

 iM DiGital Banker Academy

비트코인 주가 예측을 위한 머신러닝 기술 활용

겨울은 윈터조


박훈석 정민관 이 현 신영섭

Date

2024년 12월 13일



TABLE OF CONTENTS

- 
- 01 프로젝트 개요
 - 02 프로젝트 팀 구성 및 역할
 - 03 프로젝트 수행 절차 및 방법
 - 04 결론
 - 05 소감 및 부록

CHAPTER 1

프로젝트 개요

01 주제 선정 배경

가상화폐 시장의 급격한 가격 변동과 투자자의 비합리적 의사 결정 문제를
해결하기 위한 객관적 분석의 필요성 대두

02 프로젝트 목표

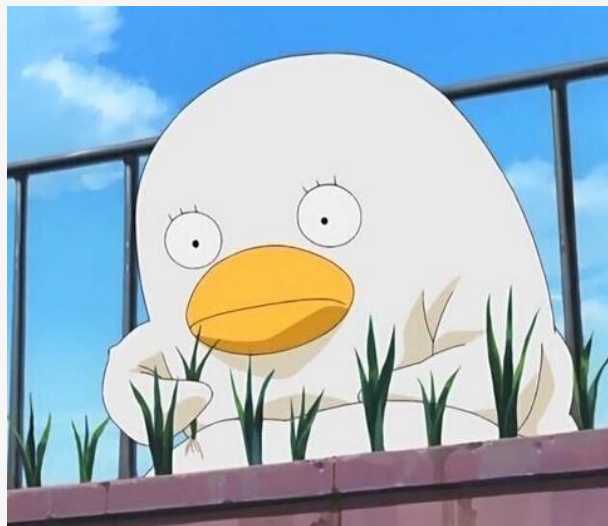
비트코인 가격 추세를 예측하는 이진분류 모델 학습

CHAPTER 1

프로젝트 팀 구성 및 역할

“겨울은 윈터조”

박훈석(팀장)



- 프로젝트 관리(PM)
- 도메인 분석
- 발표

정민관



- 데이터 베이스 설계 및 관리
- 업비트 API 연동 및 관리

이 현



- 데이터 베이스 설계 및 관리
- 머신러닝 모델링

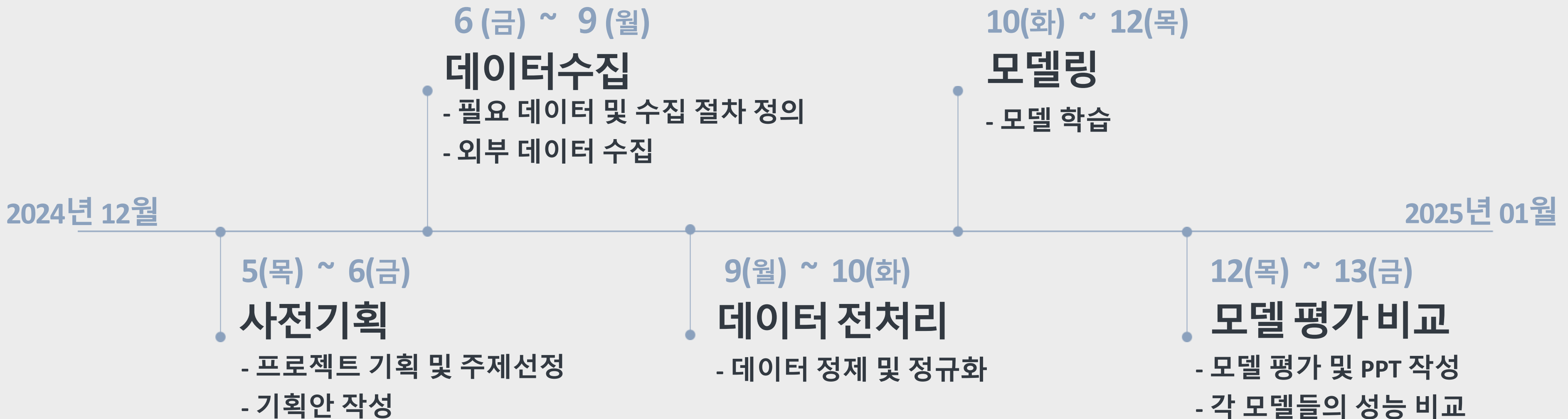
신영섭



- 데이터 분석 및 전처리
- 머신러닝 모델링

CHAPTER 1

프로젝트 일정



주제 선정 배경

가상화폐 시장의 특성

- 24시간 연중무휴 글로벌 시장
- 예측하기 어려운 급격한 가격 변동

투자자들의 비합리성

- FOMO(Fear of Missing Out) 현상
- 과신외 편향

시장의 문제점

- 높은 변동성으로 인한 투자자 스트레스 증가
- 감정적, 충동적 매매 결정



2024년 비트코인 현황

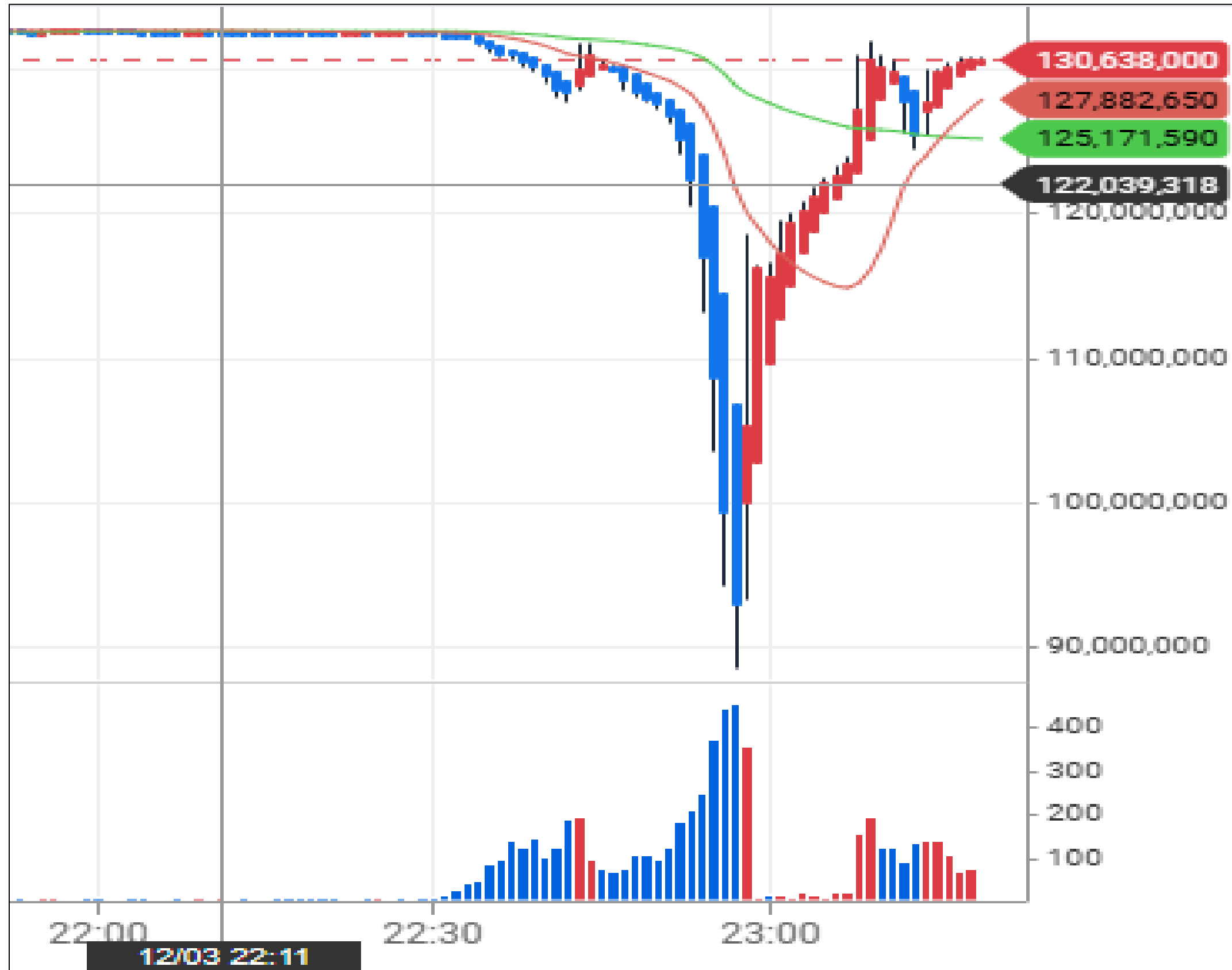
BTC/KRW, 1일, UPBIT 시143111000.0 고144490000.0 저142300000.0 종143200000.0 +89000.0 (+0.06%)
 거래량 (Volume) 1.261K

2024.12.03 계엄령 선포

2024.11.06 트럼프 당선



2024년 비트코인 현황



2024.12.03 계엄령 선포 당시 비트코인 시세



프로젝트 목표

01

데이터 분포 분석을 통한
비트코인 가격 특성 이해

02

머신러닝 기반의 가격
방향성 예측 모델 학습

03

학습 모델을 바탕으로
실제 비트코인 거래에 적용

CHAPTER 3

프로젝트 수행 절차 및 방법

01 EDA 학습 데이터 소개

02 머신러닝 모델 학습 및 평가



Pandas



CHAPTER 3

EDA

데이터 확인

- 업비트 API
- 2017-09-26 ~ 2024-12-11
- 일봉 / 09시 기준
- 컬럼 : 날짜, 시가, 종가, 최고가, 최저가, 거래량, 거래 날짜, 거래 시간, 거래 발생 시점



	type	code	opening_price	high_price	low_price
0	ticker	KRW-BTC	4322000.0	4677000.0	4318000.0
1	ticker	KRW-BTC	4657000.0	4772000.0	4519000.0
2	ticker	KRW-BTC	4586000.0	4709000.0	4476000.0
3	ticker	KRW-BTC	4657000.0	4896000.0	4651000.0
4	ticker	KRW-BTC	4889000.0	4978000.0	4682000.0

trade_price	trade_volume	trade_date	trade_time	trade_timestamp
4657000.0	32.269662	2017-09-27	09:00	1506502800
4586000.0	80.588243	2017-09-28	09:00	1506589200
4657000.0	59.352373	2017-09-29	09:00	1506675600
4895000.0	19.998483	2017-09-30	09:00	1506762000
4962000.0	27.323332	2017-10-01	09:00	1506848400



