Prediction with Boosting Machine Learning

A Case Study on Exxon Mobil Corp (XOM)

Nijat Hasanli
Daryush Ray

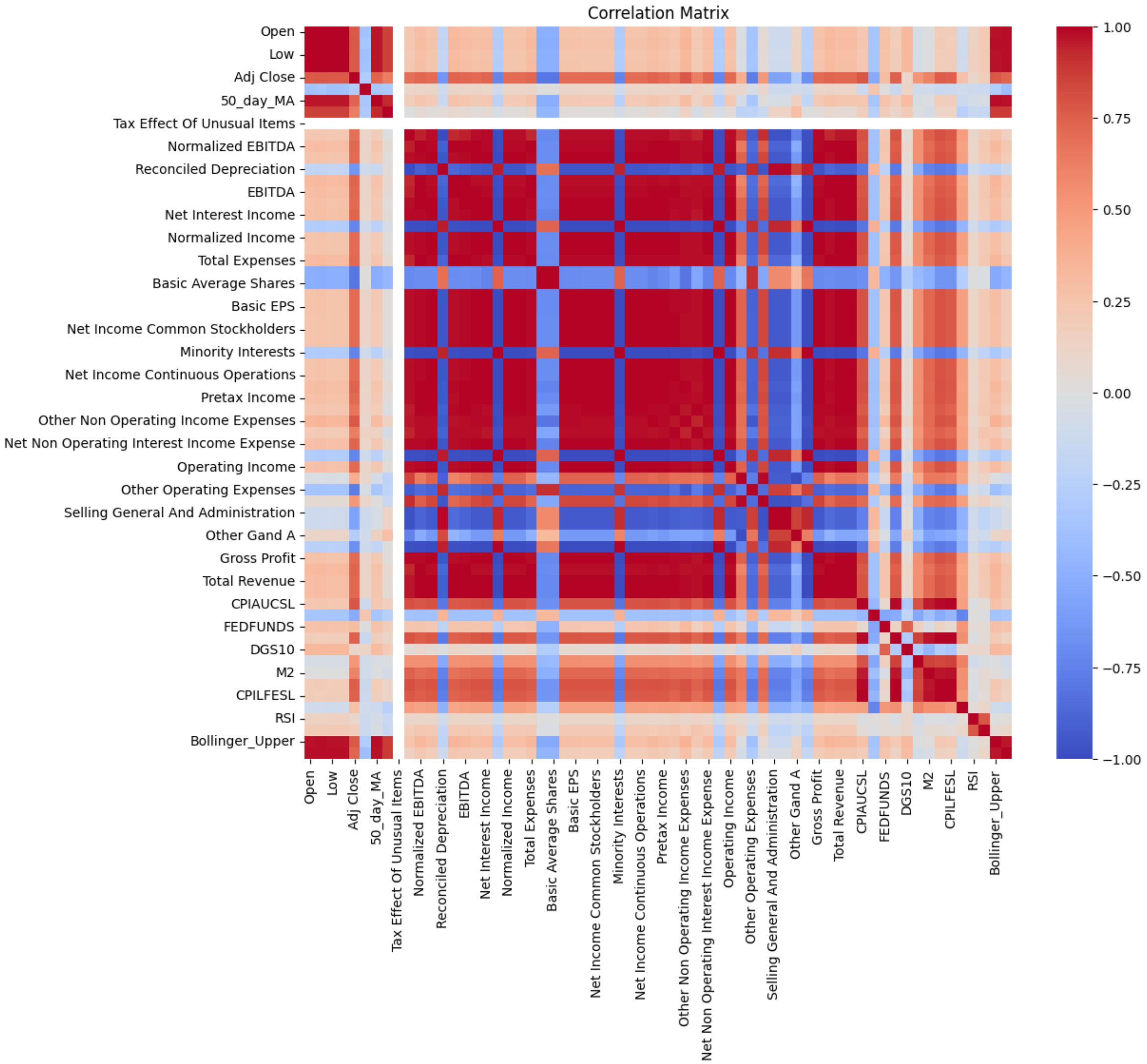
In this presentation, we will explore the methodologies and techniques used in stock price prediction, using **Exxon Mobil Corporation** as a case study. **Exxon Mobil**, a leading multinational oil and gas corporation, provides an ideal subject due to its significant market presence and the complexity of factors influencing its stock prices.

