빅데이터 분석 기법을 활용한 암호화폐 가격 예측 모델링

A Big Data Analytics Approach to Cryptocurrency Price Prediction Modeling

- 소속 : 서강대학교 AI & SW 대학원
- 전공 : 데이터사이언스 & AI 전공
- 과목 : 빅데이터 분석 예측 / 지도교수 정화민 교수님
- 5팀 : 설현일(팀장), 전명준, 유지연, 김수민

Contents

분석 목적

암호화폐 시장의 급격한 변동성과 정책적 요인을 통합적으로 분석하여,
데이터 기반의 예측 가능성을 높이고 투자 및 정책 결정의 효율성을 제고

분석 목적

분석 필요성

시장적 관점:

- 암호화폐 시장은 급격한 변동성과 복잡성을 가지며, 투자자와 기업이 정확한 데이터를 기반으로 한 의사결정을 요구.
- 단순한 가격 예측이 아닌, 시장의 주요 요인(거래량, 환율 등)을 반영한 분석 필요성 증대.

기술적 관점:

- 빅데이터 및 머신러닝 알고리즘(Random Forest, LSTM, XGBoost 등)을 활용한 고도화된 분석 기법이 필요한 상황.
- 전통적 통계 분석 방법으로는 대규모 데이터의 상관성과 패턴을 효과적으로 분석하기 어려움.

분석 목적

시장적 목표:

- PCA 및 K-Means 알고리즘을 통해 주요 암호화폐를 군집화하고 대표 코인을 선정하여 시장의 이해도를 높임.
- 암호화폐 시장 데이터를 기반으로 최적의 머신러닝 모델을 구축하여 가격 변동성을 예측하고 투자 및 전략적 결정을 지원.

기술적 목표:

- 다양한 데이터 소스(가격, 환율, 거래량 등)를 통합적으로 분석하여, 시장의 주요 패턴과 요인을 시각화.
- 예측 성능 지표(MAE, RMSE, R^2)를 통해 최적의 모델을 선정하고, 실질적인 의사결정 도구로 활용.

문헌 연구

No.	논문명	저자	출처	주요 내용
1	방법론 한국 주식시장에서의 군집화 기반 페어트레이딩 방법론	조풍진, 이민혁, 송재욱(2022)	*Journal of Korean Society of Industrial and Systems Engineering*	• 한국 주식시장에서 군집화 기반의 페어트레이딩 전략이 수익성을 가지는지 검증하고, 기존 공적분 기반 방법론과 비교하여 성과를 평가 • K-평균 및 계층적 군집 알고리즘을 적용해 유사한 주가 흐름을 보이는 종목 간 페어를 탐색 • 군집화 기반의 페어트레이딩 전략이 유효하며, 특히 계층적 군집은 안정적인 수익성과 낮은 위험도를 기록, 기존 공적분 기반 방법론보다 높은 수익성을 보임
2	암호화폐 종가 예측 성능과 입력 변수 간의 연관성 분석 방법론	박재현, 서영석 (2022)	*정보처리학회논문지*	• 암호화폐 종가 예측에서 입력 변수들의 조합이 예측 성능에 미치는 영향을 분석하고, 최적의 변수 조합을 도출 • Bitcoin, Ethereum 등 6개 암호화폐 데이터 수집 (2018~2021년), LSTM 모델을 사용해 다양한 입력 변수 조합(개장가, 고가, 저가, 종가, 거래량, 종가 변동률)을 평가, 상관분석 및 VIF(다중 공선성) 분석으로 변수의 독립성과 예측 성능에 미치는 영향 파악, 예측 성능 지표(MSE, RMSE, MAE, R^2)로 최적 변수 조합 도출 • 개장가, 고가, 저가, 거래량, 종가의 조합이 가장 우수한 성능을 기록.
3	어텐션 메커니즘을 활용한 채소 가격 예측	임혜진 (2020)	*세종대학교 석사학위 논문*	• 어텐션 메커니즘을 적용해 채소 가격 예측의 정확도를 높이고, 해당 모델이 기존 기법보다 우수한지 확인 • 어텐션 메커니즘이 포함된 딥러닝 모델 설계 및 학습, 존 모델(LSTM, GRU 등)과의 비교 실험, 예측 정확도 및 성능 지표(RMSE, MAE) 분석 • 어텐션 메커니즘 기반 모델이 기존 모델에 비해 예측 정확도 및 성능이 우수하며, 중요 변수를

방향성 설정

Question 1

암호화폐 데이터를 기반으로 대표 코인을 선정하는 최적의 방법은 무엇인가?

Key Question

빅데이터 분석기법을 활용한 암호화폐 가격 예측 모델링

Question 2

선정된 대표 코인의 데이터를 활용하여 암호화폐 가격을 예측하는 최적의 모델은 무엇인가?

Approach

1. 대표 암호화폐 선정
2. 가격 예측 및 모델 평가

- 수집 데이터 : 가격 데이터 (고가, 저가, 개장가, 종가, 거래량), 환율 데이터
- PCA와 K-Means를 사용하여 대표 암호화폐 선정 → 대표 암호화폐 (BSV, DASH, LUNC)
- Random Forest, LSTM, XGBoost 등 다양한 모델로 암호화폐 가격 예측 수행하고자 함
- 예측 성능 지표 (MAE, RMSE, R^2)를 기반으로 최적 모델 도출하고자 함

자료수집 및 방법론 설명

목표: BTC, ETH, SOL 가격 및 온체인 데이터 수집

방법:

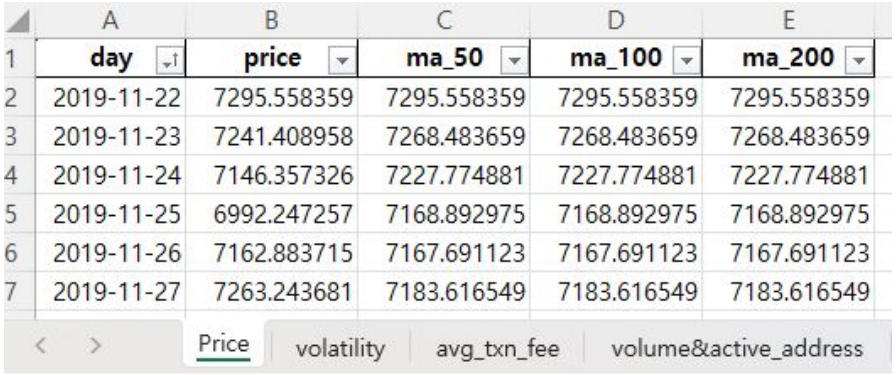
- Dune Analytics API를 통해 특정 쿼리의 결과 데이터를 CSV 형식으로 처리
- 여러 쿼리의 결과를 하나의 Excel 파일로 저장하는 Python 코드 작성

코드 구성 및 설명:

- Dune API 설정
- CSV 데이터를 데이터프레임으로 변환
- 여러 쿼리 결과를 Excel로 저장
- Excel 파일 저장 및 실행

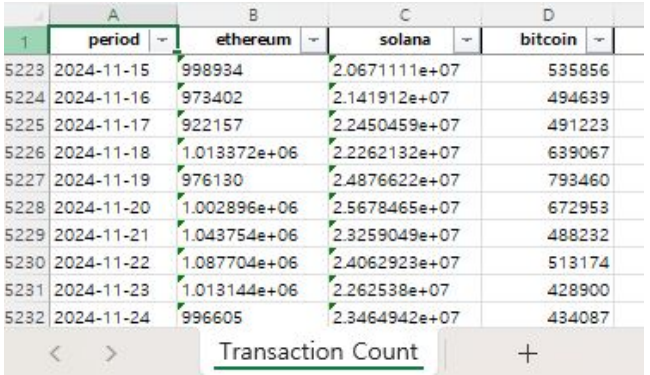
최종 수집 데이터:

- BTC, ETH, SOL (Price, Volatility, Avg_txn,fee, Active_address)
- Transaction Fees
- Transfers Volume
- Transaction Count



	A	B	C	D	E
1	day	price	ma_50	ma_100	ma_200
2	2019-11-22	7295.558359	7295.558359	7295.558359	7295.558359
3	2019-11-23	7241.408958	7268.483659	7268.483659	7268.483659
4	2019-11-24	7146.357326	7227.774881	7227.774881	7227.774881
5	2019-11-25	6992.247257	7168.892975	7168.892975	7168.892975
6	2019-11-26	7162.883715	7167.691123	7167.691123	7167.691123
7	2019-11-27	7263.243681	7183.616549	7183.616549	7183.616549

Price | volatility | avg_txn_fee | volume&active_address



	A	B	C	D
1	period	ethereum	solana	bitcoin
5223	2024-11-15	998934	2.0671111e+07	535856
5224	2024-11-16	973402	2.141912e+07	494639
5225	2024-11-17	922157	2.2450459e+07	491223
5226	2024-11-18	1.013372e+06	2.2262132e+07	639067
5227	2024-11-19	976130	2.4876622e+07	793460
5228	2024-11-20	1.002896e+06	2.5678465e+07	672953
5229	2024-11-21	1.043754e+06	2.3259049e+07	488232
5230	2024-11-22	1.087704e+06	2.4062923e+07	513174
5231	2024-11-23	1.013144e+06	2.262538e+07	428900
5232	2024-11-24	996605	2.3464942e+07	434087

Transaction Count

자료수집 및 방법론 설명

목표: 500 여종 코인 가격 데이터(5년 동안의 일간 데이터)