EDA

데이터 소개

Column	Non-Null Count	Dtype
type	2626 non-null	object
code	2626 non-null	object
opening_price(시가)	2626 non-null	float64
high_price(최고가)	2626 non-null	float64
low_price(최저가)	2626 non-null	float64
trade_price(종가)	2626 non-null	float64
trade_volume(거래량)	2626 non-null	float64
trade_date(거래 발생 날짜)	2626 non-null	object
trade_time(거래 발생 시간)	2626 non-null	object
trade_timestamp(거래 발생 시점)	2626 non-null	int64

- 데이터 개수 : 2626
- 결측치 없음

CHAPTER 3

EDA

데이터 소개

- 일봉을 선택한 이유
 - 그래프 시각화의 어려움, 모델 적용 어려움
- 09시가 기준이 되는 이유
 - success(돌파 성공률)이 제일 높은 시간

hour	minute	volatility	success	next_return	target			
0	0	243348	0.4293	0.0003	2623			
0	30	222113	0.3988	0.0001	2623			
1	0	232946	0.4674	0	2621			
1	30	203698	0.3444	-0.0002	2625			
2	0	194641	0.3895	0.0001	2621			
2	30	185818	0.4003	0.0001	2603			
3	0	189589	0.4287	0.0001	2603			
3	30	182572	0.4057	0.0003	2593			
4	0	190018	0.4518	0.0002	2594			
4	30	184285	0.4364	0.0003	2594			
5	0	202041	0.4721	0.0002	2601			
5	30	192248	0.4013	0.0004	2609			
6	0	198455	0.4422	0.0003	2619			
6	30	192315	0.4152	0.0003	2623			
7	0	203184	0.4539	0.0002	2624			
7	30	191373	0.3986	0	2627			
8	0	196369	0.4261	-0.0001	2626	0.3769	-0.0001	2627
8	30	179450	0.381	0.0001	2627	0.379	-0.0003	2628
9	0	295874	0.6372	-0.0001	2627	0.4087	-0.0002	2628
9	30	241045	0.3225	-0.0002	2626	0.36	-0.0004	2628
12	0	183424	0.3724	-0.0001	2628			
12	30	168971	0.3846	-0.0002	2629			
13	0	176152	0.4277	-0.0002	2628			
13	30	164583	0.3708	0	2627			
14	0	169302	0.425	0	2628			
14	30	162222	0.4129	0.0003	2628			
15	0	169039	0.4597	0	2628			
15	30	165648	0.4231	0.0001	2628			
16	0	178341	0.4494	0.0001	2628			
16	30	172121	0.396	0.0001	2626			
17	0	196927	0.4861	-0.0002	2623			
17	30	183749	0.3632	-0.0001	2624			
18	0	187392	0.4099	0.0002	2625			
18	30	181837	0.405	0	2627			
19	0	189345	0.4355	0	2629			
19	30	180416	0.404	0	2629			
20	0	188378	0.4513	0.0003	2628			
20	30	185220	0.4189	0.0001	2628			

EDA

컬럼에 대한 설명

	type	code	opening_price	high_price	low_price	trade_price	trade_volume	trade_date	trade_time	trade_timestamp
0	ticker	KRW-BTC	4322000.0	4677000.0	4318000.0	4657000.0	32.269662	2017-09-27	09:00	1506502800

- opening_price : 시가, 거래가 시작될 때의 가격
- high_price : 최고가, 해당 기간 동안 기록된 가장 높은 가격.
- low_price : 최저가, 해당 기간 동안 기록된 가장 낮은 가격.
- trade_price : 종가, 거래가 종료된 시점의 가격
- trade_volume : 거래량, 해당 기간 동안 거래된 총 거래량(비트코인의 개수).
- trade_date : 거래 발생 날짜
- trade_time : 거래 발생 시간,
- trade_timestamp : 거래 발생 시점, 거래가 발생한 시점을 밀리초 단위로 기록한 Unix 타임스탬프.

EDA

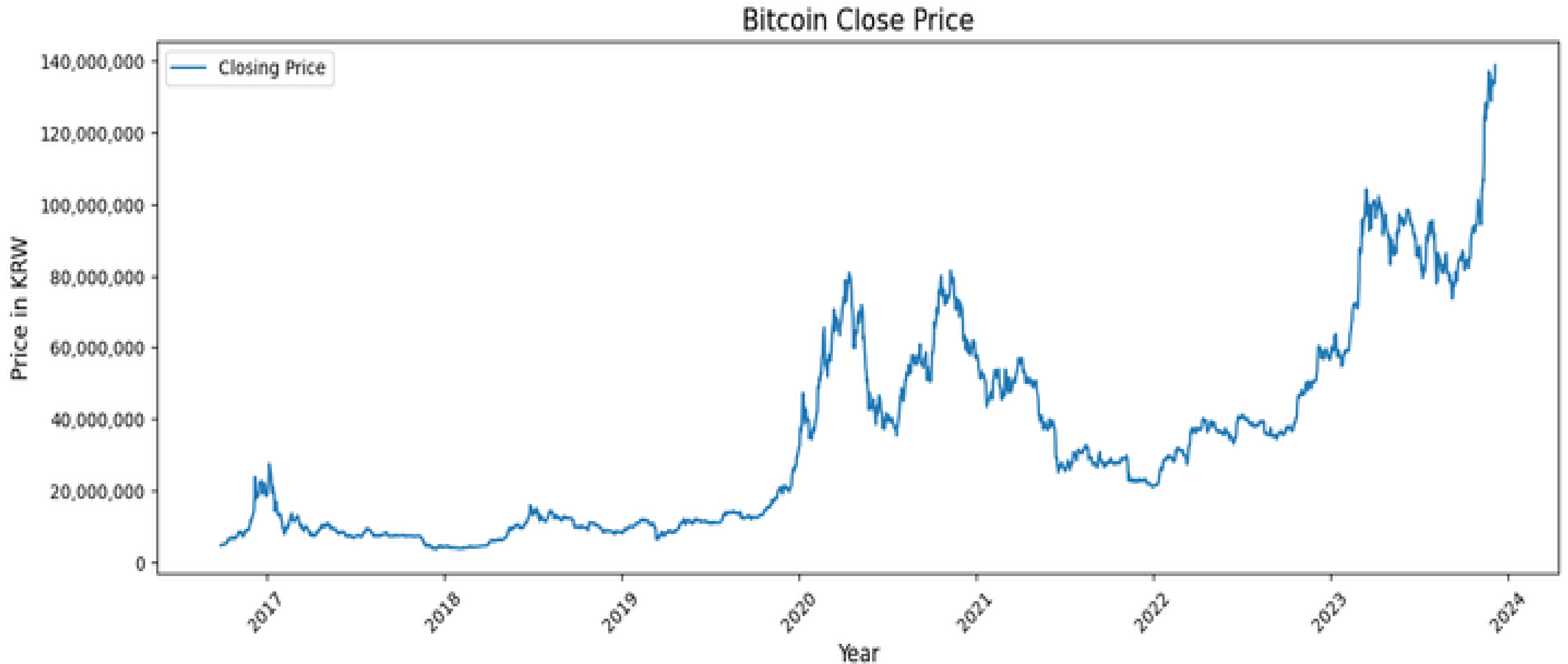
데이터 소개

	type	code	opening_price	high_price	low_price	trade_price	trade_volume	trade_date	trade_time	trade_timestamp
0	ticker	KRW-BTC	4322000.0	4677000.0	4318000.0	4657000.0	32.269662	2017-09-27	09:00	1506502800
1	ticker	KRW-BTC	4657000.0	4772000.0	4519000.0	4586000.0	80.588243	2017-09-28	09:00	1506589200
2	ticker	KRW-BTC	4586000.0	4709000.0	4476000.0	4657000.0	59.352373	2017-09-29	09:00	1506675600
3	ticker	KRW-BTC	4657000.0	4896000.0	4651000.0	4895000.0	19.998483	2017-09-30	09:00	1506762000
4	ticker	KRW-BTC	4889000.0	4978000.0	4682000.0	4962000.0	27.323332	2017-10-01	09:00	1506848400

- 종가 : 하루 동안의 모든 거래활동이 종합된 최종 합의가격, 시장의 실질적인 방향성을 반영하는 지표
- 종가는 **시장 심리의 종합적 반영, 기술적 분석의 기준점 제공, 그리고 투자 성과 측정**이라는 세 가지 핵심적인 역할을 수행하면서, 분석하는데 있어 가장 신뢰할 수 있는 기준점으로 자리잡음.