- **Global Presence:** Exxon Mobil operates in over 70 countries
- **Diversified Operations:** The company is involved in all aspects of the oil and gas industry, including upstream, downstream, and chemical manufacturing.
- **Financial Strength:** Exxon Mobil is known for its strong financial performance, with significant revenues, profits, and a solid balance sheet.

Data Collection and Preprocessing

- **Time Period:** January 1990 - Present
  - **Train:** January 1990 - December 2022
  - **Test:** January 2023 - Present
- **Data Sources:** Yahoo Finance, FRED
- **Data Types:**
  - Stock Prices (Open, Close, High, Low)
  - Moving Averages (50-day, 200-day)
  - Financial Statements (Income, Expenses, etc.)
  - Macroeconomic Indicators (CPI, Unemployment Rate, etc.)
- **Handling missing values**
- **Merging datasets**
- **Technical indicators (RSI, MACD, Bollinger Bands)**
- **Scaling features**

- **Technical Indicators:**
    - 50-Day Moving Average
    - 200-Day Moving Average
    - Relative Strength Index
    - Moving Average Convergence Divergence
    - Upper Bollinger Band
    - Lower Bollinger Band
- **Financial Data:**
    - Open:
    - Low:
    - Tax Effect Of Unusual Items
    - Normalized EBITDA
    - Reconciled Depreciation
    - EBITDA
    - Net Interest Income
    - Normalized Income
    - Total Expenses
    - Basic Average Shares
    - Basic EPS: Basic Earnings Per Share
    - Net Income Common Stockholders
    - Minority Interests
    - Net Income Continuous Operations
    - Pretax Income
    - Other Non Operating Income Expenses
    - Net Non Operating Interest Income Expense
    - Operating Income
    - Other Operating Expenses
    - Selling General And Administration
    - Other Grand A
    - Gross Profit
    - Total Revenue
- **Macroeconomic Variables:**
    - Consumer Price Index for All Urban Consumers
    - Effective Federal Funds Rate
    - 10-Year Treasury Constant Maturity Rate
    - M2 Money Stock
    - Consumer Price Index for All Urban Consumers: All Items Less Food & Energy
    - Personal Savings Rate
    - Durable Goods Orders

