Fintech 545 Project Report:
Predicting AAPL Stock Return using Non-Linear Method

1. Objective:

The primary goal of this project is to predict the future closing price of Apple Inc. (AAPL) stock using machine learning techniques. By analyzing historical stock prices along with market indices, the project aims to construct a predictive model that can help inform investment strategies.

2. Data Preparation:

Stock market data from January 2012 to December 2023 was collected for AAPL and the NASDAQ Composite Index (replacing the initially considered S&P 500 due to less feature importance). The following features were computed and included for model training:

1. Adjusted Close Price (Adj Close)
2. Opening Price (Open)
3. Daily High Price (High)
4. Daily Low Price (Low)
5. Moving Averages over 5, 10, 20, and 50 days (MA5, MA10, MA20, MA50)
6. NASDAQ Composite Index closing prices (nasdaq)

3. Feature Selection and Importances:

The feature importance analysis for the Random Forest model revealed that Adj Close, MA10, MA20, and MA5 were the most influential features. The NASDAQ index had some influence but to a lesser degree.

Feature	Importance
Adj Close	24.86%
Open	8.50%
High	7.01%
Low	4.68%
MA5	15.34%
MA10	17.25%
MA20	16.52%
MA50	2.56%
nasdaq	3.29%

4. Training Models and Data Selection:

1. **Random Forest Regressor (RF):** An ensemble model consisting of 1000 decision trees with a maximum depth of 30.
2. **Linear Regression (LR):** A simple model assuming a linear relationship between features and the target variable.

The dataset was split based on the date, using records until the end of 2022 for training and those from 2023 for testing.

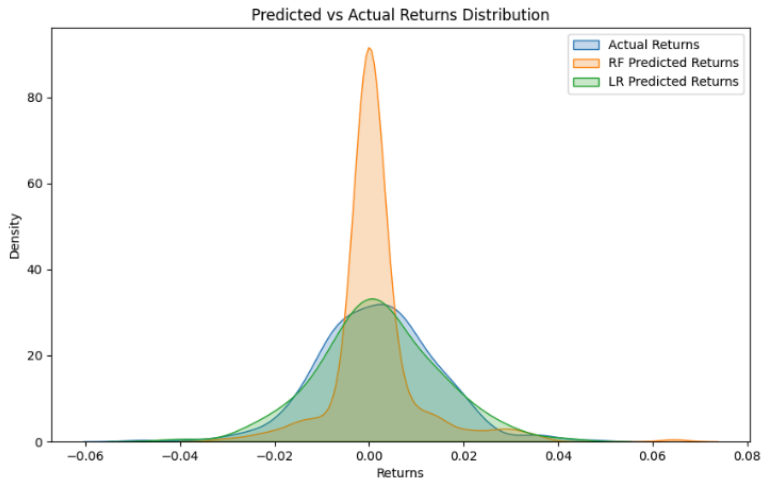
5. Model Evaluation and Metrics:
Both models were assessed based on their Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R2), and the volatility of their predictions. The evaluation yielded the following metrics:

Model/Metrics	RMSE	MAE	R2	Volatility
RF Model	9.1697	6.3311	0.7113	1.2793
LR Model	2.2422	1.7193	0.9827	0.0131

6. Metric Comparison of RF and LR Models:
The Linear Regression (LR) model demonstrates superior performance with a significantly lower RMSE and MAE compared to the Random Forest (RF) model, suggesting closer alignment with actual values. The LR's high R-squared value indicates a strong fit to the data. However, its low volatility metric raises questions about its risk sensitivity. The RF model's higher errors and greater prediction volatility suggest less accuracy but potentially more realistic risk estimation.