방법:

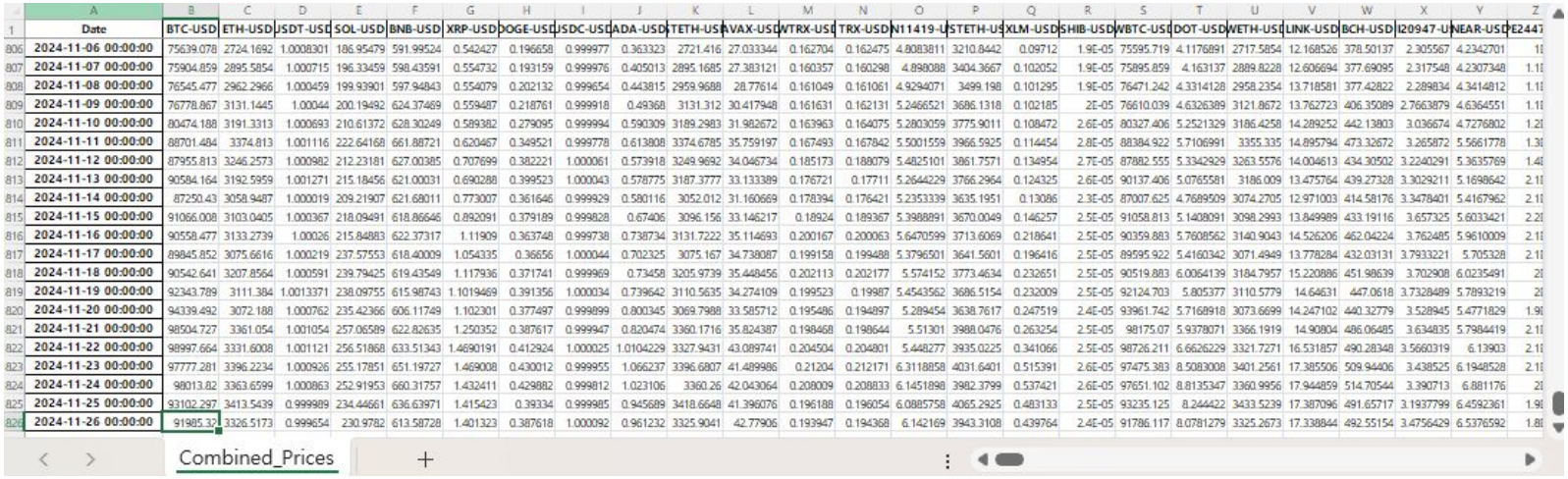
- Yahoo Finance API를 사용하여 지정한 티커(주식 또는 가상자산 등)의 금융 데이터 확보
- 전체 데이터를 Excel 파일로 저장하는 Python 코드 작성

코드 구성 및 설명:

- yfinance 라이브러리를 활용하여 지정한 티커 리스트에 대해 금융 데이터를 수집
- 기간 설정 (5년)
- 티커 목록 읽기
- 결과 저장 (Excel 파일)

최종 수집 데이터:

- 450 여종 코인 가격 데이터



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Date	BTC-USD	ETH-USD	USDT-USD	SOL-USD	BNB-USD	XRP-USD	DOGE-USD	USDC-USD	ADA-USD	STETH-USD	AVAX-USD	WTRX-USD	TRX-USD	N11419-USD	STETH-USD	XLM-USD	SHIB-USD	WBTC-USD	DOT-USD	WETH-USD	LINK-USD	BCH-USD	I20947-USD	NEAR-USD	E2447
806	2024-11-06 00:00:00	75639.078	2724.1692	1.0008301	186.95479	591.99524	0.542427	0.196658	0.999977	0.363323	2721.416	27.033344	0.162704	0.162475	4.8083811	3210.8442	0.09712	1.9E-05	75595.719	4.1176891	2717.5854	12.168526	378.50137	2.305567	4.2342701	1
807	2024-11-07 00:00:00	75904.859	2895.5854	1.000715	196.33459	598.43591	0.554732	0.193159	0.999976	0.405013	2895.1685	27.383121	0.160357	0.160298	4.898088	3404.3667	0.102052	1.9E-05	75895.859	4.163137	2889.8228	12.606694	377.69095	2.317548	4.2307348	1.1
808	2024-11-08 00:00:00	76545.477	2962.2966	1.000459	199.93901	597.94843	0.554079	0.202132	0.999654	0.443815	2959.9688	28.77614	0.161049	0.161061	4.9294071	3499.198	0.101295	1.9E-05	76471.242	4.3314128	2958.2354	13.718581	377.42822	2.289834	4.3414812	1.1
809	2024-11-09 00:00:00	76778.867	3131.1445	1.00044	200.19492	624.37469	0.559487	0.218761	0.999918	0.49368	3131.312	30.417948	0.161631	0.162131	5.2466521	3686.1318	0.102185	2E-05	76610.039	4.6326389	3121.8672	13.762723	406.35089	2.7663879	4.6364551	1.1
810	2024-11-10 00:00:00	80474.188	3191.3313	1.000693	210.61372	628.30249	0.589382	0.279095	0.999994	0.590309	3189.2983	31.982672	0.163963	0.164075	5.2803059	3775.9011	0.108472	2.6E-05	80327.406	5.2521329	3186.4258	14.289252	442.13803	3.036674	4.7276802	1.2
811	2024-11-11 00:00:00	88701.484	3374.813	1.001116	222.64168	661.88721	0.620467	0.349521	0.999778	0.613808	3374.6785	35.759197	0.167493	0.167842	5.5001559	3966.5925	0.114454	2.8E-05	88384.922	5.7106991	3355.335	14.895794	473.32672	3.265872	5.5661778	1.3
812	2024-11-12 00:00:00	87955.813	3246.2573	1.000982	212.23181	627.00385	0.707699	0.382221	1.000061	0.573918	3249.9692	34.046734	0.185173	0.188079	5.4825101	3861.7571	0.134954	2.7E-05	87882.555	5.3342929	3263.5576	14.004613	434.30502	3.2240291	5.3635769	1.4
813	2024-11-13 00:00:00	90584.164	3192.5959	1.001271	215.18456	621.00031	0.690288	0.399523	1.000043	0.578775	3187.3777	33.133389	0.176721	0.17711	5.2644229	3766.2964	0.124325	2.6E-05	90137.406	5.0765581	3186.009	13.475764	439.27328	3.3029211	5.1698642	2.1
814	2024-11-14 00:00:00	87250.43	3058.9487	1.000019	209.21907	621.68011	0.773007	0.361646	0.999929	0.580116	3052.012	31.160669	0.178394	0.176421	5.2353339	3635.1951	0.13086	2.3E-05	87007.625	4.7689509	3074.2705	12.971003	414.58176	3.3478401	5.4167962	2.1
815	2024-11-15 00:00:00	91066.008	3103.0405	1.000367	218.09491	618.86646	0.892091	0.379189	0.999828	0.67406	3096.156	33.146217	0.18924	0.189367	5.3988891	3670.0049	0.146257	2.5E-05	91058.813	5.1408091	3098.2993	13.849989	433.19116	3.657325	5.6033421	2.2
816	2024-11-16 00:00:00	90558.477	3133.2739	1.00026	215.84883	622.37317	1.11909	0.363748	0.999738	0.738734	3131.7222	35.114693	0.200167	0.200063	5.6470599	3713.6069	0.218641	2.5E-05	90359.883	5.7608562	3140.9043	14.526206	462.04224	3.762485	5.9610009	2.1
817	2024-11-17 00:00:00	89845.852	3075.6616	1.000219	237.57553	618.40009	1.054335	0.36656	1.000044	0.702325	3075.167	34.738087	0.199158	0.199488	5.3796501	3641.5601	0.196416	2.5E-05	89595.922	5.4160342	3071.4949	13.778284	432.03131	3.7933221	5.705328	2.1
818	2024-11-18 00:00:00	90542.641	3207.8564	1.000591	239.79425	619.43549	1.117936	0.371741	0.999969	0.73458	3205.9739	35.448456	0.202113	0.202177	5.574152	3773.4634	0.232651	2.5E-05	90519.883	6.0064139	3184.7957	15.220886	451.98639	3.702908	6.0235491	2
819	2024-11-19 00:00:00	92343.789	3111.384	1.0013371	238.09755	615.98743	1.1019469	0.391356	1.000034	0.739642	3110.5635	34.274109	0.199523	0.19987	5.4543562	3686.5154	0.232009	2.5E-05	92124.703	5.805377	3110.5779	14.64631	447.0618	3.7328489	5.7893219	2
820	2024-11-20 00:00:00	94339.492	3072.188	1.000762	235.42366	606.11749	1.102301	0.377497	0.999899	0.800345	3069.7988	33.585712	0.195486	0.194897	5.289454	3638.7617	0.247519	2.4E-05	93961.742	5.7168918	3073.6699	14.247102	440.32779	3.528945	5.4771829	1.9
821	2024-11-21 00:00:00	98504.727	3361.054	1.001054	257.06589	622.82635	1.250352	0.387617	0.999947	0.820474	3360.1716	35.824387	0.198468	0.198644	5.51301	3988.0476	0.263254	2.5E-05	98175.07	5.9378071	3366.1919	14.90804	486.06485	3.634835	5.7984419	2.1
822	2024-11-22 00:00:00	98997.664	3331.6008	1.001121	256.51868	633.51343	1.4690191	0.412924	1.000025	1.0104229	3327.9431	43.089741	0.204504	0.204801	5.448277	3935.0225	0.341066	2.5E-05	98726.211	6.6626229	3321.7271	16.531857	490.28348	3.5660319	6.13903	2.1
823	2024-11-23 00:00:00	97777.281	3396.2234	1.000926	255.17851	651.19727	1.469008	0.430012	0.999955	1.066237	3396.6807	41.489986	0.21204	0.212171	6.3118858	4031.6401	0.515391	2.6E-05	97475.383	8.5083008	3401.2561	17.385506	509.94406	3.438525	6.1948528	2.1
824	2024-11-24 00:00:00	98013.82	3363.6599	1.000863	252.91953	660.31757	1.432411	0.429882	0.999812	1.023106	3360.26	42.043064	0.208009	0.208833	6.1451898	3982.3799	0.537421	2.6E-05	97651.102	8.8135347	3360.9956	17.944859	514.70544	3.390713	6.881176	2
825	2024-11-25 00:00:00	93102.297	3413.5439	0.999989	234.44661	636.63971	1.415423	0.39334	0.999985	0.945689	3418.6648	41.396076	0.196188	0.196054	6.0885758	4065.2925	0.483133	2.5E-05	93235.125	8.244422	3433.5239	17.387096	491.65717	3.1937799	6.4592361	1.9
826	2024-11-26 00:00:00	91985.32	3326.5173	0.999654	230.9782	613.58728	1.401323	0.387618	1.000092	0.961232	3325.9041	42.77906	0.193947	0.194368	6.142169	3943.3108	0.439764	2.4E-05	91786.117	8.0781279	3325.2673	17.338844	492.55154	3.4756429	6.5376592	1.8

