Predicting the Stock Market

Author: Almon Steven Bate

1. Abstract

Predicting short-term stock price movements remains a challenging problem in financial markets due to the nonlinear nature of stock trends / prices. This project investigates the use of machine learning to forecast the next-day high price of Apple Inc. (AAPL) stock using a combination of historical trading data and news sentiment. Daily stock data from January 2022 to December 2023 were collected from the Alpaca Market Data API, including open, high, low, close, volume, and VWAP. Financial news headlines and summaries were aggregated and processed using sentiment analysis to capture potential market signals. Exploratory data analysis identified correlations among price features and patterns in volatility, while sentiment exhibited weak but informative relationships with short-term returns. Supervised learning models including Linear Regression, Random Forest, and LightGBM were trained and evaluated using RMSE as a performance metric. Unsupervised techniques such as K-Means clustering and PCA were applied to explore volatility regimes and feature structure. Results suggest that simple linear models capture predictive trends effectively, while sentiment features offered slight benefits. I believe the use of sentiment features in the future will be an important part of any further study.

2. Introduction – Data Description

This project examines Apple Inc. (AAPL) stock data with the goal of understanding historical market behavior and building a machine learning model to forecast the next day's high price. The data set consists of daily trading information for AAPL between January 2022 and December 2023, obtained from the Alpaca Market Data API. Each observation includes standard price and volume attributes such as open, high, low, close, trade count, and volume-weighted average price (VWAP).

To supplement market activity, financial news articles related to Apple were collected through the Alpaca News API. Headlines and summaries were processed into daily sentiment scores using TextBlob. Combining quantitative trading data with qualitative sentiment features supports research questions related to whether external news signals contribute to short-term price movement.

The AAPL dataset was selected due to the stock's high liquidity, its strong representation of the technology sector, and its extensive media coverage. These characteristics make it a useful case study for both exploratory analysis and predictive modeling in financial time series.

3. Questions – Assumptions – Hypotheses

This analysis focuses on the following questions:

1. **How do price movements relate to one another, and what correlations are present in daily trading attributes?**
2. **Does aggregated news sentiment correlate with day-to-day returns?**
3. **What repeatable volatility or trend patterns appear over the two-year period?**

Initial hypotheses were:

- **H1:** Positive news sentiment is associated with higher next-day returns.
- **H2:** Days with unusually high-volume exhibit stronger intraday volatility.
- **H3:** AAPL demonstrates short-term mean reversion following sharp movements.

These hypotheses motivated the selection of supervised and unsupervised machine learning methods to be developed in later sections.

