Background: Duke Kunshan University (DKU) is an interdisciplinary institution that grants dual undergraduate degrees, an MOE Chinese degree and a degree from Duke University in Durham, United States. The principal structure of DKU majors is robustly interdisciplinary. No student confines their study to a single discipline (for example, biology or economics). Instead, all students engage in broad inquiry related to a subject or question (for example, political economy or global health) and take a wide variety of courses related to that area (for example, in public policy, history, ethics, or economics). As a result, our graduates are prepared to engage in a wide variety of inquiries using multiple methodologies to address complex issues that require interdisciplinary approaches. This has implications for our vision and expectations of undergraduate theses and design projects, which reflect this broad interdisciplinary training. At DKU, every student completes a two-year project known as signature work which consists of multiple interconnected parts including thematic courses, experiential learning, capstones, and a final product. It seeks to integrate students' interdisciplinary educational experience and culminates in the creation of a product in a scholarly, creative, or applied nature in leu of an undergraduate thesis or design required by JED. Because DKU encourages students to cultivate their independence and creativity as one of its institutional student learning outcomes, the student-led signature work projects often reflect students' own particular interdisciplinary interests and training. In addition, signature work has an intensive emphasis on problem-solving and skill-development which is much needed for any interdisciplinary inquiry; thus, students' final products are evidence of transferrable skills that students have acquired and demonstrated through the 2-year program, rather than content knowledge narrowly defined by disciplinary training. In sum, while the Chinese major declared with any given student might be construed narrowly, the experience of our students is much broader—and intentionally so. This is a distinctive feature of our curriculum, and this distinctiveness results in broadly interdisciplinary submissions from our graduates' submitting theses or design projects. We have designed this to prepare our students for a wide variety of graduate programs in China and the West, where interdisciplinary training is a competitive advantage.

LEVERAGING FINBERT SENTIMENT AND MACRO FACTORS FOR CROSS-SECTIONAL STOCK RETURN PREDICTION AROUND FEDERAL RESERVE EVENTS

by

Yihan Wang

Signature Work Product, in partial fulfillment of the
Duke Kunshan University Undergraduate Degree Program

June 2025

Signature Work Program
Duke Kunshan University

APPROVALS

Mentor: Shixin Xu, Division of Natural and Applied Sciences

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, Prof. Shixin Xu, for his continuous guidance, encouragement, and patience throughout this Signature Work project. His ideas and feedback shaped both the design of the models and the way I think about event-driven prediction in financial markets.

I am also very grateful to Prof. David Ye, with whom I previously worked on FinBERT-based sentiment analysis for financial text. The experience and insights from that earlier project directly inspired the use of sentiment features in this work and provided an important foundation for my current research.

Finally, I would like to thank my parents, my mother and my father, for their unwavering support, understanding, and trust. Their encouragement has been essential for me to complete this project and my undergraduate studies.

ABSTRACT

This Signature Work studies whether information surrounding U.S. Federal Reserve policy announcements can help predict cross-sectional stock performance over medium horizons. Using a panel of 498 U.S. equities linked to Federal Open Market Committee (FOMC) decision dates between 2020 and 2025, the study constructs three groups of features: (i) Fed and macro variables, including policy rate changes, gold, oil, and U.S. dollar index levels, together with FinBERT-based sentiment from policy communications; (ii) stock-level price and volatility statistics computed over 5–20 day rolling windows; and (iii) structural indicators such as sector and capitalization bucket.

The prediction target is whether a stock outperforms its capitalization peers over the following 60 trading days. The study compares logistic regression, gradient boosting, XGBoost, and an LSTM sequence model using a walk-forward evaluation design. Performance peaks in early 2025 and deteriorates sharply in later regimes, with tree-based models and a simple ensemble achieving strong AUC in 2025Q1 and weaker, regime-dependent results in 2024Q4 and 2025Q2 as cross-sectional co-movement broadens. The results indicate that Fed-related and macro features contain modest but regime-dependent predictive information for relative stock selection.

Keywords: Federal Reserve events; cross-sectional equity prediction; FinBERT sentiment; macro-financial factors; machine learning ensemble.

Chapter 1

INTRODUCTION

1.1 Motivation

Decisions by the Federal Reserve's Federal Open Market Committee (FOMC) affect discount rates, risk premia, and growth expectations, often generating pronounced market reactions. A large body of literature documents these effects using short-horizon event studies around announcement windows, with a primary focus on aggregate indices or broad asset classes [3, 19]. However, investors are equally concerned with *cross-sectional* outcomes: which stocks or sectors outperform following policy announcements?

This study evaluates predictability across three consecutive quarters (2024Q4–2025Q2) using a walk-forward design. It examines whether Fed-related information—policy rate changes, macro-financial variables, and sentiment from official communications—together with recent stock behavior, can identify relative winners and losers over a 60-trading-day horizon.

1.2 From Prior Sentiment Work to This Project

This Signature Work builds on earlier FinBERT-based sentiment research with Prof. David Ye, which focused on extracting sentiment from financial text and relating aggregated sentiment measures to index-level market returns. Transformer-based language models such as FinBERT have been shown to capture context-dependent tone in financial text more effectively than traditional dictionary-based approaches [1]. That project provided experience building an end-to-end NLP pipeline from text to usable sentiment factors.

In the present capstone project with Prof. Shixin Xu, the study extends this framework to a more demanding cross-sectional setting. Instead of a single market time series, a stock–date panel is constructed in which each observation corresponds to an individual equity on a specific date. Focusing on periods around FOMC meeting dates between 2020 and 2025, the panel combines: (i) Fed-related variables (policy actions and surprises), (ii) macro-financial proxies such as gold, oil, and the U.S. dollar index, (iii) FinBERT-based sentiment extracted from Fed statements and related news, and (iv) rolling stock-level price and volatility features.

Methodologically, the project shifts from time-series analysis to cross-sectional machine learning. Machine learning can capture non-linear interactions among firm characteristics and improve cross-sectional return prediction [13]. The study compares logistic regression, gradient boosting, XGBoost, and an LSTM sequence model on the same leakage-free panel, extending the earlier sentiment work to a distinctly cross-sectional question.

1.3 Research Questions and Scope

The central goal of this study is to examine whether Fed-related information, augmented with macro and stock-specific features, contains predictive power for *cross-sectional* equity returns. Specifically, the analysis addresses the following three questions:

- **Q1.** Can a cross-sectional machine learning model trained on Fed-related sentiment, macro variables, and recent price and sector information identify stocks that *outperform their size peers* following FOMC decisions?
- **Q2.** Are these predictive relationships stable across macro regimes—for example, across 2024Q4–2025Q2—or do they deteriorate as market conditions change?
- **Q3.** Which feature blocks contribute the most to the prediction and in which sectors or capitalization buckets do errors concentrate?

To study these questions, the study constructs a daily U.S. equity panel centered on FOMC decision dates. The universe spans multiple sectors (financial, information technology, consumer discretionary, energy, and industrials) and three capitalization buckets (small, mid, and large cap). For each stock–date, features are computed from rolling windows of past prices and returns (5–20 day statistics), together with contemporaneous macro variables and Fed-related sentiment signals.

The prediction target is a medium-horizon relative-performance label. Specifically, a stock is labeled as an outperformer if its cumulative return over the subsequent 60 trading days exceeds the average return of stocks in the same capitalization bucket. Evaluation follows a walk-forward design with three validation windows: 2024Q4 (October–December 2024), 2025Q1 (January–March 2025), and 2025Q2 (April–June 2025). Data through 2025Q3 are available; **2025Q3 is reserved for out-of-sample testing only**—no threshold tuning, model selection, or parameter choice uses Q3 data, so all reported results are not conditioned on it. Models are always trained on historical data and tested only on future observations.

1.4 Contributions

This study contributes along three key dimensions.

Methodological integration. The study develops a leakage-free panel that integrates Fed policy events, macro-financial variables, FinBERT-based sentiment, rolling price and volatility features, and sector and size characteristics into a unified cross-sectional prediction framework with explicit walk-forward validation.

Empirical evidence under regime variation. Using three non-overlapping walk-forward validation windows (2024Q4, 2025Q1, 2025Q2), the analysis shows that tree-based models and their ensemble achieve strong AUC in the most favorable regime (2025Q1) but deteriorate materially in 2024Q4 and 2025Q2, providing structured evidence that Fed-event-driven cross-sectional predictability is present but regime-dependent.

Error diagnostics and interpretation. Beyond headline performance metrics, the study conducts sector-level, size-bucket, and temporal error analysis to identify where models systematically fail, highlighting the importance of diagnostic transparency in empirical financial prediction.

