2. 연구 동기

주식 시장과 마찬가지로, 비트코인 역시 (뉴스, 시장 심리, 글로벌 거시경제 변수)등에 의해 영향을 받는다. 우리는 이 논문에서 왜 비트코인이 linearity를 가정하면 안 되는지에 대해 서술하고자 한다.

linear regression을 가정하기 위해서, 우리는 기본적으로 5가지 가정을 한다.

1. (선형성) 2.(다중공선성) 3.(외생성) 4.(자기상관성) 5.(등분산성) 주가 데이터가 과연 이 5가지 조건을 엄밀히 만족하는지부터 살펴보자.

3. Data

Variable	Mean	Std	Skewness	Kurtosis	ADF Statistic	p-value	Formula
PriceUSD	0.0022	0.0406	-0.1060	6.2599	-32.8852	0.0000	$\Delta PriceUSD_t = \frac{PriceUSD_t - PriceUSD_{t-1}}{PriceUSD_{t-1}}$
BkCount	147.8582	16.5530	-0.3723	1.1361	-6.4035	0.0000	$BkCount_t$
BkSize	0.0101	0.1437	1.1260	4.2925	-6.6801	0.0000	$\Delta BkSize_t = \frac{BkSize_t - BkSize_{t-1}}{BkSize_{t-1}}$
DiffM	0.0023	0.0171	1.1096	42.8772	-7.9966	0.0000	$\Delta DiffM_t = \frac{DiffM_t - DiffM_{t-1}}{DiffM_{t-1}}$
FeeUSD	3.8228	6.6907	4.0090	20.2875	-3.3593	0.0124	$FeeUSD_t$
Hash_Rate	0.0097	0.1261	0.7776	2.5203	-26.0803	0.0000	$\Delta Hash\_Rate_t = \frac{Hash\_Rate_t - Hash\_Rate_{t-1}}{Hash\_Rate_{t-1}}$
RevUSD	0.0090	0.1294	0.8198	2.8512	-37.6867	0.0000	$\Delta RevUSD_t = \frac{RevUSD_t - RevUSD_{t-1}}{RevUSD_{t-1}}$

Variable	Mean	Std	Skewness	Kurtosis	ADF Statistic	p-value	Formula
TxValAdjUSD	0.0660	0.4274	2.8765	17.7569	-6.8575	0.0000	$\Delta TxValAdjUSD_t = \frac{TxValAdjUSD_t - TxValAdjUSD_{t-1}}{TxValAdjUSD_{t-1}}$
URTH	0.0003	0.0099	-0.9888	24.6577	-9.4896	0.0000	$\Delta URTH_t = \frac{URTH_t - URTH_{t-1}}{URTH_{t-1}}$
GSPC	0.0003	0.0105	-0.6172	21.9839	-9.7565	0.0000	$\Delta GSPC_t = \frac{GSPC_t - GSPC_{t-1}}{GSPC_{t-1}}$
RiskFree	0.0019	0.0275	1.9750	27.2993	-6.2836	0.0000	$\Delta RiskFree_t = RiskFree_t - RiskFree_{t-1}$
OAS	4.1297	1.0263	2.5562	8.7330	-3.4424	0.0096	OAS_t
TenYear	0.0006	0.0286	1.6946	44.9141	-8.0629	0.0000	$\Delta TenYear_t = \frac{TenYear_t - TenYear_{t-1}}{TenYear_{t-1}}$
TwoYear	0.0017	0.0472	0.9434	13.1693	-24.9820	0.0000	$\Delta TwoYear_t = \frac{TwoYear_t - TwoYear_{t-1}}{TwoYear_{t-1}}$
VIX	19.5261	8.4034	2.0822	8.4596	-4.1856	0.0007	VIX_t
USDIndex	0.0000	0.0026	0.2479	5.3907	-10.5806	0.0000	$\Delta USDIndex_t = \frac{USDIndex_t - USDIndex_{t-1}}{USDIndex_{t-1}}$
ExpInflation	0.0003	0.0198	5.7912	210.3612	-12.2697	0.0000	$\Delta ExpInflation_t = \frac{ExpInflation_t - ExpInflation_{t-1}}{ExpInflation_{t-1}}$
USNewsSent	-0.0432	0.2005	-0.8955	1.1627	-2.8680	0.0492	$USNewsSent_t$

4. Methodology

이 연구는 Bitcoin의 return을 예측하는 것을 목표로 하며, 가격 변수만을 사용할 때와 exogenous variables를 사용할 때 각각 linearity와 nonlinearity의 차이를 비교하고자 한다. 다음 날의 가격 또는 수익률을 예측하는 방식으로 진행되었다.
본 연구에서는 AR, ARIMA, Random Forest (RF), XGBoost 방법을 사용하여 성능을 비교한다.