Combined_Prices

자료수집 및 방법론 설명

목표: Factor loading으로 도출된 대표 코인 가격 데이터(5년 동안의 일간 데이터)

방법:

- Yahoo Finance API를 사용하여 지정한 티커(주식 또는 가상자산 등)의 금융 상세 데이터 확보
- 전체 데이터를 Excel 파일로 저장하는 Python 스크립트 작성

코드 구성 및 설명:

- yfinance 라이브러리를 활용하여 지정한 티커 리스트에 대해 금융 데이터를 수집
- 기간 설정 (5년)
- 가격 데이터 추가 (Open, High, Low, Close, Volume) + 환율
- 티커 목록 읽기
- 결과 저장 (Excel 파일)

최종 수집 데이터:

- 대표 코인의 상세 가격 데이터

	A	B	C	D	E	F
1	Date	Open	High	Low	Close	Volume
2	2019-12-07 00:00:00	0.265327007	0.273310989	0.265325993	0.270853013	275068
3	2019-12-08 00:00:00	0.270853013	0.272062987	0.26374799	0.267208993	556684
4	2019-12-09 00:00:00	0.267208993	0.269558996	0.258938015	0.262854993	500125
5	2019-12-10 00:00:00	0.262847006	0.262867004	0.253048986	0.255160004	348055
6	2019-12-11 00:00:00	0.255176008	0.262962997	0.251807988	0.262255996	372026
7	2019-12-12 00:00:00	0.262975991	0.268202007	0.257182002	0.265677989	284307
8	2019-12-13 00:00:00	0.265670002	0.276652008	0.264353991	0.271052003	554766
9	2019-12-14 00:00:00	0.271066993	0.274069011	0.260605007	0.265673995	380886

LUNC-USD | DASH-USD | BSV-USD | KRW-Exchange-Rate

자료수집 및 방법론 설명

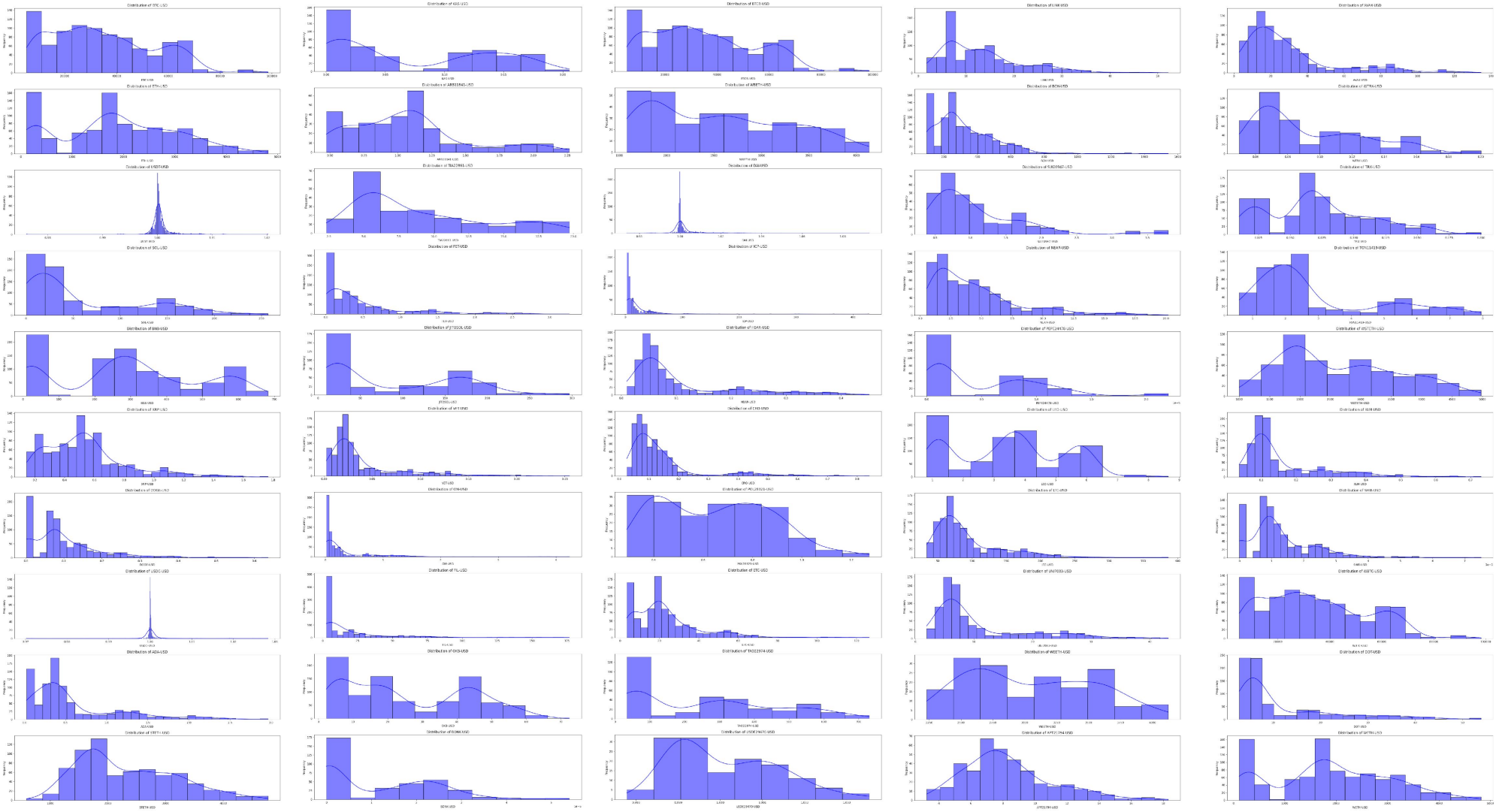
데이터 탐색 및 전체 데이터 분포 시각화

```
=== 데이터 불러오기 완료 ===

=== 데이터 구조 확인 ===
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1825 entries, 0 to 1824
Columns: 456 entries, Date to ACA-USD
dtypes: datetime64[ns](1), float64(455)
memory usage: 6.3 MB
None

=== 데이터 기초 통계량 ===
                                Date       BTC-USD  ...  COREUM-USD      ACA-USD
count                           1825   1825.000000  ...  515.000000  1122.000000
mean   2022-05-29 00:00:00.000000256  34609.428958  ...    0.110308     0.069904
min              2019-11-29 00:00:00   4970.788086  ...    0.056149     0.000142
25%              2021-02-27 00:00:00  19172.468750  ...    0.078659     0.041411
50%              2022-05-29 00:00:00  30432.546875  ...    0.095181     0.062503
75%              2023-08-28 00:00:00  48896.722656  ...    0.132619     0.105394
max              2024-11-26 00:00:00  98997.664062  ...    0.298278     0.208641
std                              NaN  19636.163172  ...    0.045084     0.048183

[8 rows x 456 columns]

=== 데이터 타입별 변수 분포 ===
float64           455
datetime64[ns]      1
Name: count, dtype: int64

=== 결측값 확인 ===
SOL-USD          133
STETH-USD        390
AVAX-USD         296
WTRX-USD         827
TON11419-USD     637
```

```
=== 데이터 크기 ===
Rows: 1825, Columns: 456

샘플링된 데이터 크기: Rows: 912, Columns: 456

=== 모든 변수의 그래프를 페이지 단위로 저장 ===
'C:\Users\INNOGRID\Downloads\all_distributions_page_1.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_2.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_3.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_4.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_5.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_6.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_7.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_8.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_9.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_10.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_11.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_12.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_13.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_14.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_15.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_16.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_17.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_18.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_19.png' 저장 완료.
'C:\Users\INNOGRID\Downloads\all_distributions_page_20.png' 저장 완료.
```