EDA

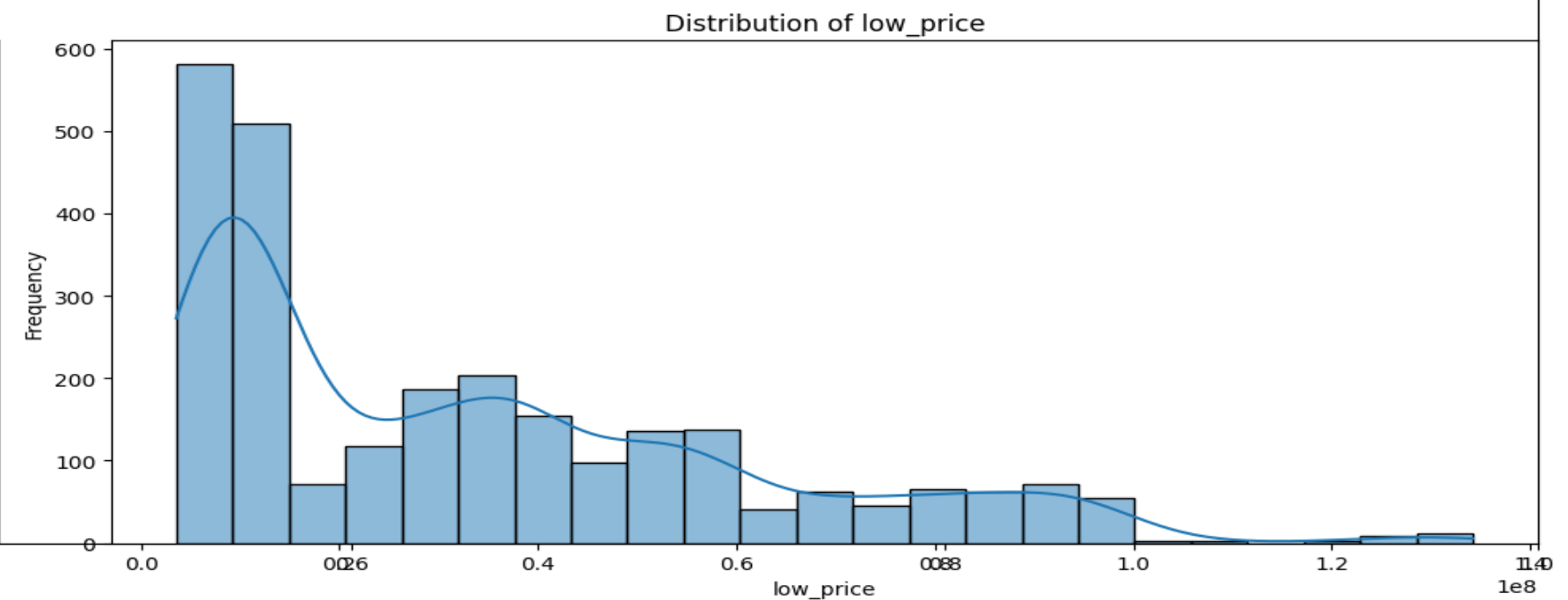
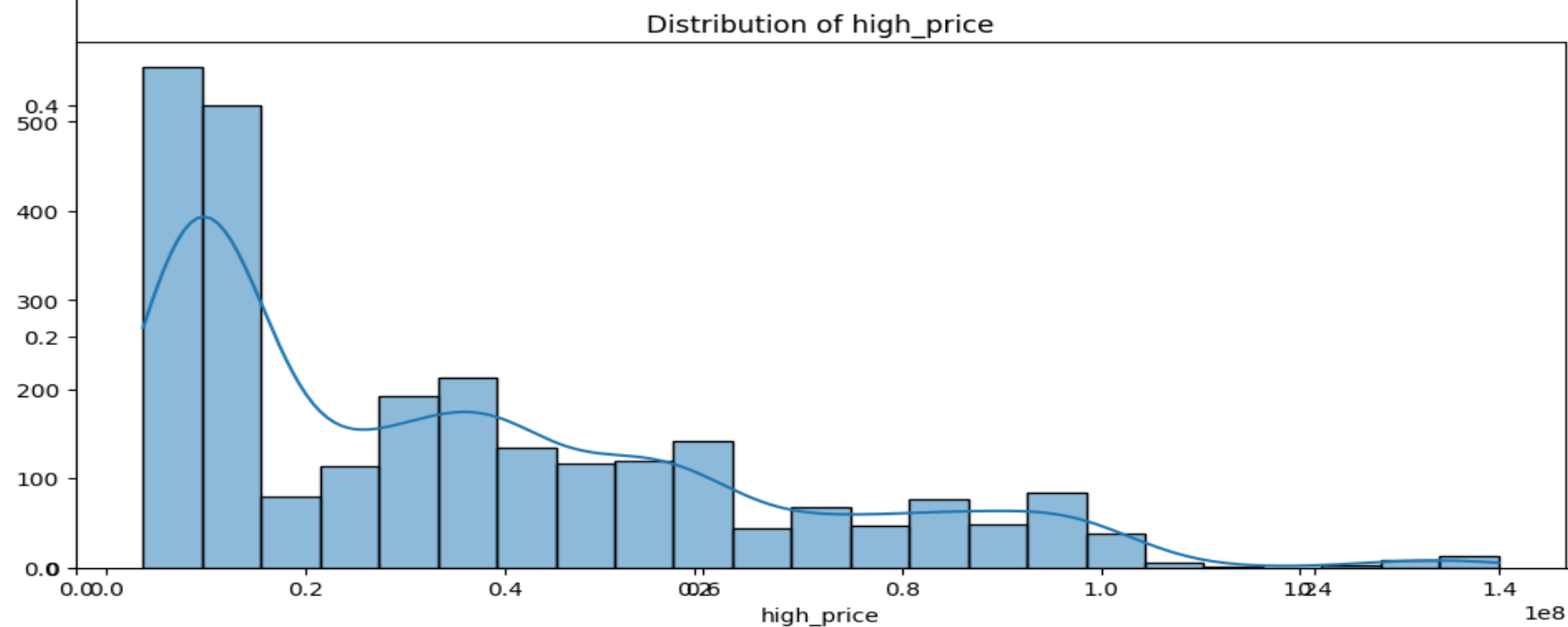
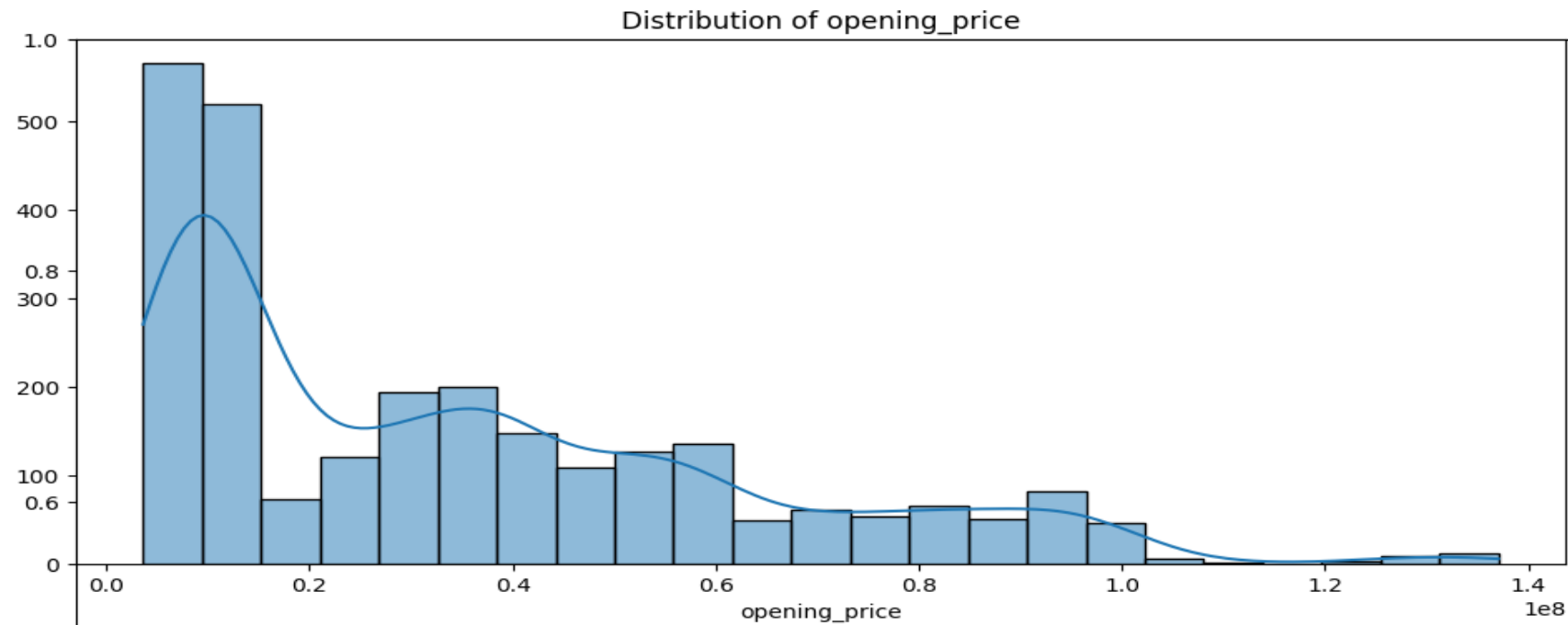
비트코인 시세 추이



