



“From Patterns to Profits: Machine Learning Models for dynamic Insider Trading Detection”

Nikola Radivojevic

Dissertation written under the supervision of Professor Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in Finance, at the Universidade Católica Portuguesa, January 1st 2026.

Abstract

This thesis investigates whether machine learning models can extract profitable trading signals from insider transactions disclosed in SEC Form 4 filings. Using 841,071 insider purchases and sales from 2010 to 2024 merged with CRSP market data, we train and compare three ensemble models: Random Forest, XGBoost, and CatBoost through time-series cross-validation with 34 folds.

Models achieve weighted precision of 78 to 79 percent, outperforming the 66 percent random baseline by 12 percentage points. However, accuracy (53 to 56 percent), recall (53 to 56 percent), and F1 (61 to 64 percent) fall below the baseline, indicating conservative classification behavior. This precision-recall trade-off suggests models prioritize prediction accuracy over coverage, minimizing false positives at the cost of missed opportunities. Whether this conservative approach is acceptable depends on the application's cost structure. Feature importance analysis reveals that portfolio turnover, stock price at filing, and market capitalization are the most informative predictors.

Backtested trading strategies show that long-short portfolios generate statistically significant risk-adjusted returns, with CatBoost achieving annualized alphas of 14.99 to 16.09 percent across factor models and a Sharpe ratio of 1.69. Long-only strategies substantially outperform market benchmarks, while short-only strategies show negative overall returns but perform well during market downturns.

However, high portfolio turnover (1,257 to 1,472 percent annually) generates substantial transaction costs. Realistic execution costs eliminate the observed alpha, reducing economic viability particularly for retail investors. The findings confirm that insider trading signals contain exploitable information but emphasize the critical importance of execution quality and cost management for practical implementation.

Title: From Patterns to Profits: Machine Learning Models for Dynamic Insider Trading Signal Detection

Author: Nikola Radivojevic

Keywords: Insider Trading, Machine Learning, Random Forest, Portfolio Management, Predictive Analytics, Market Efficiency, Ensemble Methods, Feature Engineering

Resumo

Esta tese investiga se modelos de aprendizagem automática conseguem extrair sinais lucrativos de transações internas divulgadas nos formulários SEC Form 4. Usando 841.071 compras e vendas internas de 2010 a 2024 combinadas com dados CRSP, treinamos e comparamos três modelos ensemble: Random Forest, XGBoost e CatBoost através de validação cruzada temporal com 34 folds.

Os modelos alcançam precisão ponderada de 78 a 79 por cento, superando o baseline aleatório de 66 por cento em 12 pontos percentuais. No entanto, acurácia (53 a 56 por cento), recall (53 a 56 por cento) e F1 (61 a 64 por cento) ficam abaixo do baseline, indicando comportamento conservador de classificação. Esse trade-off entre precisão e recall sugere que os modelos priorizam a precisão das previsões em detrimento da cobertura, minimizando falsos positivos ao custo de oportunidades perdidas. Se essa abordagem conservadora é aceitável depende da estrutura de custos da aplicação. A análise de importância revela que rotatividade de portfólio, preço da ação e capitalização de mercado são os preditores mais informativos.

Estratégias de negociação backtested mostram que carteiras long-short geram retornos ajustados ao risco estatisticamente significativos, com CatBoost alcançando alfas anualizados de 14,99 a 16,09 por cento e Sharpe de 1,69. Estratégias long-only superam substancialmente benchmarks de mercado, enquanto estratégias short-only mostram retornos negativos globais mas bom desempenho em quedas de mercado.

Contudo, alta rotatividade de portfólio (1.257 a 1.472 por cento anualmente) gera custos de transação substanciais. Custos de execução realistas eliminam o alfa observado, reduzindo a viabilidade econômica, particularmente para investidores de varejo. Os resultados confirmam que sinais de negociação interna contêm informação explorável, mas enfatizam a importância crítica da qualidade de execução e gestão de custos.

