Predicting the Unpredictable: Deep Learning for Bitcoin Price Dynamics

Sam Lai jl12560@nyu.edu
Zexuan Yang zy3035@nyu.edu
Yichao Yang yy5020@nyu.edu
Ao Xu ax2183@nyu.edu

Abstract

This project explores the development of a deep learning model for Bitcoin price prediction by integrating traditional market data with social media sentiment data. By leveraging advanced time-series forecasting models, this research spans from 2016 to July 2024, aiming to improve prediction accuracy through innovative feature engineering and hybrid model architectures.

1 Introduction

Cryptocurrency markets exhibit high volatility and complex dynamics, making accurate price prediction a challenging task. This study develops an advanced prediction framework by combining traditional financial indicators (e.g., Nasdaq index, VIX, Gold prices) with social media sentiment metrics. By leveraging state-of-the-art (SOTA) deep learning models, we aim to evaluate their ability to capture market trends over different time horizons.

1.1 Why Bitcoin?

Bitcoin is a digital currency that has attracted significant attention since its introduction. It was created in 2009 by an anonymous individual or group known as Satoshi Nakamoto [6]. The technology underlying Bitcoin is blockchain, a distributed ledger that records transactions across many computers to ensure that the record cannot be altered retroactively. One of the most important features of Bitcoin is its decentralized nature. This means that it operates without a central authority, unlike traditional currencies that are typically controlled by a governmental body or central bank. The decentralized nature of Bitcoin can contribute to its high volatility. Because there is no central control, its value is subject to rapid changes based on market demand, regulation changes, and external events [1].

Unlike other cryptocurrency, Bitcoin also possesses certain properties such as its limited supply that make it a subject of frequent discussion and analysis. It is often searched on platforms like Google and discussed extensively on social media sites such as X (formerly known as Twitter)and demonstrates a significant surge in Bitcoin's search interest during November 2024, surpassing that of USD on several occasions[3].

Given these properties of Bitcoin, our project aims to develop a machine learning model to analysis Bitcoin prices. By applying the dynamics of social media sentiment and market behavior, our model seeks to offer asset management insights into potential price movements. By collecting data from multiple sources, including social media trends and historical pricing, the model will assess patterns that could help predict future Bitcoin price changes. In doing so, we hope to address the challenges posed by Bitcoin's volatility while leveraging its high returns to inform strategic investment decisions.

Preprint. Under review.

2 Data Preparation

2.1 Dataset Overview

Our dataset integrates multiple sources of data to support both regression and classification tasks which ranges from **2016-11-01 to 2024-08-22**. Below are the details of the dataset components:

Market Data:

- **Bitcoin closing prices** (hourly).
- **Nasdaq OHLCV data** (daily).
  -The Nasdaq Composite Index represents a significant portion of the technology sector, which has a notable influence on the cryptocurrency market due to the overlap in investor profiles and risk appetite.
- **Gold OHLCV data** (daily).
  -Gold is often considered a safe-haven and asset. Its scarcity shares similar essense with Bitcoin as a hedge against inflation or currency devaluation. Including gold prices as a feature can improve predictions by capturing these risk-on/risk-off dynamics.
- **VIX index** (daily).
  -VIX measures the market's expectation of future volatility based on options trading in the SP 500. It serves as a fear gauge for market uncertainty. High levels of the VIX often correlate with risk-averse behavior among investors, which can influence cryptocurrency prices, as Bitcoin and other cryptocurrencies are typically considered high-risk assets.

Sentiment Data:

- Hourly social media sentiment scores collected from platforms such as Twitter, Reddit, and Bitcointalk.

2.2 Data Preprocessing

Daily data was aligned to hourly granularity using forward filling. Missing data, such as weekends, was handled using backward filling, while gaps due to moving averages were addressed with forward filling.

3 Exploratory Analysis

- **Correlation Analysis:** We have graphed the correlation heatmap and computed the time-varying correlation for long term and short term temporal dependence to assist our decisions in feature engineering.

  - **Feature Relationships:** The correlation heatmap (Figure 1) reveals strong positive correlations (0.92) between `listing_close` (Bitcoin closing price) and Nasdaq metrics (`NasClose/Last`, `NasOpen`, `NasHigh`, `NasLow`). This indicates Bitcoin's alignment with broader stock market trends.
  - **Sentiment Impact:** Negative sentiment metrics on platforms like Twitter and Bitcointalk show mild-to-moderate negative correlations with `listing_close`. Optimistic sentiment metrics have weaker positive correlations, showing limited influence.
  - **Time-Varying Dynamics:** demonstrates the dynamic nature of the correlation between Nasdaq closing prices and Bitcoin closing prices. **COVID-19 Period (2020):** A sharp dip in correlation indicates that Bitcoin behaved differently from traditional markets during the pandemic-induced crash.