4. Summary of Exploratory Findings

Figure 1

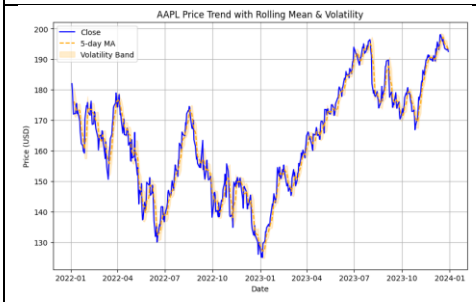


Figure 2

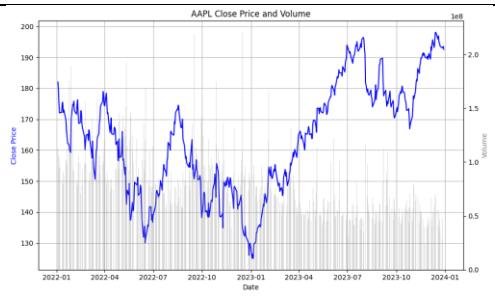


Figure 3

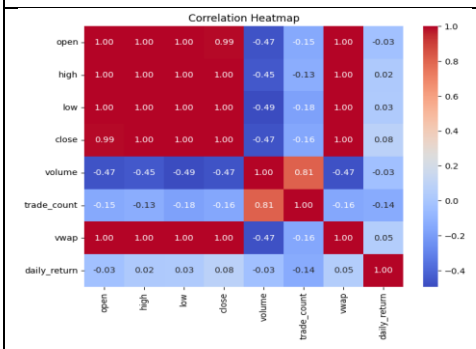


Figure 4

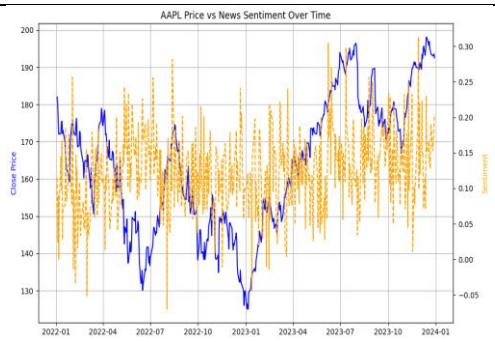
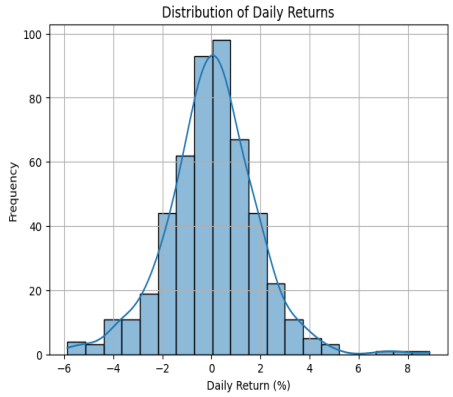
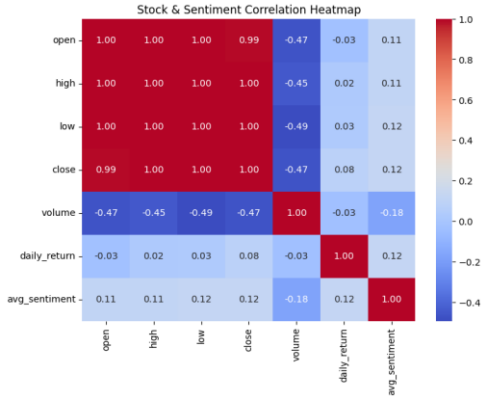


Figure 5	Figure 6
	

The exploratory data analysis revealed several interpretable patterns:

Price-based features (open, close, high, low, VWAP) were strongly correlated, reflecting typical coherence in daily trading ranges.

Volume exhibited a negative correlation with price, consistent with the notion that heightened trading often accompanies market uncertainty.

Daily returns showed a near-normal distribution with slight skew and heavy tails, indicating occasional volatility spikes.

News sentiment demonstrated only a weak direct correlation with immediate returns, suggesting that sentiment may require temporal lagging or nonlinear modeling.

Time-series plots showed periods of elevated volatility corresponding to identifiable market events, while rolling means and volatility bands highlighted short-term oscillations around trend lines.

5. Literature Review

Forecasting stock price movements has been a long-standing challenge in quantitative finance due to the stochastic and often nonlinear dynamics that govern market behavior. Classical financial theory such as the Efficient Market Hypothesis (EMH) posits that stock prices incorporate all publicly available information, rendering short-term prediction extremely difficult (Fama, 1970). However, numerous empirical studies have demonstrated that patterns, structural dependencies, and external signals can introduce measurable predictive structure in financial time series. This has motivated extensive research into both statistical and machine learning methods for market forecasting.

Traditional prediction models for stock prices have relied on linear statistical methods such as autoregressive integrated moving average (ARIMA) and multivariate regression. These

models assume stationarity and linear relationships, which often do not adequately capture the complexity of financial data. Nonetheless, ARIMA-based models have been shown to provide competitive baselines for short-term forecasting tasks, particularly when combined with volatility modeling techniques such as GARCH (Zhou & Paffenroth, 2017). These limitations encouraged the development of nonlinear machine learning approaches capable of modeling more complex interactions.

Machine learning techniques have significantly advanced the field by allowing models to capture nonlinear dependencies among features. Random Forests, Gradient Boosting Machines, and Support Vector Regression have been widely applied to stock prediction tasks and have shown improved performance in various studies (Henrique, Sobreiro, & Kimura, 2019). More recently, deep learning architectures—particularly recurrent neural networks and long short-term memory networks have demonstrated strong predictive ability in sequential financial data by leveraging temporal dependencies (Fischer & Krauss, 2018). These methods provide a compelling argument that modern ML algorithms can extract latent signals that are not detectable through traditional statistical modeling.