자료수집 및 방법론 설명

각 변수의 분포를 히스토그램으로 표시 (각 변수 데이터 50% 샘플링)



탐색적 데이터 분석 : PCA와 K-Means

[1] 분석 목적

- 암호화폐 데이터의 상관 관계를 분석.
- 주요 암호화폐 그룹을 정의하고 대표 코인 도출.
  - *대표 코인은 각 그룹의 일반적 특성을 반영하므로, 클러스터 전체를 분석하는것 보다 대표 코인을 분석하는 것이 더 효율적*

[2] 데이터 전처리

- **상관 행렬 생성**
  - 암호 화폐 간의 간격 변동 상관 관계를 계산하여 상관 행렬 생성(PCA 및 클러스터링의 입력에 필요)
- **스케일링 확인**
  - 상관 행렬 분석이기에, 데이터가 표준화된 상태(0을 중심으로 스케일링)인지 확인
- **암호 화폐간 상관 관계**
  - 특정 암호화폐간에 매우 높은 상관 관계를 가짐, 이는 유사한 시장 움직임을 보인다는것을 의미
  - 낮거나 음의 상관 관계를 가지는 암호화폐는 시장에서 서로 독립으로 움직이는 경향을 가짐

탐색적 데이터 분석 : PCA와 K-Means

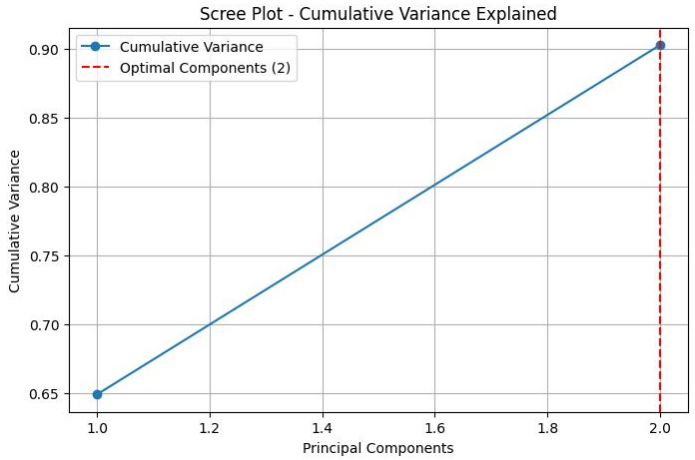
[1] PCA Analysis

Scree Plot 분석

- Scree Plot은 누적 분산 비율을 보여줌.
- **PC1과 PC2가 전체 분산의 약 73%를 설명**하며, 주요 정보를 충분히 보존.
- PC1과 PC2를 기준으로 클러스터링에 활용할 차원 축소 데이터를 구성.

Explained Variance Table

- **PC1: 57.87%**, **PC2: 15.43%** → 총 **73.30%**의 분산 설명.
- PC3부터는 기여도가 급격히 감소하여, 주성분으로 선택할 필요가 없다고 판단.

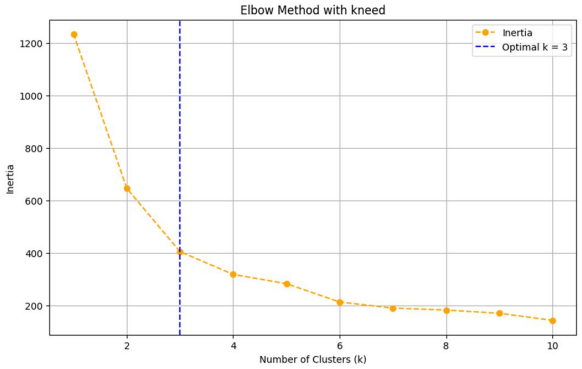


Principal C	Explained Variance Ratio
PC1	0.578724646
PC2	0.154392361
PC3	0.069895051
PC4	0.029694273
PC5	0.020768712
PC6	0.017077615
PC7	0.01444146
PC8	0.012566389
PC9	0.010669291
PC10	0.007577396
PC11	0.007124616
PC12	0.006837541
PC13	0.005641921
PC14	0.00525983
PC15	0.004990838

탐색적 데이터 분석 : PCA와 K-Means

[1] K-Means

- **Elbow Method**
  - 관성(Inertia) 값을 기반으로 최적의 클러스터 수(K)를 결정.
  - Elbow Point: 관성 값의 감소율이 완만해지는 지점.
  - 최적의 K 값으로 **3**이 선정됨.




```
Inertia for each K:
- K=1: 1233.33
- K=2: 645.84
- K=3: 404.57  ← 최적 K
- K=4: 317.79
- K=5: 282.49
- K=6: 212.71
- K=7: 189.47
- K=8: 182.18
- K=9: 169.67
- K=10: 142.87
```

- **클러스터링 수행**
  - PCA로 차원 축소된 데이터를 기반으로 K-Means 클러스터링 수행.
  - 최적 K 값에 따라 데이터를 3개의 클러스터로 그룹화.

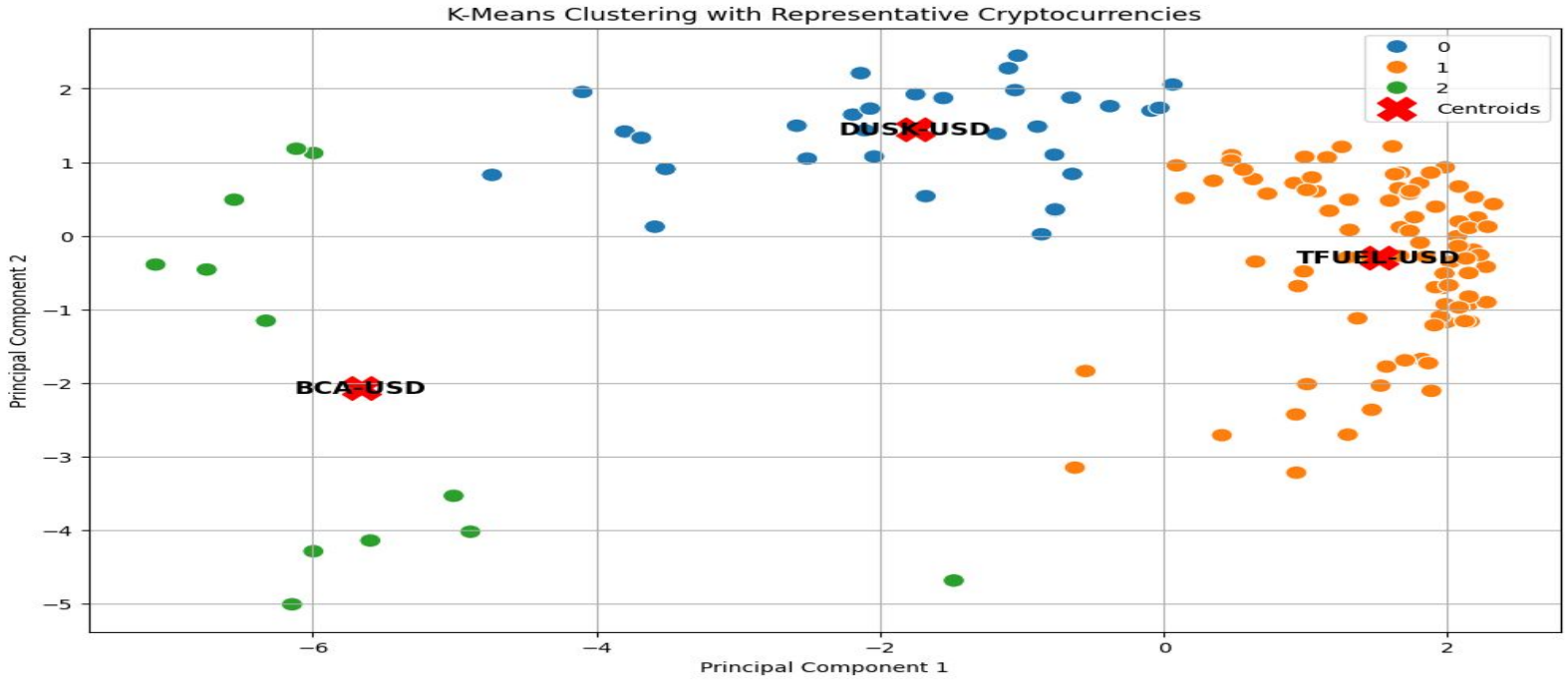
탐색적 데이터 분석 : PCA와 K-Means

[1] K-Means

- **클러스터링 결과 (K-Means)**
  각 클러스터는 암호화폐의 유사한 특성을 기반으로 그룹화 진행.

- **대표 암호화폐 선정**:
  클러스터 0: DUSK-USD
  클러스터 1: TFUEL-USD
  클러스터 2: BCA-USD

- **대표 암호화폐 선정 기준**:
  각 클러스터의 중심점(Centroid)에 가장 가까운 암호화폐를 대표로 선정.



탐색적 데이터 분석 : PCA와 K-Means

[1] Factor Loading

- **클러스터링 결과 (Factor Loading)**

  PC1에 주요 정보가 포함되어 있고 대표 암호 화폐만 선별하는 과정이라 PC2를 제외

  - 클러스터 0: LUNC-USD (PC1: 0.053541)
  - 클러스터 1: DASH-USD (PC1: 0.134100)
  - 클러스터 2: BSV-USD (PC1: 0.095326)

추가 설명

- PC1 값이 가장 큰 암호화폐는 해당 클러스터의 **특성을 가장 잘 반영**하는 암호화폐로 선정.
- Factor Loading 분석은 **클러스터의 대표성을 직관적으로 해석**할 수 있도록 도움.

