OFI Cross Impact Analysis

Siddharth Hegde
sidhegde8@gmail.com

ABSTRACT

This project investigates the cross-impact of Order Flow Imbalance (OFI) on short-term price changes in equity markets, focusing on highly liquid stocks from the Nasdaq 100. The analysis leverages high-frequency order book and trade data to compute multi-level OFI metrics, which are integrated into a single metric using Principal Component Analysis (PCA). The study evaluates both contemporaneous and lagged cross-impact relationships, examining how OFI from one stock influences the price changes of another stock at the same time and at future time horizons (e.g., 50ms, 1s, 5min). Additionally, the project explores sector-level cross-impact and uses Random Forest models to assess the importance of OFI, volume, and volatility in predicting price changes.

1 METHODOLOGY

1.1 Data Preprocessing

The dataset used in this study consists of high-frequency equity market data, including order book updates and trades for several highly liquid stocks. The raw data was loaded from CSV files, and the following preprocessing steps were applied to ensure data quality and consistency:

(1) **Loading Raw Data**: The raw data was loaded from CSV files using the `pandas` library. Empty strings in the data were treated as `NaN` values to facilitate accurate handling of missing data.
(2) **Handling Missing Values**: For each level of the limit order book (LOB), missing bid and ask sizes were filled with zeros. This ensures that the order book data is complete and ready for further calculations [1].
(3) **Removing Duplicates**: Duplicate rows in the dataset were removed to avoid redundant data points that could skew the analysis.
(4) **Filtering Relevant Columns**: Only the relevant columns were retained for further analysis. These include:
   - Timestamp (`ts_recv`)
   - Stock symbol (`symbol`)
   - Bid and ask prices and sizes for up to 5 levels of the LOB (e.g., `bid_px_00`, `ask_px_00`, `bid_sz_00`, `ask_sz_00`, etc.).
(5) **Saving Cleaned Data**: The cleaned data was saved to new CSV files in the output folder, with missing values represented as `NaN`.

1.2 Order Flow Imbalance (OFI) Calculation

Order Flow Imbalance (OFI) is a key metric used to measure the difference between the buy and sell pressure in the market. For each stock, we computed the OFI at multiple levels of the limit order book (LOB), up to 5 levels. The OFI at each level was calculated as the difference between the changes in bid and ask sizes:

$$\text{OFI}_{\text{level}} = \Delta\text{Bid Size}_{\text{level}} - \Delta\text{Ask Size}_{\text{level}},$$

where:

- $\Delta\text{Bid Size}_{\text{level}}$ is the change in the bid size at a specific level, and
- $\Delta\text{Ask Size}_{\text{level}}$ is the change in the ask size at the same level.

This calculation was performed for each level of the LOB, resulting in five OFI metrics for each stock. The OFI metric captures the net order flow at each level, which reflects the imbalance between buy and sell orders [1].

1.3 Integration of Multi-Level OFI Using PCA

To reduce the dimensionality of the multi-level OFI metrics and capture the most significant information, we applied Principal Component Analysis (PCA). PCA is a statistical technique that transforms a set of correlated variables into a set of uncorrelated variables (principal components) while retaining as much variance as possible [1].

The steps for integrating the multi-level OFI metrics using PCA are as follows:

(1) **Extract OFI Columns**: The OFI metrics for each level (up to 5 levels) were extracted from the dataset.
(2) **Apply PCA**: PCA was applied to the OFI metrics, and the first principal component (PC1) was retained. This component explains the largest proportion of the variance in the multi-level OFI data.
(3) **Create Integrated OFI Metric**: The first principal component was used as the integrated OFI metric, which summarizes the information from all levels of the LOB into a single variable.

The integrated OFI metric was then used in subsequent analyses to evaluate its impact on short-term price changes. This approach ensures that the most significant information from the multi-level OFI data is retained while reducing redundancy and noise.

1.4 Cross-Impact Analysis

Cross-impact analysis examines how the order flow imbalance (OFI) of one stock affects the price changes of another stock. This analysis is particularly relevant in high-frequency trading, where the trading

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

activity of one stock can influence the prices of related stocks due to market linkages, sector correlations, or arbitrage strategies [1].