7. Plot Analysis:

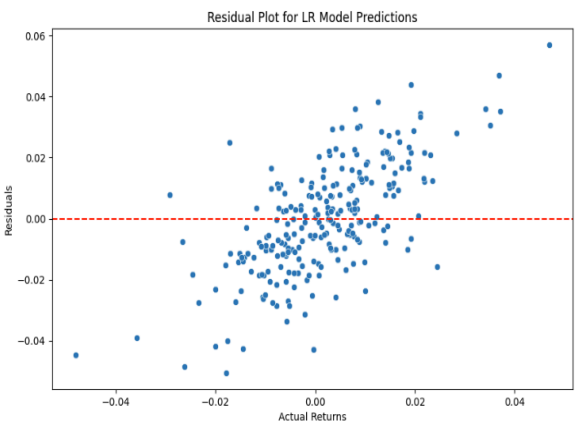
1. **Density Plot for Returns Plot**

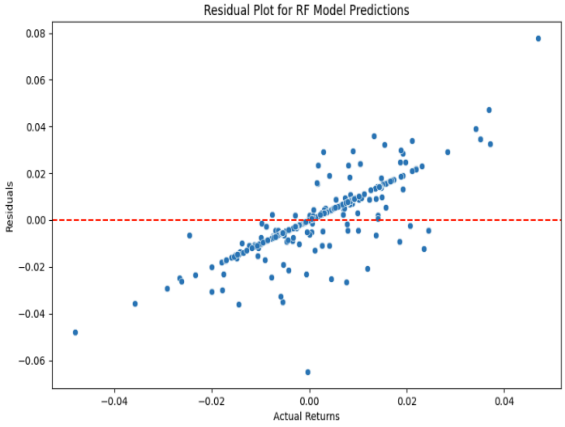


The density plot illustrates the Linear Regression model's predictions are far more concentrated than those of the Random Forest model, suggesting a higher level of confidence—or overconfidence. Compared to the actual returns, the Linear Regression model severely underestimates the variability, implying a model that might not fully account for market volatility. In contrast, the Random Forest provides a slightly broader prediction range, though still not aligning with the actual data's spread. This discrepancy could stem from the models' inherent assumptions—Linear Regression's simplicity may miss complex patterns the Random

Forest catches, yet the latter might still be too rigid or not adequately trained to capture the full scope of market dynamics.

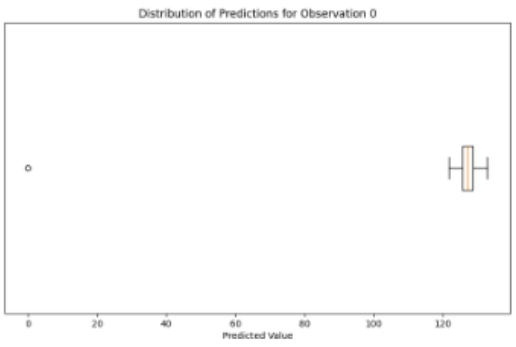
2. Residual Plots for RF and LR Models

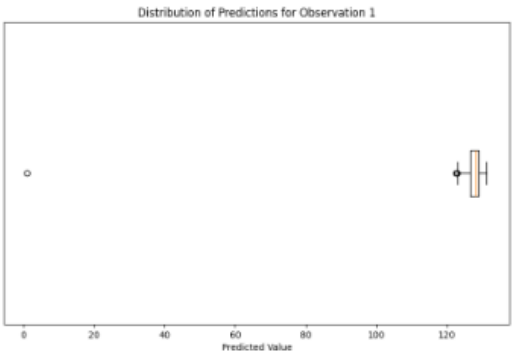


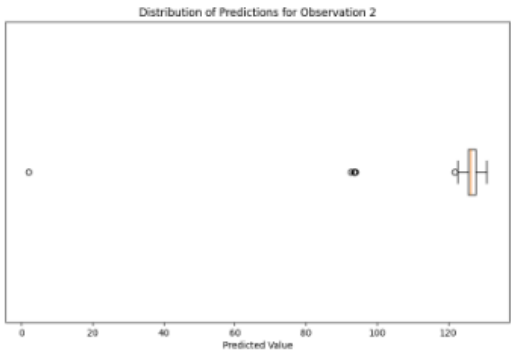


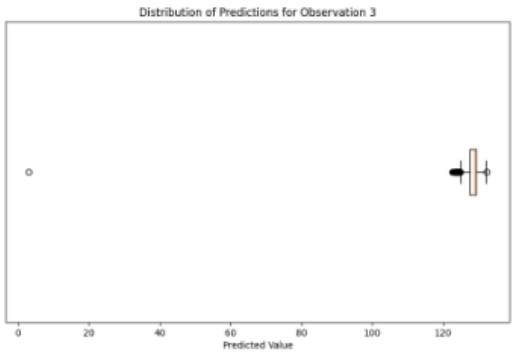
Residual plots showcase that the Linear Regression model's predictions are closer to the zero line and more uniformly distributed than the Random Forest's, hinting at better accuracy and consistency. The Random Forest model shows a discernible pattern of increased residuals with the size of actual returns, suggesting heteroscedasticity and potentially less adaptability to market conditions that vary widely. This might result from overfitting within the Random Forest model or insufficient complexity within the Linear Regression model to fully understand the nuanced financial trends.

3. Distribution of Predictions Analysis



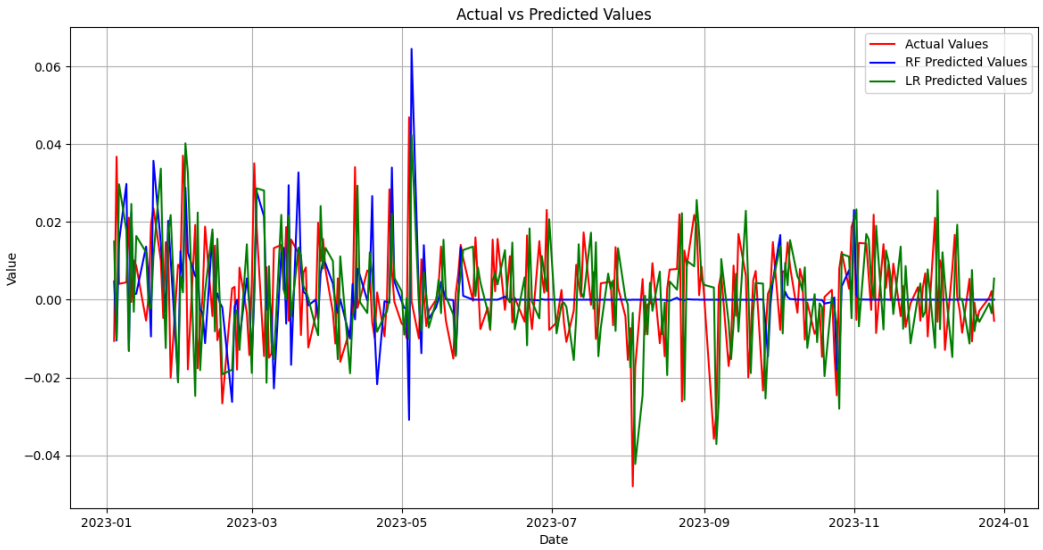






The boxplots reveal a notable variance in the Random Forest model's predictions, with outliers indicating occasional extreme predictions by individual trees. The relative uniformity of the interquartile range, however, suggests a general agreement within most of the model's predictions. The outliers could be a result of the model's sensitivity to certain atypical data points or noise within the dataset that some trees in the ensemble may overinterpret, leading to these sporadic predictions.

4. Actual vs Predicted Values Analysis



The line plot comparison indicates that neither model perfectly tracks the actual values, with both exhibiting deviations at various points. The Linear Regression model seems to follow the actual trend with less deviation than the Random Forest model, which may oscillate more dramatically. This could be attributed to the Linear Regression model’s inability to capture complex patterns, whereas the Random Forest might be reacting to noise in the data. The better performance of Linear Regression in this comparison could be due to its simplicity, making it less prone to overfitting compared to the Random Forest model.

8. Conclusion:

The analysis and metrics suggest that the Linear Regression model is the more accurate in predicting AAPL stock prices, offering lower prediction errors and a higher R-squared value. However, its narrower prediction range might underestimate the inherent volatility of the stock market. The Random Forest model, while not as precise, may provide a more cautious approach by reflecting a broader range of potential outcomes. The choice between models should consider the trade-off between precision and risk sensitivity according to the investment strategy's priorities.