Cluste	Crypto	PC1	Cluster	Crypto	PC1	Cluste	Crypt	PC1
0	LUNC-USD	-0.053541497	1	DASH-USD	-0.134099831	2	BSV-USD	-0.09532619
0	LRC-USD	-0.050789544	1	HT-USD	-0.131381484	2	TUSD-USI	-0.01222926
0	MANA-USD	-0.048784117	1	DCR-USD	-0.128970546	2	DAI-USD	-0.00170307
0	CRO-USD	-0.048190623	1	EOS-USD	-0.128516133	2	USDP-USI	0.00303719
0	STPT-USD	-0.046995257	1	FIL-USD	-0.127130395	2	USDT-USI	0.005799193
0	POWR-USD	-0.043389557	1	DGB-USD	-0.12709482	2	USDC-USI	0.009682136
0	XYO-USD	-0.041269432	1	BTT-USD	-0.126213465	2	FET-USD	0.052764973
0	GLM-USD	-0.038396094	1	ICX-USD	-0.126199052	2	BCA-USD	0.056163461
0	CHR-USD	-0.036605478	1	ONT-USD	-0.12610178	2	LEO-USD	0.067468477
0	MATIC-USD	-0.034239141	1	SXP-USD	-0.12573641	2	OKB-USD	0.071752283
0	BORA-USD	-0.03284224	1	KAVA-USD	-0.125087189	2	ABT-USD	0.07565219
0	FTM-USD	-0.029846186	1	XVG-USD	-0.124963398	2	MX-USD	0.076820493
0	IOTX-USD	-0.028492078	1	XTZ-USD	-0.123809936	2	PAXG-USI	0.082570803
0	KCS-USD	-0.024033451	1	ZEC-USD	-0.123371504			

Random Forest 대표 코인 가격 예측

[1] 목표 설정

- 암호화폐(BSV, LUNC, DASH)의 과거 데이터를 바탕으로 가격 예측 모델을 구축하고 미래의 가격 변동성을 예측

[2] 학습을 위한 데이터 전처리

- **데이터 선택**

  =>대표 암호화폐 2019년 1월부터 2024년까지의 암호화폐 데이터를 수집.

- **결측치 처리**

  =>데이터의 결측치를 **선형 보간법(Interpolate)**으로 채움.

- **기술 지표 추가**

  RSI (Relative Strength Index):시장의 과매수/과매도 상태를 판단을 위한 기술 지표.

  MACD:**추세 강도**와 방향성을 나타내는 지표로,
  장단기 이동평균선 차이를 기반으로 계산됨.

  Bollinger Bands:변동성을 **표준 편차**로 측정해 상단/중단/하단 밴드를 제공.
  가격 변동 범위를 시각화함.

  Price Range: 하루 High와 Low의 차이로 가격 변동성을 나타냄.

  환율(Exchange Rate):Close_KRW / Close로 계산하여 환율 변동성을 반영.

과거 데이터만으로는 사장의 추세나 패턴을 정확히 파악하기 힘듬. 기술 지표는 이러한 점을 고려하여 수학적으로 계산하여 추세와 시장 상태를 더 잘 이해할 수 있도록 해줌.

```python
# 기술 지표 계산 함수 정의
def compute_RSI(data, window):
    delta = data.diff()
    gain = (delta.where(delta > 0, 0)).rolling(window=window).mean()
    loss = (-delta.where(delta < 0, 0)).rolling(window=window).mean()
    RS = gain / loss
    RSI = 100 - (100 / (1 + RS))
    return RSI

def compute_MACD(data, short_window=12, long_window=26, signal_window=9):
    short_ema = data.ewm(span=short_window, adjust=False).mean()
    long_ema = data.ewm(span=long_window, adjust=False).mean()
    MACD = short_ema - long_ema
    signal = MACD.ewm(span=signal_window, adjust=False).mean()
    return MACD, signal, MACD - signal

def compute_BB(data, window=20, num_sd=2):
    rolling_mean = data.rolling(window=window).mean()
    rolling_std = data.rolling(window=window).std()
    upper_band = rolling_mean + (rolling_std * num_sd)
    lower_band = rolling_mean - (rolling_std * num_sd)
    return upper_band, rolling_mean, lower_band
```

Random Forest 대표 코인 가격 예측

[3] Random Forest 모델 학습

독립 변수 설정:

- Open, High, Low, Volume, Price_Range, RSI, MACD, Bollinger Bands, Exchange Rate, Close_KRW

종속 변수 설정:

- Close

데이터 정규화:

- **MinMaxScaler**를 사용해 독립 변수를 0~1 범위로 정규화.

하이퍼파라미터 최적화:

- **GridSearchCV**를 통해 최적의 하이퍼파라미터 탐색:
  - `n_estimators`: [100, 200, 300]
  - `max_depth`: [10, 15, 20]
  - `min_samples_split`: [2, 5, 10]

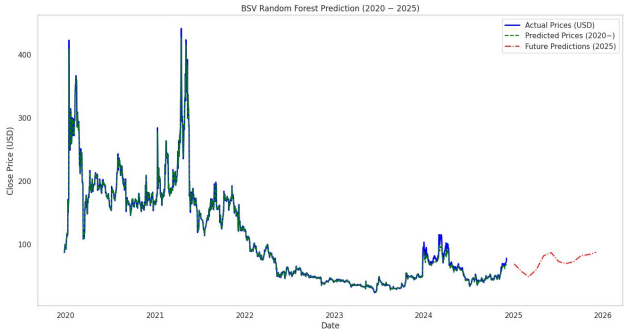
하이퍼파라미터	설명
n_estimators	트리의 개수를 조정 (100, 200, 300)
max_depth	트리의 최대 깊이 설정 (10, 15, 20)
min_samples_split	노드를 분할하기 위한 최소 샘플 수 (2, 5, 10)

성능 평가 지표:

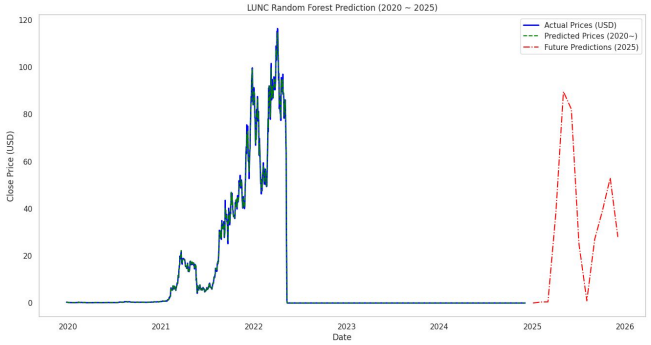
- **MAE (Mean Absolute Error)**
- **MSE (Mean Squared Error)**
- **RMSE (Root Mean Squared Error)**
- **R^2 (설명력)**

Random Forest 대표 코인 가격 예측

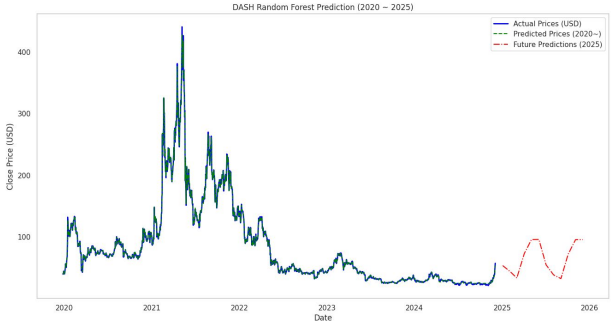
Result



BSV 평가 지표:
- MAE: 3.0566
- MSE: 22.8888
- RMSE: 4.7842
- R²: 0.9351



LUNC 평가 지표:
- MAE: 0.0000
- MSE: 0.0000
- RMSE: 0.0000
- R²: 0.9466



DASH 평가 지표:
- MAE: 0.7111
- MSE: 1.0053
- RMSE: 1.0027
- R²: 0.9635

Key Findings

BSV

- **R² 값이 0.9351**로 높은 설명력을 보이며, 모델이 코인의 데이터 패턴을 잘 학습하고 있음을 나타냄.
- **RMSE: 4.7842**로 예측값과 실제값 간의 오차가 다소 존재하지만, 비교적 안정적인 예측 성능을 보임.
- 모델의 성능을 더욱 개선하기 위해 **시계열 특성**이나 **외부 요인**(거래량 급등, 시장 뉴스 등)의 추가 반영이 필요할 수 있음.

LUNC

- **R² 값이 0.9466**로 높은 설명력을 보여주지만, **결측치**로 인해 일부 기간에 대한 데이터 손실이 심각하게 발생함.
- **MAE와 MSE 값이 0.0000**인 이유는 예측값이 실제값과 동일하거나, 데이터가 **불균형적**이기 때문임.
- 정확한 예측을 위해 **데이터 품질 개선**과 **데이터 재수집**이 필요함. 특히 **이상치**와 **결측치 처리**가 중요함.

DASH

- **R² 값이 0.9635**로 높은 설명력을 보이며, 모델이 **가격 변동 패턴**을 효과적으로 학습했음을 나타냄.
- **RMSE: 1.0027**로 비교적 낮은 오차를 보이며, 예측 안정성이 뛰어남.
- 일부 구간에서 모델이 과도하게 학습(과적합)된 가능성이 있으며, **추가 검증 데이터**를 활용해 모델의 일반화 성능을 개선할 수 있음.