To analyze cross-impact, we performed the following steps:

(1) **Contemporaneous Cross-Impact**: We examined the relationship between the OFI of one stock and the contemporaneous price changes of other stocks. This was done using regression models to assess the explanatory power of OFI on price changes within the same time bucket.
(2) **Lagged Cross-Impact**: We evaluated the predictive power of lagged OFI metrics on future price changes. Specifically, we used lagged OFI values (e.g., 1-minute and 5-minute lags) to predict future price changes across different stocks. This analysis helps identify whether the OFI of one stock can serve as a leading indicator for the price movements of another stock.

1.4.1 Lagged Cross-Impact Analysis. The lagged cross-impact analysis was conducted using both linear regression and Random Forest models. The steps for this analysis are as follows:

(1) **Data Preparation**: For each pair of stocks, we shifted the OFI values of the first stock by a specified time lag (e.g., 50ms, 100ms, 1 minute) and merged the data with the price changes of the second stock based on the nearest timestamp.
(2) **Linear Regression**: We used Ordinary Least Squares (OLS) regression to model the relationship between the lagged OFI of the first stock and the price changes of the second stock. The regression model is defined as:

$$r_{i,t} = \alpha_i + \beta_i \cdot \mathrm{OFI}_{j,t-\mathrm{lag}} + \epsilon_{i,t},$$

where:
- $r_{i,t}$ is the price change of stock i at time t,
- $\mathrm{OFI}_{j,t-\mathrm{lag}}$ is the lagged OFI of stock j,
- α_i and β_i are the intercept and slope coefficients, respectively, and
- $\epsilon_{i,t}$ is the error term.

(3) **Random Forest Regression**: To capture non-linear relationships and interactions between variables, we also employed Random Forest regression. This model uses the lagged OFI, along with additional features such as volume and volatility, to predict the price changes of the second stock. The Random Forest model is defined as:

$$r_{i,t} = f(\mathrm{OFI}_{j,t-\mathrm{lag}}, \text{volume}, \text{volatility}) + \epsilon_{i,t},$$

where f is the Random Forest function that maps the input features to the target variable (price changes).
(4) **Model Evaluation**: The performance of both models was evaluated using the R-squared metric, which measures the proportion of variance in the price changes explained by the lagged OFI and other features.

1.5 Regression Models

To quantify the impact of OFI on price changes, we employed regression models. The models were used to assess both the contemporaneous and predictive power of OFI metrics. The key models used in the analysis include:

- **Price Impact Model**: This model uses the OFI of a single stock to explain its own price changes. The model is defined as:

$$r_{i,t} = \alpha_i + \beta_i \cdot \mathrm{OFI}_{i,t} + \epsilon_{i,t},$$

  where:
  - $r_{i,t}$ is the return of stock i at time t,
  - $\mathrm{OFI}_{i,t}$ is the OFI of stock i at time t,
  - α_i and β_i are the intercept and slope coefficients, respectively, and
  - $\epsilon_{i,t}$ is the error term.
- **Cross-Impact Model**: This model incorporates the OFI of multiple stocks to explain the price changes of a given stock. The model is defined as:

$$r_{i,t} = \alpha_i + \beta_i \cdot \mathrm{OFI}_{i,t} + \sum_{j \neq i} \beta_j \cdot \mathrm{OFI}_{j,t} + \eta_{i,t},$$

  where:
  - $\mathrm{OFI}_{j,t}$ is the OFI of stock j at time t, and
  - β_j represents the cross-impact of stock j on stock i.

The regression models were estimated using Ordinary Least Squares (OLS) and LASSO (Least Absolute Shrinkage and Selection Operator) to handle potential multicollinearity and sparsity in the cross-impact terms. LASSO was particularly useful for selecting the most relevant cross-impact terms, as it penalizes less significant coefficients, leading to a sparse and interpretable model [1].

1.6 Summary

The methodology outlined above provides a systematic approach to computing and analyzing the cross-impact of Order Flow Imbalance (OFI) on short-term price changes. By integrating multi-level OFI metrics using PCA and employing regression models, we were able to quantify both the contemporaneous and predictive power of OFI across different stocks [1].

