Algorithmic Trading

With Deep Learning

Ritesh Kumar

2024SP_MS_DSP_498-DL_SEC61: Capstone

Module 3

Exploratory Data Analysis

Kimberly Chulis and Srabashi Basu

July 7, 2024

Table of Contents

Abstract

In the dynamic landscape of financial markets, deep learning techniques have revolutionized algorithmic trading strategies. This project aims to develop an advanced algorithmic trading system for BankNifty, a prominent Indian banking index. The system will leverage historical data, including Cumulative Open Interest (COI), Price, Volume, India Volatility Index, and Technical Indicators such as Moving Averages, RSI, and MACD, to predict market trends and make informed trading decisions.

The primary objective is to design a sophisticated trading system that analyzes historical data and informs trading decisions using deep learning models. Key components of the project include Feature Engineering, Backtesting, Risk Management, Performance Metrics, Scalability, and Adaptation. High-frequency data collection at 1-minute intervals will enable the system to capture short-term market movements and trends.

Data preprocessing steps will include handling missing data, feature engineering to create new features, data normalization, and splitting the dataset into training, validation, and testing sets. The deep learning models employed will include Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and Transformer Models, each chosen for their ability to capture sequential patterns, identify complex features, and handle long-range dependencies in time-series data.

Effective feature engineering will be crucial, with techniques like recursive feature elimination (RFE) and feature importance ranking identifying the most relevant features. Technical indicators will enhance the feature set. Rigorous testing will evaluate the trading strategy's performance on historical data, and optimization will fine-tune parameters to maximize profitability and minimize risk.

Risk management strategies, including stop-loss orders and position sizing, will protect capital and minimize losses. Performance metrics such as the Win-Loss Ratio and Profit Factor will provide a comprehensive assessment of the system's effectiveness.

Designed for scalability and adaptability, the system can be applied to other indices and financial instruments. Periodic retraining and continuous monitoring will maintain model accuracy and relevance, ensuring the system remains effective and responsive to market dynamics. This project aims to create a powerful tool for navigating the complexities of BankNifty trading, with principles extendable to other financial markets.

Variables Overview

In the rapidly changing landscape of financial markets, deep learning techniques have revolutionized algorithmic trading strategies. This dataset is specifically curated to aid in the development of a sophisticated algorithmic trading system for BankNifty. It aims to harness historical data to accurately predict market trends and make informed trading decisions by utilizing high-frequency data collection and advanced technical indicators. The comprehensive data provided will support the implementation of cutting-edge deep learning models, ultimately enhancing trading performance and strategy.

Date and Time Variables:

- date: Represents the date of the observation.

- time: Represents the specific time of the observation. Together, 'date' and 'time' provide a precise timestamp for each data point.

Price and Volume Variables:

- open: The price of the asset at the start of the trading interval.

- high: The highest price of the asset during the trading interval.

- low: The lowest price of the asset during the trading interval.

- close: The price of the asset at the end of the trading interval.

- vix: The Volatility Index, which measures market expectation of near-term volatility.

- ce_vol: The call option volume, representing the number of call options traded.

- pe_vol: The put option volume, representing the number of put options traded.

- total_vol: The total trading volume, combining both call and put options.

Open Interest Variables:

- ce_oi: Call option open interest, indicating the total number of call options outstanding.

- pe_oi: Put option open interest, indicating the total number of put options outstanding.

- ce_oi_chg: Change in call option open interest.

- pe_oi_chg: Change in put option open interest.

- tot_oi_chg: Total change in open interest.

Put/Call Ratio Variables:

- pcr_vol: Put/Call ratio based on volume.

- pcr_oi: Put/Call ratio based on open interest.

Moving Averages:

- SMA_5: 5-period Simple Moving Average.

- SMA_10: 10-period Simple Moving Average.

- SMA_20: 20-period Simple Moving Average.

- EMA_12: 12-period Exponential Moving Average.

- EMA_20: 20-period Exponential Moving Average.

- EMA_26: 26-period Exponential Moving Average.

Technical Indicators:

- RSI: Relative Strength Index, a momentum oscillator that measures the speed and change of price movements.

- MACD: Moving Average Convergence Divergence, calculated as the difference between the 12-period and 26-period EMAs.

- Signal_Line: 9-period EMA of the MACD.

- MACD_Histogram: The difference between MACD and Signal Line.

- Middle_Band: Middle band of Bollinger Bands, usually a 20-period SMA.

- Upper_Band: Upper band of Bollinger Bands.

- Lower_Band: Lower band of Bollinger Bands.

- %K: Stochastic Oscillator %K, indicating the current price relative to the high and low range over a set period.

- %D: Stochastic Oscillator %D, the 3-period SMA of %K.

- ATR: Average True Range, measuring market volatility.

- OBV: On-Balance Volume, using volume flow to predict changes in stock price.

- VWAP: Volume Weighted Average Price.

- ROC_10: Rate of Change over 10 periods.

- ROC_20: Rate of Change over 20 periods.

- CCI: Commodity Channel Index, measuring the variation of an asset's price from its statistical mean.

- plus_DM: Positive Directional Movement.

- minus_DM: Negative Directional Movement.

- TR: True Range, the greatest of the following: current high - current low, current high - previous close, or current low - previous close.

- plus_DI: Positive Directional Indicator.

- minus_DI: Negative Directional Indicator.

- DX: Directional Movement Index, indicating trend strength.

- ADX: Average Directional Index, smoothing the DX over a specified period.

Pivot Points and Support/Resistance Levels:

- exp_day: Expiration day, indicating the day options expire.

- pivot: Pivot point, calculated as the average of the high, low, and closing prices from the
  previous trading period.

- s1, s2, s3: First, second, and third support levels derived from the pivot point.

- r1, r2, r3: First, second, and third resistance levels derived from the pivot point.

Target Variable:

- nxt_move: The variable of interest, representing the predicted next movement in price or
  the number of points to the next close.

Data Summary:

- Minimum datetime: 2024-06-24 10:00:53

- Maximum datetime: 2024-07-02 15:31:04

- Total rows: 2561

This dataset provides an extensive and detailed view of market behavior, encompassing a
wide range of technical indicators and price/volume metrics. It includes essential variables
such as open, high, low, and close prices, along with trading volumes and open interest data
for both call and put options. Advanced technical indicators like moving averages, RSI,
MACD, Bollinger Bands, and stochastic oscillators offer insights into market trends,
momentum, and volatility.

Additionally, the dataset features pivotal points, support and resistance levels, and volatility
measures like the India Volatility Index (VIX) and Average True Range (ATR), which are
crucial for understanding market dynamics. The inclusion of both simple and exponential

moving averages (SMA and EMA) over various periods allows for the analysis of short-term and long-term trends.