1.5 Road-map of the Paper

Chapter 2 reviews the related literature on monetary policy, sentiment extraction, and cross-sectional return prediction. Chapter 3 describes the data construction process and defines the relative-performance labeling framework. Chapter 4 outlines the feature engineering strategy and model specifications, while Chapter 5 presents the walk-forward evaluation design. Chapter 6 reports the empirical findings and corresponding backtesting results. Chapter 7 provides a detailed examination of prediction errors across sectors, market-cap segments, and time periods. Finally, Chapter 8 discusses limitations and highlights directions for future research.

Chapter 2

BACKGROUND AND RELATED WORK

This study lies at the intersection of three strands of literature: monetary policy and asset prices, text-based sentiment analysis in finance, and machine learning methods for cross-sectional return prediction. Rather than treating these areas in isolation, the project integrates insights from all three to study stock-level heterogeneity around Federal Reserve policy events.

2.1 Monetary Policy and Asset Prices

Unexpected changes in Federal Reserve policy affect discount rates, risk premia, and expectations about future economic activity. A large empirical literature documents asset-price reactions around FOMC announcement windows, often using high-frequency or short-horizon event-study designs [see 4, 8, for a review]. Early work shows that monetary policy surprises have immediate effects on equity indices and bond yields [3, 19], and policy uncertainty itself can independently move markets [2].

More recent studies extend this analysis to sectoral and firm-level responses, emphasizing heterogeneity in interest-rate sensitivity across industries and balance-sheet characteristics [21]. However, much of the existing evidence remains focused on aggregate outcomes or narrow event windows. Relatively less attention has been paid to whether Fed-related information can systematically distinguish *cross-sectional winners and losers* over medium horizons. This project adopts that micro-level perspective, asking whether policy decisions, macro conditions, and Fed-related sentiment help predict relative stock performance beyond short-term announcement effects.

2.2 Sentiment Analysis in Finance

Textual information has become an increasingly important input in financial research. Early approaches relied on dictionary-based sentiment measures to quantify tone in news articles and corporate disclosures [20, 22, 23]. While interpretable, such methods struggle with context dependence and domain-specific language.

Recent advances in natural language processing, particularly transformer architectures [7], enable richer representations of financial text. FinBERT, a BERT-based model fine-tuned on financial corpora, classifies sentiment as positive, negative, or neutral while accounting for context and syntax [1]. Empirical studies show that FinBERT-based sentiment measures contain incremental information for asset prices and market dynamics [16].

In prior work with Prof. David Ye, FinBERT was applied to index-level market analysis by aggregating sentiment over time. In the present project, sentiment is instead treated as one feature block within a broader cross-sectional framework, combined with macro variables, sector indicators, and firm-level price dynamics. This design reflects the view that textual sentiment is informative, but only as part of a larger information set rather than a standalone signal.

2.3 Machine Learning for Cross-Sectional Prediction

Traditional factor models, beginning with Fama and French [9] and extended by Fama and French [10], are typically built on linear factor structures [6]. Recent research demonstrates that machine learning methods—including tree-based ensembles and neural networks—can capture non-linear interactions among firm characteristics and improve cross-sectional return prediction [11, 13, 17]. These methods are particularly well suited to high-dimensional settings where the relevant interactions are unknown ex ante.

At the same time, the literature emphasizes several challenges. Financial data are highly non-stationary, cross-sectional dependence is strong, and careless evaluation can lead to severe information leakage. Performance gains reported in one period often fail to generalize across regimes. Motivated by these concerns, this project prioritizes clean panel construction, walk-forward validation, and detailed error analysis over architectural complexity. The goal is not to maximize predictive accuracy in a single window, but to understand when and why Fed-event-driven signals succeed or fail across different market environments.

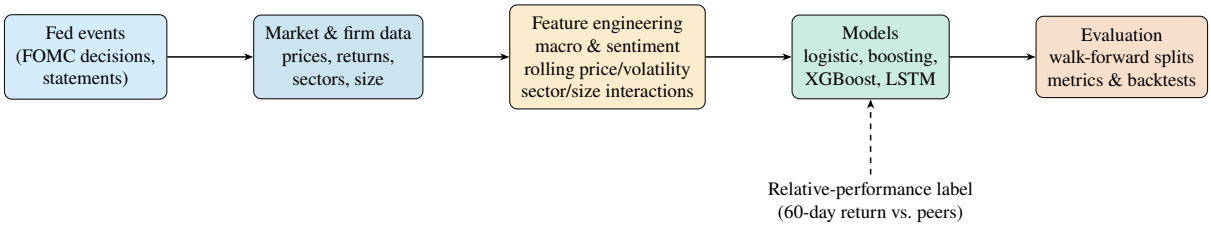

Figure 2.1: Overview of the Fed-event-driven cross-sectional prediction pipeline.

Chapter 3

DATA

This chapter summarizes the construction of the equity panel: raw data and time coverage, sector and size structure, the relative-performance label, and descriptive statistics on label distribution across sectors, cap buckets, and quarters.

3.1 Raw Data and Time Coverage

The empirical analysis uses a *daily panel* of U.S. equities augmented with macro-financial variables and Fed-related sentiment features. The processed panel contains **157,481 stock–date observations** across **315 trading days**, spanning **1 July 2024** to **1 October 2025**. Each observation corresponds to a single stock on a given trading day and includes contemporaneous macro and sentiment variables, recent price-based features, sector and capitalization indicators, and (when available) the forward-looking relative-performance label defined in Section 3.3.

The panel is a continuous daily series over the full sample; FOMC dates enter only via *event-related features* (policy rate changes, surprises, sentiment), not through sample selection.

Data sources. Equity prices from Bloomberg (proprietary); FOMC and policy-event data from Federal Reserve sources; macro and auxiliary series (gold, oil, U.S. dollar) from `yfinance` and FRED. Non-proprietary series can be regenerated from the codebase.

3.2 Stock Universe and Sector / Size Structure

The stock universe is designed to be balanced across sectors and capitalization buckets, both to reflect the diversity of the U.S. equity market and to avoid having the models dominated by a single industry or size segment. The panel contains **498 distinct tickers** in total,[1] drawn from **five** GICS-style sectors:

- consumer discretionary (`cd`),
- energy,
- financials,
- industrials,
- information technology (`it`).

By construction, each of the five sectors contributes approximately **100 stocks** to the universe (Table 3.1):

$$\text{tickers per sector} \approx 100 \text{ for each of cd, energy, financials, industrials, and it.}$$

On the size dimension, every stock is assigned to one of three capitalization buckets: *large cap*, *mid cap*,

[1] The nominal universe comprises $5 \times 100 = 500$ tickers (Table 3.1); two were dropped owing to insufficient price history, yielding 498 in the final panel.

Table 3.1: Nominal universe: five sectors × 100 tickers each; two tickers dropped for insufficient history, yielding 498.

Dimension	Count
Sectors	5 (cd, energy, financials, industrials, it)
Nominal tickers per sector	100
Total distinct tickers	498

or *small cap*. Counting distinct tickers by bucket, the panel contains:

150 large-cap, 175 mid-cap, 173 small-cap stocks.

This balanced structure ensures that the cross-sectional prediction problem is not dominated by mega-cap technology stocks alone and allows the study of how predictive performance and error patterns vary across both sectors and size buckets.

3.3 Relative-Performance Label

The prediction target is a medium-horizon *relative-performance* label defined at the stock–date level. For each stock i and date t, the cumulative return over the next $H = 60$ trading days is computed as

$$R_{i,t}^{(60)} = \prod_{h=1}^{60}(1 + r_{i,t+h}) - 1,$$

where $r_{i,t+h}$ is the simple daily return. On the same date t, the average cumulative return over the same horizon for all stocks in the *same capitalization bucket* as i is computed as

$$\bar{R}_{b(i),t}^{(60)} = \frac{1}{N_{b(i),t}} \sum_{j \in b(i)} \left[\prod_{h=1}^{60}(1 + r_{j,t+h}) - 1 \right],$$

where $b(i)$ denotes the size bucket (large, mid, or small cap) containing stock i, and $N_{b(i),t}$ is the number of stocks in that bucket that have valid returns over the next 60 days.

The binary label $Y_{i,t}$ is defined as

$$Y_{i,t} = \begin{cases} 1, & \text{if } R_{i,t}^{(60)} > \bar{R}_{b(i),t}^{(60)}, \\ 0, & \text{if } R_{i,t}^{(60)} \le \bar{R}_{b(i),t}^{(60)}, \\ \text{missing}, & \text{if the 60-day window is not fully observed.} \end{cases}$$

Thus $Y_{i,t} = 1$ indicates that stock *i beats its size peers* over the subsequent 60 trading days, while $Y_{i,t} = 0$ indicates underperformance relative to the size-bucket benchmark.

In the final panel, the label column (`label`) has the following distribution:

- 60,125 observations with $Y = 0$,
- 67,476 observations with $Y = 1$, and
- 29,880 observations with missing labels.

Missing labels occur where the 60-day forward return window is not fully observed (e.g., for the most recent 60 trading days in the sample or when a stock exits the universe). Out of 157,481 total stock–date rows, **127,601** carry a valid label, and the overall positive rate among labeled observations is **0.5288**. All model training and evaluation are conducted on the subset with non-missing labels.

3.4 Descriptive Label Statistics

This subsection documents how the positive rate varies by sector, capitalization bucket, and calendar quarter.

By Sector and Capitalization Bucket

Table 3.2 reports the average label value $\mathbb{E}[Y_{i,t}]$ by sector and by capitalization bucket.

Table 3.2: Positive label rates by sector and capitalization bucket.

Panel A: By sector

Sector	Pos. rate
Consumer discretionary (cd)	0.508
Energy	0.453
Financials	0.600
Industrials	0.548
Information technology (it)	0.535

Panel B: By capitalization bucket

Bucket	Pos. rate
Large cap	0.553
Mid cap	0.536
Small cap	0.501

By Calendar Quarter

Table 3.3 reports the positive label rate by calendar quarter.

Table 3.3: Positive label rates by calendar quarter.

Year-quarter	Pos. rate
2024Q3	0.626
2024Q4	0.422
2025Q1	0.307
2025Q2	0.748
2025Q3	0.599

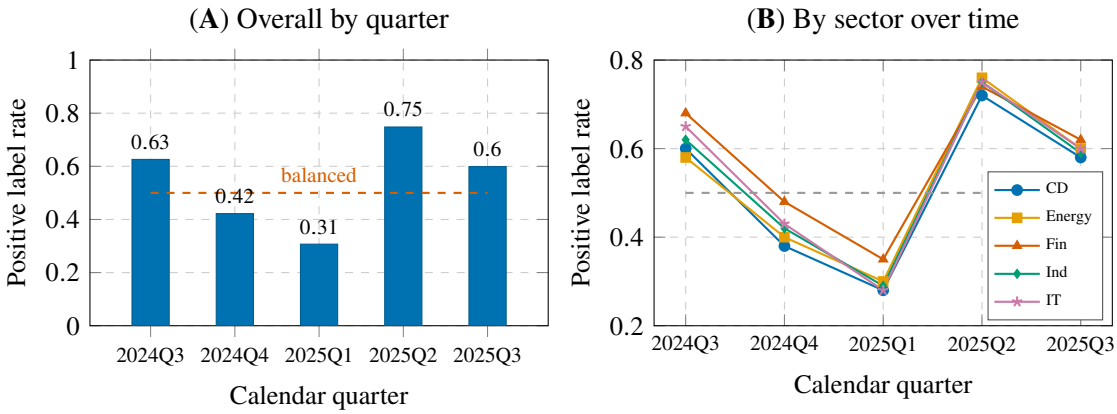

Figure 3.1: (**A**) Positive label rate by calendar quarter (overall). (**B**) Time evolution of the positive label rate by sector (cd, energy, financials, industrials, it). Both panels show strong non-stationarity; sector-level rates move with the aggregate but with distinct cross-sectional dispersion across quarters.

The positive rate varies substantially over time, reflecting changes in the market environment and cross-sectional dispersion (Figure 3.1, panel A). Panel B plots the *time evolution* of the positive label

rate by sector (cd, energy, financials, industrials, it). A striking feature is the **pronounced trough in 2025Q1**: all five sectors reach their minimum rate in that quarter (roughly 0.28–0.35), well below the balanced baseline of 0.5. Such a synchronized, across-the-board decline may reflect shifts in the broader macro or policy environment during that period—for example, changed expectations about interest rates, fiscal stance, or regulatory priorities—which would alter cross-sectional return dispersion and the share of stocks that go on to outperform their size peers. From 2025Q1 the rates rise sharply into 2025Q2 and then ease in 2025Q3, while the *level* of the rate continues to differ by sector in each quarter (e.g., financials tend to sit above the average, consumer discretionary and information technology closer to or below it). This sector-level and time-varying non-stationarity implies that a single threshold or a model trained on one regime may mis-calibrate in another, and motivates both the walk-forward evaluation design and the sector-specific error analysis in Chapter 7. The overall non-stationarity also helps explain the regime-dependent performance patterns observed in Chapter 6.

Chapter 4

FEATURES AND MODELS

The panel in Chapter 3 is converted into model inputs: tabular models use **56 features** per stock–date; the LSTM uses a **15×43** sequence. Each model outputs $\hat{p}_{i,t} \in (0, 1)$ for the 60-day relative-performance label (Section 3.3). The choice of tree-based and sequence models follows evidence that they capture non-linear interactions and improve cross-sectional prediction [11, 13].

4.1 Feature Blocks for Tabular Models

For logistic regression, gradient boosting, and XGBoost, each observation (i, t) is represented by $x_{i,t} \in \mathbb{R}^{56}$. The 56 features are organized into three blocks.

Table 4.1: Feature blocks used by tabular models. The full design yields a 56-dimensional feature vector per stock–date.

Block	Description
Fed and macro / sentiment	Indicators for FOMC hike/cut/hold decisions, magnitudes of policy rate changes and cumulative surprises, levels of gold, oil, and U.S. dollar indices, and daily FinBERT-based sentiment scores aggregated from Fed statements and related financial news.
Stock price and volatility history	Rolling averages, minima, maxima, realized volatility, and simple momentum statistics of each stock's recent prices and returns over 5–20 day windows, capturing how the stock has behaved leading up to date t.
Structural characteristics	One-hot sector indicators (consumer discretionary, energy, financials, industrials, information technology), capitalization buckets (large, mid, small cap), and simple interactions between these characteristics and Fed-related variables (for example, sector dummies multiplied by policy surprise indicators).

The feature set is compact (macro levels and simple transforms; rolling 5/10/20-day windows for price/volatility). Logistic regression provides a linear benchmark; the LSTM uses a related but more compact daily feature set below.

4.2 Sequence Input for the LSTM

To capture short-run dynamics in returns and volatility, the LSTM replaces the single vector $x_{i,t}$ with a short history

$$\left(z_{i,t-14}, z_{i,t-13}, \ldots, z_{i,t}\right), \qquad z_{i,s} \in \mathbb{R}^{43}.$$

Each $z_{i,s}$ collects daily quantities such as returns (simple and log), short-horizon volatility/range measures, a subset of aligned Fed/macro variables, and time-invariant indicators (sector and size bucket).

The LSTM input is a 15×43 tensor per stock–date [11]; output $\hat{p}_{i,t}$ uses the same label and evaluation as tabular models. Performance comparisons across families should account for the different input dimensionality (43 vs. 56).

4.3 Model Families

All models are trained to approximate $\mathbb{P}(Y_{i,t} = 1 \mid \text{features at } (i,t))$ on the labeled subset of the panel. They differ mainly in functional form and how non-linearities and interactions are represented.

Logistic Regression

Logistic regression provides a transparent baseline and corresponds to a generalized linear model widely used in empirical asset pricing. For $x_{i,t} \in \mathbb{R}^{56}$,

$$\log \frac{\mathbb{P}(Y_{i,t} = 1 \mid x_{i,t})}{\mathbb{P}(Y_{i,t} = 0 \mid x_{i,t})} = \beta_0 + \beta^\top x_{i,t}.$$

Parameters are estimated by minimizing binary cross-entropy with ℓ_2 regularization. Class weights (based on training-set frequencies) mitigate class imbalance.

Gradient Boosting on Decision Trees

Gradient boosting fits an additive ensemble of shallow trees to minimize binary cross-entropy. Tree-based ensembles have become standard tools in high-dimensional cross-sectional prediction problems because they flexibly model non-linear interactions without explicit specification [13, 14]. Each boosting step adds a small CART-style tree fit to the negative gradients [12]; the ensemble score is passed through a logistic link to yield probabilities. Limited tree depth and minimum leaf sizes regularize the model and reduce overfitting.

XGBoost

XGBoost [5] adds ℓ_1/ℓ_2 regularization and subsampling to gradient boosting. It is trained on the same 56 features with `scale_pos_weight` for label imbalance. Outputs are evaluated by threshold-based metrics and ranking-based backtests.

LSTM Sequence Model

The LSTM uses a standard many-to-one architecture [15]: the 15×43 sequence is processed by an LSTM to produce a final hidden state, followed by a small fully connected layer and a sigmoid output. Training minimizes binary cross-entropy. Capacity is kept modest (one LSTM layer) to reduce overfitting under the short sample and walk-forward design.

4.4 Training Details and Model Summary

All models follow the same stage-wise walk-forward protocol (Chapter 5): for each quarter, training uses all labeled data up to the cutoff date and validation uses only the subsequent quarter, preventing look-ahead bias.