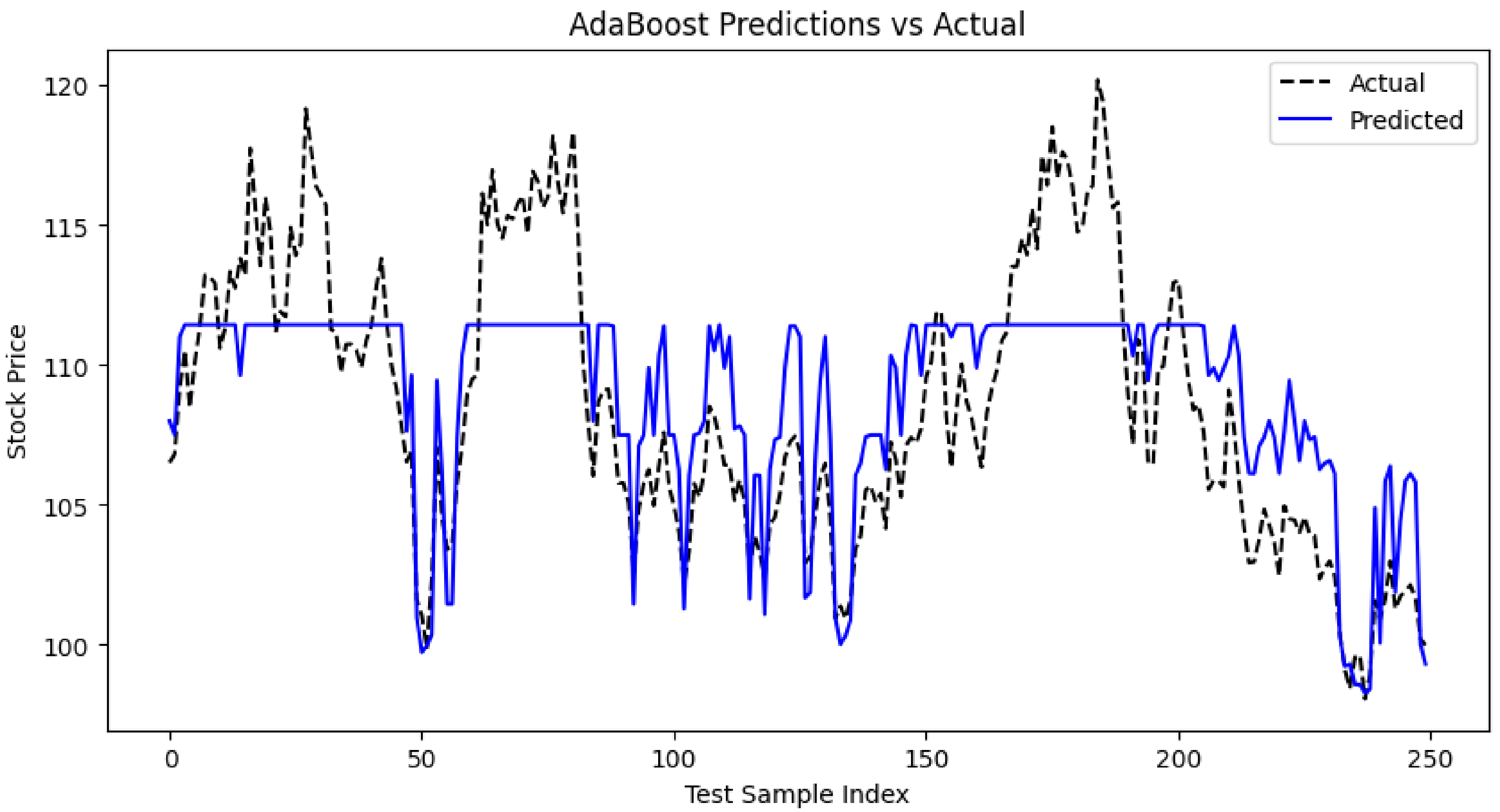


- **Highly Positive Correlations:**
  - **Open, Low, Adj Close:** These price features are highly correlated with each other.
  - **Total Revenue and Gross Profit:** High correlation indicates that total revenue directly impacts gross profit.
  - **Basic EPS and Net Income Common Stockholders:** Suggests that earnings per share are highly influenced by net income attributed to common stockholders.
- **Highly Negative Correlations:**
  - **CPIAUCSL (Consumer Price Index) and Stock Prices:** High CPI could be negatively correlated with stock prices, indicating inflation impacts.
  - **FEDFUNDS (Federal Funds Rate) and Stock Prices:** Higher interest rates might negatively impact stock prices, as borrowing costs increase.
- **Feature Importance:**
  - **Total Revenue, Gross Profit, Basic EPS, Net Income:** Strong correlations with target variables suggest these are important predictors for the stock price.