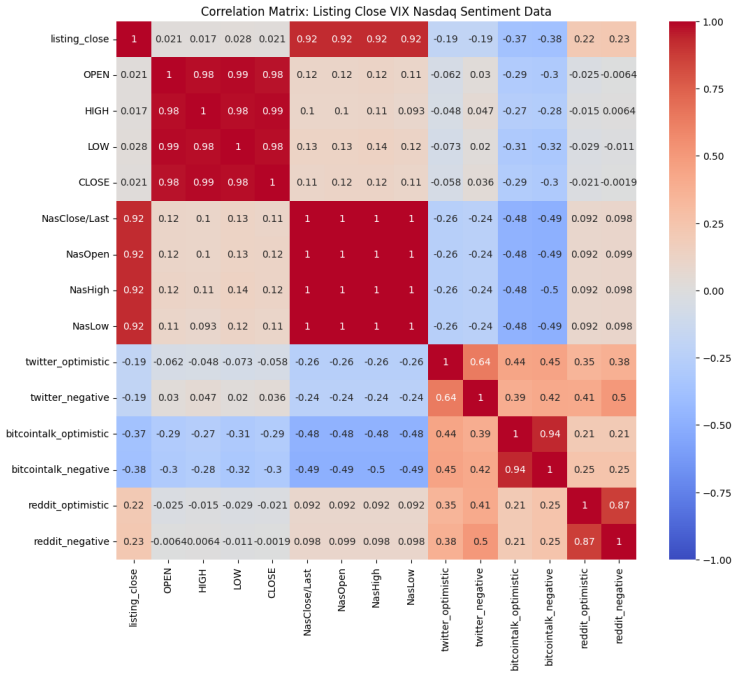

Figure 1: Overall Correlation heatmap

- **ACF and PACF Analysis:** We want to determine the number of lags for our features that should be included in the model. ACF and PACF plots also help identify the appropriate order of autoregressive (AR) and moving average (MA) terms in ARIMA models.
  - Based on the graph, the combination of the ACF's slow decay and PACF's sharp cutoff at lag 1 suggests that the time series might be modeled effectively using an ARIMA(1,1,1)
- **Volatility Analysis:** Many time series exhibit heteroscedasticity, meaning their variance changes over time. Therefore, analyzing volatility helps us identify structural breaks and regime changes and help us identify the anomaly in our model predictions, which is crucial for understanding the dynamics of the data

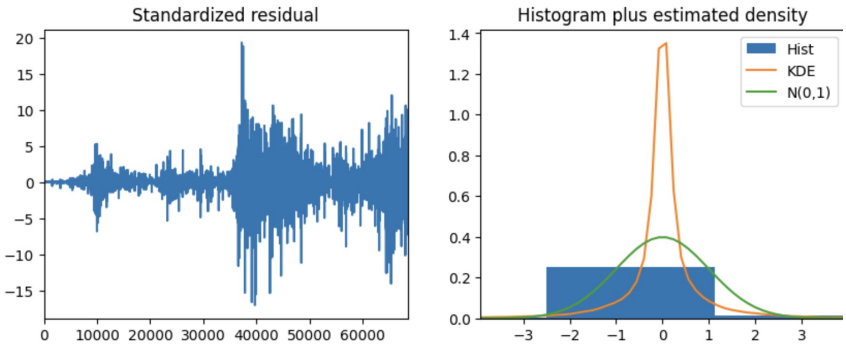

Figure 2: Volatility distribution

4 Feature Selection

Feature selection was performed using LightGBM and XGBoost, leveraging gain-based metrics to identify the most influential predictors. These tree-based models effectively captured nonlinear relationships and provided clear importance rankings.

Key Features

- **Nasdaq Features:** Close/Last, Open, High, Low, and short-term moving averages (e.g., ma_2, ma_6). These reflect market conditions and trends that serve as benchmarks for sentiment-sensitive assets.
- **Hourly Lagged Sentiment:** Historical optimistic and negative sentiment from Bitcointalk, Reddit, and Twitter, at 1-hour, 5-hour, 12-hour, and 24-hour lags. These features capture the influence of past market mood and discussions on subsequent price movements.

5 Model Developments

5.1 Objectives

Target Variables:

- **Regression Tasks:**
  - Bitcoin price predictions for the next 100 time steps.
  - Returns over the prediction horizon.

Training and Prediction Setup

The training data consists of **62,809 time points**, covering the period prior to **2024-01-01**. Predictions are generated for the next **100 time steps**, including both price and returns forecasts.

5.2 Baseline Models

The baseline models used for forecasting combine statistical methods and machine learning approaches, offering diverse perspectives on time-series analysis.

- **Statistical Models:**
  - *ARIMA:* A classical statistical model designed for modeling and forecasting stationary time series by capturing dependencies between observations at different time lags.
- **Machine Learning Models:**
  - *XGBoost:* A gradient-boosting framework that uses decision trees to model complex relationships in time-series data. Known for its efficiency and scalability, XGBoost is well-suited for predictive modeling tasks.
  - *LightGBM:* Another gradient-boosting framework optimized for speed and memory efficiency, LightGBM leverages histogram-based algorithms and provides feature importance scores for model interpretability.

5.3 Proposed Models

The proposed models were obtained from the most recent research papers published on the top conferences such as ICLR and NeurIPS in time series analysis field. We would like to leverage these advanced architectures for effective time-series forecasting.