For logistic regression and the LSTM, continuous inputs are standardized using training-set means and standard deviations only, and the same transformation is applied to validation. Tree-based models are trained on raw feature scales.

Table 4.2 provides a compact overview of inputs and key settings; performance results are reported in Chapters 6 and 7.

Table 4.2: Summary of model families and input formats.

Model family	Input type	Modeling characteristics
Logistic regression	56-dim. tabular	Linear baseline with ℓ_2 regularization and class-weighted loss to account for label imbalance
Gradient boosting	56-dim. tabular	Additive ensemble of shallow decision trees with a small learning rate, capturing non-linearities while controlling overfitting
XGBoost	56-dim. tabular	Regularized gradient-boosted trees with sub-sampling and imbalance-aware weighting via `scale_pos_weight`
LSTM sequence model	15×43 sequence	Many-to-one recurrent architecture modeling short-horizon temporal dynamics, trained with cross-entropy loss

Chapter 5

EXPERIMENTAL DESIGN AND EVALUATION

This chapter describes the walk-forward design, preprocessing, training protocol, evaluation metrics, and a simple backtest linking predicted rankings to realized performance.

5.1 Walk-Forward Train/Validation Splits

To avoid look-ahead bias, a walk-forward design is adopted in which models are trained only on past data and evaluated on strictly later periods. Three non-overlapping validation windows are considered:

- **2024Q4:** Train ≤ 2024-09-30; validate 2024-10-01–2024-12-31 (21,946 train, 31,930 validation rows).
- **2025Q1:** Train ≤ 2024-12-31; validate 2025-01-01–2025-03-31 (53,876 train, 29,940 validation rows).
- **2025Q2:** Train ≤ 2025-03-31; validate 2025-04-01–2025-06-30 (83,816 train, 30,938 validation rows).

Data extend through 2025Q3. This report uses 2024Q4–2025Q2 for evaluation and threshold tuning; **2025Q3 is used solely as an out-of-sample test set**: no hyperparameters, thresholds, or model choices are selected using Q3, so 2025Q3 serves as a strict once-only evaluation.

Feature construction (rolling statistics, macro alignment, and label definition) uses only information strictly prior to each prediction date t. The temporal split is applied after feature engineering, ensuring no leakage from validation into training.

Software and computing environment. Experiments were run in Python 3.12 on an Apple M3 MacBook Pro using `scikit-learn`, `xgboost`, and `PyTorch`. Seeds were fixed per stage; standardization used training-set statistics only and was applied unchanged to validation.

Table 5.1 summarizes the three evaluation stages.

Table 5.1: Walk-forward evaluation stages and sample sizes.

Stage	Train period	Validation period	$(n_{\text{train}}, n_{\text{val}})$
2024Q4	≤ 2024-09-30	2024-10-01 – 2024-12-31	(21,946, 31,930)
2025Q1	≤ 2024-12-31	2025-01-01 – 2025-03-31	(53,876, 29,940)
2025Q2	≤ 2025-03-31	2025-04-01 – 2025-06-30	(83,816, 30,938)

5.2 Preprocessing

Models use either the 56 tabular features (logistic regression, gradient boosting, XGBoost, and the ensemble) or 15×43 sequences (LSTM).

Within each stage:

- Rows with missing labels are dropped.
- Logistic regression and LSTM inputs are standardized using training-set means and standard deviations only; the same transformation is applied to validation.
- Tree-based models are trained on original feature scales.

For the LSTM, 15-day rolling sequences are constructed within each stock. Sequences are assigned entirely to either the training or validation set according to the final date in the window, so that no sequence spans the train/validation boundary.

5.3 Model Training

All models are trained on the same stage-specific training set and evaluated on the same validation set.

Logistic regression. An ℓ_2-regularized logistic model is fit to the 56 features with class weights reflecting training-set label frequencies. Hyperparameters are tuned coarsely on the earliest validation window (2024Q4) and then held fixed for the subsequent stages (2025Q1 and 2025Q2).

Gradient boosting. Shallow decision trees (depth 3–4) with a small learning rate are used. Settings are chosen to balance fit and stability and kept constant across stages.

XGBoost. A gradient-boosted tree classifier with moderate depth, subsampling, and `scale_pos_weight` is trained. Hyperparameters are fixed after initial tuning, and performance is reported using the out-of-sample validation windows.

LSTM. A many-to-one LSTM processes 15-day sequences and outputs a probability via a sigmoid layer. A single LSTM layer with moderate hidden size is trained using Adam [18] and a small learning rate for a limited number of epochs to reduce overfitting under the short sample span and regime shifts.

Ensemble. The ensemble averages predicted probabilities from the three tabular models (logistic regression, gradient boosting, and XGBoost). The mean probability is used for both classification and ranking. The LSTM is reported as a separate model rather than being included in the ensemble, because it uses a different sequence-based input representation and a partially different feature set; mixing it with the tabular models would make probability calibration and model-to-model comparability less direct within the scope of this study.

5.4 Evaluation Metrics

Let $\hat{p}_{i,t}$ denote the predicted probability and $Y_{i,t}$ the true label. For threshold τ,

$$\hat{Y}_{i,t}(\tau) = \mathbf{1}\{\hat{p}_{i,t} \geq \tau\}.$$

Metrics reported on validation sets include:

- Accuracy
- Precision, recall, and F1-score (positive class)
- Balanced accuracy
- Area under the ROC curve (AUC)

AUC evaluates ranking ability independent of threshold. In addition to $\tau = 0.5$, operating thresholds are tuned *within each validation window* to maximize balanced accuracy. For the ensemble, the balanced-accuracy-optimal thresholds are approximately $\tau = 0.38$ (2024Q4), $\tau = 0.40$ (2025Q1), and $\tau = 0.60$ (2025Q2). Confusion matrices are further examined by sector and size bucket (Chapter 7).

5.5 Backtesting

To assess economic relevance, a simple long-only strategy is implemented. On each validation date t:

1. Rank stocks by $\hat{p}_{i,t}$.
2. Select the top α fraction (e.g., 20%) as the long portfolio $\mathcal{L}_t$.
3. Compute 60-day cumulative returns $R_{i,t}^{(60)}$.
4. Define excess return relative to the contemporaneous labeled universe $\mathcal{B}_t$:

$$\Delta_t^{(60)} = \frac{1}{|\mathcal{L}_t|} \sum_{i \in \mathcal{L}_t} R_{i,t}^{(60)} - \frac{1}{|\mathcal{B}_t|} \sum_{j \in \mathcal{B}_t} R_{j,t}^{(60)}.$$

Average $\Delta_t^{(60)}$ across validation dates measures the portfolio's relative performance. This backtest ignores transaction costs, turnover, and risk constraints; it is used only to connect probabilistic rankings to realized 60-day relative performance.

Chapter 6

RESULTS

This section reports out-of-sample performance for four base models (logistic regression, gradient boosting, XGBoost, and an LSTM sequence model) and a stacking-style ensemble (Chapter 4). Metrics are computed under the walk-forward validation stages in Chapter 5: 2024Q4 (2024-10-01–2024-12-31), 2025Q1 (2025-01-01–2025-03-31), and 2025Q2 (2025-04-01–2025-06-30).

Unless stated otherwise, classification metrics are reported at the default probability threshold $\tau = 0.5$. To make results comparable across quarters with different label base rates (Chapter 3), a *validation-tuned* operating threshold is also reported. Throughout the paper, tuned thresholds are chosen to **maximize balanced accuracy** on the validation set. Balanced accuracy weights the two classes equally and is therefore more appropriate than optimizing F1 when the positive rate shifts materially across regimes.

6.1 Base Model Comparison Across Stages

Table 6.1 reports the four base models' performance on the three validation windows.

Table 6.1: Out-of-sample performance of base models on 2024Q4, 2025Q1, and 2025Q2 validation sets (reported at $\tau = 0.5$).

Stage	Model	AUC	Bal. acc.	Acc.	F1 (pos)
2024Q4	Logistic regression	0.507	0.511	0.511	0.430
	Gradient boosting	0.558	0.540	0.540	0.534
	XGBoost	**0.578**	**0.557**	**0.557**	0.556
	LSTM sequence	0.558	0.542	0.542	**0.624**
2025Q1	Logistic regression	0.551	0.534	0.534	0.478
	Gradient boosting	0.603	0.572	0.572	**0.596**
	XGBoost	**0.626**	**0.590**	**0.590**	**0.616**
	LSTM sequence	0.603	0.573	0.573	0.546
2025Q2	Logistic regression	0.464	0.473	0.473	0.337
	Gradient boosting	0.507	0.505	0.505	0.489
	XGBoost	0.525	0.510	0.510	0.504
	LSTM sequence	**0.588**	**0.551**	**0.551**	**0.611**

Three patterns stand out. First, in **2024Q4**, performance is modest but consistently above chance for the stronger non-linear models: XGBoost achieves AUC 0.578 and balanced accuracy 0.557, while gradient boosting and the LSTM cluster around AUC $\approx$ 0.56.

Second, in **2025Q1**, tree-based models again outperform the linear baseline: XGBoost attains AUC 0.626 and balanced accuracy 0.590, with gradient boosting close behind. The LSTM performs comparably to gradient boosting in this window (AUC 0.603, balanced accuracy 0.573).

Third, in **2025Q2**, tabular models deteriorate toward chance-level ranking (AUC 0.464–0.525), while the LSTM remains comparatively stable (AUC 0.588, F1 0.611). This contrast is consistent with the substantial quarter-to-quarter shift in label base rates documented in Chapter 3, which changes both the difficulty of separating winners/losers and the behavior of fixed thresholds.

Figure 6.1 visualizes balanced accuracy across model families in the three validation windows (2024Q4, 2025Q1, 2025Q2) and in the out-of-sample 2025Q3 test set (no tuning on Q3). The figure highlights that models with similar AUC can exhibit different threshold-dependent performance, and that the ensemble remains above chance on the held-out Q3 test.

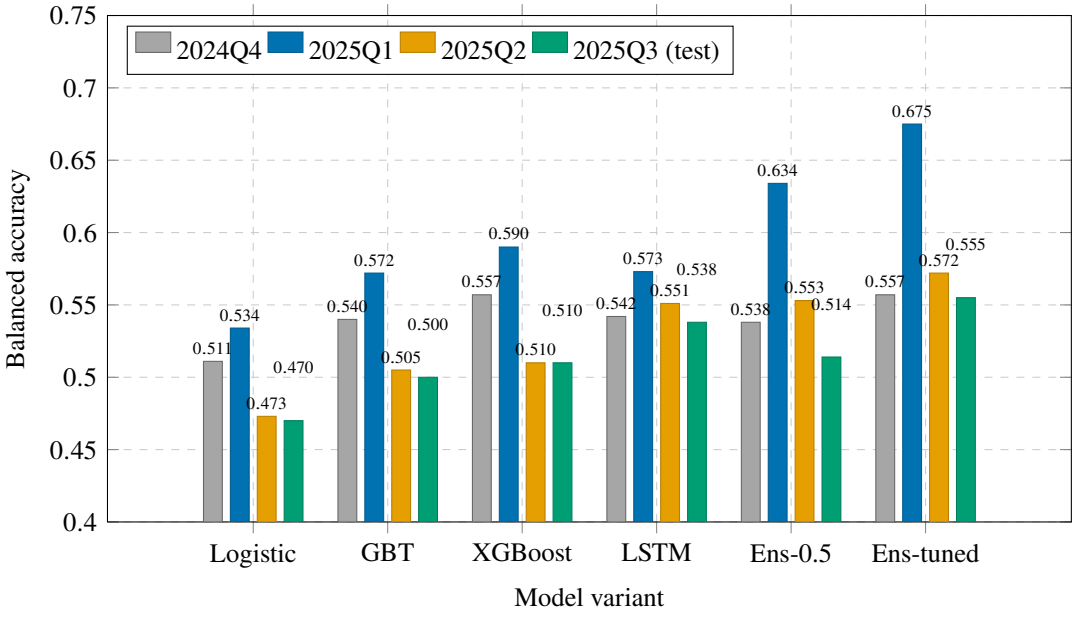

Figure 6.1: Balanced accuracy by model family across 2024Q4, 2025Q1, 2025Q2 validation windows and the out-of-sample 2025Q3 test set. The ensemble is evaluated at $\tau = 0.5$ and at a stage-specific tuned threshold (2024Q4–2025Q2 only; 2025Q3 uses no tuning).

6.2 Ensemble Performance and Threshold Tuning

The next subsection evaluates a stacking-style ensemble that aggregates predicted probabilities from the base models. Because balanced accuracy and F1 depend on the operating threshold, results are reported at (i) the default $\tau = 0.5$ and (ii) a *validation-tuned* threshold. In line with the error analysis in Chapter 7 and the appendix settings, the tuned threshold is chosen to **maximize balanced accuracy** within each quarter; it should be interpreted as a stage-specific operating point rather than as a universal choice.

Table 6.2: Ensemble performance on 2024Q4, 2025Q1, and 2025Q2 validation sets. Tuned thresholds maximize balanced accuracy within each stage.

Stage	Threshold	AUC	Bal. acc.	Acc.	F1 (pos)
2024Q4	0.5 (default)	0.577	0.538	0.575	0.376
	0.38 (tuned for Bal. acc.)	0.577	**0.557**	0.554	**0.521**
2025Q1	0.5 (default)	**0.715**	0.634	**0.715**	0.477
	0.40 (tuned for Bal. acc.)	**0.715**	**0.675**	**0.679**	**0.560**
2025Q2	0.5 (default)	0.599	0.553	0.658	**0.769**
	0.60 (tuned for Bal. acc.)	0.599	**0.572**	0.517	0.589

In **2024Q4**, the ensemble achieves AUC 0.577, indicating modest ranking skill. Threshold tuning is important: at the default $\tau = 0.5$, the classifier is conservative on the positive class (F1 0.376), while lowering the threshold to $\tau = 0.38$ improves balanced accuracy to 0.557 and raises F1 to 0.521.

In **2025Q1**, the ensemble achieves AUC 0.715, exceeding all single base models and indicating strong ranking quality in this regime. However, at the default threshold $\tau = 0.5$, the classifier is relatively conservative in predicting the positive class. As shown in Table 3.3, the positive-label rate in 2025Q1 is low, so a fixed threshold of 0.5 leads to many negative predictions. This inflates overall accuracy while depressing both balanced accuracy and the positive-class F1 score. Lowering the threshold to $\tau = 0.40$ moves the operating point to better reflect the class balance in this quarter: balanced accuracy increases to 0.675 and F1 rises to 0.560, at the cost of a modest reduction in raw accuracy.

In **2025Q2**, the ensemble AUC drops to 0.599, indicating weaker ranking quality overall. Threshold choice again plays a critical role. Because the positive base rate in 2025Q2 is very high (Table 3.3), the default threshold $\tau = 0.5$ yields a high F1 score but only moderate balanced accuracy. Increasing the threshold to $\tau = 0.60$ partially corrects this imbalance: balanced accuracy improves to 0.572, while accuracy and F1 decline as the classifier becomes less biased toward the majority positive class.

Tuned thresholds are chosen on the same validation set used for reporting, so threshold-dependent metrics (balanced accuracy, F1) may be mildly optimistic; AUC is unaffected.

Overall, the ensemble provides modest gains in 2024Q4, strong ranking performance in 2025Q1, and threshold-sensitive but weaker performance in 2025Q2. On the **out-of-sample 2025Q3 test set** (no threshold or model selection on Q3), the ensemble attains AUC ≈ 0.57 and balanced accuracy ≈ 0.55, consistent with modest but above-chance performance (Figures 6.1 and 8.1). Interpreting these results through the lens of quarter-specific label base rates clarifies why high accuracy or F1 can coexist with weaker balanced accuracy, and underscores the importance of regime-aware threshold selection when evaluating Fed-event-driven cross-sectional prediction models.

6.3 Regime Dependence and Robustness

Across all models, performance is strongly regime-dependent across the three validation windows (2024Q4–2025Q2) and on the held-out 2025Q3 test set. In **2024Q4**, results are weak but stable: the ensemble and stronger base models remain above chance (ensemble AUC 0.58, balanced accuracy 0.56), suggesting that the signal is present but less pronounced in this earlier regime. In **2025Q1**, tree-based models and the ensemble extract the strongest predictive structure, with AUC values reaching 0.71 for the ensemble and balanced accuracies well above 0.50. In **2025Q2**, tabular models deteriorate toward chance-level ranking, and only the LSTM and the ensemble retain AUC near 0.59. This systematic weakening aligns with the unusually high positive-label rate in 2025Q2 (Chapter 3), which reduces cross-sectional dispersion and makes peer-relative winners harder to separate using pre-t information. On the **out-of-sample 2025Q3 test set** (no tuning on Q3), the ensemble and base models show performance similar to or slightly below 2025Q2 (Figure 8.1), consistent with continued regime dependence rather than stable generalization.

Strategies based on these signals should be stress-tested across regimes and treat threshold selection as regime-specific. The next chapter examines where the ensemble succeeds or fails by sector, cap bucket, and time.

