QF634. Applied Quantitative Research Methods

Professor Lim Kian Guan

Bryan Wilson
Nadezhda Khusnetdinova
Yan Jun

Table of contents

Introduction

Modern hardware and technological progress that enables the execution of tens, if not hundreds, trades per second have led to the emergence and development of high-frequency trading that has been significantly influencing the financial markets ever since. HFT is an algorithmic trading strategy that implies a large number of trades executed within a very short time span (microseconds). The profit then is compiled out of many small deltas of bid-ask spreads, that is why many HFT companies are considered market-makers – they ensure liquidity and narrow the spreads. Machine learning and different statistical algorithms predict these small price movements, and execution of large amounts of orders is enabled by automatisation – generally, the main goal that HFT is aiming to achieve is to be faster than the market. Automatic risk controls are set, too, to prevent catastrophic losses.

This project is trying to gain insights into trading patterns for AAPL stock through the lens of machine learning, employing different methods that will be discussed further in this paper. This project seeks to establish patterns and correlation and find an optimal prediction algorithm. AAPL is one of the major technological companies, which makes it worthy of research interest and enables us to apply predictive analytics. We aim to determine the path towards more informed and conscious trading decisions in terms of HFT, using this renowned company's example.

Project purpose: gain insights into HFT patterns of Apple stock data using predictive machine learning algorithms

Project objective: conduct a comprehensive analysis and establish the most efficient model. As an efficiency measure accuracy was chosen. To achieve the objective, the following will be done:

1. Retrieve and preprocess the underlying data (AAPL stock)
2. Engineer relevant features
3. Develop and implement predictive ML models
4. Tune the parameters and perform cross-assessment of models
5. Compile research data and form conclusion in terms of most efficient HFT strategy for the particular case of AAPL stock

Data

Data source – public data from Wharton Research Data Services (WRDS). Apple stock was selected for the following reasons:

- Market capitalization – Apple is one of the largest companies traded on markets
- Volatility - AAPL is a large-cap stock, but it still has a sufficient level of volatility, and these price movements can be exploited by HFT
- Liquidity - AAPL is one of the most actively traded stocks in the world, thus short timespan trades would not make a significant price impact

- Sensitivity to news - AAPL stock is sensitive to news related to its field, and such events influence price fluctuations that can be capitalized using HFT strategy. There are also certain technology trends that can be exploited.
- Trading volume - AAPL has consistent trades throughout the day, which allows continuous algorithmic trading in the context of HFT.

Time period of data is 5 uncorrelated trading days of 2023 to ensure relevancy and avoid potential autocorrelation. Each day's dataframe is 518 449 lines long, which makes it to circa 10 trades per second on average, which is equivalent to over 2000 years of daily trades.

Practical Challenges and Considerations

There are numerous challenges and considerations associated with high-frequency trading. Below we outlined a few:

Transaction Fees/Costs: Engaging in frequent and rapid trades can accrue significant transaction fees and costs, impacting the overall profitability of HFT strategies.

Impact Costs/Widening Bid-Ask Spread: High-frequency trading activity can contribute to impact costs, influencing the widening of bid-ask spreads. Increased demand (more buyers) may cause the ask price to rise, while increased selling activity (more sellers) can lead to a decrease in the bid price.

Liquidity Risk: Despite the frequency of trading in training and testing using historical data, there is a liquidity risk in real-time scenarios. Liquidity risk arises when it becomes challenging to execute trades as frequently or efficiently as demonstrated in historical data.

Slippages in Market Orders: The rapid execution of market orders in HFT can sometimes result in slippages, where the actual transaction price differs from the expected price. Slippages can occur due to the dynamic nature of market conditions and the speed at which orders are executed.

Literature review

HFT has been a subject of interest and financial research ever since it emerged; in their studies scholars focused mainly on regulatory requirements, technological development and predominant trends. Issues of stability, safety and ethics concerns were also discussed. This literature review sheds light on some of the most prominent scientific works in the field.

In the realm of high-frequency trading (HFT), *"High-frequency trading strategies"* by Goldstein, Kwan, and Philip (2016) provides a comprehensive analysis of HFT strategies and their influence on financial markets. This research particularly focuses on the order book depth imbalance as a critical information channel. The authors find that HFTs effectively utilize this imbalance to

predict short-term future price movements, exhibiting superior performance in volatile market conditions. Their strategies involve trading in the direction of the order book imbalance and adapting their orders when the imbalance shifts, thus playing a dominant role in liquidity provision. Notably, during periods of high volatility, HFTs leverage their speed advantage to execute market orders more successfully and modify or cancel limit orders to mitigate adverse selection costs.