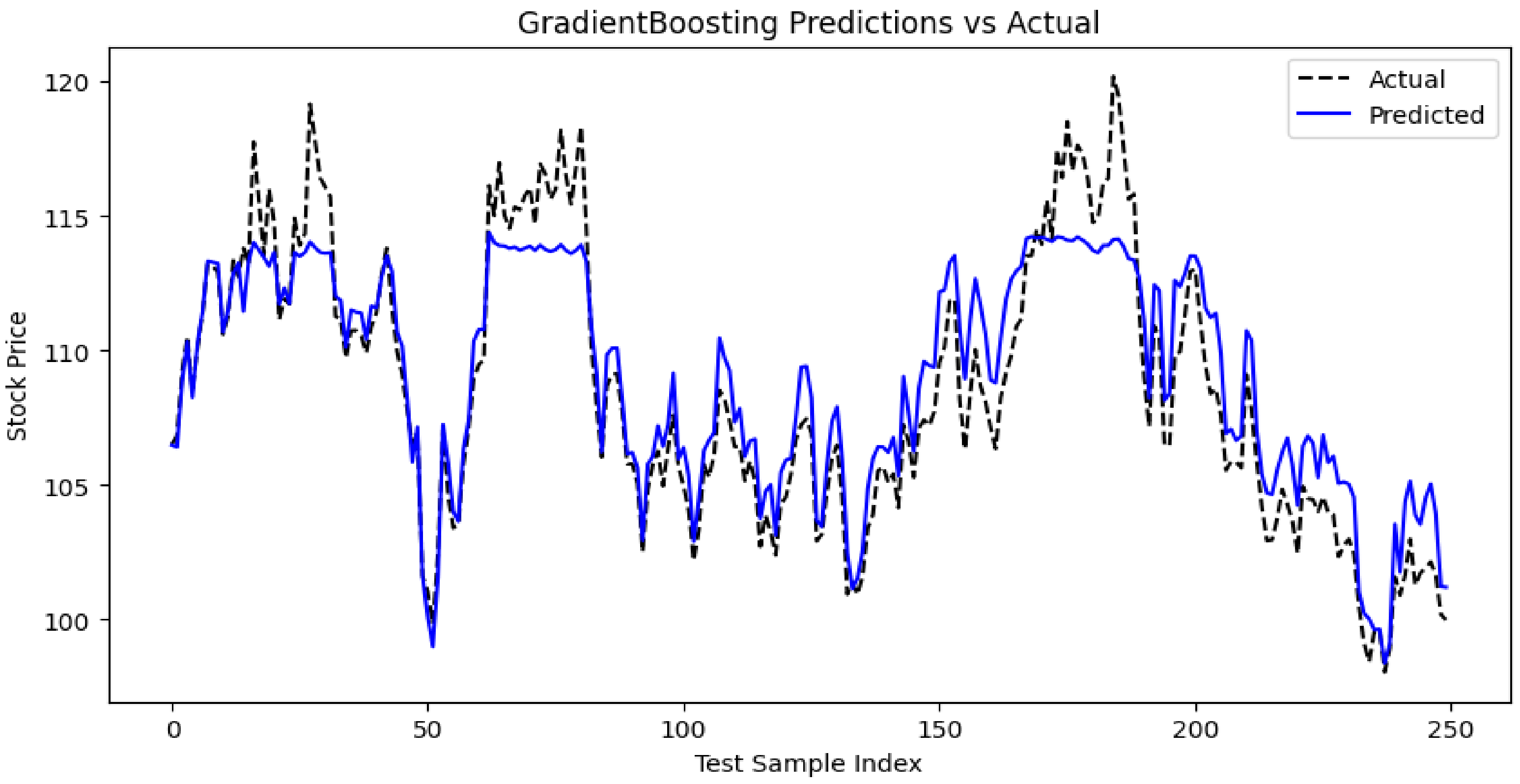
Machine Learning Models

- **AdaBoost:** Combines multiple weak learners to create a strong predictive model by focusing on errors of previous models.
- **Gradient Boosting:** Sequentially builds models by correcting errors of previous models using gradient descent optimization.
- **XGBoost:** An optimized version of Gradient Boosting designed for speed and performance, especially with large datasets.
- **LightGBM:** A highly efficient Gradient Boosting framework that uses a leaf-wise tree growth algorithm for faster training.
- **CatBoost:** Handles categorical features automatically and reduces overfitting through ordered boosting.
- **Hyperparameter Tuning:** Grid Search with Cross-Validation

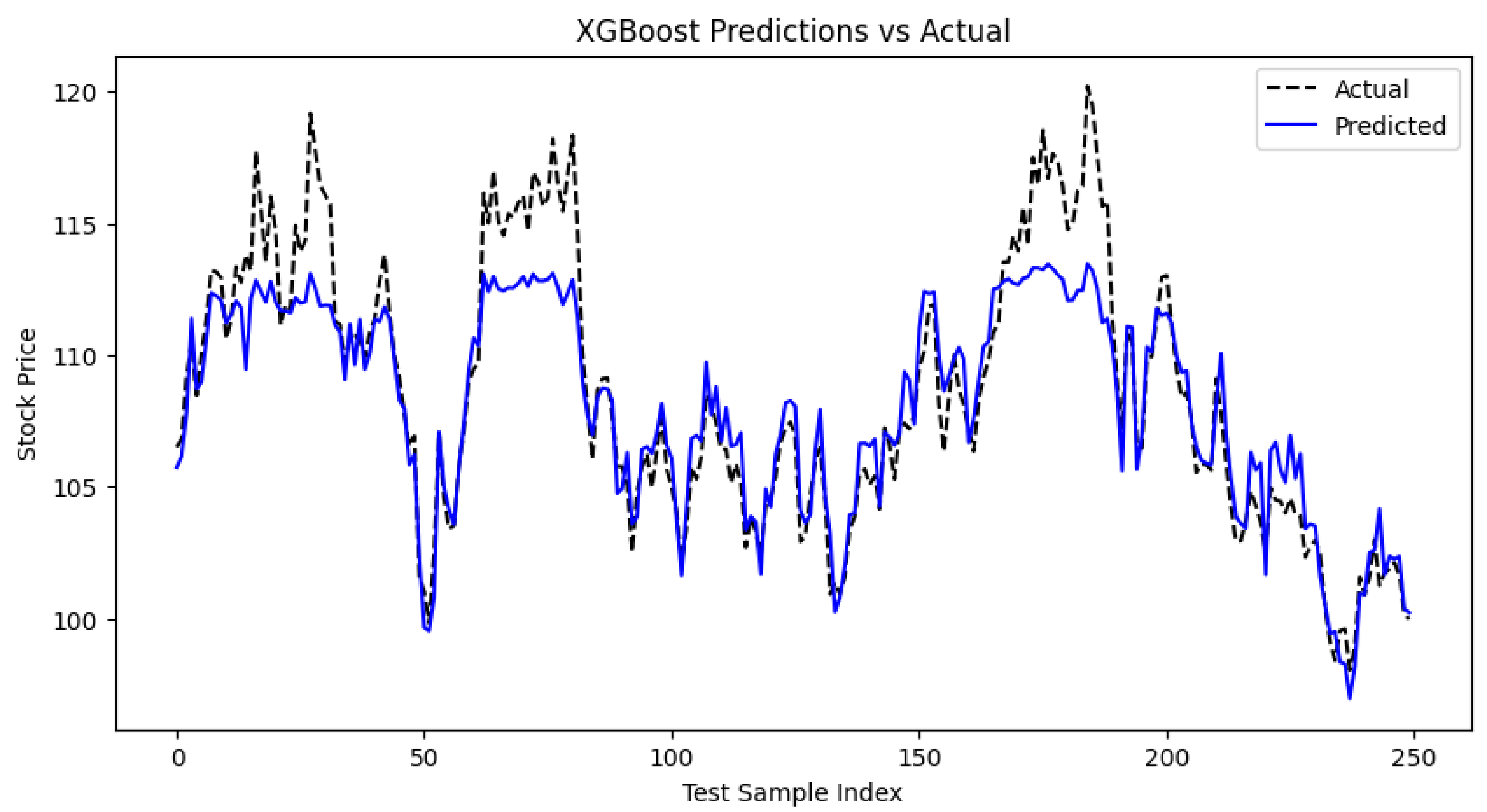
Model Performance and Predictions



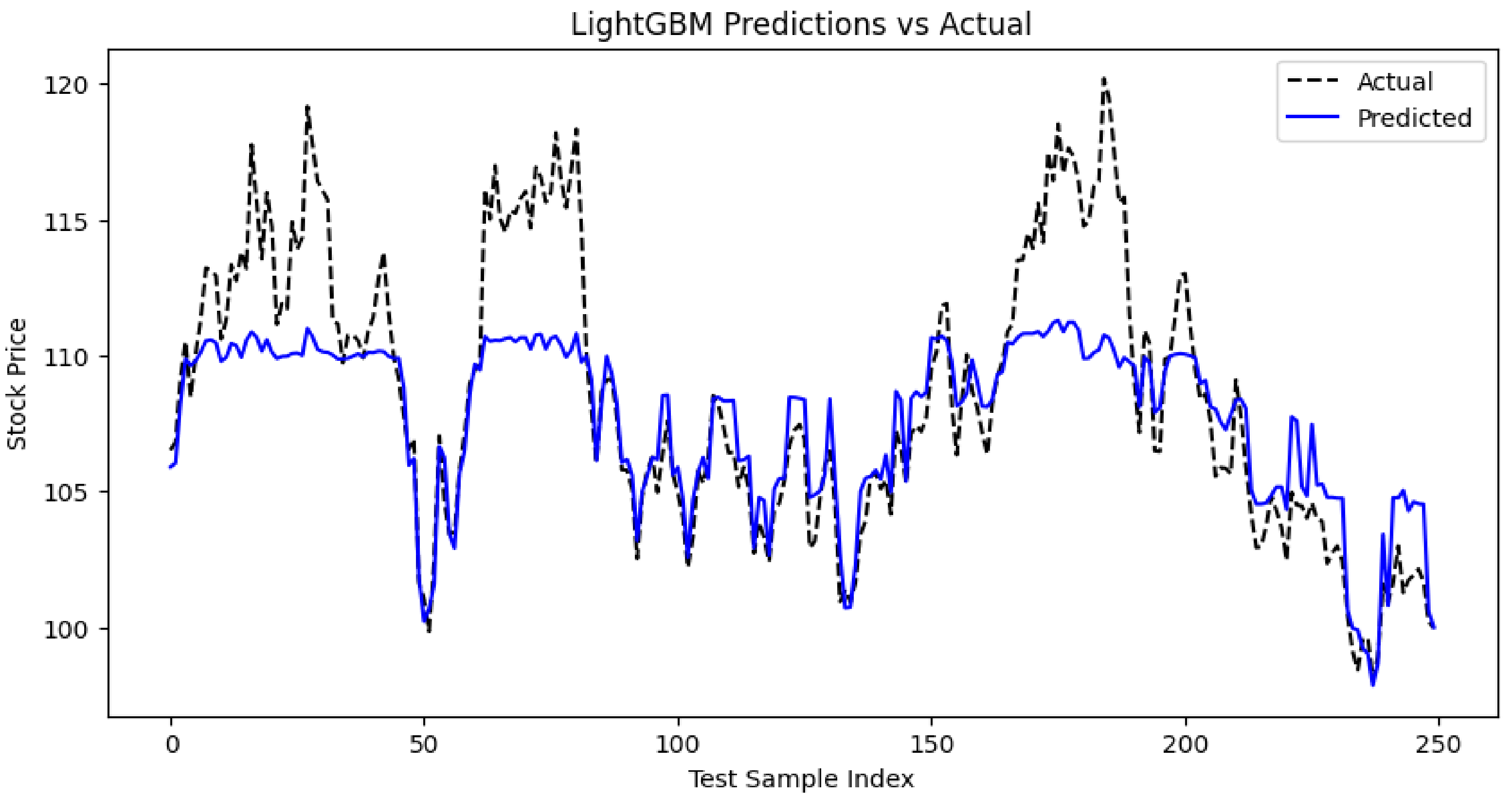
- The model captures some of the general trends in the stock price movements but appears to miss many of the smaller fluctuations and rapid changes.
- The smoother prediction line suggests that the AdaBoost model might be over-smoothing or not fully capturing the complexity of the stock price variations.



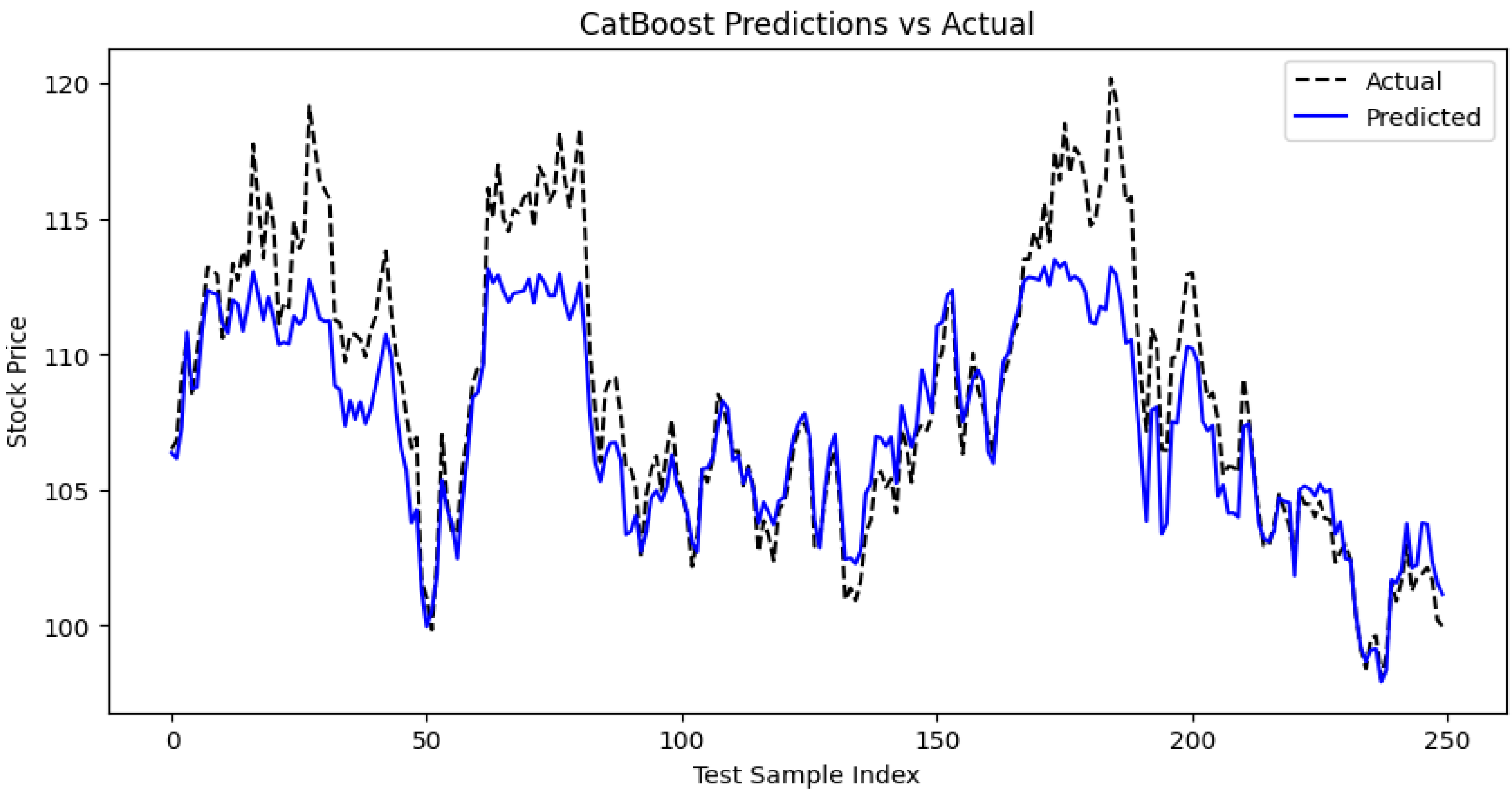
- The Gradient Boosting model provides a more accurate and detailed prediction of stock prices compared to the AdaBoost model.
- The closer fit to actual prices suggests that Gradient Boosting is more effective in capturing the complexities of stock price movements.



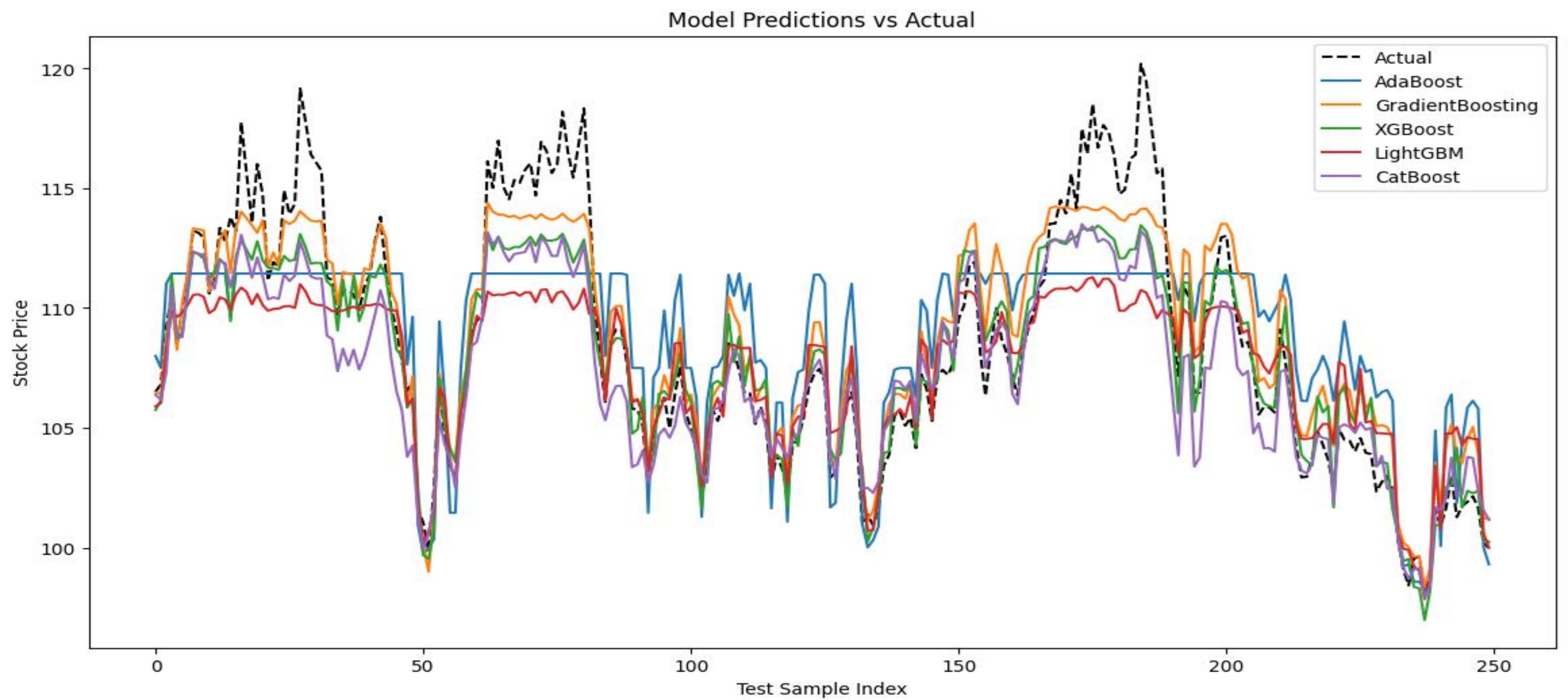
- The XGBoost model provides highly accurate and detailed predictions of stock prices, closely aligning with actual observed prices.
- The model's ability to capture the complexity of stock price movements suggests it is well-suited for applications requiring precise short-term predictions.



- The LightGBM model provides a reasonably accurate prediction of stock prices, capturing the general trends and some of the variability.
- The smoother predictions suggest that the model might be averaging out some of the noise, potentially sacrificing some accuracy in capturing the rapid fluctuations.



- The CatBoost model provides highly accurate and detailed predictions of stock prices, closely aligning with actual observed prices.
- The model's ability to capture both overall trends and finer details suggests it is well-suited for applications requiring precise short-term predictions.




```
Model Performance:
AdaBoost MSE: 10.25611430129391, Best Params: {'learning_rate': 0.1, 'n_estimators': 200}
Gradient Boosting MSE: 3.2661374397622964, Best Params: {'learning_rate': 0.1, 'max_depth': 5, 'min_samples_split': 10, 'n_estimators': 200}
XGBoost MSE: 3.9070729759437963, Best Params: {'colsample_bytree': 0.8, 'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.8}
LightGBM MSE: 8.693740382984231, Best Params: {'learning_rate': 0.2, 'max_depth': 10, 'min_child_samples': 20, 'n_estimators': 50, 'num_leaves': 100}
CatBoost MSE: 5.212198976300194, Best Params: {'depth': 3, 'iterations': 1000, 'learning_rate': 0.1}
```

Conclusion

- **Model Variety:** AdaBoost, Gradient Boosting, XGBoost, LightGBM, and CatBoost.
- **Hyperparameter Tuning:** Optimized model performance through extensive hyperparameter tuning using Grid Search with Cross-Validation.
- **Best Performing Model:** Gradient Boosting achieved the lowest mean squared error, indicating superior predictive accuracy.
- **Model Comparison:** Each model's performance was evaluated, revealing strengths and weaknesses in different market conditions.
- **Practical Implications:** Accurate predictions can aid investors in making informed decisions and enhance trading strategies.

- **Future Enhancements:**
  - Potential for improving models by incorporating more granular data and exploring alternative algorithms.
  - Introducing baseline models.
  - Complex grid search to avoid overfitting
  - Multicollinearity