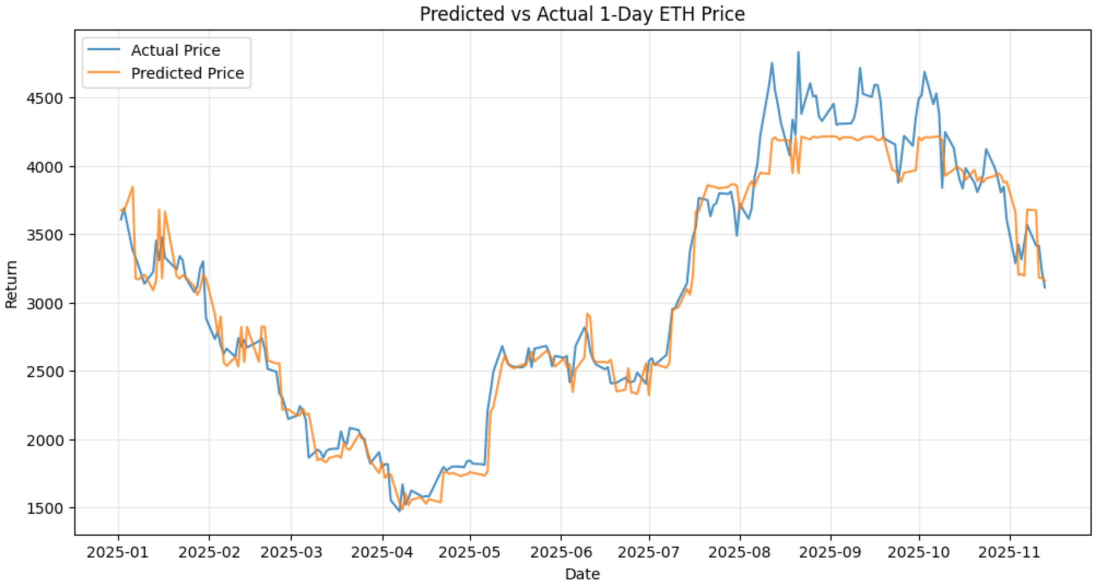
ETH Price & Volatility Forecasting using Gradient Boosting

The raw dataset contains ten daily market variables, including BTC open–high–low–close–volume, ETH close, and several macro indicators (SP500, Gold, DXY, VIX). To enrich this limited information set, we construct a compact set of handcrafted predictors based strictly on historical data. These additions include same-day returns and one-day-lagged returns for all assets (12 series total), along with a 5-day realized volatility measure for ETH that captures volatility clustering. In total, the baseline model uses 23 features.

We use LightGBM, a modern machine-learning model based on decision trees, to predict next day ETH price. LightGBM works by building many small trees one after another; each boosted tree ensembles address these challenges by iteratively building decision trees that model complex conditional relationships while controlling overfitting through regularization, subsampling, and leaf-wise tree growth. At its core, LightGBM minimizes a loss function—here, the squared error between predicted and realized volatility—using gradient boosting.

Because of this iterative boosting process, the final model can capture patterns that are non-linear, irregular, and hard to detect with traditional linear models or GARCH-style volatility models. Tree-based models like LightGBM naturally handle nonlinear effects, interactions between variables, sudden regime changes, and noisy financial data. Another important advantage is that LightGBM works extremely well with tabular features. It can easily decide which features matter and ignore the rest, making it robust even when the feature set is large.



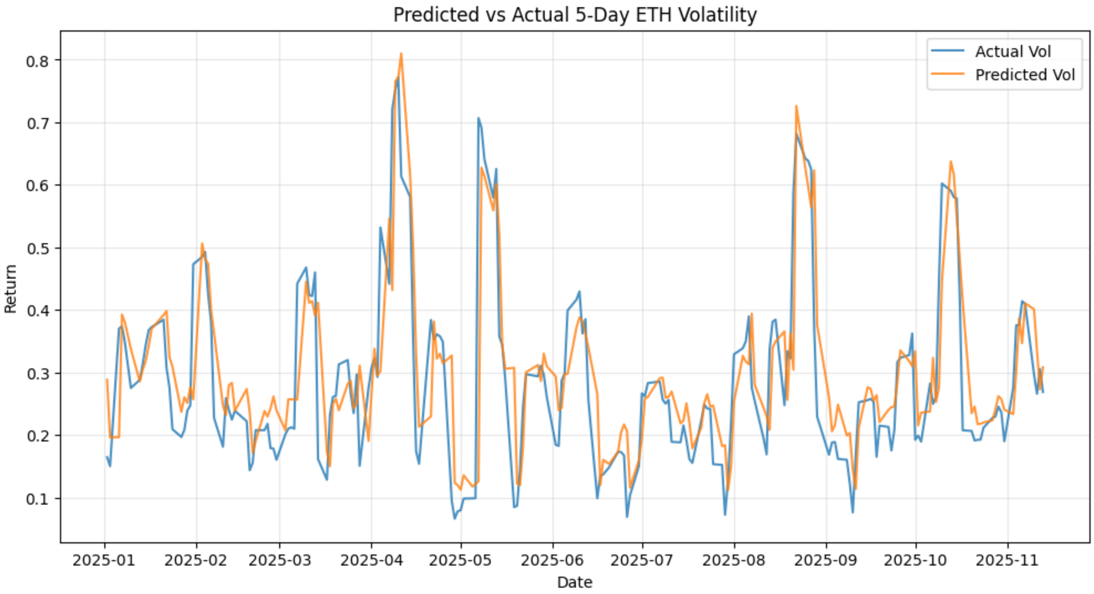
Metric	Naïve Carry-Forward Baseline	LightGBM Price Model
RMSE	143.08	194.02
MAE	105.20	143.11
R²	0.9763	0.9564