The technologies play a crucial role in HFT-related literature, and such works as *"High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems'' (2010)* by I. Aldridge is one of the fundamental studies that covers statistical and mathematical tools as well as machine learning applications and algo trading strategies. She discusses the ever-growing 'arm race' going on in the HFT field, and this aspect was studied in more detail by Haas, Khapko, Zoican in *Speed and learning in high-frequency auctions (2021).* The growing role of AI in finance is examined further in Creamer, Germán, Kazantsev and Aste *Machine learning and AI in finance* (2021).

Aldridge also talks about the necessity of market regulations, and this matter has been covered first in Menkveld's *"High-Frequency Trading and the New-Market Makers" (2013)* in terms of discussion on how information asymmetry enables opportunities for HFT trading, its influence on market microstructure. A further debate over its effects on liquidity is explored in later works such as *"High-Frequency Trading, Liquidity, and Price Discovery: Evidence from the Pinnacle ETF Platform*" by Comerton-Forde et al. (2015).

Insufficient regulation and lack of transparency create information asymmetry. That raises ethical concerns voiced by Sobolev in *Insider information: the ethicality of the high frequency trading industry* (2020), where he discusses the problem of insider trading, and Cooper, Davis, Kumiega & Van Vliet *Ethics for automated financial markets* 2020 insist that the area should be heavily regulated and suggest their own vision of a fair, effective and ethical trading space.

In conclusion, high-frequency trading is a very dynamic and rapidly evolving field that requires a holistic interdisciplinary approach in terms of study and research. As financial markets become increasingly interconnected, future studies may explore the implications of emerging technologies like blockchain or quantum computing, regulatory shifts, and societal expectations.

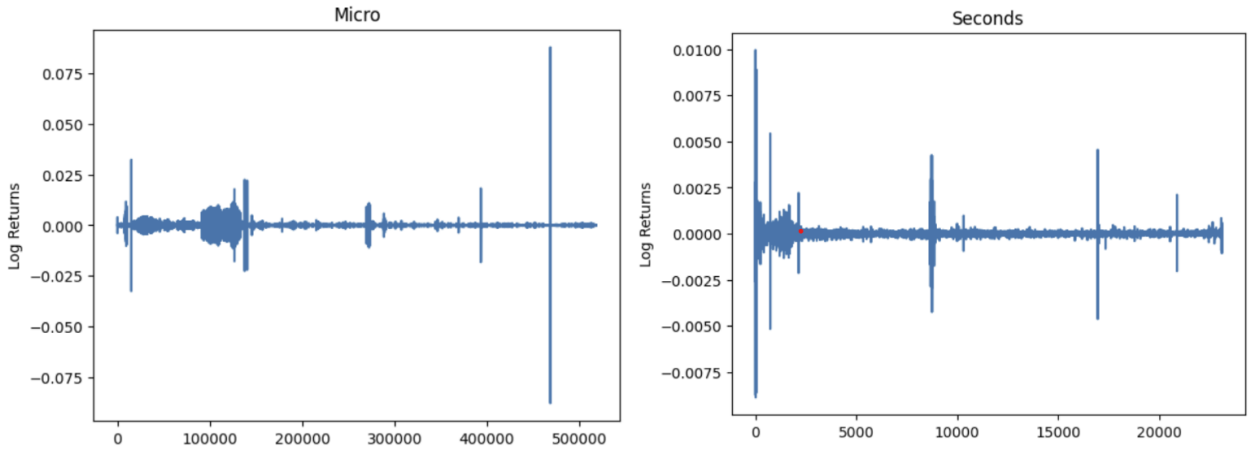
3. Methodology

3.1 Data source, collection methods, preprocessing

The data analyzed in this study originated from the Wharton Research Data Services (WRDS) and specifically concentrated on the stock of Apple Inc. The NYSE Trade and Quote (TAQ) database within WRDS provides detailed tick-by-tick trade and quote data encompassing all activities within the U.S. National Market System. Two distinct datasets were extracted from this database: CTM, which included volume and price information, and CQM, which comprised bid, ask, bid size, and ask size details. The extraction of this high-frequency trading data was accomplished through Python scripting. Subsequently, the datasets were exported in CSV format for further analysis.