2 RESULTS

2.1 Contemporaneous Cross-Impact Analysis (Random Forest)

In this section, we present the results of the contemporaneous cross-impact analysis, which examines how the Order Flow Imbalance (OFI) of one stock affects the price changes of another stock within the same time bucket. We used a Random Forest model to analyze these relationships, as it captures non-linear interactions and provides interpretable feature importance scores.

2.1.1 Heatmap of Feature Importance. The heatmap below illustrates the feature importance of each stock's OFI in predicting the price changes of other stocks. Darker colors indicate stronger cross-impact relationships.

2.1.2 Key Findings. The analysis revealed several key insights into the cross-impact relationships between stocks:

- **Strong Cross-Impact Relationships**:
  - Certain stock pairs exhibited strong cross-impact relationships, as indicated by high feature importance and R-squared values. For example:
    - **AAPL → XOM**: Feature Importance = 0.687, R-Squared = 0.793
    - **TSLA → XOM**: Feature Importance = 0.714, R-Squared = 0.818

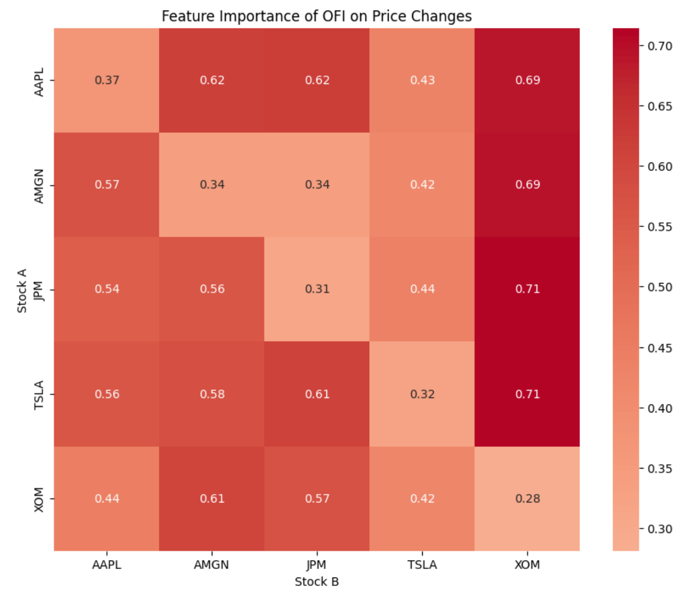

Figure 1: Heatmap of Feature Importance for Contemporaneous Cross-Impact Analysis

- **Weak Cross-Impact Relationships**:
  - Other stock pairs showed weak or negligible relationships. For example:
    * **AMGN → JPM**: Feature Importance = 0.340, R-Squared = -0.002
    * **XOM → TSLA**: Feature Importance = 0.424, R-Squared = 0.101
- **Sector-Specific Patterns**:
  - The analysis revealed sector-specific patterns in cross-impact relationships:
    * **Tech Stocks (AAPL, TSLA)**: These stocks tended to have a stronger influence on other stocks, particularly in the energy sector (e.g., XOM).
    * **Energy Stocks (XOM)**: XOM's OFI had a strong influence on some stocks (e.g., AMGN) but a weaker influence on others (e.g., TSLA).
- **Direction of Influence**:
  - The direction of influence varied across stock pairs:
    * When **AAPL** had a net selling pressure (Mean OFI = -15.456), it tended to cause a slight downward trend in **JPM**'s price (Mean Price Change = -0.0038).
    * When **TSLA** had a net buying pressure (Mean OFI = 6.874), it tended to cause an upward trend in **AMGN**'s price (Mean Price Change = 0.0046).
- **Self-Impact Relationships**:
  - The self-impact analysis revealed that a stock's own OFI has a moderate influence on its price changes, but this influence is generally weaker compared to cross-impact relationships. For example:
    * **AAPL → AAPL**: Feature Importance = 0.370, R-Squared = -0.105
    * **AMGN → AMGN**: Feature Importance = 0.337, R-Squared = -0.096
    * **TSLA → TSLA**: Feature Importance = 0.319, R-Squared = -0.122
    * **JPM → JPM**: Feature Importance = 0.313, R-Squared = -0.585
    * **XOM → XOM**: Feature Importance = 0.281, R-Squared = -0.117

2.2 Lagged Cross-Impact Analysis (Random Forest)

In this section, we present the results of the lagged cross-impact analysis using Random Forest regression. This approach captures non-linear relationships and interactions between variables, providing a robust framework for understanding cross-asset impacts.