EDA

비트코인 시세 관련 컬럼 분포

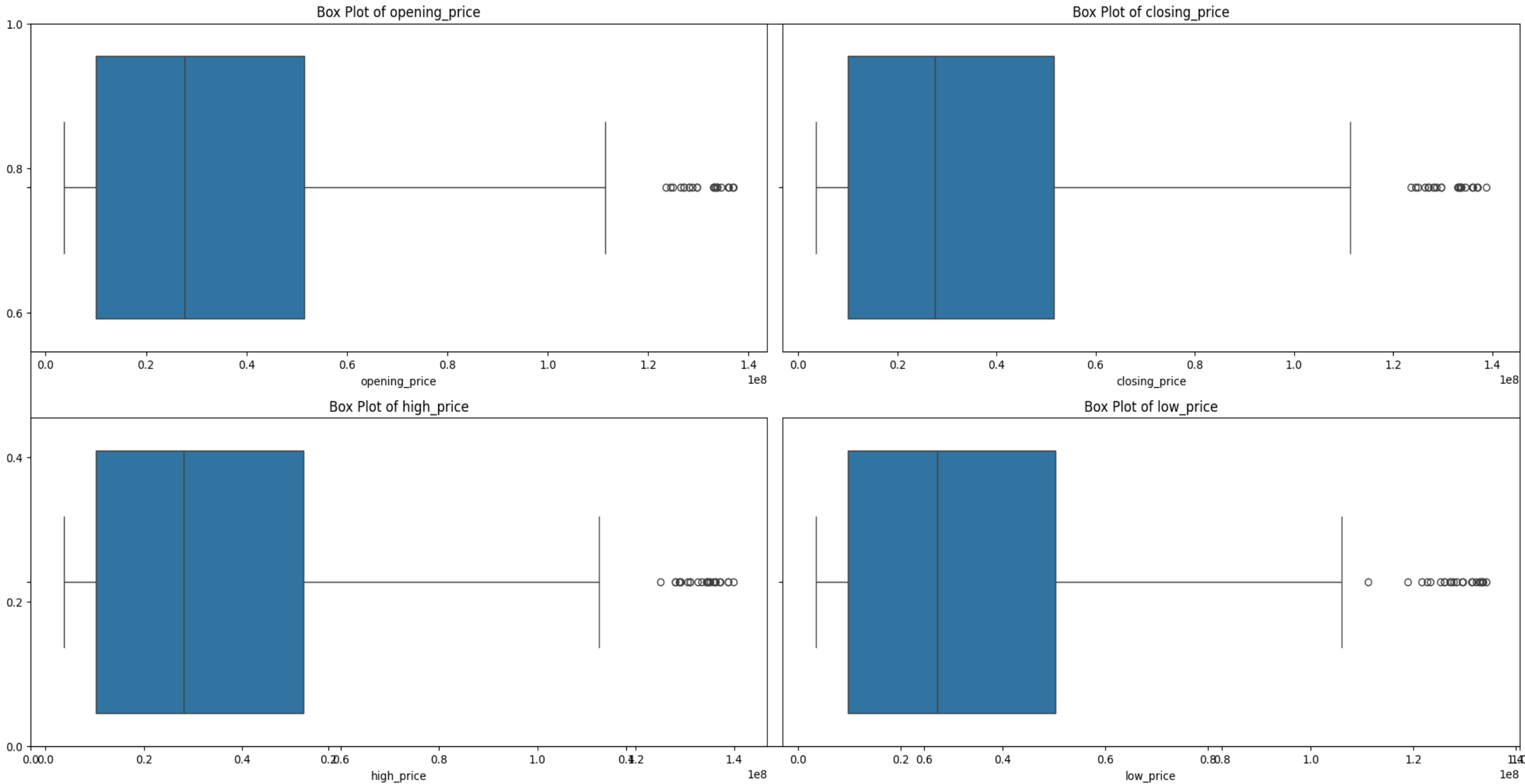
시가, 종가,최고가,최저가 분포



시계순으로 시가, 종가, 최고가, 최저가 대체적으로 **비슷한 양상**을 보임

EDA

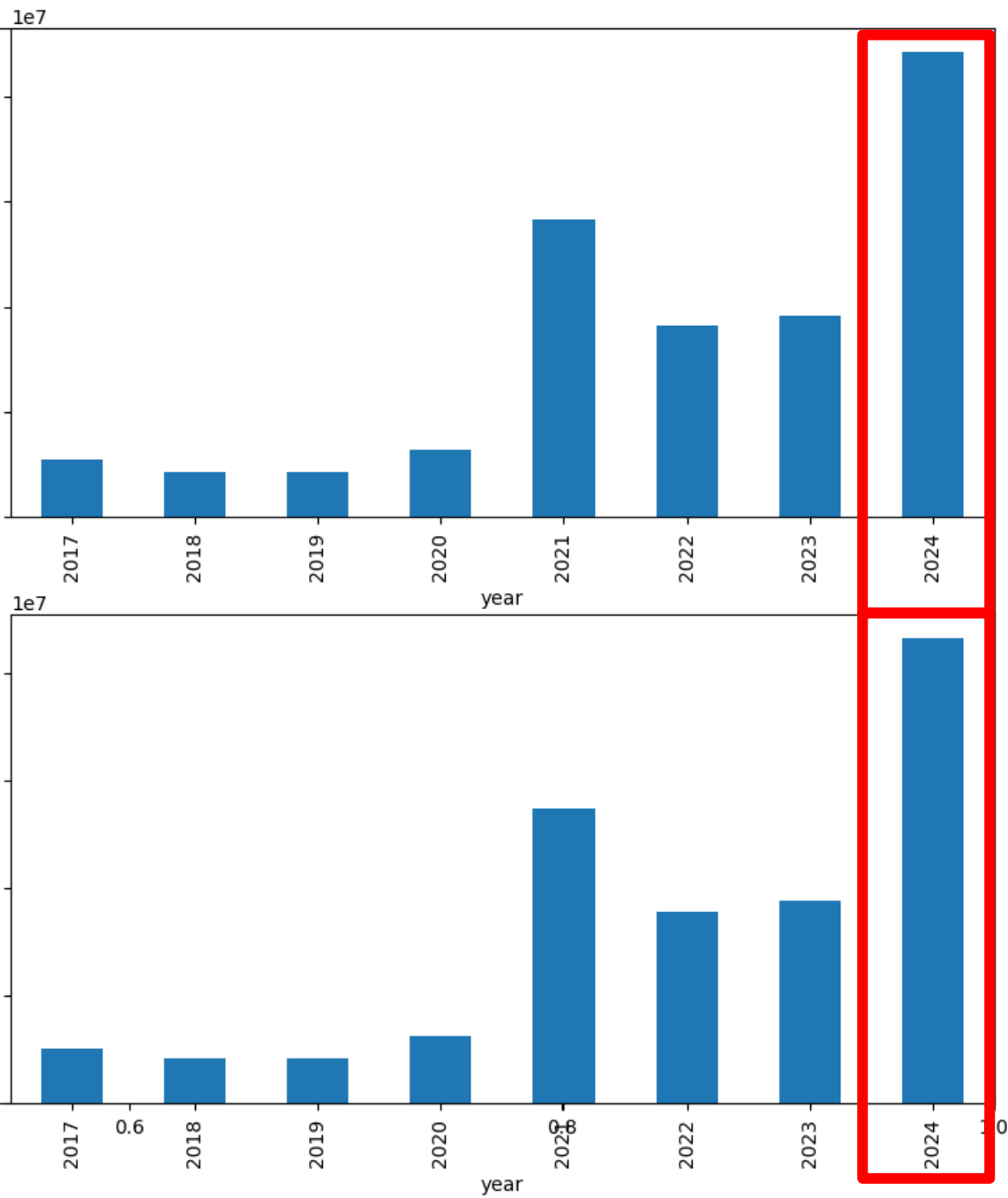
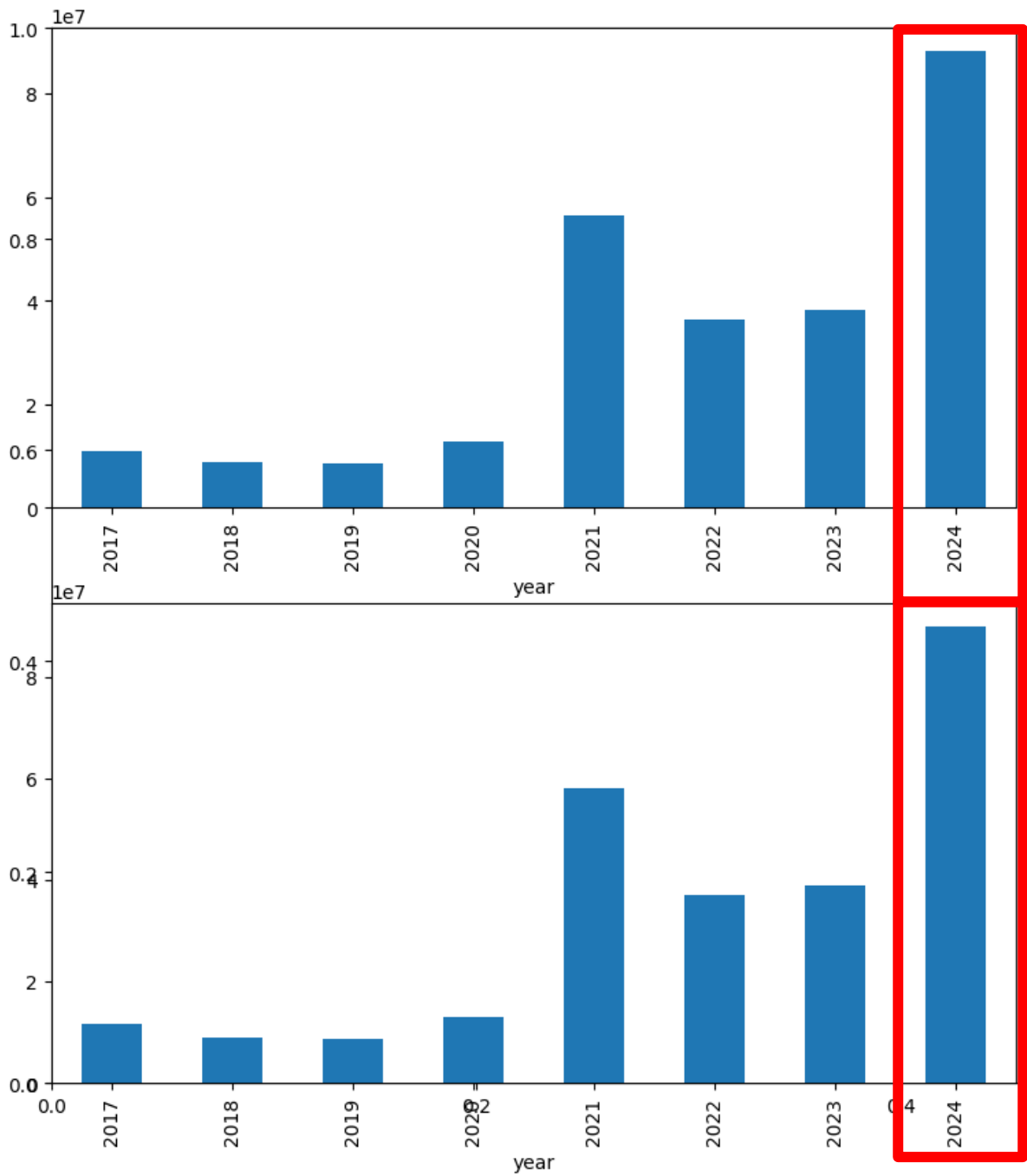
비트코인 시세 이상치 확인



이상치 개수
 시가 : 23
 종가 : 24
 최고가 : 24
 최저가 : 23

EDA

비트코인 연도별 평균 시세



2024년 비트코인 시세가 높은 이유

- 트럼프의 대선 승리 와 친암호화폐 정책
- 기관투자자들의 관심 증가
- 비트코인 반감기 효과 (비트코인 공급량 감소)
- 비트코인 ETF의 성공

데이터 전처리 과정



데이터 정제

- 중복된 데이터 제거
- 날짜 및 시간 통합
- 데이터 정렬
- 열 순서 재배치

파생변수 생성

- 날짜 분리
- 가격 차이 계산
- 분기말 여부 확인
- 타겟 변수 생성

데이터 정규화

- StandardScaler 사용

데이터 전처리 과정



파생변수 설정

- 'open close' : 시가 - 종가
- 'low-high' : 최저가 - 최고가
- 'is_quarter_end' : 각 분기의 마지막 달인지 여부
- 'target' : 전날 종가와 오늘 종가를 비교해서 상승할지, 하락할지 예측하는 종속 변수 설정