The preprocessing of the dataset involved several critical steps to make the high-frequency data more manageable and relevant for the analysis:

Time Interval Reduction: The original dataset obtained from WRDS encompassed trading data recorded at a microsecond interval, providing detailed insights into high-frequency stock market transactions. In the course of our data analysis, particularly when utilizing bid and ask spread information, it became necessary to merge the two datasets. However, a challenge arose as the two datasets featured different time intervals. To address this misalignment, we applied a resampling process, converting the data to a uniform one-second interval. This adjustment facilitated the seamless integration of the datasets for our analytical purposes.



The resampling process has a drawback in that it diminishes volatility by smoothing out returns. As depicted in the graph above, the maximum volatility decreased from 0.075 to 0.01. This reduction in volatility can impact the potential returns in backtesting, as lower volatility typically correlates with less potential for substantial price movements and, consequently, lower returns.

Averaging Prices and Summing Sizes: Within each second interval, the bid, ask, and trade prices were averaged to provide a consolidated view of the stock's trading price. Additionally, the sizes

of these trades were summed up within each one-second interval. This approach provides a clearer, aggregated perspective of trading activity within each second, making the data more interpretable for trend analysis and decision-making. However, this approach also introduces slight inaccuracies and potential data loss. An alternative strategy could involve a weighted average, which might better reflect the actual data.

Focus on Regular Trading Hours: The analysis was confined to data from regular trading hours only. This decision was based on the recognition that stocks are often thinly traded in pre-market and post-market sessions, leading to less reliable and more volatile data. By focusing on regular trading hours, the analysis benefits from a more stable and representative dataset, reflecting the core trading activities associated with Apple's stock.

Through these preprocessing steps, the dataset was transformed into a format more conducive for detailed analysis, allowing for a more accurate and insightful exploration of the trading patterns and behaviors associated with Apple's stock.

3.2 Data analysis and results

Our data analysis comprises two segments: one involves solely the price, while the other incorporates bid and ask spread alongside more intricate feature engineering. The objective is to assess whether the inclusion of these supplementary features enhances the performance of the Machine Learning models.

3.2.1 Price only

The analysis exclusively employed the stock price, incorporating lagged values (5, 15, 30, 60 trading days), simple moving averages (21, 63, 252) and exponential moving averages (10, 30, 200). Regression models were then applied for the machine learning tasks, including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree Regressor (CART), K Neighbors Regressor (KNN), Support Vector Machine (SVR) Random Forest Regressor (RFR), Extra Trees Regressor (ETR), Gradient Boosting Regressor (GBR), and AdaBoostRegressor (ABR).

Repeating the process five times for different days (January 30, March 27, June 22, August 10, and November 1) serves the purpose of reducing volatility and obtaining a more generalized perspective on the results. This approach helps mitigate the impact of specific market conditions on the outcomes, providing a broader and more robust assessment of the models' performance.

Utilizing the predictions generated by the models, which could be either positive or negative, returns were categorized as positive, negative, or zero. The predicted return was computed by multiplying the prediction with the return at time t+1. Accuracy is considered true when the predicted return is positive and false when the predicted return is negative. The accuracy

formula is defined as the number of positive predictions divided by the total predictions minus the count of zero predictions. Accuracy was chosen over recall or precision since false positives and false negatives carry equal weight in our scenario. This process was iterated five times for five different days of data, and the results were averaged to mitigate variability.

Below is a table which includes the accuracy:

	Benchmark	LR	LASSO	RIDGE	CART	KNN	SVR	RFR	ETR	GBR	ABR
130	49.56	56.87	50.43	51.88	54.14	50.35	49.56	57.08	56.8	51.29	49.56
327	49.69	56.17	50.3	50.51	53.39	49.48	49.69	54.3	53.89	55.81	49.74
622	50.89	59.16	50.89	50.79	53.83	49.98	50.89	57.9	57.16	59.23	49.23
810	49.5	54.7	50.5	49.64	54.21	50.18	49.5	54.42	53.64	55.22	49.64
1101	50.73	55.55	50.73	50.94	51.85	49.85	50.73	54.18	53.53	54.93	50.73
AVG	50.07	56.49	50.57	50.75	53.48	49.97	50.07	55.58	55.00	55.30	49.78

In the context of cumulative returns, a positive prediction signals a buy position, prompting the action to purchase the stock at time t+1 until the prediction shifts to negative. At that point, the strategy involves selling the stock and shorting it until the prediction transitions back to a positive state.