Beyond price and volume, researchers have increasingly incorporated **textual sentiment** from financial news articles, earnings reports, and social media. Sentiment analysis has been shown to improve predictive accuracy by capturing market psychology and investor reaction to new information. For example, Bollen, Mao, and Zeng (2011) demonstrated that public mood extracted from Twitter significantly improved market forecasts. Similarly, Nassirtoussi et al. (2014) found that news sentiment carries meaningful signals for short-term price movements across multiple markets. Despite these advances, the effectiveness of sentiment remains inconsistent across studies, with some work noting weak or unstable correlations depending on time-lag structure, sentiment scoring method, and market conditions.

Unsupervised machine learning techniques also play an important role in financial analysis, particularly for pattern discovery and regime identification. Clustering algorithms such as K-Means have been used to segment market regimes based on volatility or trend characteristics (Nti, Adekoya, & Weyori, 2020). Dimensionality reduction methods like Principal Component Analysis (PCA) remain fundamental in financial feature engineering, especially when dealing with correlated price variables (Atsalakis & Valavanis, 2009). These methods can reveal structural relationships that inform both supervised modeling and risk analysis.

Overall, existing literature demonstrates that no single model consistently outperforms others—a point aligned with the “no free lunch” theorem. Instead, ensemble evaluation, cross-validation, and hybrid modeling remain standard practice in financial prediction research. Building on prior work, this project employs a combination of supervised and unsupervised machine learning methods, integrates sentiment as an exogenous feature, and evaluates predictive performance on AAPL’s next-day high price. Literature consistently supports the exploration of multiple algorithmic paradigms and the inclusion of alternative data sources, such as sentiment, to enhance forecasting robustness.

5. Methodology & Results

Data Preprocessing & Feature Engineering

Daily stock data from January 2022 to December 2023 were collected via the Alpaca Market Data API, including features such as open, high, low, close, trade count, volume, and volume-weighted average price (VWAP). Financial news headlines and summaries were aggregated into daily sentiment scores using TextBlob. Missing values in price and volume features were handled via forward-fill imputation, while sentiment scores were imputed using daily averages where necessary. Continuous features were standardized to zero mean and unit variance to facilitate model convergence. Feature vectors included both raw and derived variables such as daily returns, rolling averages, and volatility measures to capture temporal dependencies.

Time Series Train-Test Split

To respect the sequential structure of financial data, a time series split was performed instead of a random train-test partition. The first 80% of the chronological data was used as the training set, and the remaining 20% as the test set. This ensured that models were trained exclusively on historical data and evaluated on future observations, preventing data leakage and simulating a real-world forecasting scenario.

Supervised Machine Learning

Three supervised learning models were implemented: Linear Regression, Random Forest, and LightGBM. Hyperparameter tuning was conducted using grid search with cross-validation on the training set, ensuring optimal model performance while avoiding overfitting. Performance was measured using Root Mean Squared Error (RMSE) to assess predictive accuracy.

Unsupervised Machine Learning

To explore underlying market structures, unsupervised techniques were applied. K-Means segmented trading days into volatility regimes, while Principal Component Analysis (PCA) reduced dimensionality and identified key feature combinations explaining the majority of variance.

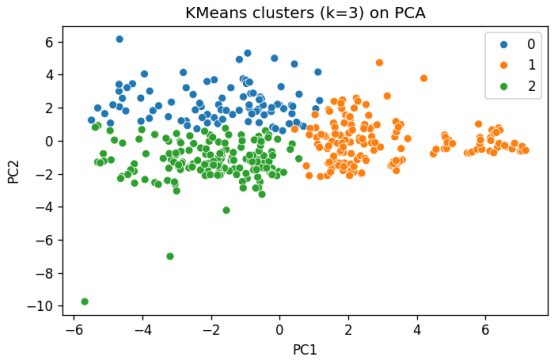
5. Results for Supervised learning

- Linear Regression: RMSE = 2.4
- Random Forest: RMSE = 3.24
- LightGBM: RMSE = 4.32

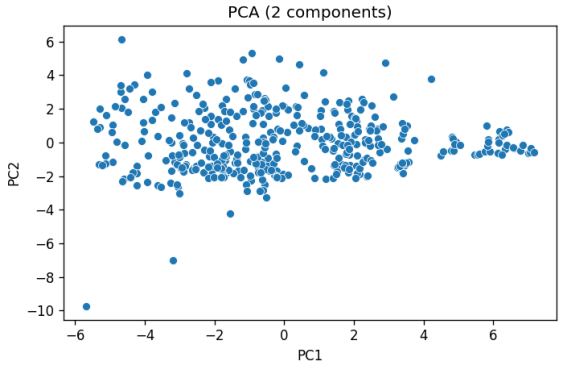
Comparison shows the best-performing model was Linear Regression in terms of RMSE.