학습 변수 선택

독립 변수

- open-close: 시가-종가 차이
- low-high: 저가-고가 차이
- is_quarter_end: 분기말 여부

종속 변수

- Target:
- 다음 날 주가 상승(1)/하락(0) 예측

변수 선정 이유

- 기존 가격 변수들(시가, 종가, 최고가, 최저가)은 높은 상관관계 (0.9 이상)로 제외
- 가격 차이와 시기적 특성을 반영하는 변수만 선택

데이터 분할 전략



학습 및 테스트 데이터

- 기간 : 2017년 ~ 2023년
- 용도 : 모델 학습 및 성능 평가

검증 데이터

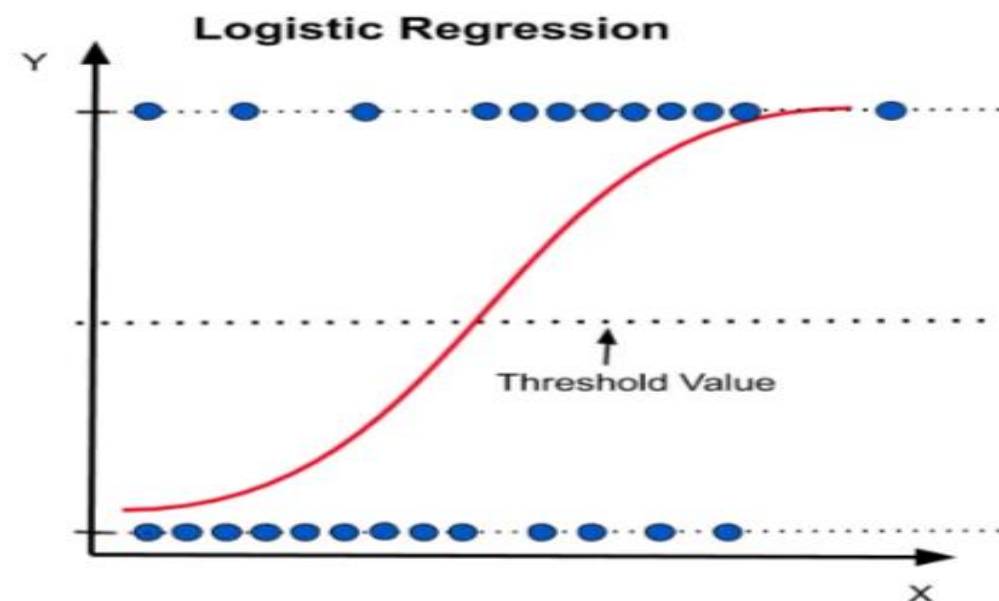
- 기간 : 2024년
- 용도 : 실제 예측 성능 검증

사용한 머신러닝 모델

모델 개요

Logistic Regression

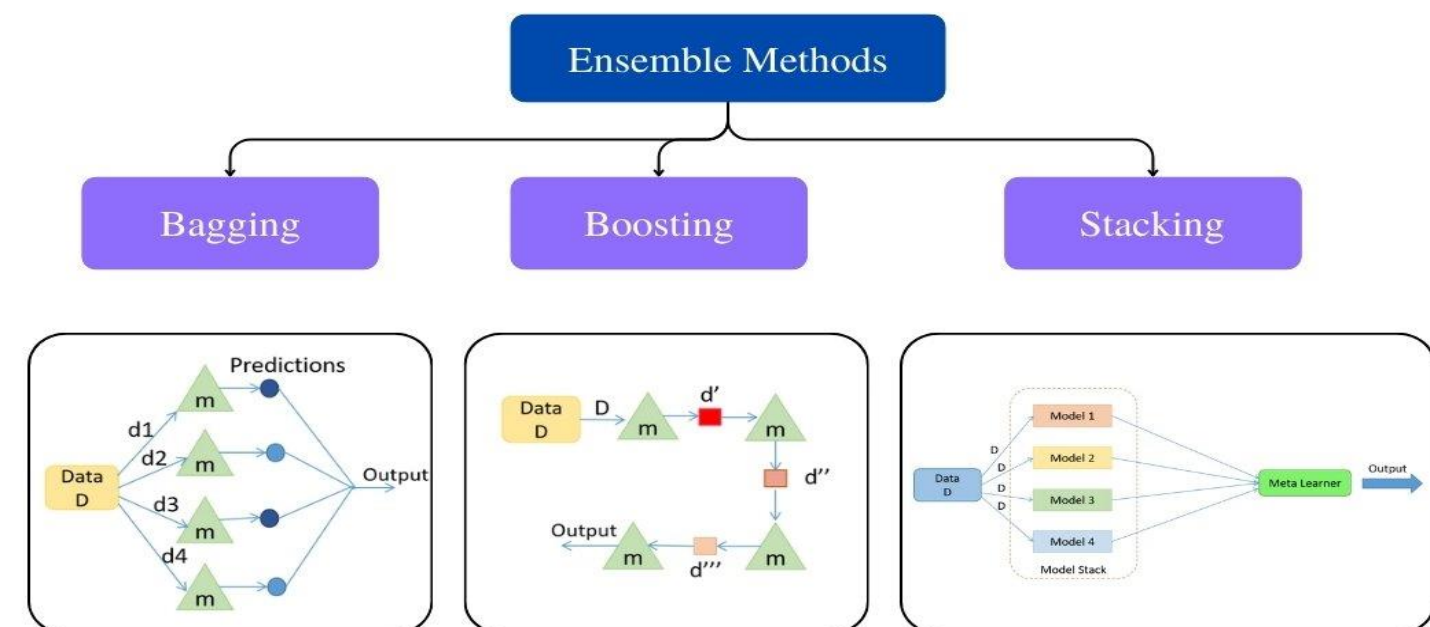
로지스틱 회귀는 선형 모델로, 출력값이 이진 또는 다중 클래스인 경우에 적합한 분류 알고리즘



Ensemble Mean

서로 다른 모델의 결과를 평균으로하여 다시 예측
장점

1. 안정성: 개별 모델의 예측 오류가 상쇄 -> 더 안정적인 예측
2. 일반화 능력 향상: 과적합 줄임
3. 변동성 감소: 특히 암호화폐시장은 변동성이 큼
-> 여러 모델의 결과를 평균내는 것이 신뢰성을 높이는데 도움이 됨



사용한 머신러닝 모델

모델 개요



XGB(XGBoost Classifier)

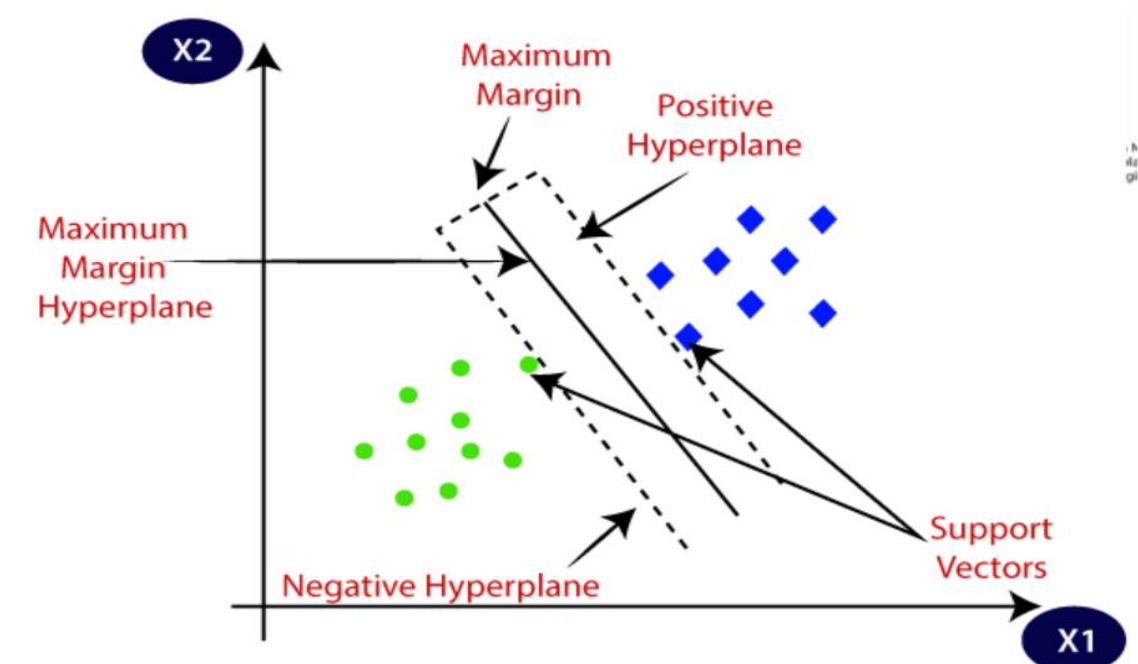
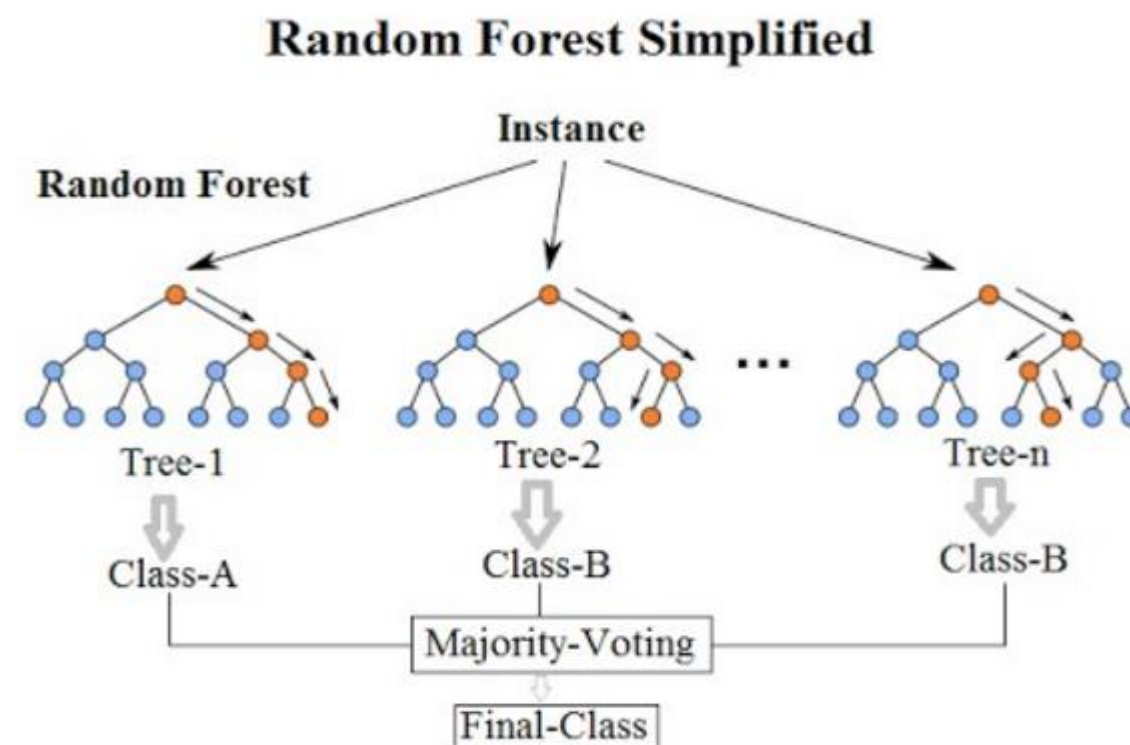
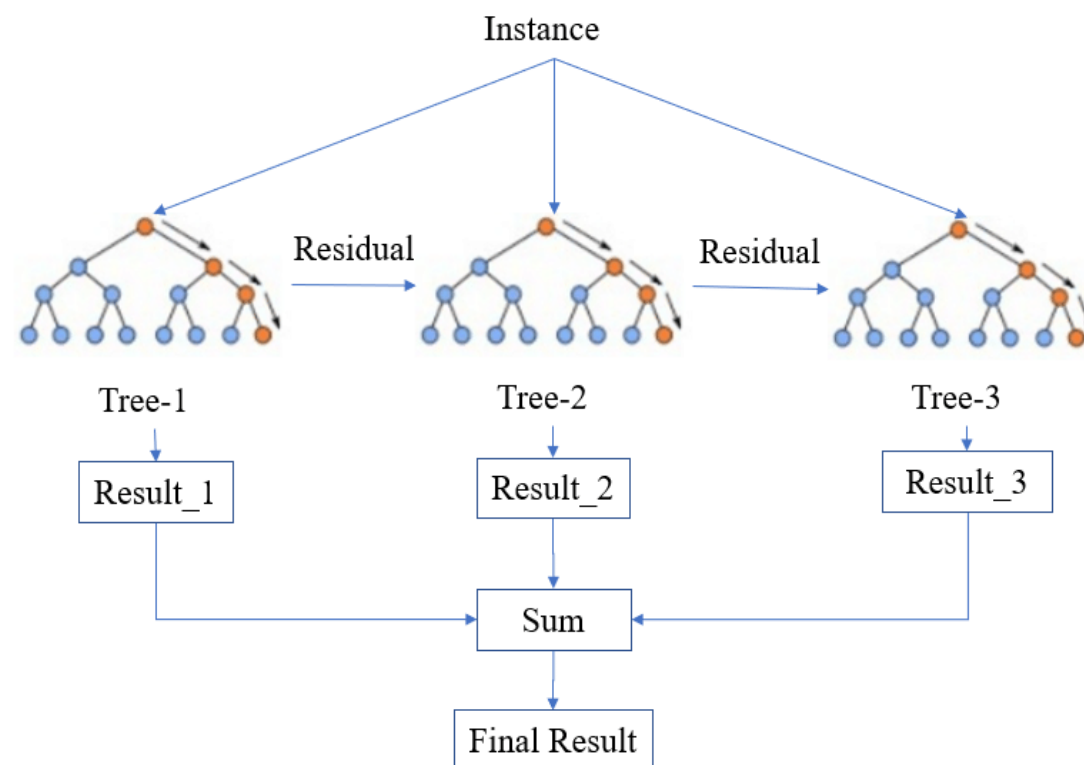
SXGBoost는 Gradient Boosting 알고리즘에 기반한 앙상블 학습 모델, 여러 모델을 결합하여 만듦.