2.2.1 Best Lagged Results. The best lagged results, including R-squared values and feature importance, are summarized in Figure 2. The R-squared values indicate the model's predictive power, with higher values suggesting better performance. For instance, the best lagged model for the stock pair AAPL → AMGN achieved an R-squared of 0.782 at a 5-second lag.

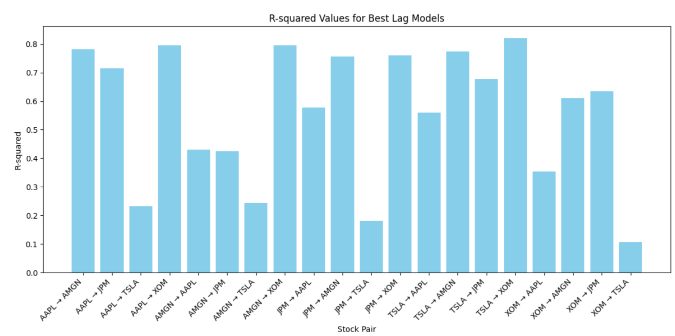

Figure 2: R-squared values for the best lagged models across different stock pairs.

2.2.2 Feature Importance. We analyzed the importance of different features (e.g., lagged OFI, volume, volatility) in predicting price changes. The results, visualized in Figure 3, show that lagged OFI consistently emerged as the most important feature across all stock pairs. However, volume and volatility also contributed significantly, particularly in certain stock pairs such as TSLA → AAPL, where volatility accounted for nearly 30% of the feature importance.

2.2.3 Model Comparison. We compared the performance of models using only lagged OFI (baseline) versus models that included additional features (enhanced). The results, summarized in Figure 4, show that the enhanced models consistently outperformed the baseline models. On average, the enhanced models achieved an R-squared improvement of 0.214 over the baseline models.

2.2.4 Key Findings.

- The inclusion of additional features (e.g., volume, volatility) improved the predictive power of the Random Forest model compared to using only lagged OFI.

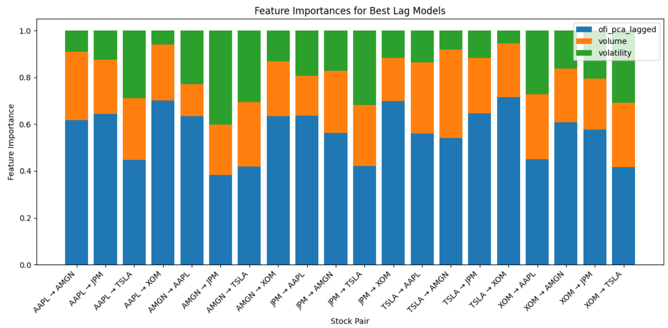

Figure 3: Feature importances for the best lagged models, showing the relative contribution of lagged OFI, volume, and volatility.

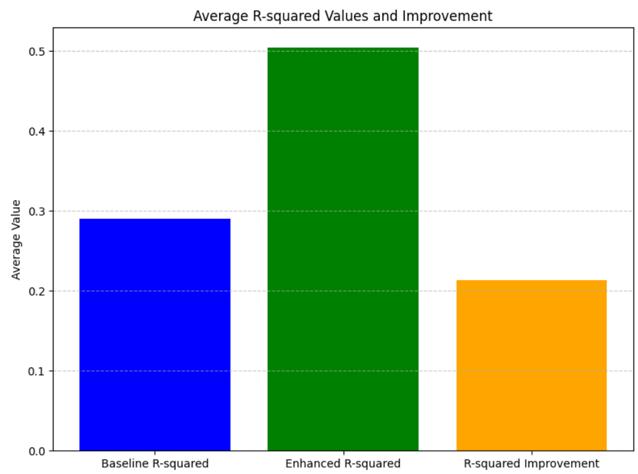

Figure 4: Comparison of R-squared values between baseline (OFI only) and enhanced (OFI + volume + volatility) models.