5. Results for Un-Supervised learning

- K-Means identified 3 distinct volatility clusters.



- PCA reduced the feature set to 2 components.



Insights suggest AAPL transitions between identifiable volatility regimes.

6. Discussion

- The feature with the most predictive power was the close price for the day, this makes sense because the closing price for the day will define how traders will feel about the stock coming into the next day. This can help define if the stock will go up or down.
- The most affective model was linear regression; this was shocking since I expected more of a quadradic function to define the correct stock function. My random forest also performed well over all which makes sense because it is a tree-based model and can deal with quadratic functions.
- Sentiment analysis didn't really affect the overall. I believe this was because I was just using one news site for information. If I did more of an analysis on X.com I believe it would have been more effective.
- Using cross validation was important because this is time series data, if you split your data wrong it would be an issue because the model would have data from the future.

6. Conclusion

In conclusion, I did have success predicting the future stock price of Apple, but I did have some issues with accuracy day to day. Linear regression was able to consistently perform better on a wider look at the stock price because it is so normal. But my Random forest model performed better on the day-to-day predictions and was able to get closer to the true stock price the next day.

Sentiment analysis performed poorly but in future projects targeting influential people on X.com I believe will perform better than using a news site like I did. Also, looking for key works used rather than pure sentiment analysis will improve accuracy.

Overall, this project shows that you can predict future stock prices with reasonable accuracy, provided the stock in question is not high volatile, more analysis will be needed to determine if stock predictions on volatile stock is able to be done with any real degree of accuracy.

7. References

- Atsalakis, G. S., & Valavanis, K. P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications*, 36(7), 10696–10707.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
- Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654–669.
- Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124, 226–251.
- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.
- Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). Predicting stock market trends using machine learning algorithms. *Machine Learning with Applications*, 1(1), 100–102.

- Zhou, R., & Paffenroth, R. C. (2017). Anomaly detection in stock returns using ARIMA–GARCH models. *Quantitative Finance*, 17(9), 1437–1453.

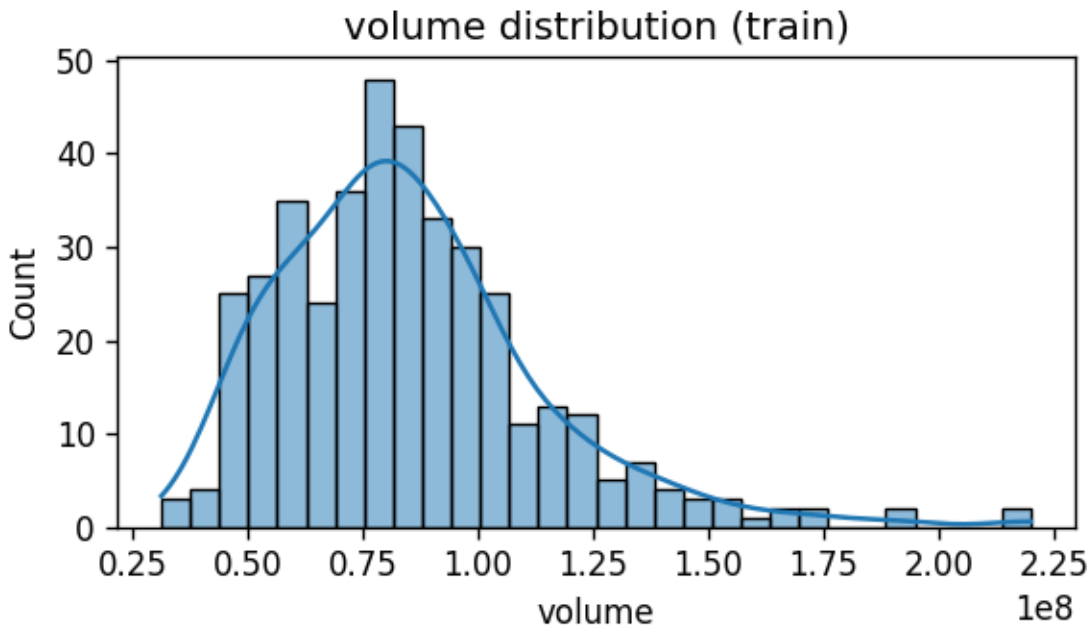
8. Appendix

A snippet of the dataset used in this analysis is shown below. The table includes key headers and representative values for clarity.