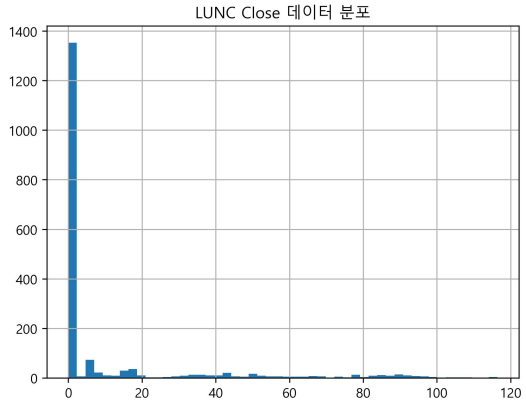
LSTM 모델 기반 대표 코인 가격 예측

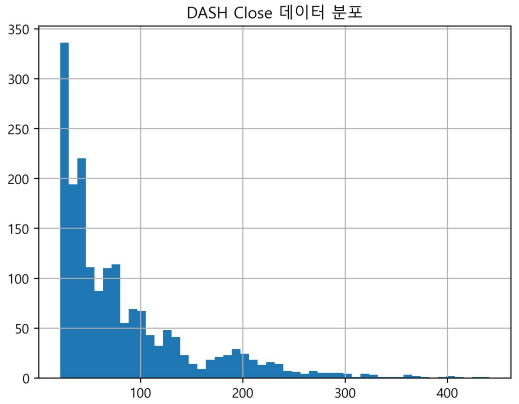
[1] 목표 설정

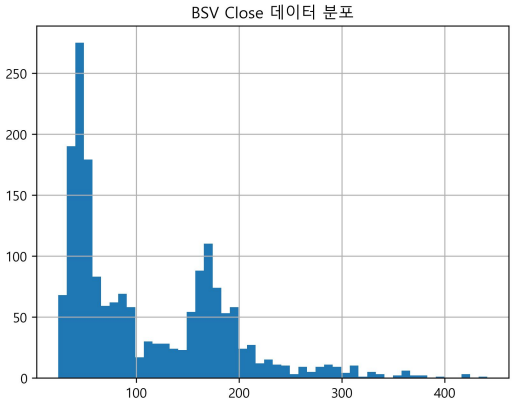
- 군집별 학습을 통해 각 군집의 대표 코인 가격 예측

[2] 학습을 위한 데이터 전처리

- **데이터 선택**
  : 각 군집의 Close 데이터, 대표 코인의 Open, Close, Low, High, Volume, 환율 Open, Close, Low, High
- **결측치 처리**
  : 코인 데이터는 1일간격, 환율 데이터는 null 값을 선형 보간 처리
- **로그 변환**
  : 값의 스케일 차이 (지나치게 0에 수렴 하는 대부분의 데이터 처리를 위함) 해결을 위함



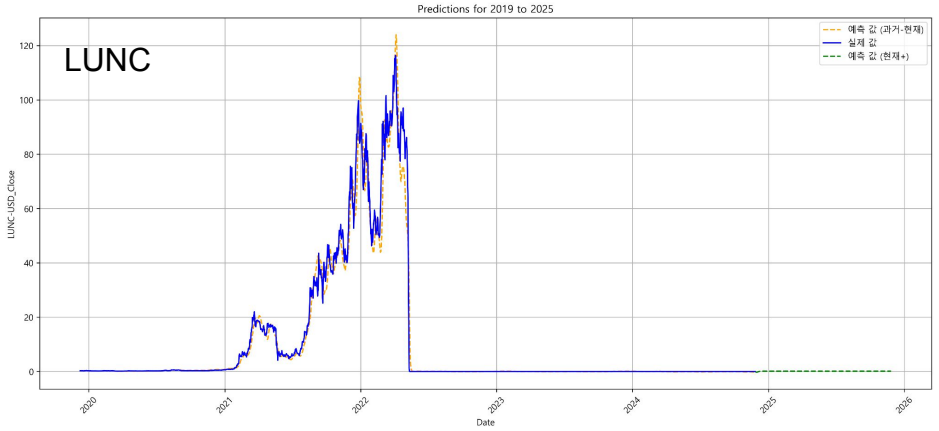


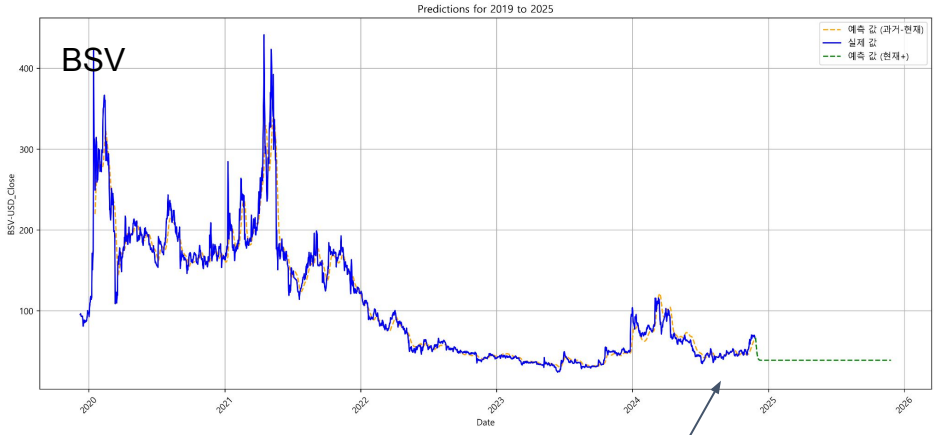


- **정규화**
  : 클러스터 내 코인 데이터의 트렌드를 이용해서 대표 코인의 종가를 예측하기 위함이므로,
  각 열을 (0,1) 범위로 스케일링하여 변화율을 반영하기 위함

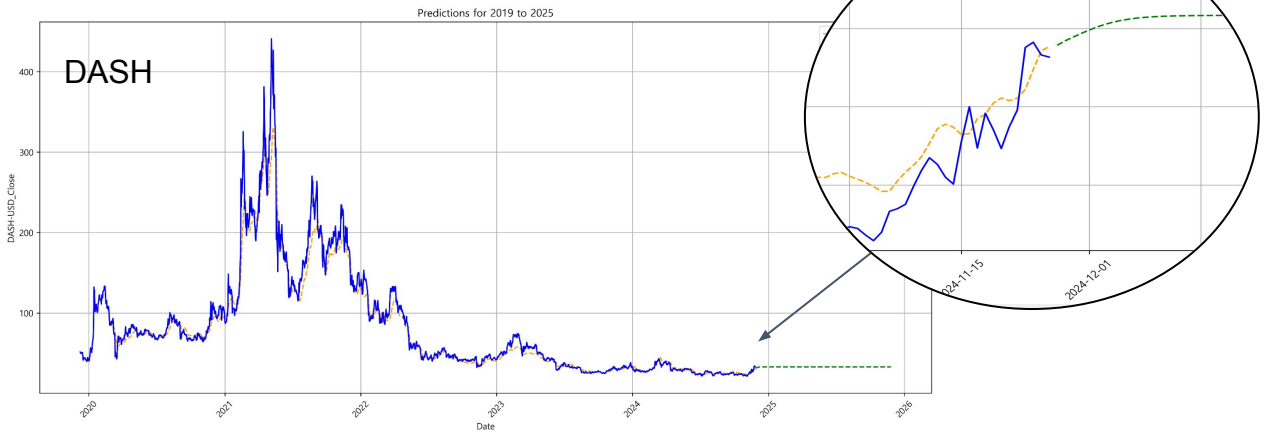
LSTM 모델 기반 대표 코인 가격 예측

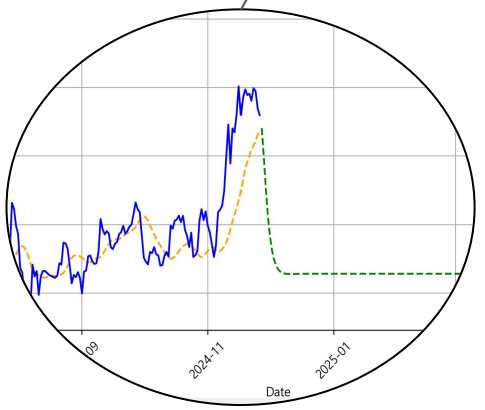
Result





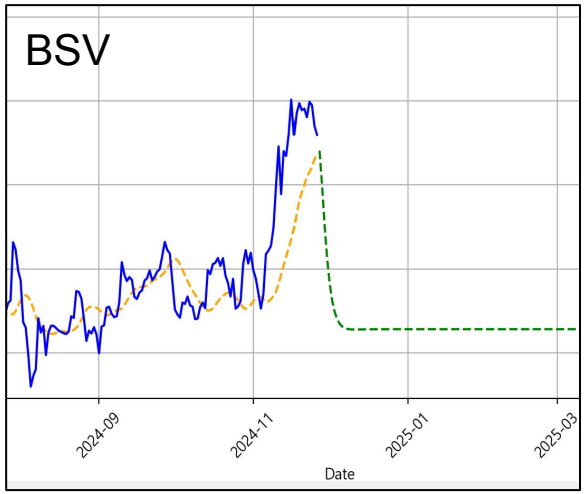
LUNC 코인은 사실상 죽은 화폐로 예측이 무의미한 것으로 보임





LSTM 모델 기반 대표 코인 가격 예측

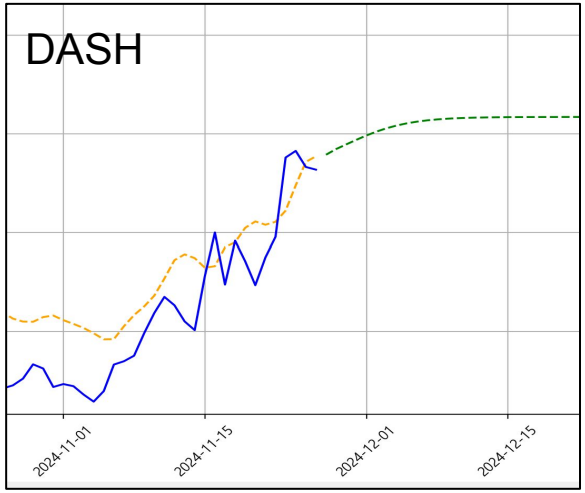
Result



최종 Test 평가 지표

Normalized Results:
- MSE: 0.0015
- RMSE: 0.0386
- MAE: 0.0299

Original Scale Results:
- MSE: 44.9174
- RMSE: 6.7020
- MAE: 4.8641
- R^2: 0.6689
- MAPE: 8.89%