RandomForest

여러 개의 결정트리를 학습시키고 그 결과를 결합하여 예측을 수행하는 머신러닝 알고리즘, 분류와 회귀문제에 적용.

SVC(Support Vector Classifier)

SVC는 분류 알고리즘으로, 데이터를 분리하는 최적의 값을 찾는 모델. 비선형 분포도 지원하는 것이 특징.



모델 학습 전략



하이퍼파라미터 튜닝

- 그리드서치를 통한 광범위한 파라미터 탐색
- 정밀도(Precision) 기준으로 최적 파라미터 선정

교차검증 방식

Time series split 방식 사용

- 일반적인 K-fold 대신 시계열 특성 고려
- 시간 순서를 보존하여 데이터 분할
- 미래 예측의 신뢰성 확보

모델 학습 전략



하이퍼파라미터 튜닝

- 그리드서치를 통한 광범위한 파라미터 탐색
- 정밀도(Precision) 기준으로 최적 파라미터 선정

교차검증 방식

Time series split 방식 사용

- 일반적인 K-fold 대신 시계열 특성 고려
- 시간 순서를 보존하여 데이터 분할
- 미래 예측의 신뢰성 확보

모델 평가



평가지표

- Accuracy
- Precision
- Recall
- F1-score
- ROCAUC
- Confusion Matrix

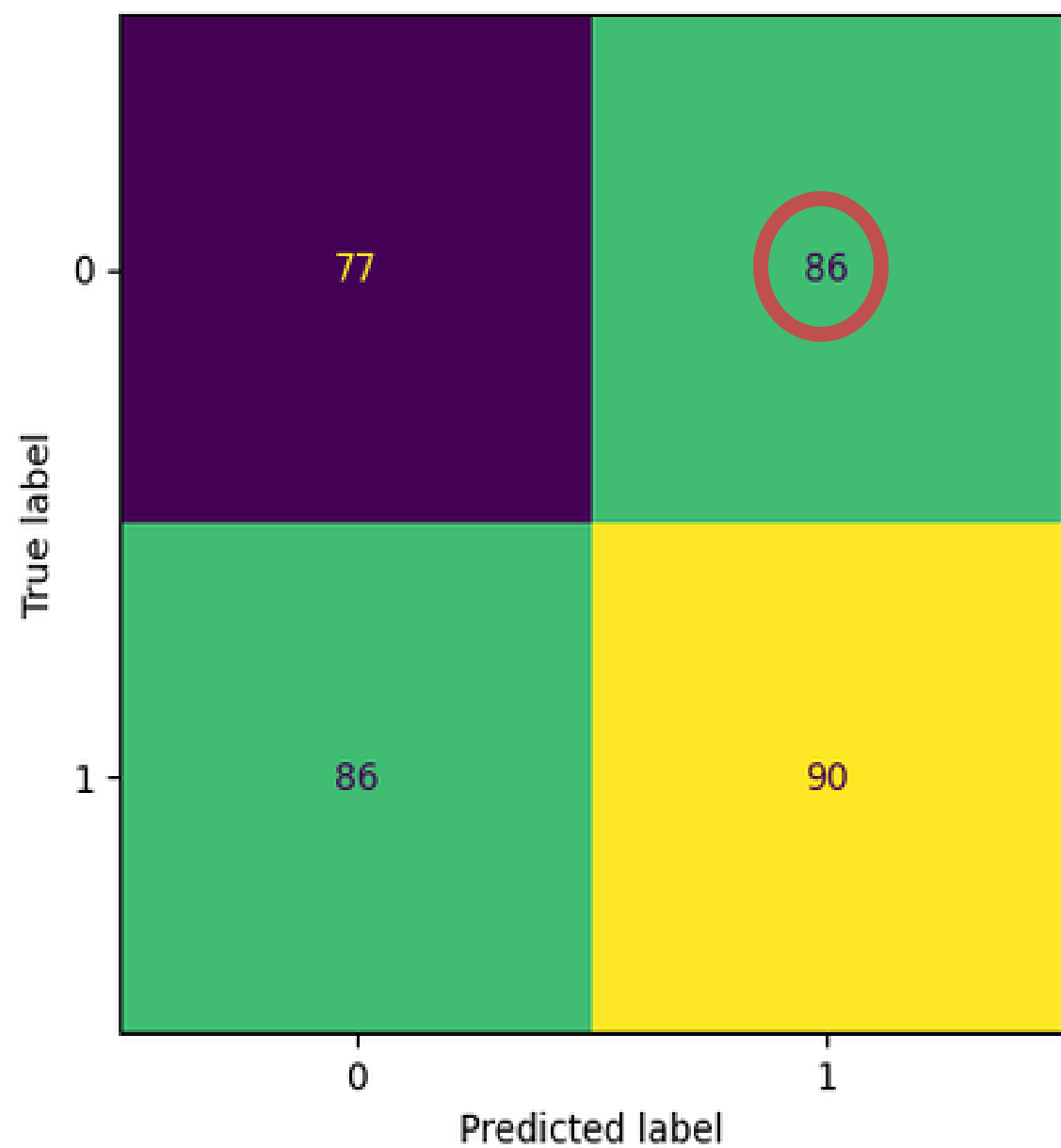
혼동 행렬 분석

- True Positive(TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

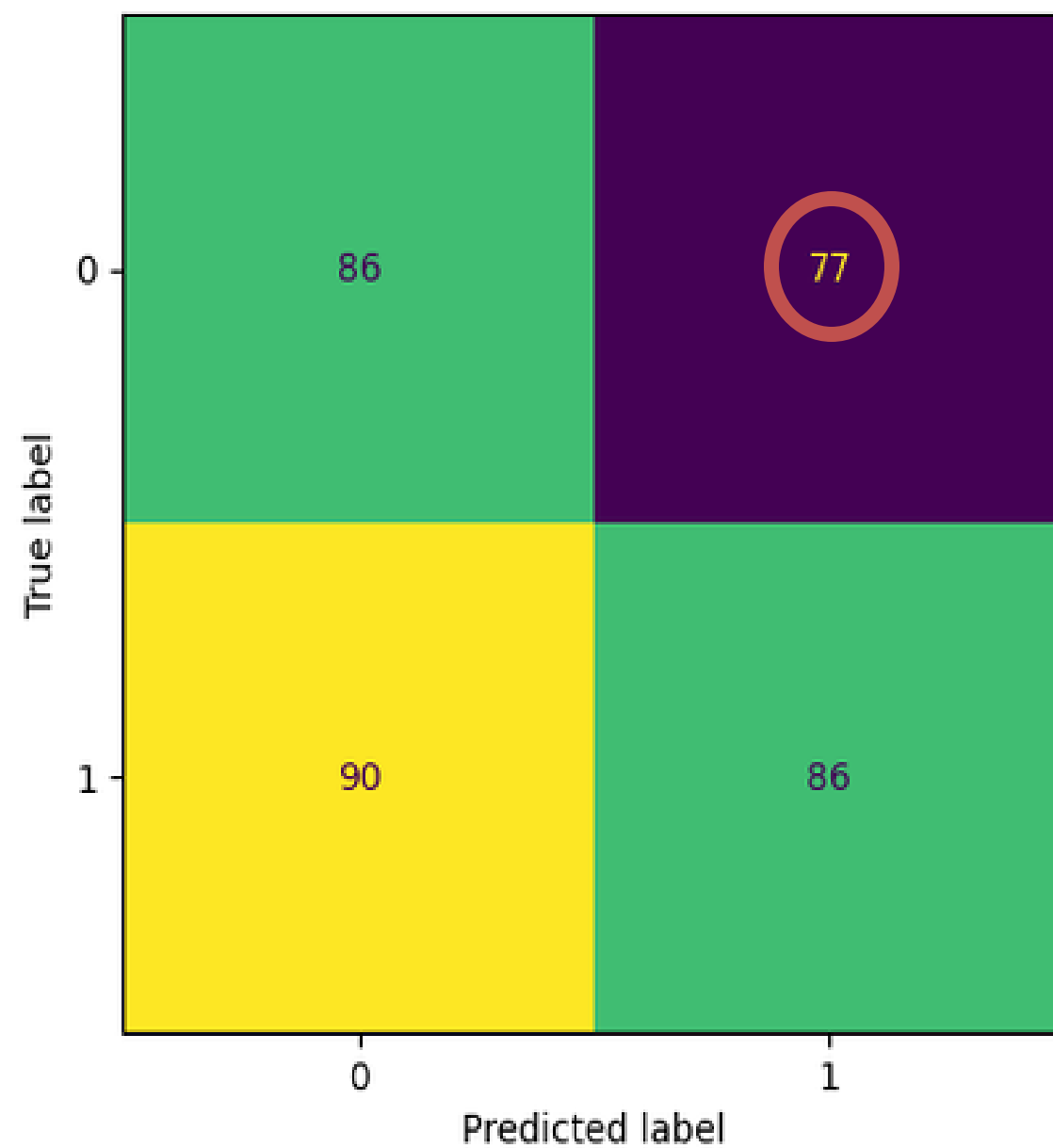