Título: De Padrões a Lucros: Modelos de Aprendizagem Automática para Detecção Dinâmica de Sinais de Negociação Interna

Autor: Nikola Radivojevic

Palavras-chave: Negociação Interna, Aprendizagem Automática, Random Forest, Gestão de Portfólio, Análise Preditiva, Eficiência de Mercado, Métodos Ensemble, Engenharia de Características

Contents

1	Introduction	1
2	Literature Review	3
2.1	Insider Trading and Market Efficiency	3
2.2	Machine Learning in Financial Prediction	4
2.3	Regulatory and Governance Context	4
2.4	Research Gap and Contribution	5
3	Data	6
3.1	Data Sources	6
3.2	Insider Classification and Transaction Types	7
3.3	Dataset Overview	8
4	Methodology	12
4.1	Problem Formulation	12
4.2	Cross-Validation Structure	12
4.3	Model Selection	13
4.4	Evaluation Metrics	15
4.5	Trading Strategy Construction	16
5	Results	18
5.1	Model Classification Performance	18
5.1.1	Predictive Accuracy Across Models	18
5.2	Trading Strategy Performance	20
5.2.1	Cumulative Returns Over Time	21
5.3	Factor Model Analysis	22
5.4	Portfolio Turnover and Transaction Costs	25
5.5	Feature Importance Analysis	28
6	Conclusion	30
6.1	Summary of Findings	30
6.2	Limitations	31
6.3	Implications for Practice	33
6.4	Final Remarks	34
A	Appendix	35

List of Figures

4.1	Time-Series Cross-Validation Fold Structure	13
5.1	Performance Heatmap Comparing Classification Metrics Across Models	19
5.2	All Models and Strategies Cumulative Returns	22
5.3	Transaction Cost Impact on Portfolio Performance	27

List of Tables

3.1	Dataset Summary Statistics	8
3.2	Insider Role and Sector Distribution	9
5.1	Multi-Model Strategy Performance Comparison	20
5.2	Multi-Model Risk-Adjusted Alpha Comparison	23
5.4	Portfolio Turnover by Market Capitalization Category	26
5.3	Cross-Model Feature Importance Consensus (SHAP-Based)	29

1. Introduction

One of the most challenging problems in financial markets is identifying when corporate insiders' trading behavior signals genuine information about future stock performance. While regulatory frameworks mandate the disclosure of insider transactions through SEC Form 4 filings, distinguishing between routine portfolio rebalancing and trades driven by material, non-public information remains a persistent puzzle for investors and researchers alike. The Securities and Exchange Commission requires insiders (defined as officers, directors, and beneficial owners holding more than 10% of a company's shares) to disclose their transactions within two business days U.S. Securities and Exchange Commission (2025), creating a publicly available record of trading activity that theoretically enables outside investors to follow insider moves. However, the volume and variety of these filings, combined with varying motivations behind insider trades, make it difficult to extract actionable investment signals without sophisticated analytical tools.

Decades of research have established that insiders earn abnormal returns Seyhun (1986); Lakonishok and Lee (2001), yet significant questions remain about whether these patterns can be systematically exploited using modern computational methods. Traditional approaches rely on linear models and aggregate measures, potentially missing complex patterns in transaction-level data Cohen et al. (2012). The gap between statistical prediction and economic profitability remains inadequately addressed, particularly when accounting for realistic implementation constraints such as transaction costs and portfolio turnover.

This thesis applies ensemble machine learning methods to predict stock returns following insider transactions, bridging the gap between prediction accuracy and trading strategy profitability. Understanding whether machine learning can extract profitable signals from publicly disclosed insider transactions has important implications for market efficiency theory Seyhun (1992), portfolio management practice, and regulatory oversight Jagolinzer (2009).

This study employs ensemble learning methods including Random Forest, XGBoost, and CatBoost to analyze over 1.5 million insider transactions from SEC Form 4 filings U.S. Securities and Exchange Commission (2025) combined with market data from CRSP covering 2010 to 2024. This period encompasses diverse market regimes, including the European debt crisis (2010–2012), the Russian financial crisis (2014–2016), and the COVID-19 pandemic (2020). The analysis constructs features capturing transaction characteristics, firm fundamentals, and market conditions, then classifies one-day forward returns into three categories (underperform, neutral, outperform). Predictions are translated into long-only, short-only, and long-short trading strategies evaluated through rigorous out-of-sample backtesting with time-series

cross-validation.

The central research question is: *Can machine learning models identify profitable insider trading signals and generate risk-adjusted returns exceeding market benchmarks?* Specifically, this study examines which models achieve the highest predictive accuracy, which features are most informative, and whether strategies deliver positive risk-adjusted returns after transaction costs. Beyond the primary research question, this thesis also investigates how trading strategies are distributed across market capitalization segments, whether insider purchases are more informative than sales, and which features prove consistently important across different model architectures: all questions with direct implications for practical portfolio construction.

The remainder of this thesis is structured as follows. Section 2 reviews the academic literature on insider trading profitability and the application of machine learning to financial prediction problems. Section 3 describes the dataset construction, feature engineering process, and cross-validation framework. Section 4 presents the machine learning methodology, including model selection, hyperparameter tuning, and evaluation metrics. Section 5 analyzes model performance, feature importance, and the risk-return profiles of model-driven trading strategies. Finally, Section 6 summarizes the key findings, discusses limitations, and suggests directions for future research.

2. Literature Review

This thesis examines the intersection of two research areas: the profitability and informativeness of insider trading, and the application of machine learning methods to financial prediction problems. This chapter reviews the relevant literature, identifies key methodological approaches, and positions the contribution of this thesis.

2.1 Insider Trading and Market Efficiency

Seyhun (1986) established that corporate insiders consistently earn positive abnormal returns, documenting predictive ability with 60,000 transactions from 1975 to 1981. Seyhun (1992) extended this to aggregate insider trading, showing that net insider purchases and sales predict up to 60% of variation in one-year-ahead market returns. Lakonishok and Lee (2001) confirmed that insider purchases predict future returns while sales show little predictive ability, reflecting different motivations: purchases signal undervaluation or positive information, whereas sales often reflect diversification or liquidity needs Jeng et al. (2003). According to Jeng et al. (2003), insider purchases earn approximately 6% annual abnormal returns.

Not all insider trades are equally informative. Cohen et al. (2012) distinguish opportunistic from routine trades, showing opportunistic insider purchases generate value-weighted abnormal returns of 82 basis points per month, far exceeding abnormal returns from routine trades which are essentially zero. Informativeness varies by insider role: Wang et al. (2011) find CFOs earn 5% higher returns than CEOs, while Ravina and Sapienza (2006) show independent directors earn substantial abnormal returns. Insiders trade strategically around earnings announcements Huddart et al. (2006). The strategic timing of insider trades suggests sophisticated exploitation of information asymmetries, with insiders concentrating purchases before positive news and sales before negative disclosures.

Insider trading profitability varies with firm characteristics. Frankel and Li (2004) finds higher profitability in firms with unclear information environments and weak external monitoring. Firms with fewer analysts, less frequent disclosure, and higher information asymmetry provide insiders greater opportunities to profit from private information. David Veenman (2012) show insider purchases help investors interpret earnings signals, while Mark E. Vargus et al. (2001) link insider selling to lower earnings quality. The magnitude of insider profits reflects both the quality of private information and the difficulty outside investors face in replicating insider knowledge.

2.2 Machine Learning in Financial Prediction

The appeal in machine learning for financial prediction is that it is able to reveal complex, non-linear patterns without requiring researchers to specify the exact functional form in advance. Junming Gao (2024) show this advantage with XGBoost models that outperform traditional OLS regression in predicting A-share returns, creating 155% higher returns in long-short strategies. Similarly, Hum Nath Bhandari et al. (2022) show that LSTM networks make better predictions for the S&P 500 index, and Masuda (2024) show that hybrid CNN-LSTM architectures can be even better than either used alone.

Despite these developments, the use of machine learning in insider trading specifically remains relatively new. Amitabh Chakravorty and Nelly Elsayed (2025) report impressive 88% accuracy using SVM models on Tesla stock, but their one-stock approach leaves open questions regarding generalizability and whether such accuracy translates into profitable trading across a larger universe. More elaborate work by Ahmed Elnahas (2024) applies LASSO, Random Forest, and XGBoost with Bayesian hyperparameter tuning, boosting predictive R^2 by 200% to 500% over OLS while achieving accuracy better than 66%. Notably, their use of SHAP values to interpret model predictions reveals that transaction sizes, insider role, and firm characteristics are responsible for much of the predictive power. This is an important methodological advance which moves beyond treating these models as black boxes.

Yet a persistent challenge remains: strong predictive accuracy does not necessarily translate into profitable trading. Vitor Azevedo et al. (2024) address this accuracy-profitability gap directly, discovering that machine learning strategies can continue to be profitable even after accounting for transaction costs, although their analysis focuses on general market anomalies rather than insider signals. This distinction is important because, as Derek Snow (2020) makes clear, portfolio construction requires careful consideration of turnover and rebalancing frequency. For insider trading strategies in particular, shorter holding periods and frequent rebalancing can quickly erode returns through transaction costs.

2.3 Regulatory and Governance Context

The Sarbanes-Oxley Act (2002) shortened Form 4 filing deadlines from 10 days to 2 business days, improving disclosure timeliness Brochet (2009). SEC Rule 10b5-1 allows pre-planned trading programs, but Jagolinzer (2009) document insiders strategically time plan initiations and terminations to exploit information advantages.

Bettis et al. (2000) document widespread corporate trading restrictions, with Darren T. Roulstone (2003) finding firms compensate executives for reduced liquidity. Lee et al. (2014) show restrictions reduce but don't eliminate trade informativeness, while Skaife et al. (2013) find greater profitability in firms with weak internal controls. Machine learning enables sophisticated detection methods Prashant Priyadarshi and Prabhat Kumar (2024).

2.4 Research Gap and Contribution

Few studies attempt to systematically combine insider trading analysis and machine learning to construct and test trading strategies for a sample of firms and time periods. Insider trading research largely uses old econometric techniques such as event studies and panel regressions that, although they offer rigorous statistical inference, may not make maximum use of the richness of transaction-level data with its complex interactions between insider characteristics, transaction attributes, firm fundamentals, and market conditions. Machine learning finance studies focus predominantly on broad market prediction using price, volume, and macroeconomic variables with limited attention to insider-specific signals that may provide superior predictive content as a result of information asymmetries. Studies applying machine learning to insider trading emphasize classification accuracy measures over trading strategy profitability Amitabh Chakravorty and Nelly Elsayed (2025), leaving open the question of whether high accuracy translates into economically exploitable returns.

This thesis fills these gaps through three main contributions. First, it integrates ensemble methods (Random Forest, XGBoost, CatBoost) using insider trading data between 2010 and 2024, building models that can capture complex non-linear relationships and feature interactions that are not possible in linear models. Second, it creates realistic trading strategies evaluated using stringent out-of-sample backtesting with time-series cross-validation, bridging the difference between the accuracy of predictions and economic profitability by explicitly accounting for portfolio construction, turnover, and transaction costs. Third, it provides a feature importance analysis to determine which transaction characteristics, firm fundamentals, and market conditions prove to be most predictive with different model architectures, giving insights into the practical viability of machine learning-based insider trading strategies.

3. Data

This chapter describes the construction of the dataset used in this study, including data sources, preprocessing steps, and feature engineering. The dataset combines insider trading transactions from SEC Form 4 filings U.S. Securities and Exchange Commission (2025) with comprehensive market and firm-level data from CRSP and Compustat, enabling the systematic analysis of insider trading signals across a broad universe of U.S. equities.

3.1 Data Sources

The analysis integrates multiple data sources to construct a rich dataset for insider trading prediction. Each source provides distinct but complementary information about insider transactions, firm characteristics, and market conditions.

SEC Form 4 Filings. SEC Form 4 filings U.S. Securities and Exchange Commission (2025) provide detailed records of insider trading transactions. Corporate insiders must report trades to the SEC, with the filing deadline shortened from 10 days to 2 business days in 2002 under the Sarbanes-Oxley Act, significantly improving disclosure timeliness Brochet (2009).

Insider trading data from SEC-api EDGAR database (2010-2024) contains 4,906,473 transactions. Each filing includes transaction date, filing date, shares traded, price, and insider identity and role (C-level executives, directors, large shareholders). Transaction types include purchases (P), sales (S), options exercises (M), awards (A), and gifts (G).

CRSP Daily Stock Data. Daily stock data from WRDS CRSP Database (Dec 2009-Dec 2024) includes returns, prices, volumes, shares outstanding, and market capitalization. Returns are adjusted for splits and dividends Lakonishok and Lee (2001).

Value-Weighted Market Index. Market returns from CRSP value-weighted index serve as the benchmark for calculating excess returns, isolating firm-specific performance from broad market movements Lakonishok and Lee (2001).

Compustat Fundamentals. Firm fundamentals (book value, assets, earnings, sector) from Compustat enable features capturing size, profitability, leverage, and valuation. Data merged with transactions ensuring only available information is used, avoiding look-ahead bias.

3.2 Insider Classification and Transaction Types

Not all insider trades are informative. Some reflect genuine conviction while others stem from routine portfolio management, compensation, or liquidity needs Cohen et al. (2012). Understanding this diversity is critical for developing effective predictive models, as treating all insider transactions equally would dilute signal quality and reduce the economic value of trading strategies based on insider activity. The challenge for investors and researchers alike is to distinguish between trades motivated by private information about firm prospects and those driven by personal reasons unrelated to expected returns.

Cohen et al. (2012) distinguish *opportunistic* from *routine* trades, providing a framework for identifying informative transactions. Routine trades occur in the same calendar month for three consecutive years, suggesting predetermined, non-informative patterns that likely reflect systematic portfolio rebalancing, diversification requirements, or tax-loss harvesting strategies. These trades exhibit little to no abnormal returns in subsequent periods, indicating they contain minimal private information. Opportunistic trades deviate from this pattern and generate higher abnormal returns, with Cohen et al. (2012) documenting that opportunistic insider purchases earn approximately 82 basis points per month versus nearly zero for routine purchases. This classification improves predictive power Biggerstaff et al. (2020), as models that explicitly account for the routine versus opportunistic distinction achieve higher out-of-sample accuracy and more profitable trading strategies. This thesis implements the Cohen classification as a binary feature to distinguish routine from opportunistic trades, allowing the models to learn differential patterns between these transaction types and potentially enhance predictive accuracy by focusing signal extraction on the more informative opportunistic trades.

The analysis focuses primarily on *purchases* (P) and *sales* (S) as discretionary trades signaling beliefs about performance. Purchases represent particularly strong signals because insiders voluntarily commit personal capital to the firm, facing downside risk and opportunity costs from alternative investments. This aligns incentives with outside shareholders and suggests positive private information about future performance Lakonishok and Lee (2001). Sales are more ambiguous: while they may indicate negative information, insiders sell for many reasons unrelated to firm prospects, including liquidity needs, diversification, estate planning, or margin calls Jeng et al. (2003). Consequently, the empirical literature consistently finds that purchases generate larger and more reliable abnormal returns than sales, with Lakonishok and Lee (2001) documenting that insider purchases outperform by approximately 4.5% annually while sales show minimal predictive power. In contrast, options exercises (M), awards (A), and gifts (G) are driven by compensation structures or tax planning rather than information Carpenter and Remmers (2000). These transactions are typically predetermined by vesting schedules or compensation agreements and therefore lack informational content about future stock returns. By filtering the dataset to focus exclusively on discretionary purchases and sales, this study enhances the signal-to-noise ratio and improves the likelihood of detecting genuine information.

3.3 Dataset Overview

Table 3.1 summarizes the key characteristics of the final dataset after filtering.

Dataset Summary Statistics			
Sample	Period	Number of Transactions	Number of Firms
Full sample	2010–2024	4,906,473	8,737
Filtered sample	All years	1,311,535	-
Final sample	2010–2024	841,071	3,348

Table 3.1. This table presents the number of insider trading transactions and unique firms at three stages of data processing. The full sample includes all 4,906,473 SEC Form 4 transactions from 8,737 firms across all years. The filtered sample (1,311,535 transactions) applies transaction type and data quality filters: (1) retaining only purchases and sales (removing 3,418,328 transactions or 69.7%); (2) removing zero-value transactions, excessive filing lags, and invalid exchanges (removing 176,610 transactions). The final sample (841,071 transactions from 3,348 firms) additionally applies a market capitalization filter (minimum \$1.0M, removing 470,464 transactions or 9.1%) and restricts the period to 2010–2024. Total retention: 17.1%. Readers should notice the dramatic reduction from full to final sample (82.9% removed), primarily driven by the transaction type filter, which reflects the focus on discretionary trades (purchases/sales) that are most likely to contain information signals.