Chapter 7

ERROR ANALYSIS AND INTERPRETATION

This section analyzes misclassification patterns by sector, cap bucket, and time for 2024Q4, 2025Q1, and 2025Q2 at the stage-specific thresholds ($\tau = 0.38, 0.40, 0.60$) that maximize balanced accuracy.

7.1 Sector-Level Patterns

Table 7.1 reports misclassification rates by sector. Across the three regimes, errors are unevenly distributed; even in the more favorable 2025Q1 window, sector differences are pronounced. Information technology and energy exhibit relatively low error rates (around 25–30%), while financials are substantially harder to predict, with an error rate near 44%.

Table 7.1: Ensemble misclassification rates by sector in 2024Q4, 2025Q1, and 2025Q2 (evaluated at the balanced-accuracy optimal thresholds $\tau = 0.38$, 0.40, and 0.60).

Sector	Error rate 2024Q4	Error rate 2025Q1	Error rate 2025Q2
Consumer discretionary (cd)	0.358	0.311	0.550
Energy	0.342	0.295	0.659
Financials	0.401	0.439	**0.268**
Industrials	0.335	0.308	0.466
Information technology (it)	**0.329**	**0.254**	0.471

In 2024Q4 and 2025Q1, errors in financials are dominated by false positives, indicating that the model frequently predicts outperformance that does not materialize. Technology and energy are comparatively easier in those regimes, suggesting that the ensemble captures more stable relationships for these sectors when cross-sectional dispersion is moderate.

Sector rankings shift markedly in 2025Q2. Error rates rise sharply in cyclical sectors: energy and consumer discretionary exceed 55%, and industrials and information technology approach 50%. Financials, by contrast, become relatively easier, with an error rate around 27%. In this quarter, false negatives dominate across most sectors, consistent with the high positive-label rate documented in Table 3.3. Under such conditions, a fixed threshold struggles to balance false positives and false negatives.

Figure 7.1 visualizes these sectoral and size-bucket patterns across the three validation windows (2024Q4, 2025Q1, and 2025Q2): the top row shows sector error rates and the bottom row shows cap-bucket error rates for each quarter. The figure highlights the strong regime dependence of the error structure.

7.2 Size-Bucket Patterns

Table 7.2 reports misclassification rates by capitalization bucket (bottom row of Figure 7.1).

Error rates are similar across size groups in 2024Q4 and 2025Q1. In 2025Q2 a clear gradient emerges

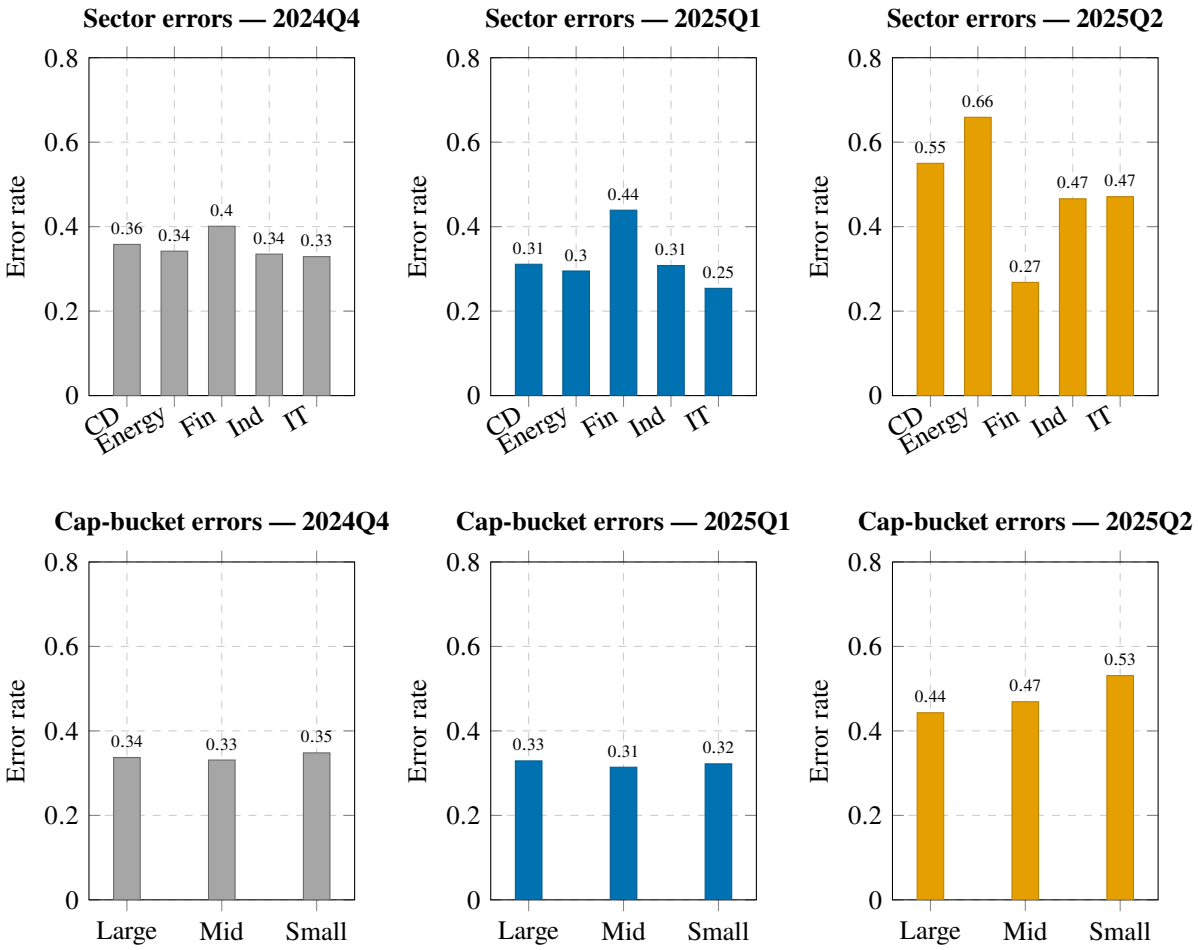

Figure 7.1: Ensemble misclassification rates by sector (top row) and capitalization bucket (bottom row) in the 2024Q4, 2025Q1, and 2025Q2 validation windows. Error rates are computed at the quarter-specific thresholds that maximize balanced accuracy.

Table 7.2: Ensemble misclassification rates by capitalization bucket in 2024Q4, 2025Q1, and 2025Q2.

Cap bucket	Error rate 2024Q4	Error rate 2025Q1	Error rate 2025Q2
Large cap	0.337	**0.329**	**0.443**
Mid cap	**0.331**	**0.314**	0.469
Small cap	0.348	0.322	0.531

(large < mid < small cap), suggesting the ensemble captures macro/sector structure better for large firms and struggles with idiosyncratic and liquidity effects in small caps.

7.3 Temporal Error Dynamics

Five-day rolling error rates stay below 25% in 2024Q4 and 2025Q1, and rise persistently in 2025Q2 (often near or above 30%), with spikes when many stocks outperform simultaneously. In those episodes, labels become harder to predict.

7.4 Interpretation

The error patterns clarify the results in Chapter 6. First, error rates below 50% for many sectors and size buckets confirm that the ensemble extracts non-trivial predictive structure from Fed-related variables, macro conditions, and recent price behavior. Second, the sharp contrast across 2024Q4, 2025Q1, and 2025Q2 demonstrates strong regime dependence: relationships learned in one quarter do not fully generalize when label balance and cross-sectional dispersion shift. Third, the concentration of errors in cyclical sectors and small caps suggests that the current feature set emphasizes broad macro and sector signals while underrepresenting finer firm-level heterogeneity.

The error analysis supports the conclusion that Fed-event-driven predictability exists but is limited and regime-dependent, motivating the extensions in Chapter 8.

Chapter 8

DISCUSSION AND CONCLUSION

8.1 Discussion and Implications

This project examined whether Fed-related and macro-financial features can predict *cross-sectional* equity performance over a medium horizon, using a stock–date panel and walk-forward evaluation.

Figure 8.1 summarizes AUC (left panel) and balanced accuracy (right panel) for the main model families across the three validation windows (2024Q4, 2025Q1, 2025Q2) and the out-of-sample 2025Q3 test set (no tuning on Q3), highlighting both the presence of predictive signal and its sensitivity to regime changes.

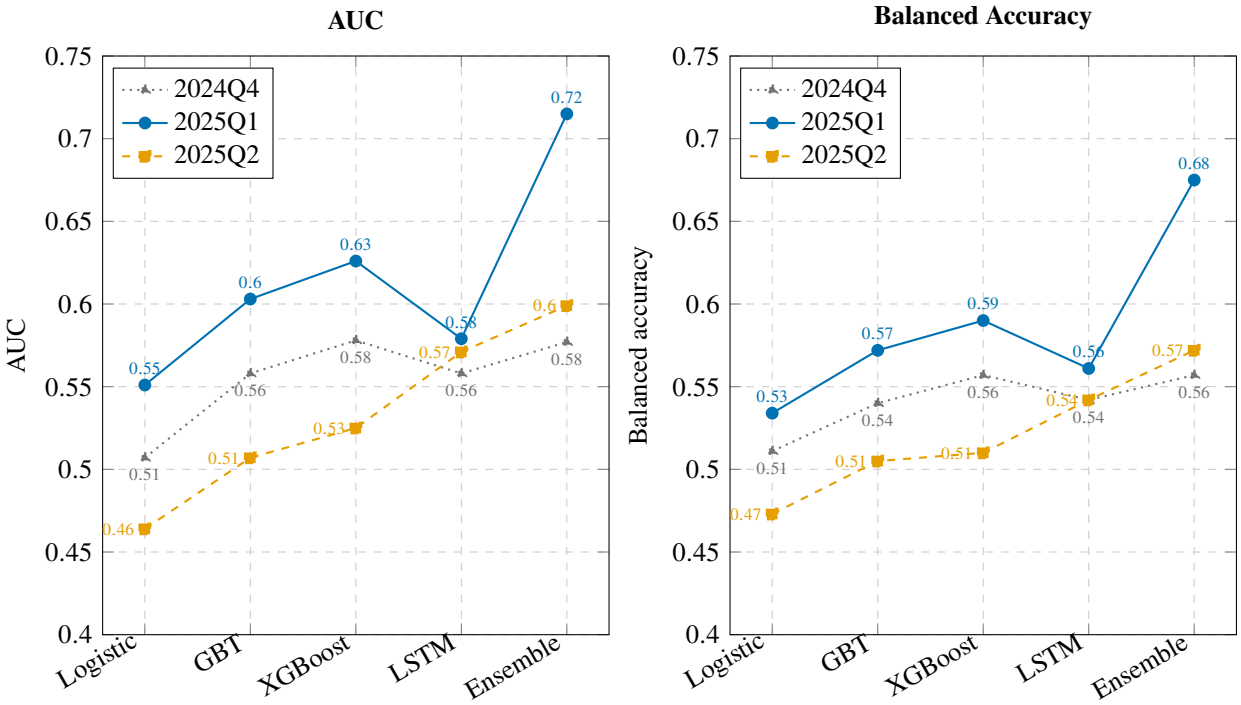

Figure 8.1: Summary of AUC and balanced accuracy by model family in the 2024Q4, 2025Q1, and 2025Q2 validation windows.

Predictive content of Fed-related and macro features. In the 2025Q1 validation window, the ensemble model achieves an AUC of approximately 0.71 and balanced accuracy well above 0.50, with misclassification rates near 30% across many sectors and size buckets. These results indicate that macro variables (gold, oil, and the U.S. dollar index), sector and size indicators, and recent price and volatility measures encode non-trivial cross-sectional structure around Federal Reserve events.

A simple long-only backtest (Chapter 5) holds the top 20% by ensemble score and computes 60-day forward excess returns. Table 8.1 summarizes results by stage.

Table 8.1: Long-only backtest (top 20% by ensemble score): average 60-day excess return (long portfolio minus rest of universe) and hit rate (% of validation dates with positive excess return) by stage. **2025Q3 is an out-of-sample test only**—no tuning or model selection was performed on Q3; only 5 validation days have 60-day forward returns in the panel.

Stage	Excess return (60d)	Hit rate
2024Q4	+2.23%	82.8%
2025Q1	+4.04%	71.7%
2025Q2	−2.87%	24.2%
2025Q3 (held-out)	+6.87%	n.a.[1]

In 2024Q4 and 2025Q1, the strategy delivers positive average excess returns and hit rates above 70%, suggesting that probability-based rankings can translate into economically meaningful differences in realized cross-sectional performance in those regimes. In 2025Q2, the same strategy yields a *negative* average excess return (−2.87%) and a hit rate of only 24.2%, illustrating economic failure under a regime shift: relationships that worked in earlier quarters do not hold when label balance and cross-sectional structure change sharply (Chapters 6 and 7). On the **out-of-sample** 2025Q3 test set (no threshold tuning, model selection, or any use of Q3 data before this evaluation), the strategy attains an average excess return of +6.87%; interpretation is limited because only five validation days in the panel have full 60-day forward returns.

The magnitude of excess returns remains modest and the strategy abstracts from transaction costs, turnover, and risk constraints; results should be interpreted as evidence of conditional signal presence rather than as a deployable trading strategy.

Regime dependence and inverted-U pattern. Predictive performance follows an **inverted-U pattern** across the validation regimes: weak but stable in 2024Q4, strongest in 2025Q1, and systematically weaker in 2025Q2. As documented in Chapters 6 and 7, the positive-label rate and cross-sectional co-movement rise sharply in 2025Q2, particularly among cyclical sectors. Under these conditions, the ensemble produces many false negatives and fails to generalize relationships learned in earlier quarters. The 2025Q3 backtest (purely out-of-sample: no parameters or thresholds were chosen using Q3) shows positive excess return (+6.87%) over a very short window (5 days), consistent with either partial recovery of signal or sampling noise; it does not alter the conclusion that Fed-related and macro signals are strongly regime-dependent and do not provide stable predictive power across all market conditions.

Sources of success and failure. Error patterns across sectors and size buckets point to a consistent interpretation. Tree-based models rely heavily on macro variables, sector and size indicators, and past volatility and momentum, which are effective when macro exposures and sector structure dominate cross-sectional variation. However, the models struggle with finer firm-level heterogeneity, especially among small-cap stocks.

The elevated error rates for small caps in 2025Q2 suggest that the current feature set does not fully capture liquidity, balance-sheet strength, or idiosyncratic risk. Similarly, the prevalence of false positives for financials in 2025Q1 indicates that coarse macro or sentiment signals may be overinterpreted at the individual stock level in certain sectors.

Overall, the results suggest that the feature set effectively encodes *macro and sector structure*, but remains insufficient for robust within-sector and within-size discrimination.

8.2 Limitations and Future Work

The findings highlight several limitations. The data span only 315 trading days and cover a limited number of macro regimes, making performance highly sensitive to short-run shifts in cross-sectional structure. The stock universe is restricted to 498 tickers in five sectors and three capitalization buckets, and macro and sentiment features are intentionally coarse.

On the modeling side, the analysis relies on a small set of baseline models, limited hyperparameter tuning, and three walk-forward validation windows. Backtesting is simplified and does not incorporate transaction costs, turnover constraints, or risk management considerations.

Natural extensions include longer panels and more macro regimes; richer sentiment (earnings calls, analyst reports) and firm-level features for small caps; regime-aware or sector-specific models; and portfolio-level evaluation (Sharpe ratios, turnover, factor-adjusted returns).

8.3 Conclusion

This Signature Work develops a leakage-free empirical pipeline that integrates Federal Reserve events, macro-financial variables, FinBERT-based sentiment, and stock-level features for cross-sectional equity prediction. The analysis shows that such information contains non-trivial but fragile predictive power over 60-day horizons.

Performance is strongly regime-dependent: models perform well in certain environments but deteriorate sharply when cross-sectional structure changes. Errors concentrate in specific sectors and among small-cap stocks, underscoring the limits of static models and coarse feature sets.

The results provide a structured case study for incorporating monetary policy and macro information into cross-sectional prediction; longer histories and regime-aware modeling could extend the framework.

Chapter 9

SIGNATURE WORK NARRATIVE

This Signature Work represents a sustained, multi-year trajectory across my undergraduate study rather than a single-semester effort. At Duke Kunshan University, academic interests gradually converged on a central theme: applying quantitative and computational methods to complex, data-rich systems—particularly financial markets—with careful attention to uncertainty, robustness, and responsible interpretation. This narrative explains how the three thematic courses and capstone experience shaped the Fed-event-driven cross-sectional study presented in this paper, and how the project prepared me for future graduate-level work in statistics, data science, and quantitative research.

9.1 Thematic Courses and Their Contributions

My three thematic courses contributed complementary perspectives on modeling, time dependence, and economic interpretation. Together, they directly informed how the research question was framed, how the empirical pipeline was built, and how model performance was evaluated in this project.

Machine Learning

In *COMPSCI 371: Machine Learning*, a formal understanding of supervised learning, loss functions, regularization, and model evaluation was developed. Implementing models such as logistic regression and tree-based methods trained me to view prediction as the estimation of conditional structure under finite samples, rather than the pursuit of a single deterministic outcome.

These ideas directly shaped the model lineup in Chapter 4. More importantly, the course emphasized that evaluation choices are part of the scientific claim. Because the goal of this project is forward-looking prediction, both metrics and data splits must respect the forecasting setting. This principle motivated the emphasis on AUC and balanced accuracy under the walk-forward evaluation framework described in Chapter 5, where random or cross-sectional splits would introduce severe information leakage in financial panels.

Stochastic Processes for Finance

MATH 411: Stochastic Processes for Finance deepened the understanding of dependence, non-stationarity, and regime shifts in financial time series, as well as the practical risks of information leakage. This perspective guided the construction of the stock–date panel in Chapter 3: all rolling features are computed strictly from information available at time t, and model evaluation uses only genuinely future periods.

As a result, large quarter-to-quarter shifts in label balance (Table 3.3) are interpreted as evidence of changing market regimes rather than as technical artifacts. This course fundamentally shaped how empirical results are interpreted in this project, reinforcing that instability across time is a property of the data generating process rather than a modeling failure alone.

Financial Economics

In *ECON 101: Principles of Economics*, core asset-pricing concepts such as discount rates, risk premia, and factor structure, together with foundational tools for empirical analysis, were introduced. This course motivated several central design choices in the project.

First, it informed the definition of the prediction target as a *relative* 60-day performance measure versus size peers (Section 3.3), rather than an absolute return. Second, it motivated the inclusion of macro-financial proxies such as gold, oil, and the U.S. dollar index, which reflect discount-rate and risk-sentiment channels through which monetary policy may affect asset prices. Finally, the course shaped how model behavior is interpreted—not only whether prediction works, but why performance may differ across sectors or market regimes, as discussed in Chapter 7.

9.2 From Earlier Projects to the Fed-Event Study

Before this capstone, prior work with Prof. David Ye on FinBERT-based sentiment analysis focused on aggregating sentiment from financial text and relating it to index-level market behavior. That experience established an end-to-end pipeline mindset, from text preprocessing and model inference to transforming outputs into usable sentiment signals.

In this Signature Work, those skills are extended to a more demanding cross-sectional setting. Rather than treating the market as a single time series, a stock–date panel was constructed linking FinBERT-based sentiment and macro variables to nearly 500 individual equities, combined with sector and size indicators and rolling return and volatility features. This shift required greater attention to data alignment, leakage control, and heterogeneity across stocks, sectors, and market regimes.

More broadly, quantitative training reinforced habits that transferred directly to empirical finance: traceability of data processing, careful handling of noise, and recognition that seemingly minor pipeline decisions can materially affect empirical conclusions.

9.3 Process, Challenges, and Learning During the Capstone

The capstone emphasized end-to-end rigor more than any single modeling choice.

First, the capstone reinforced that clean panel construction is itself a central research task. Merging stock prices, macro series, Fed calendars, and sentiment outputs required consistent date alignment, principled handling of missing data, stable sector and size definitions, and systematic exclusion of observations with incomplete 60-day horizons. This process highlighted that reproducibility and leakage prevention depend on disciplined data engineering rather than on modeling complexity alone.

Second, the capstone reinforced how to interpret "moderate" predictive metrics in context. In financial prediction, AUC values in the range of 0.62–0.72 can be meaningful yet still correspond to substantial error rates. The regime contrast and error analysis in Chapter 7 reinforced that predictability is fragile and that careful diagnostics are necessary to understand where and when models fail.

Third, page and time constraints forced prioritization. Rather than maximizing methodological breadth, the project focused on a transparent pipeline, a limited set of standard models, leakage-free evaluation, and structured error analysis. This experience strengthened the ability to make and clearly justify research trade-offs.

9.4 Connections to Broader Undergraduate Experience

Broader undergraduate experience also influenced how this work was approached. As a Resident Assistant, planning events and reflecting on outcomes strengthened the ability to communicate clearly, iterate under

constraints, and anticipate stakeholder concerns. These skills carried over into how the motivation, limitations, and implications of this project were framed in Chapters 1 and 8.

More broadly, Duke Kunshan University's liberal-arts environment reinforced the responsibility that accompanies quantitative tools. Because predictive models can influence risk-taking and resource allocation, explicitly documenting limitations, regime dependence, and error concentration is an essential part of responsible modeling. This principle guided how weaknesses and extensions were presented throughout the paper.

9.5 Preparation for Future Goals

Looking ahead, the author plans to pursue graduate study in statistics, data science, or a closely related quantitative field, and to continue working at the intersection of machine learning and applied domains such as finance. This Signature Work strengthened several foundations: designing end-to-end pipelines, enforcing realistic evaluation, and connecting statistical performance to domain interpretation.

Most importantly, the project provides a concrete synthesis of the thematic training. Concepts from *COMPSCI 371: Machine Learning* informed model selection and evaluation; ideas from *MATH 411: Stochastic Processes for Finance* informed leakage control and regime-aware thinking; and intuition from *ECON 101: Principles of Economics* informed the relative-performance target and macro feature design. Integrating these components into an independent study has been a defining part of the undergraduate experience and provides a strong foundation for deeper research in the future.

Code and Data Availability

Code uses Python 3.12 with `scikit-learn`, `xgboost`, and `PyTorch`.

Daily equity close prices were downloaded from Bloomberg (proprietary) and therefore cannot be redistributed. All other inputs (macro series and auxiliary market data) were collected programmatically via public APIs and web scraping (including `yfinance` and FRED) and can be regenerated from source.

Code for data collection (including Fed-related scraping), panel construction, model training, error analysis, and backtesting is available at:

https://github.com/qqgjyx/hanhan_SignatureWork.git

REFERENCES

[1] Doruk Araci. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:1908.10063*, 2019.

[2] Scott R. Baker, Nicholas Bloom, and Steven J. Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

[3] Ben S. Bernanke and Kenneth N. Kuttner. What explains the stock market's reaction to federal reserve policy? *The Journal of Finance*, 60(3):1221–1257, 2005.

[4] John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[6] John H. Cochrane. *Asset Pricing*. Princeton University Press, revised edition, 2005.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[8] Tom Engsted and Thomas Q. Pedersen. Event studies in economics and finance. *The New Palgrave Dictionary of Economics*, 2018.

[9] Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[10] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

[11] Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

[12] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29 (5):1189–1232, 2001.

[13] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[16] Bryan Kelly, Seth Pruitt, and Yinan Su. Textual analysis and machine learning in asset pricing. *Annual Review of Financial Economics*, 11:1–20, 2019.

[17] Bryan Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Finance*, 77(6):3261–3303, 2022.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.

[19] Kenneth N. Kuttner. Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics*, 47(3):523–544, 2001.

[20] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[21] Lubos Pástor and Pietro Veronesi. Uncertainty about government policy and stock prices. *Journal of Finance*, 67(4):1219–1264, 2012.

[22] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[23] Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

Appendix A

ADDITIONAL TABLES

This appendix collects supplementary tables that support, but are not essential to, the main text. They are included here for completeness and to document implementation details that may be useful for replication.

A.1 Hyper-parameter Settings

Table A.1 summarizes the main hyper-parameter choices for the baseline models introduced in Chapter 4. These settings correspond to the configurations used to produce the validation results in Chapters 6 and 7.

Table A.1: Hyper-parameter settings for the main model families.

Model family	Hyper-parameter	Value / description
Logistic regression	Penalty	ℓ_2
	Regularization strength	$C = 1.0$
	Class weights	`class_weight=balanced`
	Solver / max iter	`lbfgs`, `max_iter`=1000
Gradient boosting	Number of trees	`n_estimators`=300
	Max depth	`max_depth`=3
	Learning rate	`learning_rate`=0.05
	Subsample	`subsample`=0.8
	Min leaf size	`min_samples_leaf`=100
	Random seed	`random_state`=42
XGBoost	Number of trees	`n_estimators`=300
	Max depth	`max_depth`=3
	Learning rate	`learning_rate`=0.05
	Subsample / `colsample_bytree`	0.8 / 0.8
	Child / split regularization	`min_child_weight`=10, `gamma`=0
	Objective / metric	`binary:logistic`, `eval_metric=auc`
	Imbalance weight	`scale_pos_weight` set per stage from label imbalance
	Parallelism / seed	`n_jobs`=4, `random_state`=42
LSTM sequence model	Sequence length	15 trading days
	Hidden size	64
	Number of layers	1 LSTM layer
	Dropout	0.1
	Batch size / epochs	256 / 5
	Optimizer	Adam, learning rate 10^{-3}, weight decay 10^{-5}
	Loss	`BCEWithLogitsLoss` with `pos_weight` (neg/pos, stage-specific)
Ensemble (soft voting)	Base models	Logistic regression, gradient boosting, XGBoost
	Combination rule	Mean of predicted probabilities
	Thresholds	$\tau = 0.38$ (2024Q4), 0.40 (2025Q1), 0.60 (2025Q2), tuned per stage