모델 선정 기준

- FP 최소화가 핵심 목표
- Precision이 높은 모델 선호

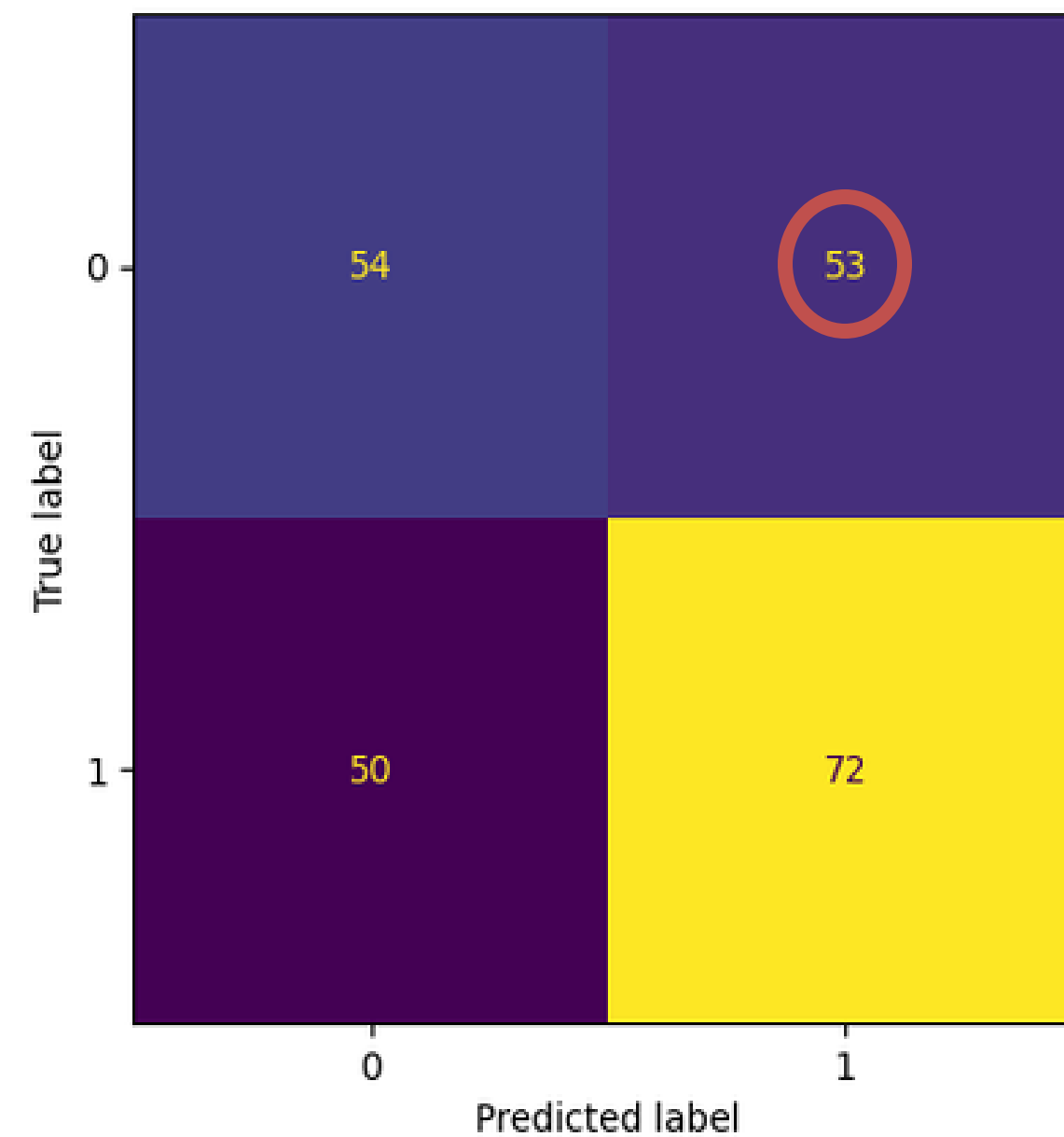
Confusion Matrix for SVC



Confusion Matrix for Random Forest



Confusion Matrix for Ensemble Model



Model	Accuracy	Precision	Recall	F1	ROC_AUC	FP
Logistic	0.5133	0.5192	0.8466	0.6436	0.5085	138
SVM	0.4926	0.5114	0.5114	0.5114	0.4965	86
XGboost	0.4956	0.5123	0.5909	0.5488	0.5051	99
RandomForest	0.5074	0.5276	0.4886	0.5074	0.4946	77
Ensemble_Mean	0.5502	0.5760	0.5902	0.5830	0.5533	53

모델 평가 결과 분석



혼동 행렬 해석

- TP/TN: 예측값과 실제값 일치
- FN: 상승을 하락으로 예측
→ 기회 손실만 발생
- FP: 하락을 상승으로 예측
→ 실제 금전적 손실 발생

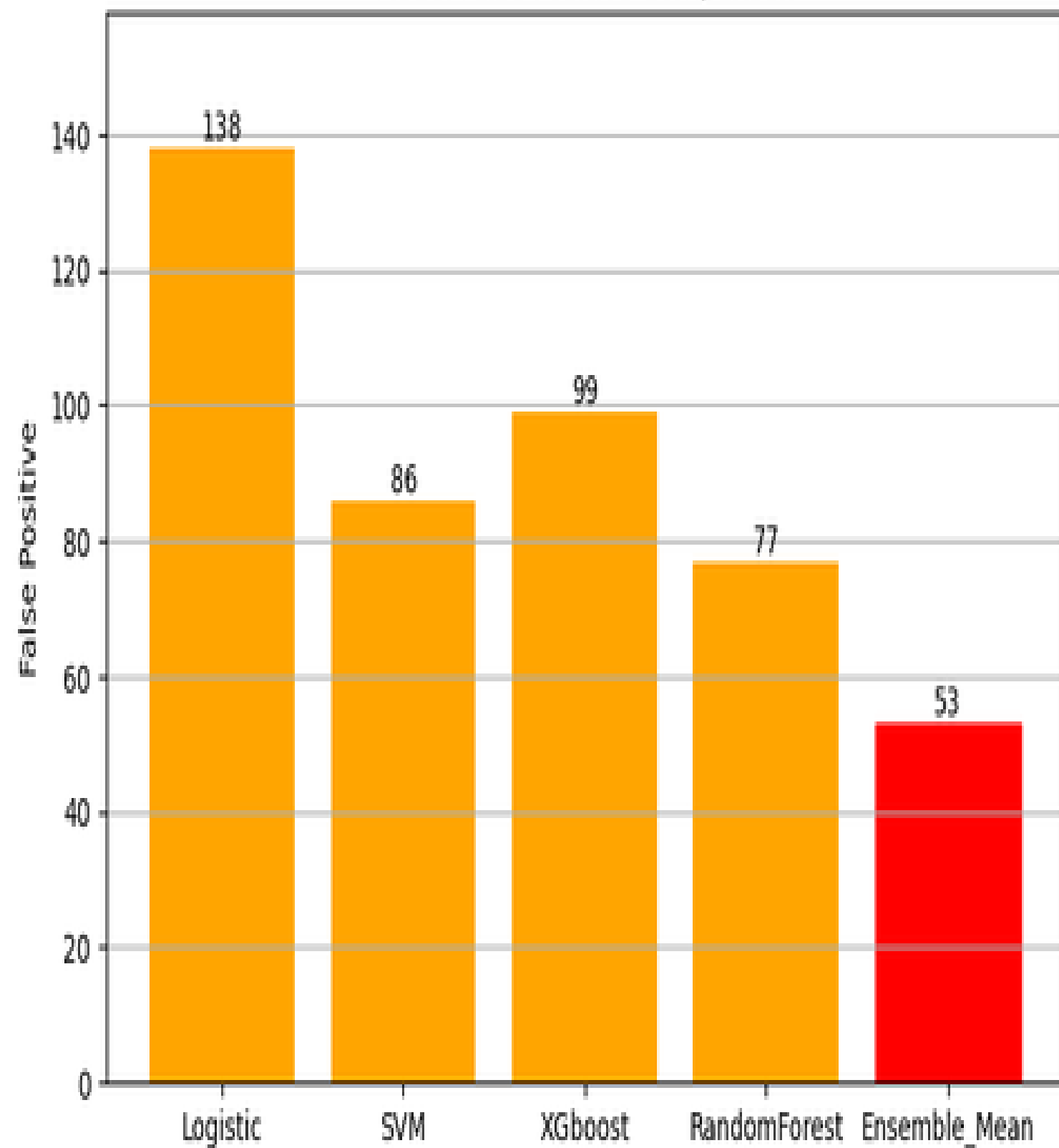
선정 근거

- $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$
지표 최적화
- FP 최소화를 통한
투자 손실 위험 관리
- 그리드서치를 통한
최적 파라미터 도출

