451 Feature Engineering: Programming Assignment 1
Google Trading Research Report from 2020 to 2025
Hongduo, SHAN

Abstract

We develop and rigorously evaluate a machine-learning pipeline to predict the next-day directional move of Alphabet Inc. ("GOOG") using only daily OHLCV data. After constructing a compact yet expressive feature set of lagged closes, trading ranges, reversal measures, and volume lags, we train an XGBoost classifier. A baseline model with default hyperparameters achieves a cross-validated accuracy of 0.784 ± 0.026. We then optimize five key hyperparameters via a 50-iteration RandomizedSearchCV under time-series cross-validation, yielding an improved CV accuracy of 0.809. The final in-sample model (trained on the full dataset with tuned parameters) attains 93.0% accuracy, 98.2% ROC AUC, and balanced precision/recall across both "up" and "down" classes. We close by discussing out-of-sample validation, feature interpretability, and avenues for future enhancement.

Introduction

Forecasting short-term price direction in equity markets remains a cornerstone problem in quantitative trading. Technical analysis posits that past patterns of price and volume contain signals of future moves. In this study, we ask:

- **How well can a tree-based learner distinguish "up" vs. "down" days for GOOG using only lagged and derived features?**
- **What lift does rigorous hyperparameter tuning provide over a default model?**
- **Which engineered features carry the most predictive weight?**

Our contributions:

- A reproducible Polars + XGBoost pipeline implementing time-series-aware CV.
- A demonstration that RandomizedSearchCV can boost CV accuracy by ~3% in this context.
- A full suite of diagnostic plots and metrics (ROC, confusion, precision-recall, feature importances).

Data & Feature Engineering

2.1 Data Source

- **Period:** January 2, 2020 – July 13, 2025
- **Frequency:** Daily trading sessions (NYSE)
- **Fields:** Open, High, Low, Close, Volume, plus corporate actions (Dividends, Stock Splits)

After loading with Polars and parsing dates, we drop “Dividends” and “Stock Splits” to focus purely on price/volume dynamics.

2.2 Feature Construction

We construct fifteen predictors, each capturing different market micro-dynamics:

Index	Feature	Description
0	CloseLag1	$\text{Close}_{\square-1}$
1	CloseLag2	$\text{Close}_{\square-2}$
2	CloseLag3	$\text{Close}_{\square-3}$
3	HML	$\text{High}_{\square} - \text{Low}_{\square}$
4	HMLLag1	$\text{HML}_{\square-1}$
5	HMLLag2	$\text{HML}_{\square-2}$
6	HMLLag3	$\text{HML}_{\square-3}$
7	OMC	$\text{Open}_{\square} - \text{Close}_{\square}$

8	OMCLag1	OMC_{t-1}
9	OMCLag2	OMC_{t-2}
10	OMCLag3	OMC_{t-3}
11	VolumeLag1	$Volume_{t-1}$
12	VolumeLag2	$Volume_{t-2}$
13	VolumeLag3	$Volume_{t-3}$

Key rationales:

- **Lagged closes** capture simple momentum.
- **HML (Range)** gauges intraday volatility.
- **OMC (Reversal)** measures whether the day closed lower than it opened (and vice versa).
- **Volume Lags** proxy trading intensity shifts.

Rows with nulls (due to lagging) are dropped, leaving ~3,800 observations. The **binary target** is set to 1 if Close_t > Close_{t–1}, else 0. Across the sample, “up” days constitute ~49%, so class balance is acceptable without reweighting.

Modeling Methodology

3.1 Time-Series Cross-Validation

We use TimeSeriesSplit(n_splits=5) with no shuffling, ensuring that each fold’s training set strictly precedes its test set chronologically. This guards against lookahead bias.

3.2 Baseline XGBoost

Hyperparameters:

default except n_estimators=1000, random_state=2025, eval_metric='logloss'.

Evaluation Metric: accuracy.

Running cross_validate on the baseline yields:

CV accuracy = 0.784 ± 0.026

This establishes a performance floor.

3.3 Hyperparameter Optimization

We define uniform and integer-range distributions for five XGBoost parameters:

Parameter	Search Range
`max_depth`	3–11
`min_child_weight`	1–9
`subsample`	0.5–1.0 (uniform)
`learning_rate`	0.01–0.30 (uniform)
`n_estimators`	100–1000

We employ `RandomizedSearchCV` with 50 draws, scoring by accuracy, and the same 5-fold time-series CV. This process takes ~5 minutes on a modern CPU with parallel jobs.

Results

4.1 Tuning Outcomes

Best hyperparameters found:

{ "max_depth": 5,

"min_child_weight": 9,

"subsample": 0.96,

"learning_rate": 0.011,

"n_estimators": 986}

Tuned CV accuracy: 0.809

Thus, tuning lifts CV accuracy by ~3 percentage points, indicating the default tree depth and shrinkage were suboptimal.

4.2 Final In-Sample Performance

Retraining on the full dataset with the optimal settings yields:

Metric	Value
Accuracy	0.930
Precision	0.932
Recall	0.934
F_1-Score	0.933

ROC AUC	0.982

- **The Confusion Matrix** shows ~92% true positive and 94% true negative rates.
- **ROC Curve** hugs the top-left corner, confirming strong discriminative power.
- **Precision-Recall Curve** remains steep, indicating the model retains high precision even at high recall.

Feature Importance & Diagnostics

5.1 Gain-Based Importance

According to XGBoost's built-in gain metric, the top five features are:

- **CloseLag3**
- **HMLLag1**
- **OMCLag2**
- **VolumeLag1**
- **CloseLag1**

This suggests that momentum from three days prior and recent volatility spikes are most informative.

5.2 Residual Analysis

Plotting misclassified days against realized volatility reveals that errors concentrate on low-range days (HML small), implying that flat markets are harder to predict.

Discussion & Limitations

- **In-Sample vs. Out-of-Sample:** Our final metrics are computed in-sample; real predictive power must be validated on a strictly held-out test period or via walk-forward backtesting.
- **Feature Scope:** All features derive from GOOG alone. Incorporating cross-asset signals (e.g., S&P 500, VIX, US Dollar index) or macroeconomic release indicators could capture broader market regimes.
- **Market Regime Shifts:** A single static model may degrade across bull and bear phases; dynamic model re-training or regime-aware methods could mitigate drift.

Conclusions & Next Steps

- **Hyperparameter tuning** yielded a clear performance uplift (+3% CV accuracy).
- **The final model** demonstrates strong directional accuracy (93%) and discriminatory capacity (AUC 0.98).

- **Immediate next step:** Implement a walk-forward backtest on 2020–2025 hold-out, measuring realistic P&L using a simple long/short strategy.
- **Extended work:** Augment feature set with multi-asset and sentiment data; explore ensemble stacking with logistic regression meta-learners; apply SHAP explanations for robust interpretability.

References:

[1] Hyndman, R. J. & Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed., OTexts, 2021.

This paper has used the GenAI(GPT4.5) to modify the words and paragraphs.