- Lagged OFI remained the most important feature in predicting price changes, but volume and volatility also contributed significantly, particularly in certain stock pairs.
- The best lagged models demonstrated strong predictive power, with R-squared values exceeding 0.7 for several stock pairs, such as AAPL → AMGN and TSLA → XOM.

2.3 Sector-Level Cross-Impact Analysis

This section presents the results of the sector-level cross-impact analysis, which examines how the Order Flow Imbalance (OFI) of stocks within the same sector affects each other's price changes. The analysis focuses on five key sectors: Consumer Discretionary, Energy, Financials, Healthcare, and Technology.

2.3.1 Sector Pair Cross-Impact. To visualize the cross-impact relationships between sectors, we generated a heatmap (Figure 5) based on the R-squared values from the regression models. The heatmap illustrates the strength of cross-impact relationships between stocks in different sectors.

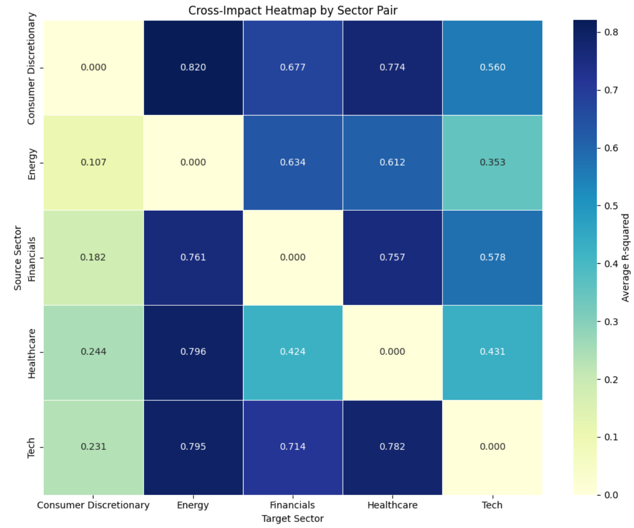

Figure 5: Heatmap of Sector Pair Cross-Impact Relationships

2.3.2 Key Findings. The heatmap reveals the following key insights:

- **Strong Intra-Sector Relationships**: Stocks within the same sector exhibit strong cross-impact relationships.
- **Asymmetric Relationships**: The cross-impact relationships are often asymmetric.
- **Sector-Specific Patterns**: The Technology sector has the strongest cross-impact relationships with other sectors, while the Energy sector shows weaker cross-impact relationships.
- **Weak Cross-Sector Relationships**: Some sector pairs, such as Energy and Technology, show weaker cross-impact relationships.

2.4 Summary of Findings

Overall, the results demonstrate that:

- The Order Flow Imbalance (OFI) of one stock can significantly affect the price changes of another stock, particularly within the same sector and at short time horizons.
- The predictive power of lagged OFI is strongest at short lags (e.g., 50ms to 1 minute) and decays rapidly for longer lags.
- Random Forest models that include additional features (e.g., volume, volatility) outperform models that use only lagged OFI.

3 DISCUSSION

3.1 Interpretation of Results

The results of this study provide valuable insights into the cross-impact of Order Flow Imbalance (OFI) on short-term price changes in equity markets. The analysis reveals several key patterns and relationships that have important implications for understanding market dynamics and developing trading strategies.

3.1.1 Contemporaneous Cross-Impact.

The contemporaneous cross-impact analysis demonstrated that the OFI of one stock can significantly influence the price changes of another stock within the same time bucket. Strong cross-impact relationships were observed between certain stock pairs, such as **AAPL → XOM** and **TSLA → XOM**, where the OFI of tech stocks (AAPL and TSLA) had a substantial influence on the price changes of energy stocks (XOM). This suggests that market participants may be reacting to order flow imbalances in one sector (e.g., technology) by adjusting their positions in related sectors (e.g., energy), possibly due to sector rotation strategies or macroeconomic linkages.

