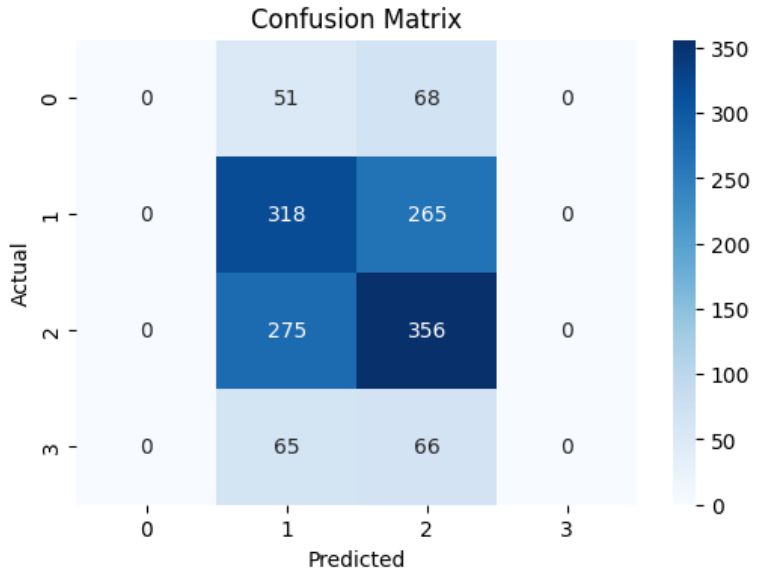
  2. **훈련 및 테스트 데이터 적용**: 훈련 데이터에 적용한 PCA 결과를 테스트 데이터에도 동일하게 적용하였다.
  3. **상관관계 분석**: CCF 플롯을 그려 비트코인의 종가와 상관관계가 높은 변수들에 대해서만 롤링과 이동 평균을 적용하였다.

이 과정을 통해 최종적으로 변수의 개수를 **약 200개**로 줄일 수 있었으며, 시각화 결과 target인 0,1,2,3을 잘 구분한다는 사실을 확인할 수 있었다.



PCA와 target을 시각화한 그래프 중 일부 첨부

- 상승/하락 확률 예측



지금까지 만든 데이터로 모델 학습을 시킨 결과 성능이 매우 낮았다. 0과 3은 거의 예측하지 못할 뿐더러 1과 2도 거의 구분하지 못하는 수준이라고 볼 수 있다. 따라서 더 잘 분류하기 위해 상승/하락 확률을 예측했다.

- 데이터 결합

  기존에 생성한 데이터프레임과 PCA에서 만들어진 데이터프레임을 동시에 활용하고자 하였으나 변수의 개수가 너무 많았다.

  따라서 두 데이터프레임을 효과적으로 결합하기 위해 **0과 3을 잘 예측하는 모델**(XGBoost)과 **비교적 높은 Accuracy를 보이는 모델**(CatBoost, LGBM)을 각 데이터 프레임에 적용하였다.

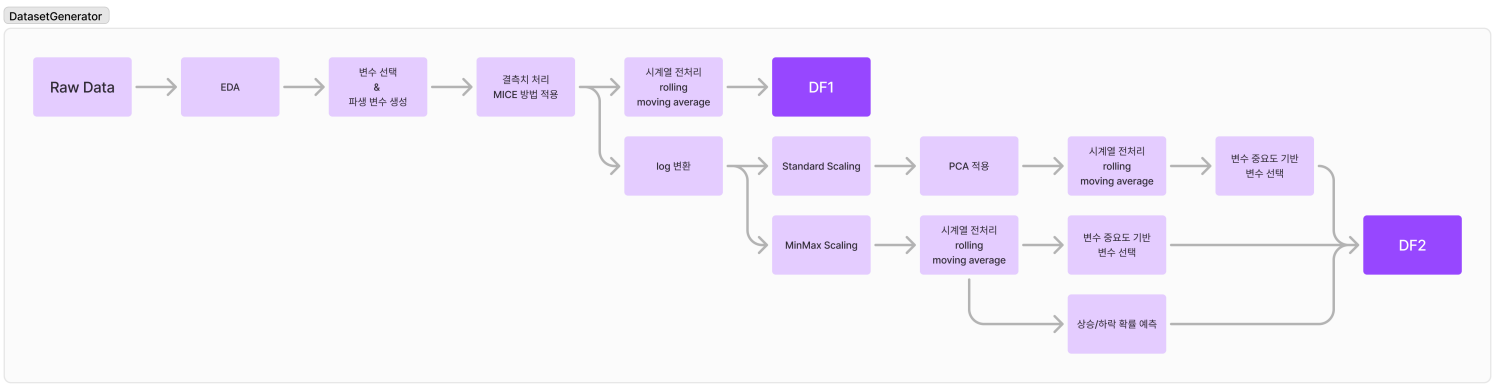
  각 모델에서 **변수 중요도 상위 20개**를 선정하여 변수명을 가져왔고, 이를 통해 두 데이터프레임에서 각각 60개 정도의 변수를 선택하였다.