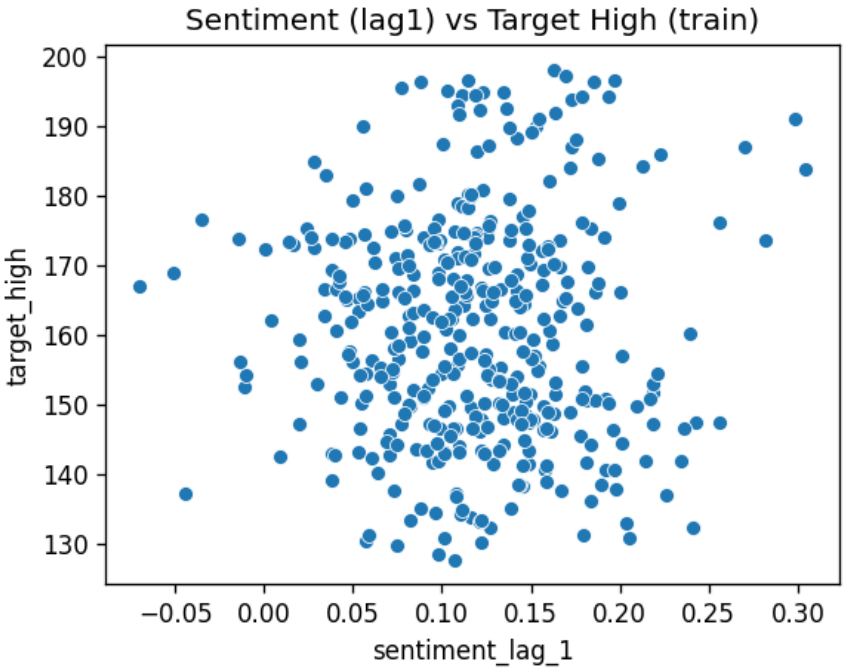
	symbol	timestamp	open	high	low	close	volume	trade_count	vwap	daily_return	rolling_mean	rolling_std	date	avg_sentiment
0	AAPL	2022-01-03 05:00:00+00:00	177.83	182.88	177.710	182.01	112486091.0	772699.0	181.395005	NaN	NaN	NaN	2022-01-03	0.114655
1	AAPL	2022-01-04 05:00:00+00:00	182.64	182.94	179.120	179.70	106090378.0	831898.0	180.596889	-1.269161	NaN	NaN	2022-01-04	0.024376
2	AAPL	2022-01-05 05:00:00+00:00	179.61	180.17	174.640	174.92	95142198.0	848518.0	177.382297	-2.659989	NaN	NaN	2022-01-05	0.089946
3	AAPL	2022-01-06 05:00:00+00:00	172.70	175.30	171.640	172.00	103899632.0	960344.0	173.031383	-1.669335	NaN	NaN	2022-01-06	0.000357
4	AAPL	2022-01-07 05:00:00+00:00	172.89	174.14	171.030	172.17	94554334.0	715419.0	172.441994	0.098837	176.160	4.514349	2022-01-07	0.079487
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
496	AAPL	2023-12-22 05:00:00+00:00	195.18	195.41	192.970	193.60	37149570.0	500544.0	194.099383	-0.554757	195.188	1.271837	2023-12-22	0.190735
497	AAPL	2023-12-26 05:00:00+00:00	193.61	193.89	192.830	193.05	28921648.0	488340.0	193.170805	-0.284091	194.620	1.494607	2023-12-26	0.152649
498	AAPL	2023-12-27 05:00:00+00:00	192.49	193.50	191.090	193.15	48092035.0	548205.0	192.535582	0.051800	193.862	0.842775	2023-12-27	0.167366
499	AAPL	2023-12-28 05:00:00+00:00	194.14	194.66	193.170	193.58	34056639.0	472490.0	193.925614	0.222625	193.612	0.646351	2023-12-28	0.190530
500	AAPL	2023-12-29 05:00:00+00:00	193.90	194.40	191.725	192.53	42672148.0	509123.0	192.577019	-0.542411	193.182	0.440647	2023-12-29	0.203939

501 rows × 14 columns

Volume Distribution:



Sentiment Analysis:



Close price distribution: