MSDS 451 — Programming Assignment 1 Report

Author: Vrishani Shah
Asset: AAPL

Abstract

This project predicts the **next-day direction** of returns for Apple Inc. (AAPL) using only **price-based features**constructed from daily OHLCV data. I engineered lags of close, range (high–low), net change (open–close), volume lags, and short-half-life EMAs, and avoided leakage by shifting all features so they contain only information available at time *t*. A compact subset was selected via **AIC-based all-subsets logistic** screening; final modeling used an **XGBoost** classifier with **time-series cross-validation** (`TimeSeriesSplit` with a 10-day gap) and randomized hyperparameter search. The best cross-validated accuracy from the randomized search was **0.4920**, while the **training** confusion matrix for a fixed XGBoost configuration showed **~79.6%** accuracy (diagnostic only, not out-of-sample). Results are typical for one-day equity direction prediction with purely price-based features: signals are weak and noisy, but the pipeline is reproducible and leakage-safe.

---

1. Problem Description

- **Objective.** Binary classification: predict whether the next day's log return is **positive (1)** or **non-positive (0)**.
- **Motivation.** At a one-day horizon, equities may exhibit weak momentum/volatility structure; the goal is to implement a rigorous pipeline (feature engineering, feature selection, time-series CV, tuning) that respects temporal ordering.
- **Target.** `LogReturn = ln(Close_t / Close_{t-1})`; `Target = 1(LogReturn > 0)`.

---

2. Data & Feature Engineering

- **Source.** Yahoo! Finance via `yfinance`; CSV committed as `data/msds_getdata_yfinance_aapl.csv`.
- **Frequency & Window.** Daily bars from **2000-01-03** to **2025-09-24** (original rows: **6,471**).
- **Leakage controls.** All features are based on past information (lags/EMAs), then `drop_nulls()` removes initial burn-in rows (effective rows used by the model in your run: **6,468**; see confusion matrix totals).
- **Features (15 total).**
    - **Close lags:** `CloseLag1`, `CloseLag2`, `CloseLag3`

- **Range (HML) lags:** `HML = High - Low`, plus `HMLLag1..3`
- **Net change (OMC) lags:** `OMC = Open - Close`, plus `OMCLag1..3`
- **Volume lags:** `VolumeLag1..3`
- **Short-half-life EMAs** of `CloseLag1`: `CloseEMA2`, `CloseEMA4`, `CloseEMA8`

---

3. Research Design

- **Cross-validation.** `TimeSeriesSplit(n_splits=5, gap=10)` (forward-chaining, no shuffling). The 10-day **gap**further reduces subtle leakage from smoothing features.
- **Feature selection (AIC).** AIC-based all-subsets logistic screening was used to rank combinations, after which I fixed a compact subset used consistently in modeling:
  - `CloseLag3`, `HMLLag1`, `OMCLag2`, `OMCLag3`, `CloseEMA8`.
- **Model & tuning.** Final estimator: **XGBoost** (`binary:logistic`). Hyperparameters tuned via **RandomizedSearchCV** over `max_depth, min_child_weight, subsample, learning_rate, n_estimators`, using the same time-series CV object.
- **Metrics.** Fold-level **accuracy** for selection; final diagnostics include a confusion matrix (and ROC if probabilities are used).

---

4. Results

4.1 Cross-Validation (Randomized Search)

- **Best CV accuracy: 0.4920**

Best parameters:

```
{
  "learning_rate": 0.08964522558051516,
  "max_depth": 6,
  "min_child_weight": 5,
  "n_estimators": 881,
  "subsample": 0.8014120643245294
}
```

4.2 Final Model (diagnostic on full sample)

Your notebook then fit a **fixed XGBoost** configuration (separate from the best CV params) and printed diagnostics on the **training** set. As expected, these are optimistic compared to CV:

Confusion matrix (Actual rows × Predicted cols):
[

• 2337	• 750
• 570	• 2811

]

[2337570 7502811]Total = 6,468; **training accuracy ≈ 0.7959**.

- **Classification report:** header printed; detailed line items weren't captured in the HTML export.
- **ROC AUC:** not computed in this run (requires using `predict_proba`).

Interpretation. Cross-validated accuracy around **0.49–0.50** indicates little out-of-sample edge with pure price features for one-day direction on AAPL. The ~0.80 **training** accuracy reflects overfitting when evaluating on data used for fitting; the CV figure is the correct measure of generalization.

---

5. Discussion

- **Feature effects.** The chosen subset combines:
  (i) **CloseLag3** (short-lag momentum),
  (ii) **HMLLag1** (range/volatility proxy),
  (iii) **OMC** lags (intraday reversal/pressure), and
  (iv) a smoothed trend via **CloseEMA8**.
  This balances short-term direction and volatility/context while avoiding leakage.
- **Signal strength.** A **single-asset, one-day horizon** with price-only features typically yields weak signals; that's consistent with the CV result.
- **Threats to validity.** Non-stationarity (structural change) and class balance shifts. Time-ordered CV with a gap mitigates leakage but doesn't solve drift.
- **What would likely help.** Probability-based policies with thresholds, cost-aware backtesting, and **context features**(e.g., SPY/VIX lags) often provide incremental gains.

6. Conclusion & Future Work

The leakage-safe pipeline is reproducible and correct for a financial time series. On **AAPL**, a compact, price-only feature set yields **no reliable out-of-sample directional edge** over a 50/50 baseline. Next steps: (1) compute **ROC AUC** and calibrate probabilities, (2) add **market/context features**, (3) test **rolling re-tunes** for drift, and (4) run a **cost-aware**trading backtest with thresholding.

Reproducibility

- **Environment.** Python ≥3.10; key libraries: `polars`, `scikit-learn`, `xgboost`, `matplotlib`, `seaborn`.
- **Random seeds.** `random_state = 2025`.
- **Pipeline safety.** All features are lagged; no pre-CV scaling outside a pipeline; **time-series CV** with a **gap** used throughout.
- **Files.** Notebook (`.ipynb`) and **HTML export** included; data CSV (`data/msds_getdata_yfinance_aapl.csv`) committed.

References

- yfinance (data), scikit-learn (CV & metrics), XGBoost (model), Polars (data ops).
- Hastie, Tibshirani, Friedman — *Elements of Statistical Learning* (boosting, regularization).
- Hyndman, Athanasopoulos — *Forecasting: Principles and Practice* (time-series evaluation).