Since the price-prediction model already achieved strong results and left limited room for further improvement compare to naive baseline: predicting

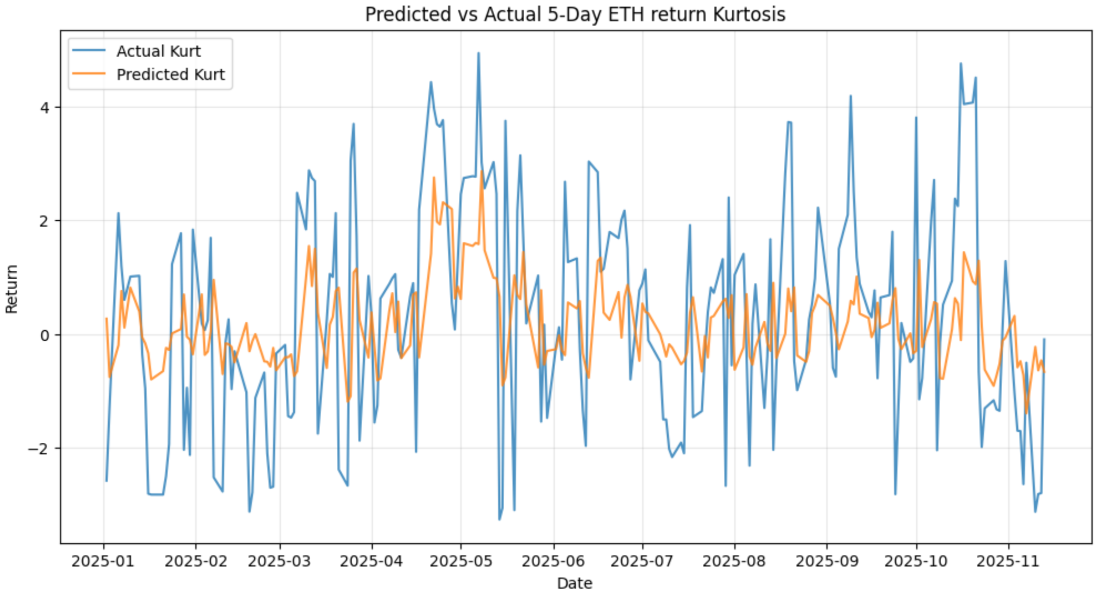
$$x_{t+1} = x_t$$

we shifted our focus to forecasting more indicators.

While forecasting crypto remains highly challenging due to extreme noise and weak serial correlation, volatility exhibits strong structural persistence (volatility clustering) and cross-asset co-movements. This project investigates 5-days ETH volatility/ 5-days return kurtosis forecasting:



Metric	Naïve Baseline	LightGBM
RMSE	0.0939	0.0889
MAE	0.0563	0.0584
R²	0.5812	0.6248



Metric	Naïve Baseline	LightGBM
RMSE	2.0219	1.7428
MAE	1.4059	1.4132
R²	-0.0699	0.2052

The study examines one-day price changes, short-horizon volatility, and realized kurtosis because they represent a natural hierarchy of predictability. One-day prices are extremely persistent and display very strong autocorrelation: tomorrow's level is overwhelmingly determined by today's level, consistent with a discrete-time Brownian-motion structure. Volatility is less persistent but still exhibits clear clustering, meaning recent fluctuations contain real information about upcoming uncertainty. Realized kurtosis, which captures the heaviness of the return distribution's tails, is the least stable of the three—its movements react sharply to shocks and show much weaker serial dependence. Moving from price to volatility to kurtosis therefore takes us from highly autocorrelated processes to increasingly noisy, less persistent ones.

All three targets are evaluated against the naïve rule, which provides a natural benchmark for any process with autocorrelation. For one-day prices, the naïve model is almost unbeatable: its extremely high R² reflects the dominant role of price persistence, and machine learning offers little additional accuracy. Volatility shows the opposite pattern—its autocorrelation allows the naïve baseline to perform reasonably well, but LightGBM still improves the fit by capturing nonlinear interactions and cross-market effects that simple persistence cannot. Kurtosis provides the clearest distinction: its weak autocorrelation causes the naïve model to perform poorly, while the machine-learning model recovers meaningful structure and achieves a substantially higher R². This progression highlights a consistent pattern: the more a variable deviates, the more room machine learning has to add predictive value.