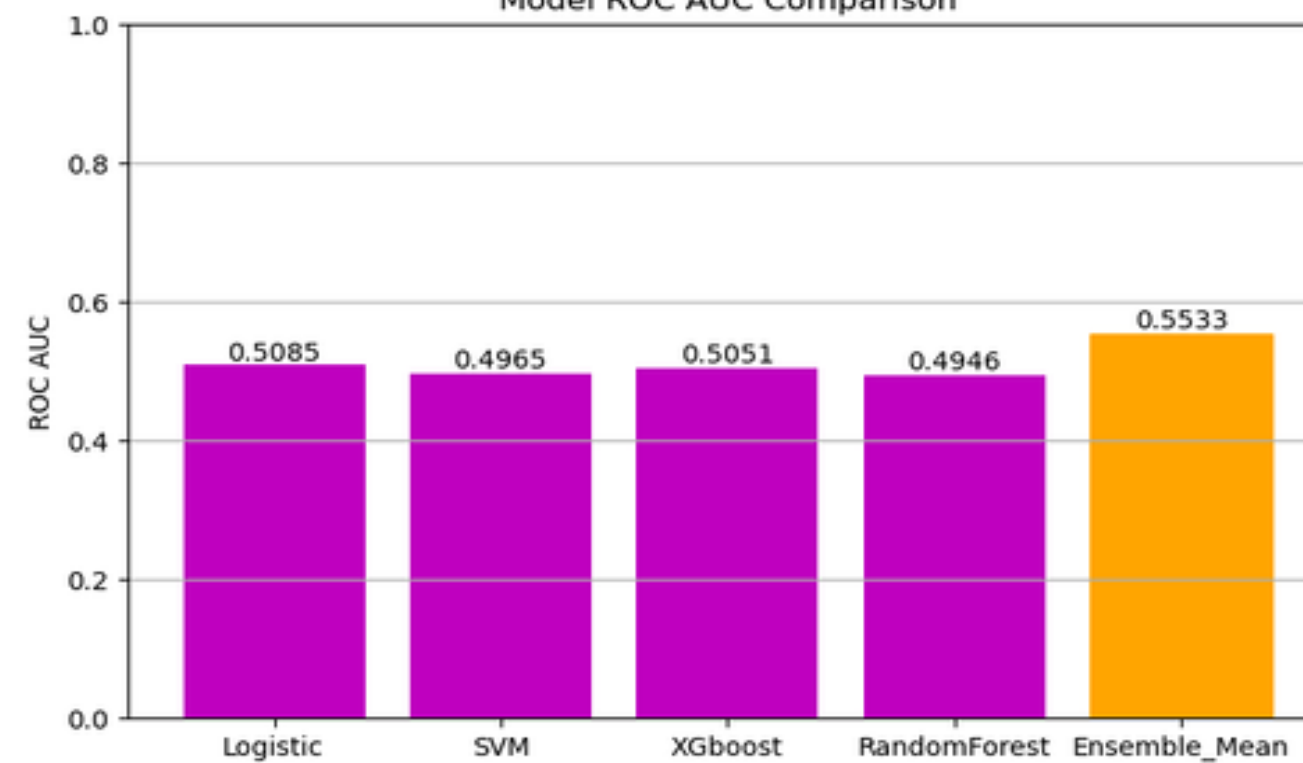
최종 모델 선정

- Ensemble Mean
- Precision 점수가 가장 우수
- 실제 투자 손실 위험 최소화

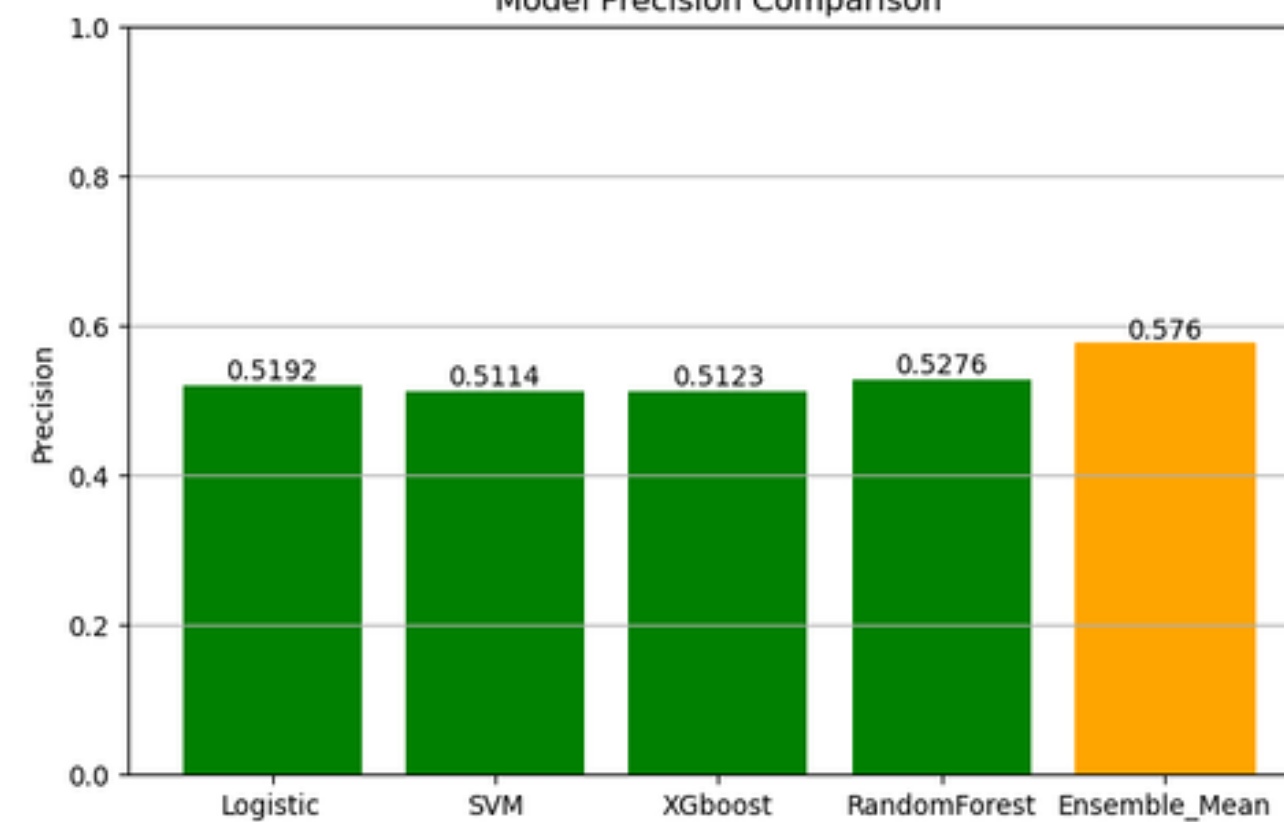
Model False Positive Comparison



Model ROC AUC Comparison



Model Precision Comparison



CHAPTER 4

결론

한계점 및 향후 개선 방향

한계점

- 데이터의 복잡성 -> 글로벌 경제상황, 정치적 요인 등 정량적인 데이터만으로 설명하기 힘든 외부 영향을 많이 받음.
- 정밀도의 한계 -> 0.57라는 긍정적인 결과에도, 투자 의사결정에 활용하기에는 부족한 수준.

개선 방향

- 추가 독립변수 설정 -> 비트코인 데이터 외에도 매크로 경제 지표, 뉴스, SNS 감성 분석을 포함하여 모델의 입력 변수를 확장.
- 데이터의 추가 확보-> 활용한 업비트의 API는 2017년 9월 27일부터 지원을 하기 때문에 머신러닝 모델에 적용하기에는 표본의 크기가 부족한 상황, 그리하여 리플, 이더리움과 같은 다른 가상화폐 데이터를 추가하여 개인 투자 성향에 맞춘 포트폴리오를 구성하는 방향.



CHAPTER 5

부록

실패의 역사

실패의 역사 1

볼린저 밴드, RSI, MA 등의 기술적 지표를 독립변수로, 변동성 돌파 여부를 종속변수로 사용했습니다. 실시간으로 비트코인 데이터를 받아오면서 돌파 여부를 분류하는 이진분류모델로 학습시켜 봤으나 돌파 여부는 모델로 예측보다는 실시간 예측으로도 바로 실행할 수 있기 때문에 보류 하였습니다.

실패의 역사 2

매수 목표 가격을 파생변수로 만들고 종속변수로 설정하여 회귀모델학습을 진행하였습니다. 하지만 매수목표가격은 바로 계산이 가능하기 때문에 종속변수로 설정한 것 자체가 잘못되었습니다. 실제로 모델 학습까지 진행하였을 때도 종속변수의 값 범위가 좁았기때문에 모델이 평균으로 예측을 하기만해도 얼추 비슷하기 때문에 R^2 값이 1에 근접하여 잘못된 모델임을 알 수 있었습니다

CHAPTER 5

소감

박훈석 

팀장으로서 팀원들이 열외 없이 프로젝트에 참여하도록 유도하였습니다. 팀원들이 잘 따라와 준 덕분에 곤란한 상황이 발생하여도 유연하게 대처할 수 있었습니다. 프로젝트 기간이 짧아 아쉬운 점이 있지만, 다음 프로젝트 때는 더 나은 결과물을 낼 수 있도록 하겠습니다.

정민관

이번 비트코인 머신러닝 프로젝트를 통해 암호화폐 시장에 대한 깊은 이해와 다양한 매매 전략을 배웠습니다. 특히 데이터 수집과 전처리 경험이 향후 프로젝트에 큰 도움이 될 것입니다. 그러나 타겟을 위한 독립변수 선택에서 어려움을 겪었습니다. 비트코인 시장의 변동성과 외부 요인들로 인해 최적의 변수를 찾는 것이 쉽지 않았습니다. 기술적 지표, 시장 심리, 뉴스 이벤트 등 여러 요소를 고려해야 했지만, 제한된 시간 안에 결정하기가 어려웠습니다.

이 현

머신 러닝을 활용한 첫 프로젝트여서 시작부터 기대가 많이 되었습니다. 하지만 짧은 시간으로 인하여 진행 중 발생한 문제에 대해서 유연하게 대처하지 못했던 점이 아쉬웠습니다. 그래도 좋은 팀원들 덕분에 머신러닝에 관하여 심도 깊게 배울 수 있었던 시간이었습니다. 팀원들에게 진심 가득 담은 감사의 말씀을 전합니다.

신영섭

도메인 공부부터 데이터 전처리 과정, 또 그 전처리된 데이터를 바탕으로 다양한 머신러닝 모델을 적용하는 시간이 의미 있었습니다. 하지만 비트코인이라는 도메인 지식이 부족하고 투자에 대한 실전 경험이 없다보니 인사이트 도출에 어려움을 겪었던 점이 아쉬웠던 것 같습니다. 사랑합니다.