최종 Test 평가 지표

Normalized Results:
- MSE: 0.0002
- RMSE: 0.0152
- MAE: 0.0105

Original Scale Results:
- MSE: 1.6824
- RMSE: 1.2971
- MAE: 0.8449
- R^2: 0.7329
- MAPE: 3.24%

Key Findings

Test MSE (Mean Squared Error)
모델의 예측값과 실제값의 평균 제곱 오차
0에 가까울 수록 실제 값과 같음을 의미한다.

Test MAE (Mean Absolute Error)
모델의 예측값과 실제값의 평균 절대 오차

모든 데이터를 로그변환 및 정규화 한 상태로 학습하였으며,
평가 시에는 학습시킨 스케일에 대해 지표를 계산하고 (Normalized),
저장해둔 스케일을 이용해 역정규화, 역로그변환하여
기존 스케일에 대해 지표를 계산했다. (Original Sacle)

Test R^2 (결정 계수)
모델이 데이터를 얼마나 잘 나타내는가를 의미
$R^2 = 1$: 모델이 완벽하게 예측.
$R^2 = 0$: 모델이 평균값만 예측.
$R^2 < 0$: 모델이 데이터를 잘못 설명.
모델이 테스트 데이터 변동의 약 **n%를 설명**한다.

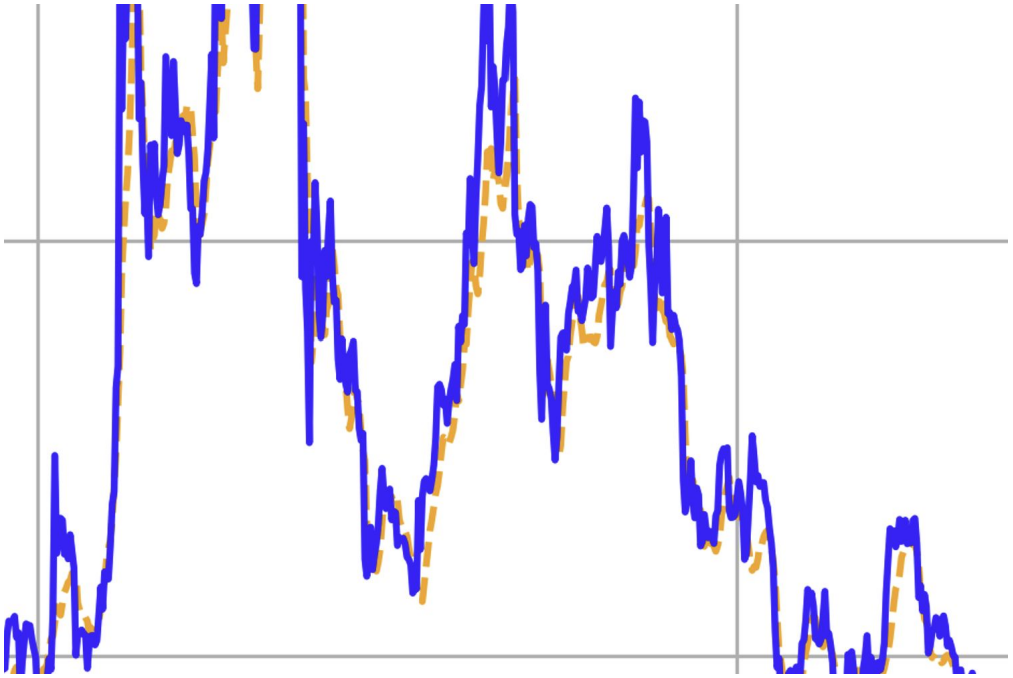
Test MAPE (Mean Absolute Percentage Error)
예측값과 실제값의 상대적 오차 백분율
예측값이 실제값과 대비하여 평균적으로n% 의 오차가 있다.

하이퍼파라미터를 조정해보며 얻은 결과이다.
LSTM은 BSV가 대표하는 군집의 코인 가격보다 DASH가 대표하는 군집의 코인 가격을 비교적 잘 예측하는 것으로 보인다.
또한 둘 다 장기적인 예측 성능은 떨어지는 것으로 보인다.

LSTM 모델 기반 대표 코인 가격 예측

Result

예측 결과를 통한 최종 분석



패턴을 예측했는가?

No. "전날의 데이터와 오늘의 데이터가 같다."

RandomWalk
이전 상태에 기반하지만 완전히 무작위로 움직이는 데이터

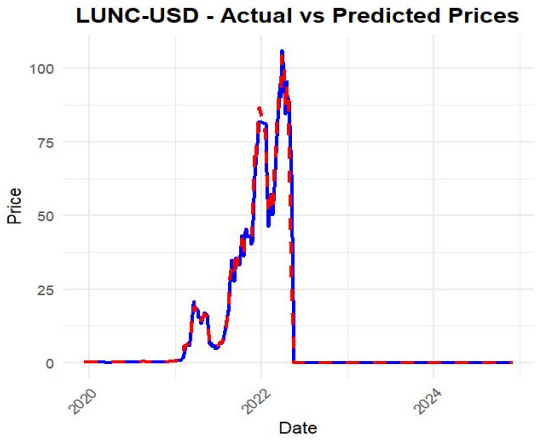
랜덤워크를 판단하는 방법
정상성 검사
: ADF (Augmented Dickey-Fuller)
: p-value > 0.05 이면 랜덤워크일 가능성이 크다.

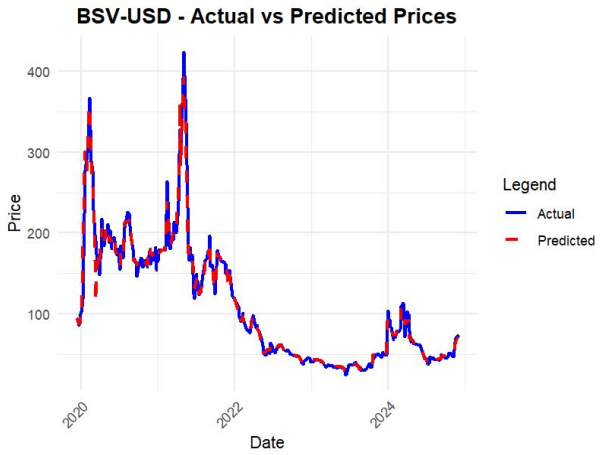
원본 DASH 데이터:
p-value: 0.24789181176065345
로그변환 및 정규화 후 DASH 데이터:
p-value: 0.5079344850129396

암호화폐의 가격은 **패턴이 없는 무작위성을 띄기 때문에 LSTM 모델을 통해 미래의 가격을 예측하기 어렵고**,이를 통해 같은 군집 내 코인들의 가격을 예측하기는 어려울 것으로 보인다.

XGBoost 알고리즘 기반

Result






```
> # 결과 출력: 상위 5개 값 + RMSE
> for (coin in names(detailed_results)) {
+   result <- detailed_results[[coin]]
+   cat("\nResults for", coin, ":\n")
+   cat("RMSE:", round(result$RMSE, 2), "\n")
+   top_5 <- head(data.frame(Date = result$Da
$Actual), 5)
+   print(top_5)
+ }
```