  선택된 변수만을 가지고 있는 두 데이터프레임과 기존에 **예측된 상승/하락 확률**을 결합하여 데이터의 차원을 줄이면서도 중요한 정보를 유지할 수 있었다.

최종 데이터 셋

데이터 셋이 만들어진 과정을 시각화하면 다음과 같다.

이 두 가지 데이터 셋은 이후 과정에서 모두 사용하였다.



모델 개요

모델명	특징
Tree Model	LightGBM, XGBoost, CatBoost, Random Forest 모델
dir_prob_pred_model	비트코인 상승하락 확률 예측값을 변수로 추가하여 최종 예측한 모델
LightGBM focal loss	기본 LightGBM에 focal loss function을 적용한 모델
SVM_RandomForest_XGB	SVM, Random Forest, XGB Hard voting 앙상블 모델
LGBM_binary_ensemble	상승하락 binary LGBM 모델 기본 LGBM 모델 Soft Voting 앙상블 모델

모델 선정

- 모델 선정 기준

    - 최종 2개의 Train Dataset에서 stratified 5-fold Cross-validation으로 accuracy 계산
    - Bayesian Search로 Hyper Parameter 튜닝
    - Validation accuracy가 높은 모델 Hard voting, soft voting 적용
    - Public accuracy와 validation accuracy를 기준으로 각각에서 가장 높은 점수의 모델 2개를 최종 모델로 선정
- 성능 평가

모델명	valid 정확도 (data 1)	valid 정확도 (data 2)
LightGBM	0.4381	0.4544
XGBoost	**0.4493**	0.4484
Random Forest	0.4505	0.4557
LightGBM_focal	0.4381	0.4481
SVM	0.4133	0.3833
Tabnet	0.4025	0.4257
LGBM_binary_ensemble	0.4354	0.4482
CatBoost	0.4478	**0.4613**
CatBoost_XGBoost	0.4327	0.4584

- 최종 선정 모델: XGBoost(data 1), CatBoost(data 2)
    - 리더보드 결과: **Public Score 3위(0.4233)**, **Private Score 7위(0.4046)**

자체 평가 의견

- **잘했던 점**
    - 공통된 실험 환경을 구축하기 위해 우리 팀만의 baseline code를 사용하였으며 코드 모듈화를 진행함
    - 팀 내에서 각자 자신 있는 부분에 대해 역할 분배를 잘함
- **시도하였으나 잘 되지 않았던 점**
    - 0,3 예측을 위해 데이터 증강을 시도해보았으나, 성능 악화로 포기함
    - buy_volume, sell_volume을 직접 예측하여 사용하고자 하였으나, 예측 성능이 좋지 않아서 포기함
- **아쉬운 점**
    - 깃허브 활용을 늦게 시작함
    - cross validation set 고정을 늦게 하여 일반화 성능 관리에 실패함

2. 개인 회고

1) 강현구_T7502

초기 목표와 달성한 정도

이런 캐글형 대회를 한번도 참여해본적이 없다보니 어떻게 진행되는지 큰 그림을 한번 굴려보고 싶었고 그걸 미약하게나마 이해한것같아 다행이다.

내가 가장 신경 썼던 점

너무 세세한 부분에 집중하기보다는 최대한 모든걸 경험해보려 애썼다.

구현하지 못한 아이디어

데이터 유출에 대해 아예 무지할 때 이야긴데 그냥 바이낸스에서 증가,감소 가격 따와서 test값 적으면 안되나? 생각함. 당연히 안되는 이야기지만.

한계와 아쉬웠던 점

어떻게 대회와 협업이 이루어지는지는 이해했지만 그렇기에 내가 좀 기여한 부분이 적은 부분이 아쉬웠다.

다음에 시도해볼 것

Wandb를 다른 팀들은 많이 쓴거같길래 한번 공부해서 써보고싶다. 그리고 모듈화도 한번 다뤄보고 싶고.