SE-GRN (Squeeze-and-Excitation Gated Recurrent Network):

- Combines GRU layers with Squeeze-and-Excitation blocks to recalibrate features dynamically [8].
- Employs attention mechanisms to effectively capture temporal dependencies in time-series data.

iTransformer[5]:

- A state-of-the-art time-series forecasting model that restructures Transformers by embedding each time point as an independent variable token.

- Designed to improve multivariate correlation modeling and capture complex temporal dynamics for both short-term and long-term predictions.
- Enhances sequence representation learning through a feed-forward network while maintaining computational efficiency.
- Demonstrates competitive performance across various time-series forecasting benchmarks.

Times-FM (Time-Series Foundation Model)[2]:

- A state-of-the-art large language model developed by Google for zero-shot or one-shot time-series forecasting.
- A 200M parameter model, implements an encoder-decoder architecture to model complex temporal relationships.
- Supports uni-variate and multivariate covariates, improving accuracy and robustness in predictions.

SOFTS (Efficient Multivariate Time Series Forecasting with Series-Core Fusion)[4]:

- Designed for efficient and accurate multivariate time-series forecasting.
- Employs a series-core fusion mechanism to model inter-series relationships effectively.
- Optimized to capture both long-term trends and short-term variations, balancing efficiency and accuracy.

CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory) [7]:

- Integrates convolutional layers for feature extraction from raw time-series data, which helps identify important patterns and trends in historical Bitcoin prices.
- Utilizes LSTM layers to capture temporal dependencies and long-term sequences, enhancing the model's ability to predict future price movements based on past data.
- Combines the strengths of CNNs and LSTMs to effectively address the challenges of volatility and trend variations in Bitcoin price prediction.

6 Model Results

Table 1: Performance of Models in Price Prediction (MAE and MSE)

Model Category	Model	MAE	MSE
Baseline Models	LightGBM	254.37	258918.8
	ARIMA	1602.19	4847913.23
	XGBoost	258.53	271929.61
Proposed Models	SE-GRN	2377.06	883273
	iTransformer	1948	8.1235M
	Times-FM	2672.28	3.60M
	SOFTS	304.23	183679.47
	CNN-LSTM	1941.32	6.73M

6.1 Visualization

The visualization consists of six key components: (1) Full timeline prediction, illustrating the complete prediction period for Bitcoin prices; (2) single prediction window, focusing on the model's recent short-term accuracy; (3) time series validation windows, showcasing robustness of the performance trends; (4) Returns comparison, evaluating actual versus predicted daily returns; (5) Distribution of prediction errors, highlighting the error variability; and (6) Timeline of error percentages, depicting the fluctuation of errors over time.

6.1.1 Plots from SOFTS

Figure 3: Actual vs Predicted and Error plots for SOFTS

7 Conclusion And Future Work

In this study, we developed a Bitcoin price prediction framework by combining traditional market indicators and social media sentiment. The **SOFTS** model demonstrated superior predictive accuracy and stability through its series-core fusion mechanism, outperforming baseline models and advanced deep learning architectures like **SE-GRN** and **CNN-LSTM**. Notably, **TimesFM**, despite its large parameter size, performed poorly, revealing that model complexity does not guarantee improved predictions.

Our findings underscore three key insights:

- Specialized architectures with effective feature fusion outperform general models.
- Social media sentiment contributes meaningfully to price dynamics when combined with market indicators.
- Model simplicity and interpretability are often more valuable than sheer complexity in financial forecasting.

To explore the relationship between external events, such as U.S. elections, and Bitcoin price movements, we extended the prediction horizon. However, due to outdated and incomplete data, we were limited to historical sources, resulting in suboptimal performance.

For future work, we propose:

- Acquiring updated, real-time datasets to analyze event-driven market shifts accurately, such as elections and major policy changes.
- Expanding our model to include alternative data sources, including news headlines, blockchain activity, and real-time social media trends.

These improvements will enable more reliable insights into Bitcoin price dynamics under evolving market conditions.

8 References

References

[1] David Lee Kuo Chuen. *Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data*. Academic Press, 1st edition, 2015.

[2] A. Das, W. Kong, R. Sen, Y. Zhou, and Google Research. A decoder-only foundation model for time-series forecasting. *ICML 2024*, 2024.

[3] Google. Google trends. `https://trends.google.com/`, 2024. Accessed: 2024-11-19.

[4] L. Han, X.-Y. Chen, H.-J. Ye, D.-C. Zhan, China National Key Laboratory for Novel Software Technology, Nanjing University, and China School of Artificial Intelligence, Nanjing University. Softs: Efficient multivariate time series forecasting with series-core fusion. *NeurIPS 2024*, 2024.

[5] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023. Last revised 14 Mar 2024 (v4).

[6] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. `https://bitcoin.org/bitcoin.pdf`, 2008.

[7] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[8] Jiawei Zhang, Limeng Cui, and Fisher B. Gouza. Segen: Sample-ensemble genetic evolutional network model. *arXiv preprint arXiv:1803.08631*, 2018.