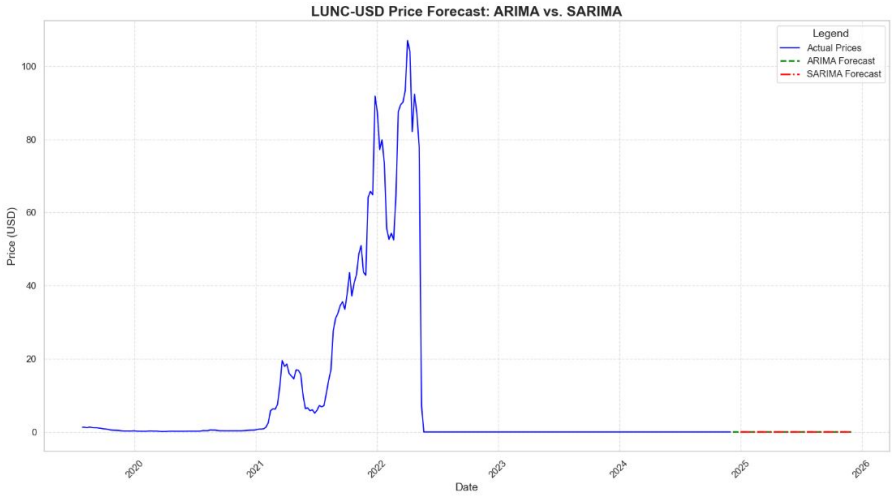
Key Findings

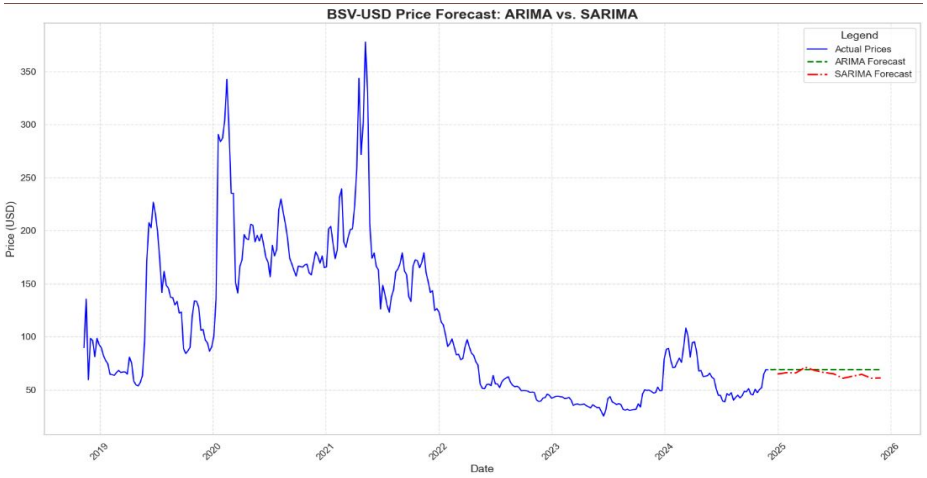
```
Results for LUNC-USD :
RMSE: 1.99
        Date Predicted   Actual
1 2019-12-11 0.2530410 0.262256
2 2019-12-13 0.2530410 0.271052
3 2019-12-21 0.2502207 0.258650
4 2019-12-26 0.2727849 0.352461
5 2019-12-29 0.2670636 0.310099

Results for DASH-USD :
RMSE: 6.15
        Date Predicted   Actual
1 2019-12-11  50.35869 49.75650
2 2019-12-13  50.35869 50.71099
3 2019-12-21  43.82304 43.04770
4 2019-12-26  42.07561 40.16201
5 2019-12-29  42.07561 44.75007

Results for BSV-USD :
RMSE: 7.89
        Date Predicted    Actual
1 2019-12-11  93.22430  94.64265
2 2019-12-13  93.22430  93.37571
3 2019-12-21  87.71725  85.33881
4 2019-12-26  90.59963  87.62069
5 2019-12-29  96.54996 100.35604
```

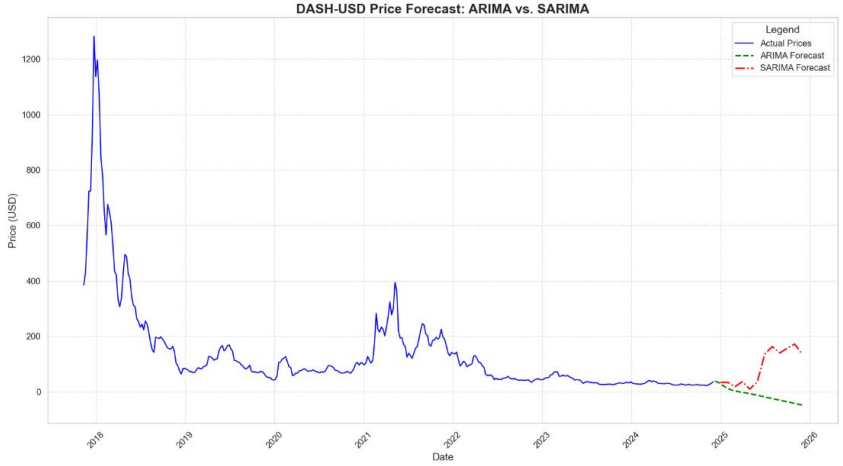
- ✔ LUNC-USD : RMSE 1.99
- ✔ DASH-USD : RMSE 6.15
- ✔ BSV-USD : RMSE 7.89

전통적 시계열 모델 예측 기반

Result








```
LUNC-USD - ARIMA RMSE: 24.88, SARIMA RMSE: 24.24
DASH-USD - ARIMA RMSE: 391.85, SARIMA RMSE: 191.18
BSV-USD - ARIMA RMSE: 80.81, SARIMA RMSE: 85.89
```

- ✔ **LUNC-USD**: ARIMA와 SARIMA 모두 유사한 성능을 보임
- ✔ **DASH-USD**: SARIMA 모델이 더 나은 성능을 보여, 계절성을 반영하는 SARIMA 모델이 적합
- ✔ **BSV-USD**: ARIMA 모델이 더 낮은 RMSE를 보이며, 단순한 ARIMA 모델을 사용하는 것이 더 효율적

결론

Model Comparison

Model	LUNC	BSV	DASH
Random Forest	0.0000	4.7842	1.0027
LSTM	-	6.7020	1.2971
XGBoost	1.99	7.89	6.15
ARIMA	24.88	80.81	191.18
SARIMA	24.24	85.89	391.85

Model Comparison

Random Forest, LSTM, XGBoost, 전통 시계열 (ARIMA, SARIMA) 모델 비교 결과,

-> **Random Forest가 더 낮은 RMSE 값을 보여 성능이 뛰어난 것처럼 보이지만**, 시계열 데이터를 기준으로 봤을때는 **LSTM** 기법이 시계열 특성을 더 잘 반영하여 장기적인 예측에 유리함을 확인할 수 있었습니다.

- RMSE : 실제 값과 모델이 예측한 값 간의 오차를 측정하며, RMSE 값이 작을수록 예측 성능이 좋은 모델임을 나타냄

자료 수집의 한계

- 비정형 데이터의 처리 부족:
  가격 데이터와 온체인 데이터 외에도 뉴스, 소셜 미디어, 투자 심리 등 비정형 데이터를 활용하지 못했습니다.
  이러한 비정형 데이터는 가상자산 가격에 중요한 영향을 미칠 수 있지만, 현재 방법론에서는 다루지 않았습니다.

느낀 점

(전명준) 가상자산 투자를 하고 있는 입장에서 방대한 코인 가격 데이터와 온체인 데이터 수집을 하는 방법을 익히고 빅데이터 학습모델을 적용하여 한계점과 개선할 점을 팀원들과 함께 고민 할 수 있는 유익한 프로젝트였습니다.

(유지연) 처음 암호화폐를 접한 입장에서, 관련 논문을 읽으며 암호화폐가 무엇인지 알 수 있었고 데이터 수집 방법을 알게 되어 흥미로웠다. LSTM 모델을 사용하여 모델을 학습하는 과정에서, 수집한 데이터 안에서도 학습에 쓸 데이터를 선택하고 정리하는 방식이 정말 다양함을 배웠고, 데이터에 대해 더 배우고 싶어진 유익한 시간이었다.

(설현일) 이번 프로젝트를 통해 암호화폐와 데이터 분석 알고리즘을 처음 접했으며, 데이터를 분석하고 예측하는 과정은 새로운 도전이었습니다. 생소한 분야라 어려움도 있었지만, 교수님의 명확한 피드백 덕분에 한계점을 파악하고 개선 방안을 고민하며 프로젝트를 성공적으로 마무리할 수 있어 보람된 경험이었습니다.

(김수민) 이번 프로젝트를 통해 다양한 머신러닝 기법을 활용하여 데이터 분석을 실질적으로 해볼 수 있어서 흥미로웠습니다. 교수님께서 몇차례에 걸쳐 멘토링을 해주셔서 방향성을 설정할 수 있었고 팀원들과 모델의 한계점과 개선 방안을 논의하며 많은 것을 배울 수 있었습니다.

참고문헌

조풍진, 이민혁, 송재욱 (2022). 한국 주식시장에서의 군집화 기반 페어트레이딩 포트폴리오 투자 연구, *Journal of Korean Society of Industrial and Systems Engineering*, 45(3), 123-130. https://doi.org/10.11627/jksie.2022.45.3.123:contentReference[oaicite:0]{index=0}

박재현, 서영석 (2022) 암호화폐 종가 예측 성능과 입력 변수 간의 연관성 분석, *KIPS Transactions on Software and Data Engineering*, 11(1), 19-28. https://doi.org/10.3745/KTSDE.2022.11.1.19:contentReference[oaicite:1]{index=1}

김준호, 성한울 (2022) 비트코인 가격 예측을 위한 LSTM 모델의 Hyper-parameter 최적화 연구, A Study on the Hyper-parameter Optimization of Bitcoin Price Prediction LSTM Model -Journal of the Korea Convergence Society | Korea Science

정화민 교수님 2024-2 빅데이터 예측분석 강의자료

감사합니다!

- 소속 : 서강대학교 AI & SW 대학원
- 전공 : 데이터사이언스 & AI 전공
- 과목 : 빅데이터 분석 예측 / 지도교수 정화민 교수님
- 5팀 : 설현일(팀장), 전명준, 유지연, 김수민