Metrics such as the Put/Call ratio based on volume and open interest further enhance the dataset's ability to gauge market sentiment. The dataset also incorporates crucial elements like the On-Balance Volume (OBV) and Volume Weighted Average Price (VWAP), providing a robust foundation for assessing the flow of funds and average trading price.

By offering a granular view of market data at 1-minute intervals, this dataset is exceptionally suited for high-frequency trading analysis. It supports the development and validation of advanced predictive models, including LSTM, CNN, and Transformer models, ensuring a comprehensive approach to algorithmic trading analysis. This dataset is instrumental in building and fine-tuning sophisticated trading strategies, ultimately contributing to more informed and effective decision-making in the financial markets.

Target Variable

The descriptive statistics of the target variable `nxt_move` are as follows:

- Count: There are 2,378 observations for the `nxt_move` variable.

- Mean: The average value of `nxt_move` is approximately 0.224.

- Standard Deviation (std): The standard deviation is 22.48, indicating a significant spread around the mean.

- Minimum (min): The minimum value is -151.55, suggesting a substantial negative movement.

- 25th Percentile (25%): The 25th percentile is -11.04, meaning 25% of the data points are less than -11.04.

- Median (50%): The median value is 0.00, indicating that half of the observations are below and half above 0.

- 75th Percentile (75%): The 75th percentile is 11.05, meaning 25% of the data points are greater than 11.05.

- Maximum (max): The maximum value is 278.55, suggesting a substantial positive movement.

The target variable `nxt_move` has a mean close to zero with a high standard deviation, indicating a wide range of values. The distribution is centered around zero, as shown by the median. The data contains both significant negative and positive values, highlighting the potential for considerable market movements in either direction. This variability and spread in `nxt_move` are critical for developing predictive models, as they must account for the broad range of possible outcomes.
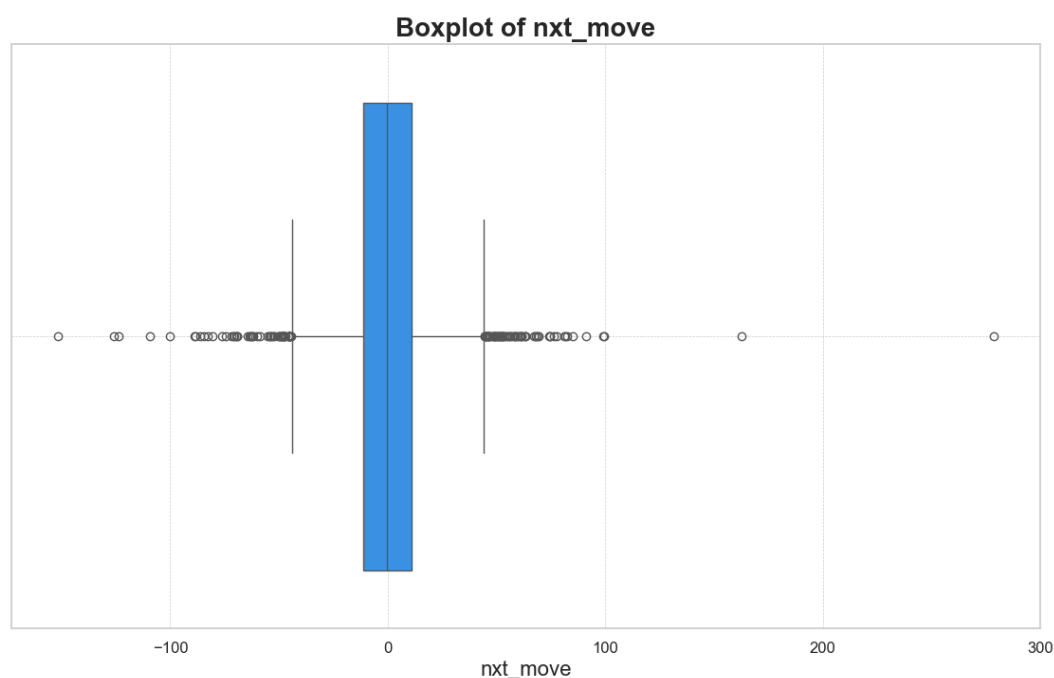
Boxplot of nxt_move

- Median and Quartiles: The boxplot shows that the median (50th percentile) of `nxt_move` is around 0, indicating that half of the observations are below and half above 0. The interquartile range (IQR), represented by the box, is relatively narrow, highlighting that 50% of the data points lie between approximately -11 and +11.

- Whiskers and Outliers: The whiskers extend to around -50 and +50, beyond which numerous outliers are observed. These outliers indicate extreme values, both on the negative and positive sides, which are significantly distant from the central bulk of the data.

Distribution Plot of nxt_move
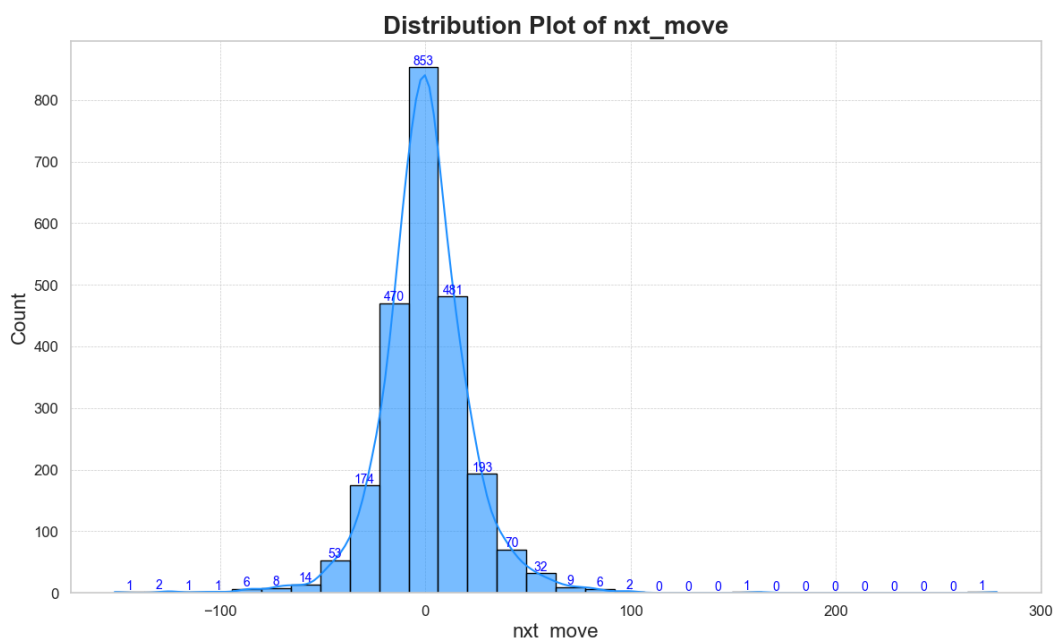
- Shape and Central Tendency: The distribution plot is bell-shaped, indicating a normal-like distribution centered around 0. The highest count of observations (853) is near the mean value.