	Benchmark	LR	LASSO	Ridge	CART	KNN	SVR	RFR	ETR	GBR	ABR
130	0.9985	1.2132	1.0015	1.0959	1.0818	1.0069	0.9985	1.217	1.2128	1.0895	1.0008
327	1.0006	1.2991	0.9994	1.1198	1.1056	1.0248	1.0006	1.2965	1.2843	1.2931	1.0907
622	1.0056	1.208	1.0056	1.036	1.0867	1.0046	1.0057	1.19	1.1464	1.2041	1.0359
810	1.0007	1.3047	0.9993	1.3384	1.3602	1.0495	1.0007	1.4105	1.4577	1.2548	1.4372
1101	1.0095	1.2493	1.0095	1.0973	1.0712	1.0049	1.0095	1.2041	1.1888	1.2339	1.0098
AVG	1.003	1.255	1.003	1.137	1.141	1.018	1.003	1.264	1.258	1.215	1.115

The benchmark scenario simulates the outcome of buying AAPL stock at the beginning of the testing set and selling it at the end. The average accuracy stands at 50.07, closely aligning with the 50% expectation from a random outcome.

Among the models, LR, RFR, and ETR exhibit the highest accuracies and returns. GBR, while achieving a high accuracy, falls short in predicting spikes, leading to lower returns.

The LASSO model, by potentially excluding crucial features, maintained only one position throughout the entire test set, resulting in diminished accuracy and returns. In contrast, the RIDGE regression penalty, which reduces correlated features without eliminating them, yields superior results compared to LASSO in our backtesting.

Regarding the Support Vector Machine, while it can be highly effective for regression, its preference lies more in classification tasks. In our backtesting, SVR consistently generates positive predictions throughout the test set, mirroring the benchmark outcome.

In response to the challenges posed by the dataset's size and the complexities of trading at a microsecond level, we tried to increase the frequency to seconds. The dataset was resampled and averaged on a per-second basis. Subsequently, the same procedures as outlined earlier

were repeated using this adjusted dataset. The primary goal is to assess the significance of high-frequency data in the context of high-frequency trading.

Accuracy:

	Benchmark	LR	LASSO	RIDGE	CART	KNN	SVR	RFR	ETR	GBR	ABR
130	49.48	50.26	50.43	49.65	51.84	51.14	49.48	50.35	49.83	51.66	49.91
327	49.89	49.52	51.09	49.96	50.57	49.61	51.09	50.22	48.03	51.18	49.17
622	50.34	52.27	49.21	50.79	50.61	49.56	50.7	54.2	51.92	52.27	49.3
810	49.9	47.24	47.68	52.93	48.99	50.04	47.68	48.38	51.18	49.78	52.84
1101	51.66	47.68	53.72	46.19	46.54	46.8	53.71	46.37	46.54	46.46	49.78
AVG	50.25	49.39	50.43	49.90	49.71	49.43	50.53	49.90	49.50	50.27	50.20

Cumulative Returns:

	Benchmark	LR	LASSO	Ridge	CART	KNN	SVR	RFR	ETR	GBR	ABR
AVG	1.002	0.997	1.004	0.999	0.998	0.999	1.002	1.001	1.000	1.000	1.003
130	0.9998	0.9957	1.0002	1.0034	1.0032	1.0041	0.9998	0.9997	0.9993	1.0029	1.004
327	0.9976	0.9997	1.0023	0.9989	0.9996	0.9959	1.0024	1.0063	0.9991	1.0065	0.9983
622	1.0039	1.0007	1.0039	0.9959	0.9963	0.9989	0.9961	1.0046	0.9984	1.0007	1.001662
810	0.9972	0.9916	1.0028	1.0054	0.9996	1.0036	1.0028	1.0006	1.0097	0.999	1.00457
1101	1.0103	0.9988	1.0103	0.9898	0.9912	0.9927	1.0103	0.9921	0.99155	0.991	1.0078

The dataset experienced a substantial reduction from 500,000 to approximately 23,000 entries when resampled to seconds. This downsizing resulted in retaining only about 5% of the original data. Consequently, the accuracy of all models witnessed a drastic decline, with an average accuracy hovering around 50%, indicating a less favorable performance. This notable reduction underscores the significance of high frequency in optimizing the models' accuracy and overall effectiveness.

