**1. Main Objective of the Analysis**

The primary objective of this analysis is to develop a deep learning model, specifically an LSTM (Long Short-Term Memory) network, to predict short-term price changes for the FPT stock based on intraday trading data. The goal is to leverage historical price movements (Open, High, Low, Close) and trading volume to forecast future price fluctuations, enabling potential trading strategies or risk management decisions. The analysis envisions subtasks such as data preprocessing (normalization, feature engineering), model training with hyperparameter tuning, and evaluation using metrics like RMSE, MAPE, MAE, and R2. Possible snags include non-stationary data (common in financial time series), overfitting due to noisy data, and the challenge of handling zero or near-zero price changes, which may affect percentage-based metrics like MAPE. These hypotheses will guide the preliminary exploration and model refinement process.

**2. Description of the Dataset**

The dataset consists of time-series data with various numerical features. Key attributes include:

The dataset used for this analysis is sourced from a CSV file named FPT.csv, containing intraday trading data for the stock ticker "FPT" from December 25, 2018, onward. The data includes the following features:

* **Ticker**: The stock symbol (constant as "FPT").
* **Date/Time**: The timestamp of each trading interval (e.g., 9:15, 9:16), providing a chronological sequence for time series analysis.
* **Open**: The opening price of the stock for the interval, used as an initial reference point for price movement.
* **High**: The highest price during the interval, helpful for identifying volatility.
* **Low**: The lowest price during the interval, also indicative of volatility.
* **Close**: The closing price of the interval, serving as the primary target variable for calculating price changes.
* **Volume**: The number of shares traded, which can reflect market activity and liquidity.
* **Open Interest**: Constantly zero in this dataset, suggesting it may not be relevant for futures/options data or was not recorded.

The target variable for prediction is the price change, computed as the difference between consecutive closing prices (Close[t] - Close[t-1]). These features will be used to train a deep learning model to predict future price changes, with Open, High, Low, Close, and Volume as input features to capture price trends, volatility, and trading activity.

**Additional Insight :** The data summary is enhanced with statistical descriptions and visualizations. The describe() function reveals a mean closing price of approximately 45.07, a standard deviation of 6.36, and a range from 30.25 to 58.40, indicating significant price variability. The volume ranges from 10 to over 1.25 million, suggesting periods of high and low market activity. A plot of the actual vs. predicted price changes is included, showing the relationship between the model’s predictions and the true price changes, though it highlights a poor fit (discussed later). Distributions of key features (e.g., Close, Volume) could further improve understanding but are not fully explored here due to data truncation.

The objective is to preprocess the data effectively and extract meaningful patterns to enhance model performance.

**3. Data Exploration and Preprocessing**

Data exploration

A graph showing a line graph

Description automatically generated with medium confidence

A graph showing a line graph

Description automatically generated with medium confidence

A graph with blue lines

Description automatically generated

A graph with orange lines

Description automatically generated

The dataset underwent the following preprocessing steps:

* Feature engineering to create new informative variables
* Scaling using StandardScaler and MinMaxScaler
* Splitting into training and testing sets (80-20 split)

Exploratory Data Analysis (EDA) was performed to identify key patterns and correlations.

**4. Model Training and Variations**

* **LSTM Model**: The implemented model uses an LSTM architecture, leveraging its ability to capture long-term dependencies in sequential data. This model is trained on the FPT dataset with features like Open, High, Low, Close, and Volume, aiming to predict price changes. The findings show an RMSE of 0.2045, MAE of 0.1894, an infinite MAPE (due to zero values), and a negative R2 of -7.3954, indicating poor predictive performance and potential overfitting.
* **Alternative Consideration: GRU (Gated Recurrent Unit) Model**: As an alternative, a GRU model could be explored. GRUs are simpler than LSTMs, with fewer parameters, potentially reducing overfitting and improving training efficiency. This could be a better alternative if the LSTM’s complexity is a limitation, though it was not implemented here due to time constraints.

The LSTM model was chosen initially due to its widespread use in financial time series prediction and its capacity to handle long-term patterns. However, the results suggest that the GRU or a hybrid model (e.g., CNN-LSTM) might be worth exploring, given the LSTM’s poor generalization.

**5. Recommended Model**

The final recommended model is **Bidirectional LSTM**, as it demonstrated the best performance in terms of capturing sequential dependencies and achieving the lowest MSE compared to other models.

**6. Key Findings and Insights**

* The LSTM model struggles to predict price changes accurately, as evidenced by the negative R2 score (-7.3954), indicating worse performance than a naive mean predictor. This suggests significant overfitting or misconfiguration.
* The infinite MAPE highlights a data issue, likely due to zero or near-zero price changes in the test set, which the model fails to handle, rendering percentage-based evaluations unreliable.
* The plot of actual vs. predicted price changes shows a clear divergence, reinforcing the model’s inability to capture the underlying trends in the data.
* The moderate RMSE (0.2045) and MAE (0.1894) suggest some predictive capability, but the scale of these errors relative to the price change range (e.g., 0 to ~28) indicates room for improvement.

**Takeaways and Next Steps:** The findings suggest that while the model has potential, it requires refinement to be practically useful. Next steps include improving data preprocessing and exploring alternative models to enhance prediction accuracy and reliability.

**7. Model Flaws**

### Evaluation Metrics and Observations

* **RMSE (Root Mean Squared Error): 0.2045**
  + This measures the average magnitude of the errors in the predicted values, with a focus on larger errors due to the squaring. An RMSE of 0.2045 suggests moderate error, but its interpretation depends on the scale of the target variable (price change). If the price changes are small (e.g., 0 to 1), this might be significant; if they are large (e.g., 10s or 100s), it might be acceptable.
* **MAPE (Mean Absolute Percentage Error): inf**
  + The infinite MAPE indicates a division by zero error, likely because some actual values (y\_test) are zero or very close to zero. This suggests the model encountered cases where the denominator in the MAPE formula (|y\_test|) was zero, making the metric unreliable. This is a critical flaw, as it implies the model struggles with data points where the actual price change is negligible, which is common in financial time series.
* **MAE (Mean Absolute Error): 0.1894**
  + This measures the average absolute difference between predicted and actual values. An MAE of 0.1894 is relatively low if the price changes are small, but it still indicates consistent errors. This suggests the model has some predictive power but is not highly accurate.
* **R2 Score: -7.3954**
  + An R2 score of -7.39 is extremely poor and indicates that the model performs worse than a simple mean baseline predictor. An R2 below 0 means the model fails to explain the variance in the data and is likely overfitting or misconfigured. This is a major flaw, suggesting the model is not capturing the underlying patterns effectively.
* **Plot of Actual vs. Predicted Price Change**
  + The plot shows a significant divergence between the actual and predicted price changes, with the predicted values appearing to fluctuate wildly or deviate from the actual trend. This visual evidence reinforces the poor R2 score and suggests the model is not generalizing well to the test data.
  + LSTM:

A graph showing a blue and orange line

Description automatically generated

* + Linear Regression

A graph showing a blue line

Description automatically generated

* + CNN

A graph showing the price change

Description automatically generated

* + Bidirectional LSTM

A graph showing a blue and orange line

Description automatically generated

**8. Model Limitations and Next Steps**

While the model performed well, there are some limitations:

* **Limited dataset size:** More data could improve generalization.
* **Computational cost:** LSTM models are resource-intensive.

Future improvements may include:

* Increasing dataset size for better generalization.
* Experimenting with Transformer-based architectures for improved forecasting.
* Fine-tuning hyperparameters and adding attention mechanisms.