- Spread and Skewness: Most data points are concentrated around the center, with the frequency sharply decreasing as we move away from the mean. There are longer tails on both sides, especially on the positive side, indicating a slight right skewness.

- Outliers: There are visible outliers on both extremes of the plot, confirming the boxplot's indication of substantial outliers. These extreme values show the variable's wide range and significant deviations.

The `nxt_move` variable has a distribution centered around zero with a high concentration of data points within a narrow range. However, it also exhibits substantial outliers and a slight skewness towards the positive side. This variability and presence of outliers are critical for

developing robust predictive models, as they must effectively handle and predict these
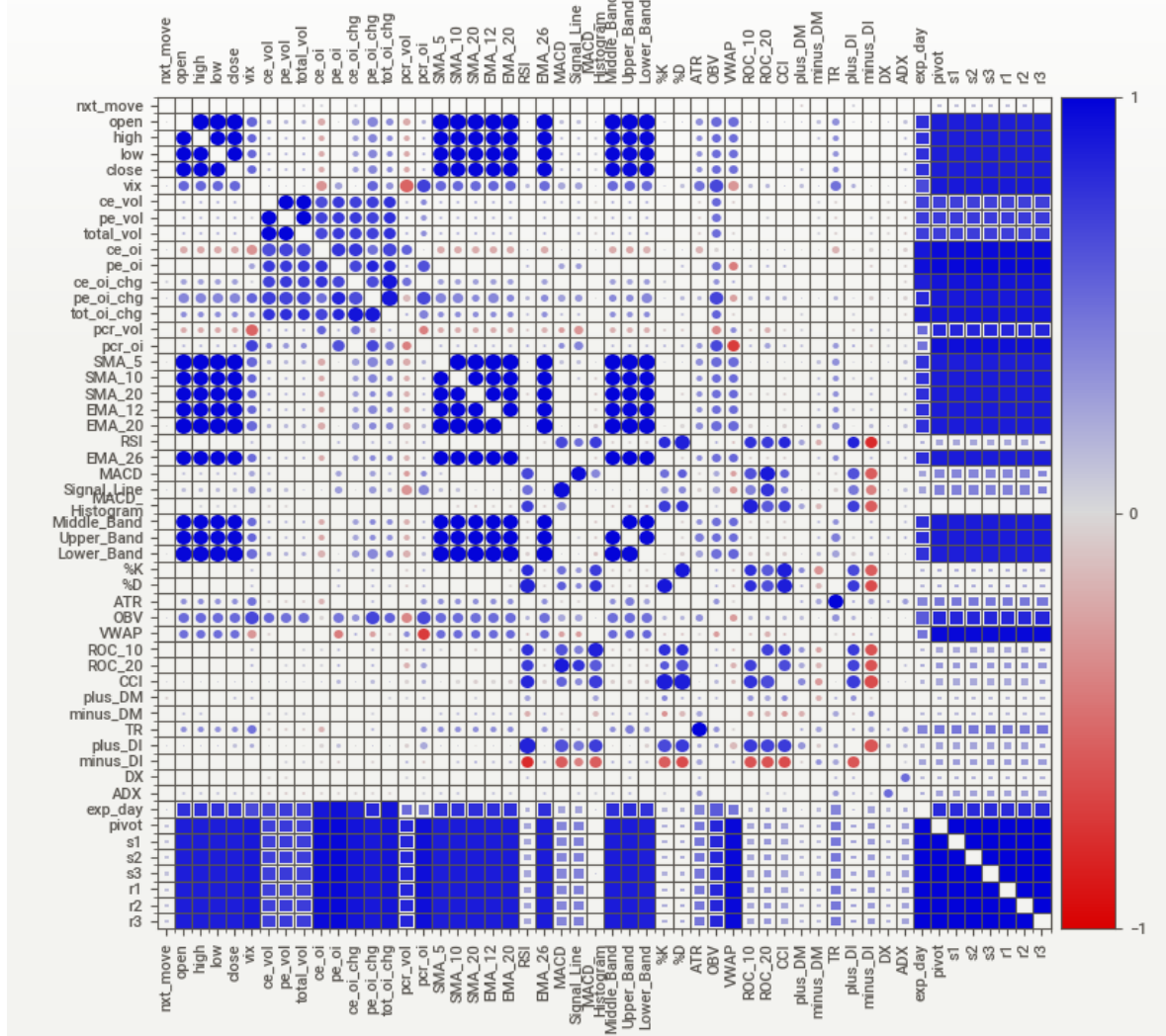extreme market movements.

Correlation with Predictor Variables

The correlation matrix reveals that `nxt_move` has weak correlations with most variables,
indicating that no single factor strongly predicts the next market movement. The variable
`pcr_oi` shows the strongest negative correlation, suggesting that the put/call ratio based on
open interest is inversely related to `nxt_move`. Other variables, including price metrics
(`open`, `high`, `low`, `close`) and moving averages (SMA, EMA), display very weak

correlations. This implies that while these factors are essential for understanding market trends, predicting `nxt_move` effectively requires a combination of multiple indicators rather than relying on individual ones.

Negative Correlations:

- Directional Movement Indicators (DMI): `plus_DM` (-0.083949) and `minus_DM` (0.058600) have notable negative and positive correlations respectively, indicating the directional movements in the market can impact `nxt_move` significantly.

- Support Levels: `s3`, `s1`, and `s2` all have negative correlations (around -0.05), showing that lower support levels are associated with decreases in `nxt_move`.

- Price Metrics: Variables like `low`, `close`, `VWAP`, `open`, and `high` also show slight negative correlations (around -0.05), suggesting that lower prices are associated with decreases in `nxt_move`.

- Moving Averages: Short-term and long-term moving averages (`SMA_5`, `EMA_12`, `SMA_10`, `EMA_20`, `EMA_26`, `SMA_20`) have negative correlations, indicating that lower moving averages correspond to decreases in `nxt_move`.

- Bollinger Bands: `Lower_Band`, `Middle_Band`, and `Upper_Band` have negative correlations, suggesting that positions within Bollinger Bands are indicative of the direction of `nxt_move`.

- Pivot Points: The `pivot` variable has a slight negative correlation, implying that lower pivot points might be associated with decreases in `nxt_move`.

- Technical Indicators: Indicators like `RSI`, `%K`, `CCI`, `TR`, `ATR`, `ROC_10`, `ROC_20`, `MACD`, `Signal_Line`, `plus_DI`, and `ADX` all show slight negative correlations, indicating that lower values in these technical indicators are associated with decreases in `nxt_move`.