Enhancing the results could be achieved by incorporating the data from the preceding 20 days at a second frequency. However, this endeavor would require a substantial investment of time, and as of now, has not been pursued.

Finally, the Long Short-Term Memory (LSTM) model was employed for both the original microsecond data and the resampled second data.

	Micro	Seconds
Train RMSE	0.010	0.029
Test RMSE	0.007	0.013

Similar to the earlier findings, the microsecond data yielded superior results compared to the second data when utilizing the Long Short-Term Memory (LSTM) model. Notably, the LSTM model projected outcomes at t+60, as opposed to the t+1 predictions made by the other models previously listed. Consequently, when subjected to the same backtesting strategy, the performance of the LSTM model was notably subpar. It's essential to acknowledge that comparing the LSTM model with the previous strategies requires adjustments for fairness.

Additionally, due to the extended processing time, the decision was made to exclude LSTM from further comparisons.

3.2.2 Bid - Ask

The analysis of Apple Inc.'s stock prices incorporated extensive feature engineering, using both traditional statistical measures and advanced machine learning techniques. The dataset, chosen randomly from five available datasets as of November 1st, offered a representative sample for a thorough examination of stock price trends and patterns.

Narrowing the focus to a single day, specifically November 1, represents a deliberate shift in the project's approach. This adjustment allows for a more in-depth and concentrated analysis, offering a detailed exploration of the dynamics and performance factors within the high-frequency trading models on that specific day. However, a downside is that specific market conditions can have a large impact on the outcome, resulting in a less robust assessment of the models' performance.

This methodology aimed to unravel the underlying dynamics in stock price movements. The dataset was enhanced with time-based features, like hours and minutes extracted from the index, for a more detailed temporal analysis. Additionally, the data was strategically split into 80% for training and 20% for testing, enabling a robust evaluation of the predictive models. This addition is crucial in high-frequency trading data where price movements can be significantly influenced by the time of the day. The spread between the bid and ask prices was calculated, offering insights into the liquidity and market depth for Apple's stock. Volatility was quantified using the standard deviation of prices over 5 and 10 minute windows. This metric is vital in assessing the risk and instability associated with the stock's price. Volume imbalance provides a quick snapshot of market sentiment at a single price level, depth imbalance offers a more comprehensive and detailed view, taking into account the distribution of buy and sell orders across different price levels. This deeper insight can be crucial for traders and investors looking to understand potential support and resistance levels, as well as the overall strength of buying or selling pressure in the market. However due to the limitation of the dataset, we are only able to reflect supply and demand dynamics. Similarly, historical volume data with 1 and 3-minute lags were added. This feature provides context on trading activity preceding the current price.

$$Volume\ imbalance_j = \frac{\sum_{j=1}^{n} BuyVolume_j - \sum_{i=1}^{n} SellVolume_j}{\sum_{j=1}^{n} BuyVolume_j + \sum_{i=1}^{n} SellVolume_j}$$

$$DI_t = \frac{\sum_{i=1}^{n} VolBid_{it} - \sum_{i=1}^{n} VolAsk_{it}}{\sum_{i=1}^{n} VolBid_{it} + \sum_{i=1}^{n} VolAsk_{it}}$$

The enriched dataset served as a foundation for applying various regression models, including: Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, Logistic Regression, Decision Tree Regressor (CART), Random Forest Regressor (RFR), Extra Trees Regressor (ETR), Gradient Boosting Regressor (GBR), XGB Regressor(XGB). Other Models like K Neighbors Regressor (KNN), Support Vector Machine (SVR), Naive Bayes were also included. These models were chosen for their ability to handle nonlinear relationships and interactions between features, which are common in financial time series data.

The Ridge Regression model appears to have the best performance in terms of cumulative return, which is a critical factor in stock prediction. Similar MSE and RMSE to Linear Regression, implying minor improvements due to regularization. R^2 is slightly higher than Linear Regression, suggesting a slightly better fit. Logistic Regression, while not directly comparable to the other models due to its classification nature, shows higher accuracy and could be useful for binary prediction tasks (e.g., price going up or down).

