MSAI 451 – Programming Assignment 1 Report

Title: Forecasting AAPL Daily and Weekly Price Direction using Technical Indicators

Author: Divyanka Thakur

---

1. Introduction and Problem Statement

This project constructs a supervised machine learning pipeline to model short-term stock price direction. The objective is to forecast the future binary movement (up/down) of **Apple (AAPL)** stock using a set of common technical indicators derived from historical price data.

The core task is to predict whether the next period's return will be positive (coded as 1) or negative/zero (coded as 0). Two distinct forecasting horizons were evaluated:

- **t+1:** Next trading day's price direction.
- **t+5:** Next trading week's (5-day) price direction.

This effort serves as a rigorous exploration of whether predictive structure can be reliably identified in an efficient financial market using purely price-derived features.

---

2. Data Preparation and Feature Engineering

The analysis utilizes **AAPL daily OHLCV** (Open, High, Low, Close, Volume) data sourced from Yahoo Finance, covering the period **2000 to 2025**.

Data Preprocessing

Preprocessing steps included:

- Date normalization and ensuring strict chronological order.
- Handling of minor missing values (primarily volume).
- Standardization of all continuous features to ensure fair treatment by the Logistic Regression model.

Feature Set

A comprehensive set of technical indicators, commonly used in quantitative finance, were engineered:

- **Momentum and Volatility:** Daily return, 10-day rolling volatility (Vol10), RSI(14), and Bollinger Band Width (BB_Width).
- **Moving Averages:** Simple moving averages (MA5, MA20), Exponential moving averages (EMA5,EMA12,EMA26), and the Moving Average Convergence Divergence (MACD).
- **Price Structure:** High-Minus-Low (HML) and Open-Minus-Close (OMC) to capture candlestick shape, along with lags (L1−L3) for short-term memory.
- **Calendar Feature:** A simple binary indicator for 'Is Friday'.

The target variables, Target(t+1) and Target(t+5), were constructed as binary outcomes (1 for positive return, 0 otherwise) for the respective horizons.

---

3. Research Design and Modeling

Time-Series Validation

To preserve the crucial temporal relationship, a **time-aware split** was used:

- **Training Set:** 80% of the historical data (chronologically first).
- **Test Set:** 20% of the data (chronologically last), representing an out-of-sample period.

Within the training set, a TimeSeriesSplit was used to create a small, sequential validation fold. This was primarily used to tune the classification probability threshold for the Logistic Regression model.

Baseline Models

Two diverse classification algorithms were selected as baselines:

1. **Logistic Regression (LogReg):** A linear, interpretable model used to establish performance against standardized features.
2. **Random Forest Classifier (RFC):** A non-linear, ensemble model configured with 500 trees, maximum depth of 8, and a minimum leaf size of 5 to mitigate overfitting.

Evaluation Protocol

Model performance was rigorously assessed using standard classification metrics:

- **Accuracy:** Overall correct predictions.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure of performance, especially important for class imbalance (though less so here).
- **ROC AUC:** Area Under the Receiver Operating Characteristic curve, a measure of separability between classes.

- **Equity Curve:** A visualization of the cumulative returns of a trading strategy based on the model's predictions versus a simple Buy-and-Hold strategy over the test period.
- **Feature Importances:** Analysis of the RFC's feature contributions.

---

4. Results and Discussion

t+1 Model (Next-Day Forecast)

Metric	Logistic Regression	Random Forest
Accuracy	0.505	0.501
F1 Score	0.601	0.598
ROC AUC	0.485	0.483

The performance for the next-day forecast horizon is statistically indistinguishable from a random coin flip (Accuracy ≈0.50, AUC ≈0.48). This result is entirely consistent with the **Efficient Market Hypothesis (EMH)**, which posits that all historical information is already priced in, leaving no persistent, short-term predictive edge. The models fail to provide a reliable signal for daily direction.

t+5 Model (Weekly Forecast)

Metric	Logistic Regression	Random Forest
Accuracy	0.548	0.552
F1 Score	0.695	0.701
ROC AUC	0.472	0.474

The weekly forecast demonstrated a marginal improvement, with Accuracy slightly above 55%. While better than the daily model, the ROC AUC remains below 0.50. This suggests that while the models might correctly classify the *majority* class more often (contributing to the higher F1 and Accuracy), they are not robustly distinguishing between positive and negative returns. The longer horizon likely dampens daily noise, yet it still fails to reveal strong, consistent predictive power.

Key Feature Interpretability (Random Forest)

The most influential features in the Random Forest models consistently included:

- **OMC** (Open-Minus-Close)
- **BB_Width** (Bolatillity/Range)
- **Vol10** (Short-term Volatility)
- **RSI(14)** (Momentum)
- **MACD** (Trend-following)

This feature importance analysis confirms that the models rely on signals related to **volatility** and **short-term momentum/trend**. However, the poor performance metrics indicate that these signals are not translating into a reliable, exploitable trading edge. The models capture subtle directional tendencies but lack the necessary precision for profitable short-term forecasting.

---

5. Project Exposition and Deliverables

The following artifacts provide a clean, reproducible record of the analysis and are included in the designated GitHub repository:

- **`assignment1_divya.py`:** The complete, documented Python pipeline for data acquisition, feature engineering, modeling, and evaluation.
- **`equity_curve_test.png`:** Visual comparison of cumulative returns for the model-based strategies (t+1) versus the Buy-and-Hold strategy on the test set.
- **`equity_curve_test_t5.png:`** Visual comparison of cumulative returns for the model-based strategies (t+5) versus the Buy-and-Hold strategy on the test set.
- **`feature_importances.png`:** Graphical representation of the top 10 contributing features from the Random Forest model.
- **`README.md`:** Instructions for setup and reproduction of the analysis results.
- **`report.pdf`:** This technical documentation.

---

6. Conclusion and Future Work

Future research should focus on incorporating features that capture broader market regimes and non-price information, such as:

- **Macroeconomic Indicators:** Returns of SPY (S&P 500 ETF) or VIX (Volatility Index).
- **Sentiment Analysis:** Features derived from news or social media sentiment related to AAPL.
- **Advanced Time-Series Models:** Exploration of recurrent neural networks (RNNs) for pattern recognition.