Positive Correlations:

- Put/Call Ratios: `pcr_oi` (0.031451) and `pcr_vol` (0.006902) have slight positive

  correlations, suggesting that higher put/call ratios might be associated with increases in

  `nxt_move`.

- Volatility Index: `vix` (0.023932) has a positive correlation, indicating that higher market

  volatility is associated with increases in `nxt_move`.

- Open Interest Changes: `pe_oi_chg` (0.023425) and `tot_oi_chg` (0.012114) show slight

  positive correlations, suggesting that changes in open interest, particularly in puts, might

  be linked to increases in `nxt_move`.

- Expiration Day: `exp_day` (0.012748) shows a slight positive correlation, indicating that

  the day of options expiration can have a slight impact on `nxt_move`.

Minimal Correlations:

- Volume Metrics: Variables such as `ce_vol`, `pe_vol`, and `total_vol` have very minimal

  negative correlations, suggesting that trading volume has little to no direct correlation

  with `nxt_move`.

- Directional Indicators and ADX: `minus_DI` and `ce_oi` have minimal positive

  correlations, while `ce_oi_chg` has a minimal negative correlation, indicating that these

  metrics have a negligible impact on `nxt_move`.

Overall, the data indicates that several price, volume, and technical indicators have varying

degrees of correlation with `nxt_move`. Most correlations are relatively weak, suggesting a

complex relationship between these indicators and market movements. The negative

correlations generally suggest that lower values in these indicators are associated with

decreases in `nxt_move`, while the few positive correlations indicate an association with

increases.

Ten Strongest Correlations

```
plus_DM        -0.083949
minus_DM        0.058600
s3             -0.053277
s1             -0.051322
s2             -0.051155
low            -0.051004
close          -0.050778
VWAP           -0.050679
open           -0.049579
high           -0.049290
```

plus_DM: Trend Identification: plus_DM helps traders identify the strength and direction of an upward trend. When plus_DM is consistently positive and large, it indicates a strong upward movement.

DMI and ADX: plus_DM is a component of the Directional Movement Index (DMI), which includes plus_DI (Positive Directional Indicator) and minus_DI (Negative Directional Indicator). The Average Directional Index (ADX) uses these indicators to assess the strength of the trend without considering its direction.

plus_DI = (Smoothed plus_DM / Average True Range) * 100

minus_DI = (Smoothed minus_DM / Average True Range) * 100
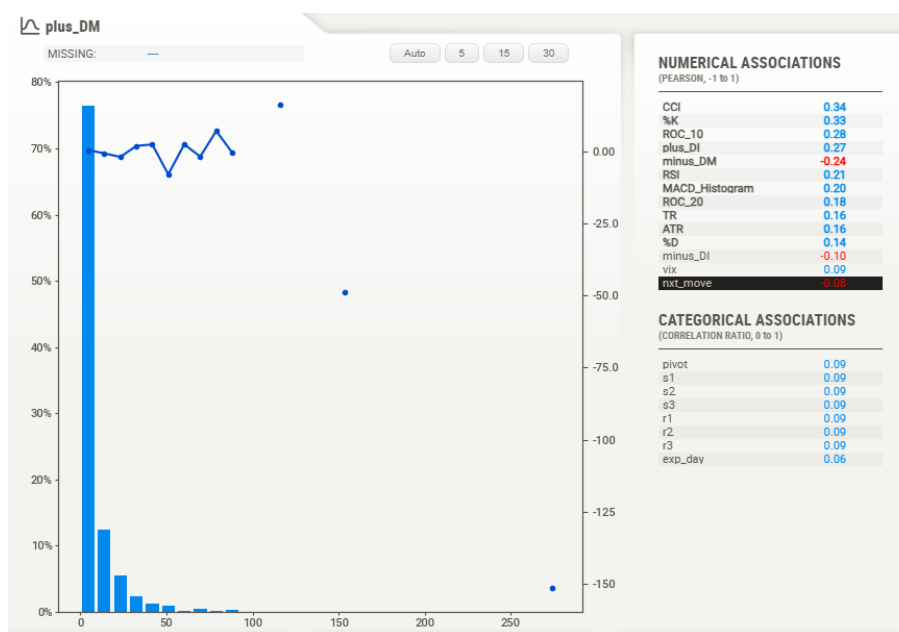
The correlation coefficient between `plus_DM` and `nxt_move` is -0.08, indicating a weak negative relationship. This suggests that as the positive directional movement (`plus_DM`) increases, the value of `nxt_move` tends to decrease slightly. However, this relationship is not strong, implying that `plus_DM` alone is not a significant predictor of `nxt_move`. The histogram reveals that most `plus_DM` values are close to zero, with occasional significant outliers. While `plus_DM` may not be a strong standalone predictor, it could still be useful when combined with other indicators for a more comprehensive analysis of market dynamics.

minus_DM:
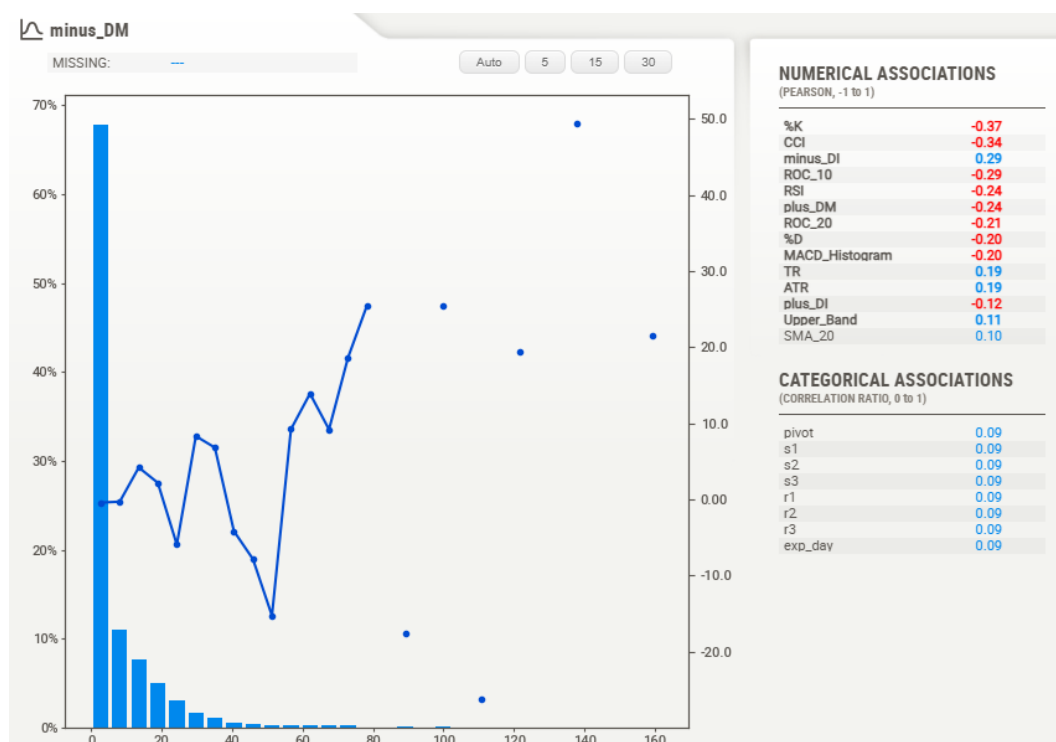
The correlation coefficient between `minus_DM` and `nxt_move` is 0.10, indicating a very weak positive relationship. This suggests that as the negative directional movement (`minus_DM`) increases, the value of `nxt_move` tends to increase slightly. However, this relationship is not strong enough to consider `minus_DM` a significant predictor of `nxt_move`. The histogram shows that most `minus_DM` values are close to zero, with some significant outliers. Despite the weak correlation, `minus_DM` can still contribute valuable