4.1 AR 모델

Autoregression (AR) 모델은 대표적인 linear regression model로, 과거 자기 자신의 값을 이용하여 미래 값을 예측하는 방식이다. 이 모델은 다음과 같이 정의된다:

$$ X_{t+1} = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{i,t} + \varepsilon_{t+1} $$

여기서 $p = 6$을 사용하였다.(p는 Paye (2012)를 참고하여 설정.)

또한 단일 변수를 사용하는 AR 모델 외에도, 여러 개의 exogenous variables를 포함하는 ARX (Autoregressive with Exogenous Variables) 모델을 적용하였다. ARX 모델은 다음과 같이 정의된다:

$$ X_{t+1} = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{i,t} + \sum_{i=1}^{q} \beta_i Y_{i,t} + \varepsilon_{t+1} $$

여기서 Y_t는 exogenous variables들의 집합이며, q는 변수의 수를 나타낸다.
(exogenous variables는 contemporaneous term만 반영하고, lag term은 주지 않았다.)

4.2 ARIMA 모델

ARIMA는 autoregression을 기반으로 differencing과 moving average 요소를 결합한 linear regression model이다. 시계열 데이터가 비정상(non-stationary)일 경우, 차분을 통해 정상성(stationarity)을 확보한 후 모델을 적용한다.

ARIMA 모델은 다음과 같은 수식으로 표현된다:

$$ \Delta X_{t+1} = \alpha_0 + \sum_{i=1}^{p} \alpha_i \Delta X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_{t+1} $$

여기서 $\Delta X_{t} = X_t - X_{t-1}$는 1차 차분값을 의미하고, p는 AR 차수, q는 MA 차수를 의미한다.

AR 모델과 마찬가지로, ARIMA에도 exogenous variables를 추가한 확장형인 ARIMAX (ARIMA with Exogenous Variables) 모델을 적용하였다. ARIMAX는 다음과 같이 표현된다:

$$ \Delta X_{t+1} = \alpha_0 + \sum_{i=1}^{p} \alpha_i \Delta X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{r} \gamma_k Y_{k,t} + \varepsilon_{t+1} $$

여기서 $Y_{k,t}$는 k번째 exogenous variable이며, r은 전체 변수 수를 의미한다.
(ARIMAX에서도 exogenous variables는 contemporaneous term만 사용했다.)

4.2.2 비선형 모델

- **Random Forest Regressor**
- **XGBoost Regressor**
- **LightGBM Regressor**

비선형 모델은 다수의 결정 트리를 이용하여 복잡한 비선형 관계를 포착하며, 예측식은 다음과 같이 나타낼 수 있다:

$$ \hat{y}*{t+1} = *frac{1}{T}* *sum*{j=1}^{T} f_j(X_t) $$

여기서 T는 트리의 총 수, f_j는 j번째 트리를 의미한다.

4.3 변수 중요도 해석

비선형 모델에서는 예측 결과에 대한 변수 기여도를 분석하기 위해 SHAP (SHapley Additive exPlanations) 값을 이용하였다. SHAP 값은 다음과 같은 선형 가산 모형으로 표현된다:

$$ f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i $$

여기서 ϕ_0는 베이스 값, ϕ_i는 각 변수 i의 기여도를 의미한다.

이를 통해 외생 변수들이 비트코인 가격 예측에 미치는 상대적 중요성을 평가하였다.

4.4 평가 지표

모델 성능은 다음과 같은 지표를 통해 평가하였다:

- **RMSE (Root Mean Squared Error)**: 평균 제곱 오차의 제곱근
- **MAPE (Mean Absolute Percentage Error)**: 백분율 기준 절대 오차
- **R^2 (결정계수)**: 설명력
- **MDA (Mean Directional Accuracy)**: 상승/하락 방향성 예측 정확도

RMSE는 다음과 같이 정의된다:

$$ \text{RMSE} = \sqrt{ \frac{1}{N} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 } $$

MAPE는 다음과 같이 정의된다:

$$ \text{MAPE} = \frac{100}{N} \sum_{t=1}^{N} \left| \frac{y_t - \hat{y}_t}{y_t} \right| $$

R^2는 다음과 같이 정의된다:

$$ R^2 = 1 - \frac{ \sum_{t=1}^{N} (y_t - \hat{y}*t)^2 }{ *sum*{t=1}^{N} (y_t - \bar{y})^2 } $$

여기서 $\mathbf{1}(\cdot)$은 인디케이터 함수이다.

4.1 ar model

모든 모델 학습은 시계열 순서를 고려한 **Rolling Forecasting Window** 방식을 사용하였다. 구체적으로, 30일간의 학습 데이터를 이용하여 다음 1일을 예측하는 구조를 반복하였다. 학습 및 검증 단계에서는 다음의 과정을 따른다:

- 학습 구간: 30일
- 예측 구간: 다음 1일
- 이동 간격: 1일
- Time lag: 6일

하이퍼파라미터 튜닝은 시계열 특성을 반영하기 위해 TimeSeriesSplit을 사용하고, 각 모델의 최적 파라미터는 Grid Search를 통해 탐색하였다.