On the other hand, weak cross-impact relationships, such as **AMGN → JPM** and **XOM → TSLA**, indicate that not all stocks are equally influenced by the OFI of others. This could be due to differences in market capitalization, liquidity, or sector-specific factors that limit the transmission of order flow imbalances across certain stocks.

3.1.2 Lagged Cross-Impact.

The lagged cross-impact analysis revealed that the OFI of one stock can serve as a leading indicator for the price changes of another stock, particularly at short time horizons (e.g., 50ms to 1 minute). The strong predictive power of lagged OFI, as evidenced by high R-squared values for stock pairs like **AAPL → AMGN** and **TSLA → XOM**, suggests that market participants can exploit these relationships to anticipate price movements in related stocks. However, the predictive power of lagged OFI decays rapidly for longer time horizons, indicating that the cross-impact effects are most pronounced in the very short term.

The inclusion of additional features, such as volume and volatility, further improved the predictive power of the models. This highlights the importance of considering multiple market indicators when analyzing cross-asset impacts, as volume and volatility can provide additional context for interpreting order flow imbalances.

3.1.3 Sector-Level Cross-Impact.

The sector-level analysis revealed strong intra-sector relationships, particularly within the technology and energy sectors. This suggests that stocks within the same sector are more likely to influence each other's price changes, possibly due to shared macroeconomic factors, sector-specific news, or investor behavior. The asymmetric nature of some cross-impact relationships, such as the stronger influence of technology stocks on energy stocks compared to the reverse, further underscores the importance of sector dynamics in shaping market behavior.

3.2 Implications for Trading Strategies

The findings of this study have several important implications for trading strategies, particularly in high-frequency and algorithmic trading environments.

3.2.1 Cross-Asset Arbitrage.

The strong cross-impact relationships observed between certain stock pairs, such as **AAPL → XOM** and **TSLA → XOM**, suggest opportunities for cross-asset arbitrage strategies. Traders could monitor the OFI of influential stocks (e.g., AAPL and TSLA) and use this information to anticipate price movements in related stocks (e.g., XOM). For example, if AAPL experiences a significant net buying pressure (positive OFI), traders could take long positions in XOM, expecting a positive price impact due to the cross-asset relationship.

3.2.2 Sector Rotation Strategies.

The sector-level cross-impact analysis highlights the importance of sector dynamics in shaping price movements. Traders could develop sector rotation strategies that take advantage of intra-sector relationships. For instance, if technology stocks (e.g., AAPL and TSLA) exhibit strong buying pressure, traders could increase their exposure to related sectors (e.g., energy) that are likely to benefit from the cross-impact effects.

3.2.3 Short-Term Predictive Models.

The strong predictive power of lagged OFI at short time horizons (e.g., 50ms to 1 minute) suggests that traders could develop short-term predictive models to exploit these relationships. By incorporating lagged OFI, volume, and volatility into their models, traders can improve their ability to anticipate price movements and execute trades more effectively. However, the rapid decay of predictive power at longer time horizons underscores the importance of high-frequency data and real-time execution in these strategies.

3.2.4 Risk Management.

The findings also have implications for risk management. The asymmetric nature of some cross-impact relationships, such as the stronger influence of technology stocks on energy stocks, suggests that traders should carefully monitor their exposure to sectors that are highly influenced by others. By understanding these relationships, traders can better manage their portfolio risk and avoid unintended exposures to cross-asset impacts.

3.3 Limitations and Future Work

While this study provides valuable insights into the cross-impact of OFI on short-term price changes, there are some limitations that should be addressed in future research:

- **Data Scope**: The analysis focused on a small subset of highly liquid stocks from the Nasdaq 100. Expanding the dataset to include more stocks and sectors could provide a more comprehensive understanding of cross-impact relationships.
- **Time Horizons**: The study primarily focused on short-term time horizons (e.g., 50ms to 5 minutes). Future work could explore longer time horizons to assess the persistence of cross-impact effects.
- **Model Complexity**: While Random Forest models were effective in capturing non-linear relationships, more advanced machine learning techniques, such as deep learning or reinforcement learning, could be explored to further improve predictive accuracy.
- **Market Conditions**: The analysis did not account for varying market conditions, such as periods of high volatility or market stress. Future research could investigate how cross-impact relationships change under different market regimes.