insights when used alongside other indicators in a comprehensive predictive model for market dynamics.

Pivot Points: Pivot points are a technical analysis indicator used to determine the overall trend of the market over different time frames. The pivot point itself is the average of the high, low, and closing prices from the previous trading day. It serves as a predictive indicator of market movements and potential support and resistance levels.

$$Pivot\ Point = \frac{High + Low + Close}{3}$$

Support Levels (`s1`, `s2`, `s3`): Support levels are price points below the pivot point that act as a floor, preventing the price from falling further. They are calculated based on the pivot point and the high and low prices from the previous day.

First Support Level (s1):

$$s1 = 2 * Pivot\ Point - High$$

Second Support Level (s2):

$$s2 = Pivot\ Point - (High - Low)$$

Third Support Level (s3):

$$s3 = Low - 2 * (High - Pivot\ Point)$$

These levels are useful for traders to identify potential buy points where the price is likely to rebound.

Resistance Levels (`r1`, `r2`, `r3`):

Resistance levels are price points above the pivot point that act as a ceiling, preventing the price from rising further. They are calculated similarly to the support levels but use the high and pivot point values differently.

First Resistance Level (r1):

$$r1 = 2 * Pivot\ Point - Low$$

Second Resistance Level (r2):

$$r2 = Pivot\ Point + (High - Low)$$

Third Resistance Level (r3):

$$r3 = High + 2 * (Pivot\ Point - Low)$$

These levels help traders identify potential sell points where the price is likely to face resistance and fall back.

Summary: Pivot points, along with their corresponding support (`s1`, `s2`, `s3`) and resistance (`r1`, `r2`, `r3`) levels, are essential tools in technical analysis. They help traders anticipate potential price movements, making them valuable for setting entry and exit points in trading strategies.

s3:
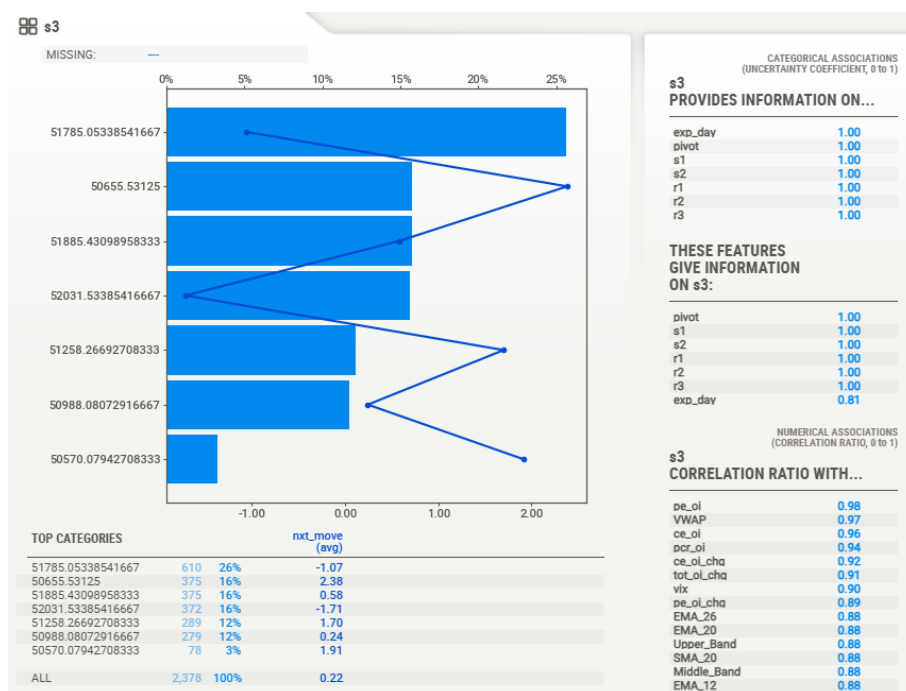
The correlation coefficient between `s3` and `nxt_move` is very low, suggesting minimal direct correlation. The average `nxt_move` values across different `s3` categories show varying but generally small impacts. For instance, the highest category `51785.05` has an average `nxt_move` of -1.07, while the lowest category `50570.08` has an average `nxt_move` of 1.91. This variance implies that `s3` has a nuanced effect on `nxt_move`, potentially influenced by other interacting variables. However, overall, `s3` does not appear to be a strong standalone predictor of `nxt_move`, but it could still contribute valuable insights when combined with other indicators.
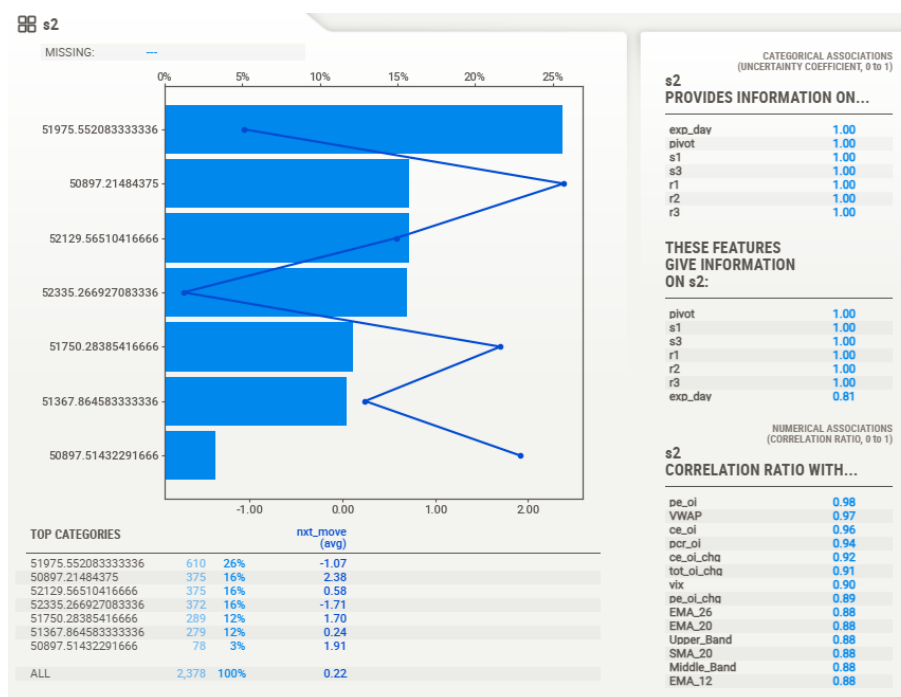
s1:
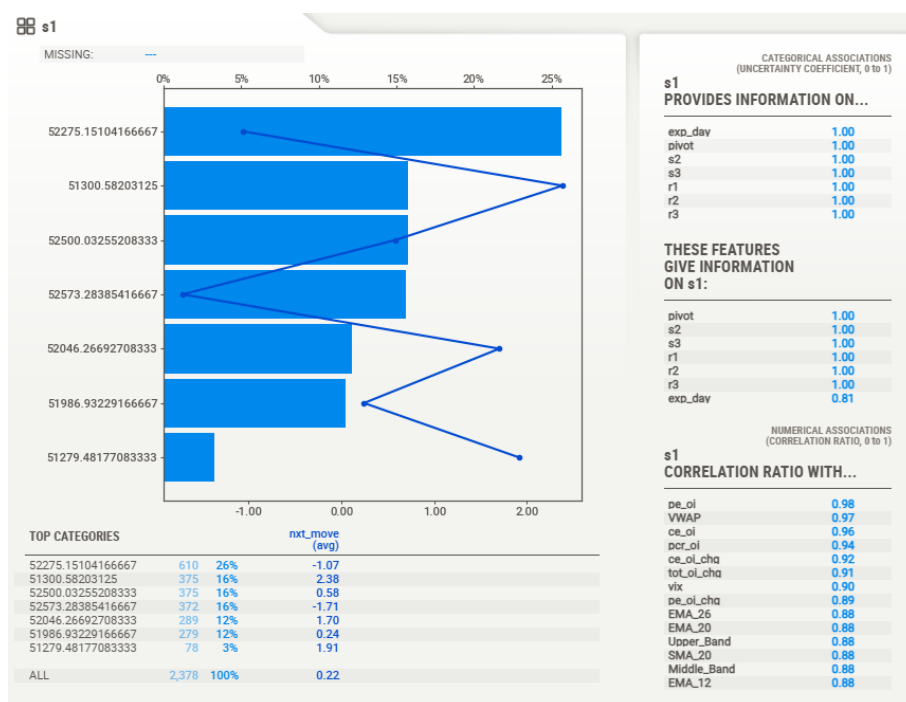
The correlation coefficient between `s1` and `nxt_move` is low, suggesting a weak direct relationship. The average `nxt_move` values across different `s1` categories vary significantly, from -2.38 to 1.91, indicating that `s1` levels can influence `nxt_move` differently. For example, the category `51300.58` has an average `nxt_move` of -2.38, while `51279.48` has an average of 1.91. Despite this variance, `s1` does not appear to be a strong

standalone predictor of `nxt_move`. However, it might provide valuable context when combined with other variables, highlighting potential support levels where price movements could change direction.

s2:

The correlation coefficient between `s2` and `nxt_move` is low, indicating a weak direct relationship. The average `nxt_move` values across different `s2` categories vary, suggesting that `s2` levels can influence `nxt_move` differently. For example, the category `50897.21` has an average `nxt_move` of 2.38, while `52335.27` has an average of -1.71. This variance implies that while `s2` alone is not a strong predictor of `nxt_move`, it can provide valuable insights when used with other indicators. The weak correlation suggests that `s2` reflects potential support levels where price movements may experience changes in direction.

Low: The variable `low` has a low correlation with `nxt_move`, suggesting a weak direct relationship. The close, open, high, and various moving averages (SMA, EMA) all show a perfect correlation with `low`, indicating that these price metrics move together. The

histogram indicates variability in `low` values, which align with other price-related indicators. Categorical associations like pivot points, support, and resistance levels also show a strong correlation with `low`, reinforcing its role in determining market trends. However, the weak direct correlation with `nxt_move` suggests that while `low` is a critical price metric, it does not significantly predict `nxt_move` on its own.
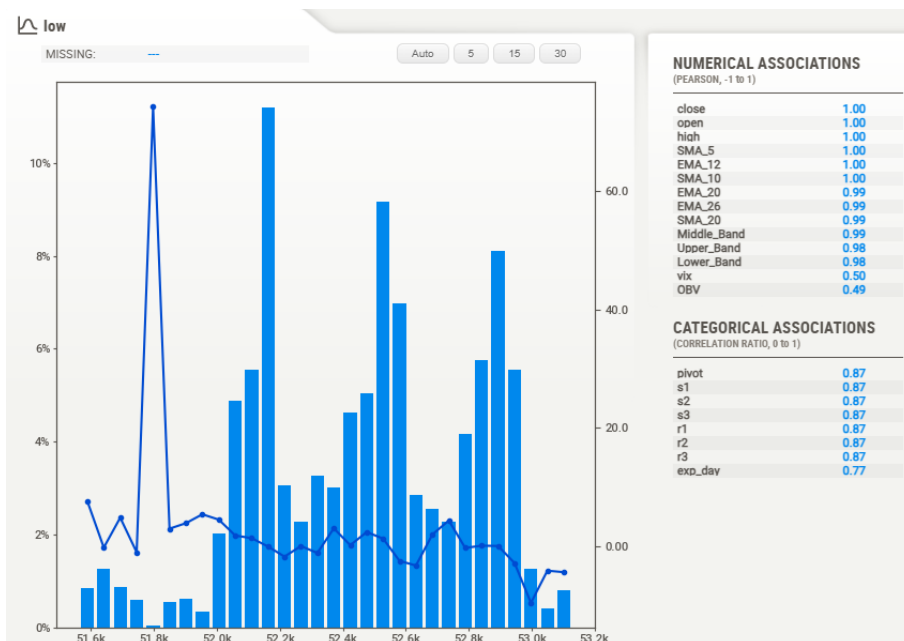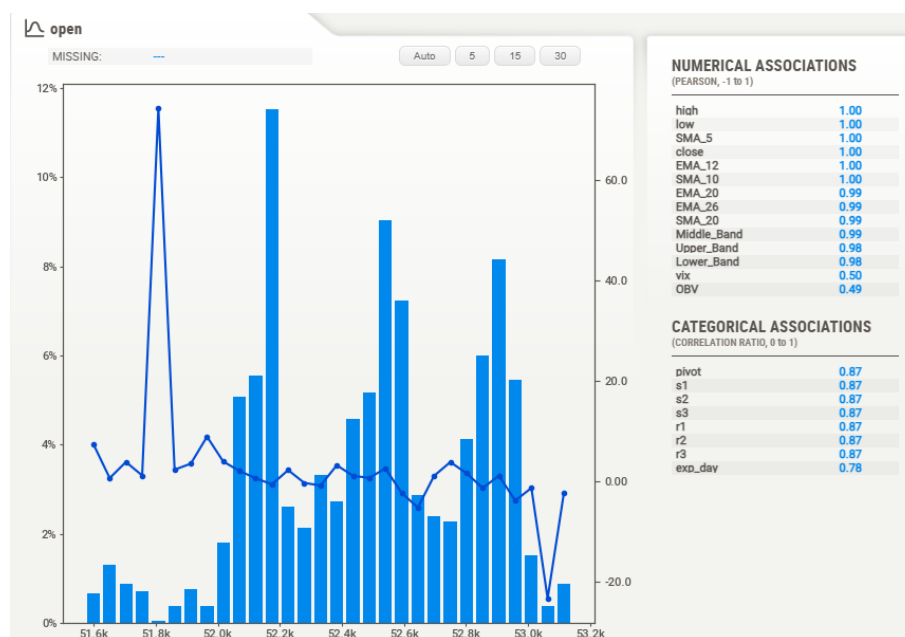