Below is a table which includes the accuracy and cumulative return for linear model :

	Linear Regression	Ridge Regression	Lasso Regressi	Elastic Net	Logistic Regression
MSE	0.1495	0.1485	0.4361	0.2547	NA
RMSE	0.3867	0.3854	0.6604	0.5047	NA
R^2	0.8003	0.8016	0.4174	0.6598	NA
Accuracy	0.4953	0.4989	0.5000	0.5000	0.5518
ROC AUC	NA	NA	NA	NA	0.5479
	Benchmark	LR	Ridge	LASSO	Elastic Net
Cum Return	1.0169	1.0025	1.0153	1.0027	1.0027

The extended analysis suggests that ensemble methods like RFR, XGB, and ETR outperform single decision trees (CART) and the earlier models (Linear, Ridge, Lasso, Elastic Net, Logistic Regression) across most metrics for regression tasks. They have significantly lower MSE and RMSE and higher R^2 values, indicating better fit and predictive accuracy. For classification tasks, as indicated by Accuracy and ROC AUC, RFR and GBM perform notably well.

However, models with high R^2 and low MSE/RMSE, such as CART and ETR, did not translate their apparent predictive accuracy into equivalent returns, the simpler models like Linear Regression still hold their ground. This discrepancy highlights the importance of evaluating machine learning models on financial metrics that directly reflect investment outcomes.

Below is a table which includes the accuracy and cumulative return for Tree based model :

	CART	RFR	GBM	XGB	ETR	
MSE	0.0026	0.0014	0.0140	0.0022	0.0013	
RMSE	0.0513	0.0378	0.1182	0.0468	0.0359	
R^2	0.9965	0.9981	0.9813	0.9971	0.9983	
Accuracy	0.5371	0.6014	0.6157	0.5941	0.5976	
ROC AUC	0.5368	0.6499	0.6637	0.6366	0.6437	
	Benchmark	CART	RFR	GBR	XGB	ETR
Cum Return	1.0169	0.9027	0.8663	0.9833	0.9345	0.8590

The SVM models, particularly the Non-linear SVM, and Naive Bayes, offer competitive returns close to the benchmark, with SVM slightly outperforming it. However, KNN, despite a high R^2 value, falls short in terms of cumulative return, suggesting that its practical application in stock prediction might be limited.

Below is a table which includes the accuracy and cumulative return for other model :

	SVM	Non-linear SVM	KNN	Naïve Bayes
MSE	0.1734	0.1095	0.0074	NA
RMSE	0.4164	0.3309	0.0863	NA
R^2	0.7684	0.8537	0.9900	NA
Accuracy	NA	NA	NA	0.5136
ROC AUC	NA	NA	NA	0.5359

	Benchmark	SVM	Non-linear SV	KNN	Naïve Bayes
Cum Return	1.0169	1.0212	1.0026	0.9414	1.0018

3.3 Interpretation of Results

When analyzing cumulative returns, the observed returns of approximately 1 in both cases align with the predictions made during data preprocessing. The resampling process, as anticipated, significantly reduced returns. An alternative approach could involve exploring different backtesting strategies or using a distinct dataset with native seconds data instead of resampling the original dataset. Unfortunately, such a dataset wasn't available in WRDS. Given this limitation, a more meaningful comparison might involve evaluating accuracy rather than returns, offering insights into the performance of the models in a way that is less affected by the specific data processing steps.

Accuracy	Benchmark	LR	LASSO	RIDGE	CART	RFR	ETR	GBR
Price	51.66	47.68	53.72	46.19	46.54	46.37	46.54	46.46
Bid Ask	51.66	49.53	50	49.89	53.68	64.99	64.37	61.57
Diff	0.00	-1.85	3.72	-3.70	-7.14	-18.62	-17.83	-15.11

In the context of linear models (LR, LASSO, RIDGE), it was observed that LASSO performed better using price than using bid-ask information alone. On the other hand, LR and Ridge Regression exhibited improved performance when incorporating bid-ask data, although the differences in performance were relatively moderate.

However, within the domain of decision tree models, the CART model exhibited a slight improvement when incorporating Bid and Ask prices compared to utilizing only stock prices. In contrast, for more complex models such as RFR, ETR, and GBR, the inclusion of Bid and Ask prices led to notably enhanced performance compared to scenarios where only stock prices were considered.

Overall, the trend suggests that incorporating bid and ask information enhances the predictive power of the models.

Accuracy	Benchmark	LR	LASSO	RIDGE	CART	KNN	SVR	RFR	ETR	GBR	ABR
Micro	50.074	56.49	50.57	50.752	53.484	49.968	50.074	55.576	55.004	55.296	49.78
Seconds	50.254	49.394	50.426	49.904	49.71	49.43	50.532	49.904	49.5	50.27	50.2
Diff (%)	-0.36	12.56	0.28	1.67	7.06	1.08	-0.91	10.21	10.01	9.09	-0.84

Cum Ret	Benchmark	LR	LASSO	Ridge	CART	KNN	SVR	RFR	ETR	GBR	ABR
Micro	1.003	1.255	1.003	1.137	1.141	1.018	1.003	1.264	1.258	1.215	1.115
Seconds	1.002	0.997	1.004	0.999	0.998	0.999	1.002	1.001	1.000	1.000	1.003
Diff (%)	0.12	20.52	-0.08	12.20	12.54	1.88	0.07	20.81	20.54	17.70	10.00

The reduction in training size was associated with a decrease in accuracy, underscoring the importance of an adequate volume of training data for model performance. Additionally, a decrease in trading opportunities and lower volatility further contributed to diminished returns. These observations highlight the interconnected nature of data quantity, trading opportunities,

and market volatility, all of which play crucial roles in the effectiveness of high-frequency trading models.

4. Conclusion with discussion of Limitations

Using the original data, it can be seen that employing only the stock price yields promising results. Next, after the resampling process, incorporating bid and ask data enhances the outcomes compared to using only the stock price.

The inherent characteristics of bid and ask information can provide additional insights and contribute to improved model performance, especially in the context of high-frequency trading analysis.

Considering this, if feasible, utilizing bid and ask data in the original dataset might indeed lead to the most favorable results. However, achieving this seamlessly is challenging due to the disparities in time intervals between the two datasets. The difference in time intervals poses a hurdle in directly comparing or combining the datasets, necessitating careful consideration and potential adjustments in the analysis approach.

The challenge of achieving optimal results by using bid and ask data in the original dataset could potentially be overcome by accessing a different dataset specifically designed or structured for such purposes. Unfortunately, the unavailability of such a dataset poses a constraint in this context. While obtaining an alternative dataset with native bid and ask information would be an ideal solution, practical limitations sometimes necessitate making the best use of the available data, despite potential challenges and compromises.

Obtaining high-quality data for high-frequency trading poses a significant challenge. Like any machine learning endeavor, the success and reliability of the models are heavily dependent on the availability and quality of the data. The intricacies of high-frequency trading require precise and detailed information, making the task of sourcing and curating suitable datasets a crucial aspect of the analysis process.

In conclusion, it's important to note that the strategies outlined in this project did not incorporate considerations for the practical challenges associated with high-frequency trading. Nevertheless, the primary focus of this project lies in the application of machine learning techniques to analyze and derive insights from the available data.

Key findings

HFT supplies liquidity to the thick side of the order book(where it is not required) and demands liquidity from the thin side of the order book (where it is most needed). This trading behavior exacerbates future order book imbalance.

HFT becomes more strategic with faster trading speed. However, HFT strategies come at the cost of crowding out non-HFT limit order from the order book

Future Research

1. **In-Depth Analysis of Market Impact**: Further research could investigate the market impact of high-frequency trading (HFT) strategies, especially in terms of market liquidity and volatility. This could include a closer examination of how HFT affects price discovery and market efficiency.
2. **Algorithmic Improvements and Machine Learning Models**: Exploring advanced machine learning algorithms and their application in HFT strategies could be another research focus. This might involve developing new predictive models or enhancing existing ones (hyperparameter tuning) to better capture market dynamics and improve trading performance. It can also involve a more robust backtesting strategy.
3. **Cross-Market Analysis**: Conducting cross-market studies to understand the differences and similarities in HFT practices across various global financial markets. This could help in understanding the universal principles of HFT and the specific nuances in different market environments.