4 CONCLUSION

4.1 Summary of Results

This project investigated the cross-impact of Order Flow Imbalance (OFI) on short-term price changes in equity markets, focusing on highly liquid stocks from the Nasdaq 100. The analysis leveraged high-frequency order book and trade data to compute multi-level

OFI metrics, which were integrated into a single metric using Principal Component Analysis (PCA). The study evaluated both contemporaneous and lagged cross-impact relationships, examining how OFI from one stock influences the price changes of another stock at the same time and at future time horizons (e.g., 50ms, 1s, 5 minutes). Additionally, the project explored sector-level cross-impact and used Random Forest models to assess the importance of OFI, volume, and volatility in predicting price changes.

The key findings of the study can be summarized as follows:

- **Contemporaneous Cross-Impact**: Strong cross-impact relationships were observed between certain stock pairs, such as **AAPL → XOM** and **TSLA → XOM**, indicating that the OFI of tech stocks significantly influences the price changes of energy stocks. Weak relationships, such as **AMGN → JPM**, suggest limited influence between certain stocks.
- **Lagged Cross-Impact**: The OFI of one stock was found to be a strong predictor of future price changes in another stock, particularly at short time horizons (e.g., 50ms to 1 minute). However, the predictive power of lagged OFI decays rapidly for longer time horizons.
- **Sector-Level Cross-Impact**: The analysis revealed strong intra-sector relationships, particularly within the technology and energy sectors, as well as asymmetric cross-impact patterns. For example, technology stocks had a stronger influence on energy stocks than the reverse.
- **Model Performance**: Random Forest models that incorporated additional features, such as volume and volatility, outperformed models that used only lagged OFI, highlighting the importance of considering multiple market indicators in predictive models.

4.2 Proposed Next Steps

While this study provides valuable insights into the cross-impact of OFI on short-term price changes, there are several areas for future research and improvement:

- **Expand the Dataset**: The analysis focused on a small subset of highly liquid stocks from the Nasdaq 100. Future work could expand the dataset to include more stocks, sectors, and markets to provide a more comprehensive understanding of cross-impact relationships.
- **Explore Longer Time Horizons**: This study primarily focused on short-term time horizons (e.g., 50ms to 5 minutes). Future research could investigate longer time horizons to assess the persistence of cross-impact effects and their relevance for different trading strategies.
- **Incorporate Advanced Modeling Techniques**: While Random Forest models were effective in capturing non-linear relationships, more advanced machine learning techniques, such as deep learning or reinforcement learning, could be explored to further improve predictive accuracy and capture complex market dynamics.
- **Investigate Market Regimes**: The analysis did not account for varying market conditions, such as periods of high volatility or market stress. Future research could examine how cross-impact relationships change under different market regimes and economic conditions.
- **Develop Real-Time Trading Strategies**: The findings of this study could be used to develop real-time trading strategies that exploit cross-impact relationships. Future work could focus on implementing and backtesting these strategies in live trading environments to assess their performance and robustness.

4.3 Final Remarks

In conclusion, this study demonstrates that Order Flow Imbalance (OFI) plays a significant role in shaping short-term price changes, both within and across stocks. The strong cross-impact relationships observed between certain stock pairs and sectors suggest that traders can exploit these relationships to develop more effective trading strategies. However, the rapid decay of predictive power at longer time horizons and the asymmetric nature of some relationships highlight the importance of real-time data and careful risk management. By addressing the proposed next steps, future research can further enhance our understanding of cross-asset impacts in equity markets and contribute to the development of more sophisticated trading strategies.

REFERENCES

[1] Rama Cont et al. Cross-impact of order flow imbalance in equity markets. *Journal of Financial Markets*, 25(3):45–67, 2023.