Research Report Checkpoint A

Title: Large-Cap Tech Directional Forecast Project (NVDA)

Author: Divyanka Thakur

Date: October 2025

1. Introduction

The goal of this project is to explore whether modest predictive signals derived from price, volatility, momentum, and regime/contextual features can provide an edge in forecasting short-term direction for technology stocks. I focus first on **NVIDIA (NVDA)** as a case study.

Potential users of this knowledge might include:

- Quantitative asset managers or hedge funds seeking incremental signals
- Active ETF issuers wanting algorithmic overlays
- Portfolio managers and analysts looking to complement fundamentals with systematic predictions

By building a reproducible pipeline and testing over multiple periods, the aim is to evolve toward a deployable model or signal system.

2. Literature Review

- **Active ETFs / actively managed vehicles**: The growth of active ETFs shows institutional interest in combining algorithmic approaches with real-world fund structures. For example, iShares has published whitepapers on how active ETFs are reshaping investor access to active strategies [https://www.ishares.com/us/literature/whitepaper/decoding-active-etfs.pdf] .
- Empirical studies on ETF performance find mixed evidence: some show active ETFs are more volatile and may not consistently beat passive benchmarks after costs [https://www.pm-research.com/content/iijindinv/4/4/10] . Others examine whether active ETFs are "really active" in their trading behavior [https://share.google/k1g2nagLoWSBkR6W6] .

- The theoretical foundation comes from **Grinold & Kahn's Active Portfolio Management**, which formalizes how manager skill, information coefficient, and breadth combine to produce alpha under risk control

- Recent surveys and reports show accelerated inflows and rising assets in the active ETF space (e.g. ETFGI's data on $1.30T in active ETF AUM globally) , and articles highlighting the structural growth in active ETFs amid market volatility and investor demand [https://www.spglobal.com/market-intelligence/en/news-insights/research/2025/06/the-rise-of-active-etfs-in-a-bull-market] .

This body of work motivates the exploration: there is growing interest in active/algorithmic funds, but success is uncertain, especially with short-term forecasting.

3. Methods

Data Acquisition & Preprocessing

- Ticker: **NVDA**

- Date range: 2000 to 2025

- Source: yfinance with fallback logic, if Adj Close is missing, we fallback to using Close column

- Clean and align data: drop rows with missing or invalid dates

Feature Engineering

- **Volatility**: rolling standard deviation for windows 5 and 20

- **Trend**: Simple Moving Averages (SMA5, SMA20)

- **Momentum**: RSI(14)

- Combine features into a single DataFrame aligned on date; drop rows with any NaNs

Exploratory Data Analysis

- Summary statistics (mean, std, quartiles)

- Histograms of individual features to inspect distribution and outliers
- Pairwise scatter plots to visualize relationships
- Correlation heatmap to detect collinearity

Baseline Modeling

- Target: next-day directional move (1 if return > 0, else 0)
- Feature subset: Vol5, RSI14, SMA5, SMA20
- Split: chronological train/test (no shuffling)
- Model: logistic regression (default settings, max_iter = 200)
- Evaluate: accuracy, confusion matrix, classification report

This checkpoint focuses on implementing the pipeline, feature set, and initial baseline – not yet full optimization.

4. Results

From running the baseline logistic model on NVDA:

Accuracy: 0.5447

Confusion matrix:

```
 [[  0 571]
  [  0 683]]
```

- The model **always predicts “up”** (never “down”)
- Precision and recall metrics reflect that behavior:
  - Precision for “down” is undefined (no predicted negatives)
  - Recall for “up” = 1.0
  - Accuracy ~ 54.5% reflects that ~54.5% of days were “up” in this sample

- Conclusion: with this limited feature set and simple model, the model cannot discriminate directionality; it defaults to the majority class.

This outcome is not unexpected given the inherent noise in daily returns and the simplicity of the baseline. It serves as a **benchmark baseline** against which future models must improve.

5. Conclusions & Concerns

At this stage:

- The pipeline is functioning, data is cleaned, features are working, and baseline modeling is executed.
- The baseline result is weak with no directional differentiation, but that is expected at this juncture.
- The key is improvement: the next versions must **beat this baseline** meaningfully through better features, modeling, and validation.

Concerns / caveats at this point:

- Overfitting risk when adding many features or using complex models
- Signal decay or non-stationarity over time
- Data snooping bias (optimism from testing many features)
- Need to ensure out-of-sample robustness via walk-forward validation

The path forward is iterative: augment features, validate carefully, compare models, and guard against overfitting. The goal is to build a disciplined, defensible predictive pipeline, not a perfect magic formula.