마지막 한마디

화이팅

2) 서동준_T7527

초기 목표와 달성한 정도

- 첫 대회이자 앞으로 많은 프로젝트가 예정되어 있어, 팀원들과의 효율적인 협업 방법을 익히는 것을 목표로 삼았다. 노션을 통해 협업 과정을 기록하고 Github로 코드 관리하는 방법을 익혀 만족스럽다.
- raw 데이터를 다루는 대회였기 때문에, 실제 협업에서 문제를 접근하는 방법처럼 생각해보는 것이 목표였다. 하지만 비트코인 도메인에 대한 생소함과 목표 달성을 위한 예측 모델의 정확도가 기대에 미치지 못해 성과가 다소 만족스럽지 못했다.

내가 가장 신경 썼던 점

- 다양한 모델과 아이디어를 실험해보려고 했다. Regression모델, TabNet, AutoML, Stacking 앙상블 등, 다양한 모델링 방법을 사용해봤다.
- 재사용 가능한 코드를 작성하기 위해 노력했다. 후반부의 모델 앙상블과 튜닝과정까지 고려하여 큰 틀에서 생각하려고 노력했다.

구현하지 못한 아이디어

- Feature를 예측하는 모델을 만들 때, 예측한 Feature값을 input으로 하는 최종 모델을 사전학습하는 모델에 대해 고민해봤다. 시계열 데이터이기 때문에 Train데이터에서는 현재 시점 이후의 데이터를 Train에 활용할 수 있다. 이 값을 통해 모델을 사전학습하고, Test데이터를 예측할 때는, 미래 시점 값을 예측해서 사전학습한 모델 feature로 넣어주는 접근 방법이다. 하지만 Tree Model을 사전 학습하는 과정에서 어려움이 존재하여 시도하지 못했다.

한계와 아쉬웠던 점

- Tree 모델에서 feature importance의 순서가 target과의 correlation 순서라고 생각한 부분이 아쉬웠다. 하지만 대회가 끝난 후, Tree 모델의 feature importance는 트리가 분기할 때 정보 엔트로피의 총합을 가장 낮추는 feature에 불과하다는 사실을 알게 되었다.
- Public score에 과도하게 의존한 결과, private score에서 낮은 점수를 얻게 되어 아쉬움이 남는다. 이는 public score에 과적합된 것으로 판단되며, 앞으로 다양한 metric을 적용해야 할 필요성을 느꼈다.

다음에 시도해볼 것

- EDA 과정에 시간을 좀 더 분배해도 되겠다는 생각이 들었다. Feature Selection 과정과, 전처리 과정을 팀원에게 맡기게 된 느낌이 있었다.
- 하이퍼 파라미터 튜닝 과정에서 Wandb나 optuna를 활용해보고 싶다. 또한 모델 학습, 성능 기록 과정을 end-to-end로 자동화를 도전해보고 싶다.

마지막 한마디

프로젝트 초반, 미래 데이터를 shift하는 편법이 점수에 큰 영향을 미친다는 사실을 먼저 발견해 조교님들께 제보했다. 그 결과, [데이터 시점 일관성 규정]이 새롭게 도입되었다. 이를 통해 대회가 더욱 공정해진 점에 나름 뿌듯함을 느꼈다!

3) 이도걸_T7540

초기 목표와 달성한 정도

- 결과보다는 과정을 중요하게 생각하자.
  - 결과도 중요하지만 이번 대회에서 얻어가는 부분이 많으면 좋겠다고 생각했다.
  - 평소 사용해보지 않았던 모델들이나 알고리즘을 사용해봤다.

- 팀원들과 싸우지 않기
  - 협업 과정에서 의견 충돌이 있을 수 있음을 인지하고 그런 과정에서 이성적으로 판단하기

내가 가장 신경 썼던 점

- 코드 재활용을 위한 코드 함수화
  - 이번 프로젝트가 끝이 아닌 앞으로도 많은 프로젝트를 진행해야 함으로 이번 프로젝트에서 전체적인 코드 틀을 잡아두면 좋을 것 같았다.
  - 이후 모델들만 추가한다면 코드를 재활용할 수 있도록 개발을 진행했다.
- 리더 보드 신경 쓰지 않기.
  - 멘토분들과 조교분들이 강조했던 리더 보드에 집착하지 않기를 지키기 위해 노력했다.
  - 다른 팀이 높은 점수의 모델을 제출했을 때, 이기려고 생각하기보다는 어떤 방법을 사용했을지 궁금해 했던 것 같다.

구현하지 못한 아이디어

- 사람들이 많이 이용하면 등락폭이 클까...?
  - 데이터 EDA를 통해 거래량이 많은 시간대에 대해 알 수 있었다. 이 부분을 토대로 거래량이 많을 때 어떤 특성을 보이는 지 분석해서 각 시간에 대한 Feature embedding을 진행했다면 결과가 어땠을 지 궁금하다.
  - 거래량이 많은 시간에 0과 3의 클래스가 생길 확률이 더 높지 않을까...
- 월급날 가격이 오를까...?
  - 사람들의 심리가 새로운 돈이 생긴다면 투자에 대한 관심이 늘지 않을까라고 생각했던 것 같다.
  - 이러한 부분들을 적용하여 시계열 모델을 만들었다면 성능이 어땠을지 궁금하다.

한계와 아쉬웠던 점

- 데이터 EDA와 Feature embedding
  - 이 부분에 강점을 가진 팀원이 있어 본인은 EDA와 Feature embedding에 적극적으로 참여하지 않은 부분이 아쉬웠다.
  - 결과적으로 데이터에 대한 이해도가 떨어져 조금 더 창의적인 모델을 만들지 못한 것 같아 아쉬웠다.
- 학습 된 모델을 평가하는 방법
  - 최종 모델을 선택하기 위한 일관되고 객관적인 평가 기준을 선정하지 못했던 것 같다.
  - 모델이 Public score에 overfitting 될 수 있었다는 점을 눈치채지 못했던 부분이 아쉬웠다.
- 각 모델에 대한 이해도
  - 많은 모델을 사용하긴 했지만 오픈 코드에서 가져온 코드들을 사용했고 모델의 구조나 사용된 알고리즘에 대해 정확하게 알지는 못했다.
  - 그러다 보니 하이퍼 파라미터 튜닝을 제외하면 모델 자체적인 성능을 높일 수 있는 방법을 적용하지 못했다.

다음에 시도해볼 것

- Github의 다양한 기능들을 사용한 협업
- 모델 자체의 성능을 높이기 위한 방법

마지막 한마디

인공지능도 대략 40%밖에 못 맞추는 비트코인은 하지 말자.

4) 이수미_T7541

초기 목표와 달성한 정도

- 꼼꼼한 EDA를 통해 데이터에 알맞는 전처리 진행하는 것이 목표였는데, 다양한 관점에서 데이터를 뜯어본 것 같아 뿌듯하다.
- 협업을 잘 활용하고자 하였는데 잘 해낸 것 같다. 모듈화도 처음 해봤는데 생각보다 훨씬 편해서 놀랐다.

내가 가장 신경 썼던 점

- 데이터 셋 생성하기: 머신 러닝 기법은 비슷할 거라 생각해서 조금 더 다양한 데이터 셋을 만들고자 하였다.
- 이상치 처리를 어떻게 해야 하는 지에 대한 고민을 많이 하였다.

구현하지 못한 아이디어

- close 변수와 difficulty 변수는 계속 증가한다는 점을 파악하였으나 '미래의 예측 값은 계속 증가할 것이다' 를 트리 모델에 어떻게 반영해야 할 지.... 모르겠어서 생각만 해봤다.

한계와 아쉬웠던 점

- 내가 했던 내용을 팀원들에게 제대로 전달하지 못한 것 같아서 아쉽다. 다음 프로젝트를 할 때는 조금 더 정확한 전달을 위해 노력해야겠다.
- 뭘 해도 점수가 오르지 않아 조금 답답했었다.... 데이터 셋을 바꾸든, 모델을 바꾸든 큰 개선점이 보이지 않아 어떠한 처리가 더 적절한가에 대한 답변을 얻지 못해서 아쉬웠다.
- 새로 알게 된 내용을 활용하기 보다는 기존에 알던 내용을 복습하며 프로젝트를 진행한 것 같다.

다음에 시도해볼 것

- 체계적인 깃허브 활용: 이번 프로젝트에서는 막바지에만 깃허브에 정리하고자 하였지만, 다음 프로젝트에서는 깃허브를 조금 더 써보고자 노력하겠다.
- 새로운 방법론 적용

마지막 한마디

프로젝트는 재밌었지만 비트코인은 하지말자.....

5) 최윤혜_T7549

초기 목표와 달성한 정도

- 체계적이며 논리적인 분석 전개
  - 공모전에 참가하여 데이터 분석을 하다 보면 주어진 시간 내에 결과물을 완성해내야 하기 때문에 항상 데이터를 차근차근 살펴볼 시간을 충분히 갖지 못했다. 그래서 이번 프로젝트에는 결과 자체보다는 학습에 초점을 맞춰서, EDA부터 모델링까지 깊이 있게 고민하며 논리적으로 분석해보고자 했다.
  - 나름대로 목표한 바는 달성했다고 생각한다. 방대한 데이터의 양에 막막함을 느끼긴 했지만 포기하지 않고 데이터를 끝까지 살펴보았다. EDA, 파생변수 생성, 변수선택, 모델링 모든 과정에서 끊임없이 고민하고 시도해보았다.
- git을 활용한 프로젝트 협업
  - 부스트캠프를 시작하기 전까지 git을 거의 사용해본 적이 없다. git을 통한 협업은 더욱이 경험해볼 기회가 없었다. 그래서 이번 프로젝트를 통해 git 사용법을 익히고 더 나아가 협업툴로 사용해보고 싶었다.
  - 그러나 아쉽게도, git을 제대로 활용해보지 못한 것 같다. git 사용이 막막하다 보니 자연스럽게 서버 내에서 코드를 공유하고, 노션으로 실험 정보를 기록하게 되었다. 나중에 개발자처럼 github 사용하기 강의를 듣고 몇가지 이슈를 생성해보는 시도를 하긴 하였으나, 많이 부족하였다.

내가 가장 신경 썼던 점

- 데이터 분석
  - 약 9천개의 학습 데이터셋으로, 딥러닝 보다는 머신러닝이 적합해보이는 데이터셋이라고 생각했고, 머신러닝의 경우, 모델에 따른 차이보다는 데이터셋에 따른 차이가 더 클 것이라고 판단했다. 따라서 다양한 모델 실험보다는 다양한 데이터 분석 실험에 초점을 맞춰서 진행하였다. 비트코인 도메인 지식도 부족할 뿐만 아니라 107개의 데이터가 합쳐진 만큼 굉장히 복잡하고 어려운 데이터였지만 약 일주일 간의 시간을 데이터 분석에 쏟았다. 유의한 변수를 추출해내기 위해 모든 데이터를 시각화하기도 하고, 다양하게 파생 변수를 생성해보고, CCF plot을 그려보기도 하였다. 결과적으로 이를 기반으로 생성한 데이터셋을 통해 성능 향상을 이뤄낼 수 있었다.

구현하지 못한 아이디어

- 증강 데이터 가치 측정

    - 0,3의 데이터가 적었기 때문에 다양한 증강 기법을 시도하였으나, 오히려 성능이 떨어지는 양상을 보였다. 결국, 0,3을 제대로 예측하는 모델 구현에 실패하였다. 하지만 만약 시간이 충분했다면 데이터를 다양하게 증강해보고, 이에 대해 데이터 가치를 측정해보고 싶다. 부스트캠프 강의에서 배운 influence function이나 data-oob 방식과 같이 데이터의 가치를 평가한다면 증강을 통해 충분한 데이터를 확보하면서도 학습에 방해되는 데이터를 삭제할 수 있을 것이라 생각한다. 만약 기회가 된다면 한번 시도해보고 싶다.

한계와 아쉬웠던 점

- 과적합 문제 간과
    - 결과적으로 리더보드 상에서 private 점수가 생각보다 많이 낮게 나왔다. 프로젝트를 진행하면서 public점수에만 집중하고, validation accuracy를 제대로 관리하지 못한 것 같다. cross validation 방식도 나중에 적용하는 바람에 모든 실험에 대한 일관성을 잃게 되었다. 다음에는 public 점수에만 집중하기 보다는 초반에 validation set을 명확히 지정하고 실험 결과를 관리해야 할 것 같다.
- 단조로운 모델링
    - 기존에 이미 알고 있던 모델들을 위주로 시도했다. Prophet과 같은 새로운 시계열 모델을 시도해보고 싶었으나 시간에 쫓기다보니 자연스럽게 알고 있던 모델을 이용하게 되었다. 다음에는 논문도 좀 찾아보고 새로운 모델들도 다양하게 시도해보면서 학습할 수 있도록 해야겠다.

다음에 시도해볼 것

- github를 통한 실험 관리
    - 지금까지는 대부분의 실험을 노션을 통해 기록했으나, 다음부터는 github를 최대한 활용하여 체계적인 실험 관리를 해봐야겠다.

마지막 한마디

미래 변수 이슈가 가장 핫했던 프로젝트. 타임머신 있는 거 아니면 비트코인 예측할 생각하지 말기..