Close:
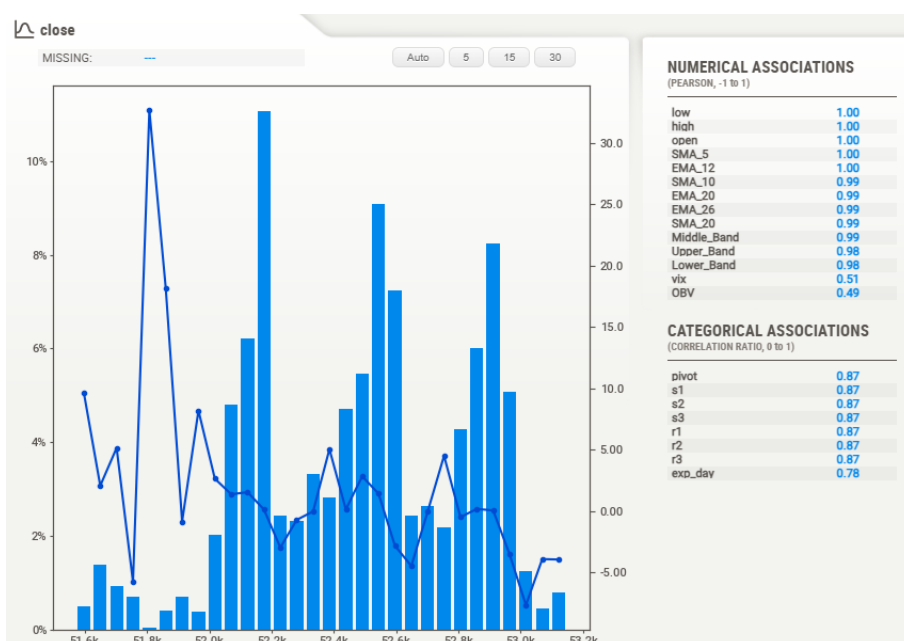
The `close` variable has a weak direct correlation with `nxt_move`, indicating it is not a strong standalone predictor of future market movements. While `close` shows perfect correlation with other price metrics such as `low`, `high`, `open`, and various moving averages, this relationship does not extend significantly to `nxt_move`. The histogram indicates variability in closing prices, reflecting the market's dynamic nature. Categorical associations, including pivot points and support/resistance levels, are also strongly correlated with `close`, emphasizing its role in market analysis. However, predicting `nxt_move` effectively requires considering additional variables and indicators beyond just the closing price.

VWAP: The Volume Weighted Average Price (VWAP) is a trading benchmark that calculates the average price a security has traded at throughout the day, based on both volume and price. It is used by traders to determine the true average price of a security over a specific period. VWAP is calculated by taking the total rupee value of all trades and dividing it by the total volume of trades. It helps traders assess the market trend and make informed decisions by comparing current prices to the VWAP. Prices above VWAP indicate a bullish trend, while prices below suggest a bearish trend.
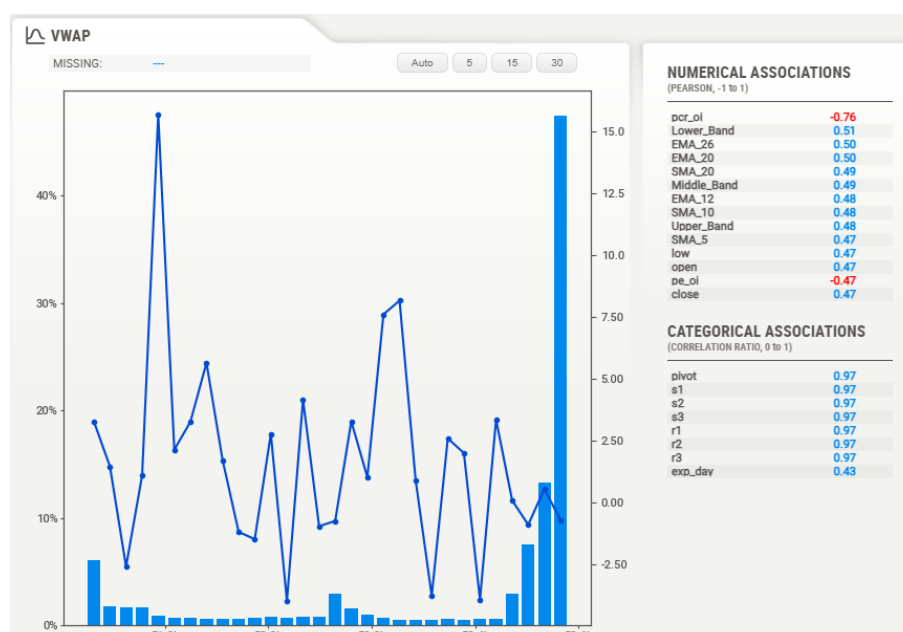
The `VWAP` shows a weak correlation with `nxt_move`, suggesting it is not a strong standalone predictor of future market movements. The numerical associations indicate `VWAP` is significantly correlated with various moving averages and price metrics but has a negative correlation with `pcr_oi` (-0.76). This implies that while `VWAP` is an essential metric for understanding average price levels based on volume, it does not directly predict `nxt_move`. Categorical associations such as pivot points and support/resistance levels are strongly correlated with `VWAP`, emphasizing its importance in technical analysis, but additional indicators are needed for accurate predictions of `nxt_move`.
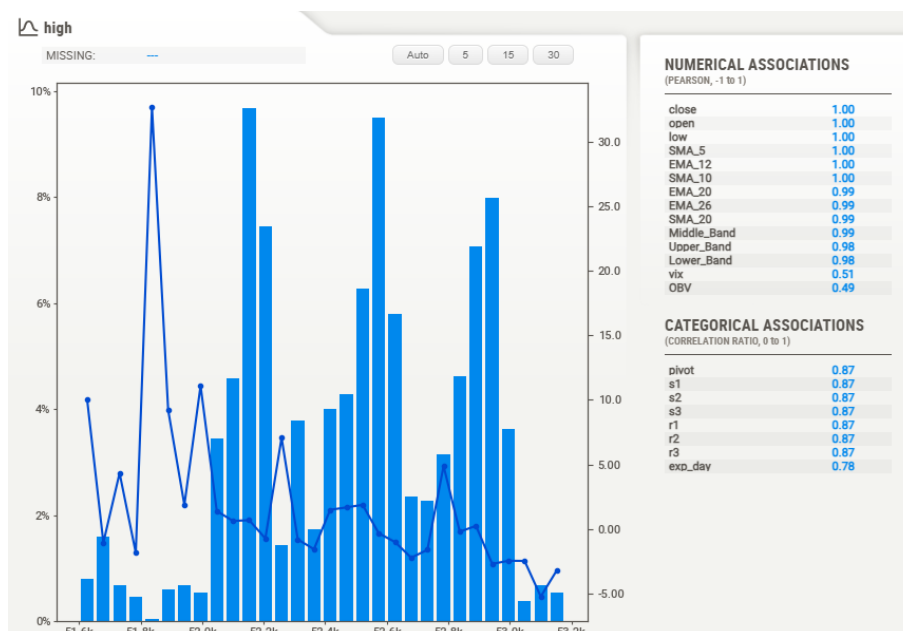
Open:

The `open` variable shows a weak correlation with `nxt_move`, indicating it is not a strong standalone predictor of future market movements. The `open` price is perfectly correlated with other price metrics such as `high`, `low`, and `close`, and various moving averages (SMA, EMA), reflecting a consistent relationship among these metrics. The histogram depicts a varied distribution of opening prices. Strong correlations with categorical associations like pivot points and support/resistance levels highlight the importance of the

`open` price in market analysis. However, predicting `nxt_move` effectively requires additional indicators beyond the `open` price alone.

High:

The `high` variable exhibits a weak correlation with `nxt_move`, indicating it is not a strong standalone predictor of future market movements. The `high` price is perfectly correlated with other price metrics such as `close`, `open`, and `low`, as well as various moving averages (SMA, EMA). This reflects a consistent relationship among these metrics. The histogram shows variability in high prices.

Conclusion

In this exploratory data analysis (EDA) phase, we examined a comprehensive dataset. Currently spanning seven trading days, this dataset includes detailed historical data on Cumulative Open Interest (COI), price metrics, trading volumes, the India Volatility Index (VIX), and numerous technical indicators such as moving averages, RSI, and MACD.

Despite the dataset's richness, our correlation analysis revealed that no single factor strongly predicts the target variable, `nxt_move`. The strongest correlations observed were still

relatively low, indicating that predicting market movements effectively will require a multifactorial approach. Although the correlations are low, this can still be leveraged effectively using advanced modeling techniques. Here are the steps to build a robust model:

1. Feature Engineering:

   - Creating new features from the existing data to better capture relevant information.

   - Applying techniques like recursive feature elimination (RFE) and feature importance ranking to identify and retain the most impactful features.

2. Data Normalization:

- Applying min-max scaling or Z-score normalization to bring all features to a consistent scale, ensuring that the model treats all features uniformly.

3. Train-Validation-Test Split:

- Dividing the dataset into training (70%), validation (20%), and testing (10%) sets. The first 70% of the data will be used for training, the next 20% for validation, and the most recent 10% for testing to ensure that the model is evaluated on the most current data.

4. Advanced Modeling Techniques:

- Long Short-Term Memory (LSTM) Networks: LSTM networks, a type of recurrent neural network (RNN), are adept at capturing sequential patterns and dependencies in time-series data. Given their success in financial time-series forecasting, LSTM networks will be employed to analyze BankNifty historical data and predict OHLC prices for the next 3 minutes.

- Convolutional Neural Networks (CNN): CNNs, traditionally used in image processing, have shown potential in identifying patterns in time-series data. By treating time-series data as a temporal sequence of images, CNNs can extract complex features that traditional models might miss.

- Transformer Models: Transformer models, known for their performance in natural language processing, can handle long-range dependencies and have shown promise in financial market predictions. These models will be explored to capture intricate patterns and trends in the BankNifty data.

5. Performance Evaluation:

- Evaluating the performance of the deep learning models and trading strategy is critical. Key performance metrics will include the Win-Loss Ratio and Profit Factor. These metrics will provide a comprehensive assessment of the trading system's effectiveness and guide continuous improvement.

6. Continuous Monitoring and Retraining:

- Continuously monitor model performance and update it with new data to maintain accuracy and relevance. Incremental learning techniques will be used to allow the model to learn from new data without being retrained from scratch.

The `nxt_move` variable exhibited a distribution centered around zero with substantial outliers and slight positive skewness, highlighting the potential for significant market movements in either direction. This variability underscores the importance of developing robust predictive models capable of handling a broad range of possible outcomes.

The current dataset, although limited to seven trading days, provides valuable insights that are expected to hold as the dataset expands. When modeling begins with a larger dataset, we anticipate the correlation values to be similar to those observed in this EDA phase. The findings from this analysis will inform the feature engineering, model selection, and optimization processes, contributing to the development of a sophisticated trading system aimed at making informed and effective trading decisions in the financial markets.

References

1. Manveer Kaur Mangat, Erhard Reschenhofer, Thomas Stark, Christian Zwatz. High-Frequency Trading with Machine Learning Algorithms and Limit Order Book Data. Data Science in Finance and Economics, 2022, 2(4): 437-463. ([Link](#))

2. Arévalo, Andrés & Nino, Jaime & Hernandez, German & Sandoval, Javier. (2016). High-Frequency Trading Strategy Based on Deep Neural Networks. 9773. 424-436. 10.1007/978-3-319-42297-8_40. ([Link](#))