The raw SEC Form 4 dataset U.S. Securities and Exchange Commission (2025) contains 4,906,473 transactions covering the period from 2010 to 2024, with approximately 35% involving nano-cap firms (market capitalization below \$10M) that suffer from extreme illiquidity and unreliable pricing, making them unsuitable for realistic trading strategies Fidrmuc et al. (2005). After rigorous filtering, the final dataset contains 841,071 insider trading events, representing 17.1% of the initial data, spanning January 2010 to June 2024 with market capitalizations above \$1 million and balanced across different market environments. The transaction composition consists of 768,363 sales (91.4%) and 72,708 purchases (8.6%), consistent with prior findings that sales are more frequent but less informative than purchases Lakonishok and Lee (2001). The exchange distribution is nearly balanced with 50.4% traded on NASDAQ and 49.6% on NYSE. Transaction values exhibit substantial variation, with a median of \$193,458 and mean of \$3.0M, while stock prices range from \$0.30 to \$7,754 with a median of \$61.75.

The raw SEC Form 4 data U.S. Securities and Exchange Commission (2025) underwent rigorous filtering to ensure data quality and relevance for predictive modeling through a pre-processing pipeline consisting of multiple sequential filters, each designed to address specific data quality issues. The first filter retained only Purchases (P) and Sales (S) transactions, removing 3,418,328 transactions (69.7%), as these reflect genuine insider conviction about future stock performance (Cohen et al., 2012), while excluding other transaction types including Gifts (G), Awards (A), Options exercises (M), and other non-market transactions that are often driven by compensation arrangements or tax planning rather than information about future performance. The second filter removed zero-value transactions, eliminating 130,670 transactions

(2.7%) that likely represent data errors or non-market transfers, with M-code transactions (exercise/conversion) examined but none found in the filtered dataset. The third filter restricted transactions to the period from 2010 to 2024, removing 993 transactions (0.0%) to focus on the modern market regime, ensuring the analysis reflects current market dynamics and regulatory environments. The fourth filter excluded transactions with problematic filing timing, removing 44,947 transactions (0.9%), including filings more than 7 days after the transaction date, as the SEC requires 2-day filing under the Sarbanes-Oxley Act, ensuring that the dataset reflects realistic market conditions where investors can act on publicly available insider trading information in a timely manner. The fifth filter restricted the dataset to NYSE and NASDAQ exchanges, removing 26,237 transactions (0.5%) to focus on the most liquid and well-regulated U.S. equity markets. The sixth and final filter excluded trades involving firms with market capitalization below \$1 million, removing 444,227 transactions (9.1%), eliminating illiquid micro-cap stocks that often have unreliable pricing and are not representative of the broader market and may exhibit extreme volatility and wide bid-ask spreads that do not reflect genuine trading opportunities available to institutional or retail investors. In total, the filtering process removed 4,065,402 transactions (82.9%) from the initial dataset, retaining 841,071 transactions (17.1%) for analysis.

Table 3.2 presents the composition of the final dataset by insider role and industry sector.

Insider Role and Sector Distribution						
Panel A: Insider Role and Transaction Type						
Role	Purchases	% of Role	Sales	% of Role	Total	% of Sample
CEO	16,644	11.0%	134,561	89.0%	151,205	20.5%
Other Officers	8,974	2.8%	310,417	97.2%	319,391	43.4%
Director	28,210	14.7%	163,486	85.3%	191,696	26.0%
10% Owner	10,480	14.1%	64,000	85.9%	74,480	10.1%
Panel B: Top 5 Industry Sectors						
Sector	Transactions	% of Sample				
Technology	233,936	27.8%				
Healthcare	149,658	17.8%				
Consumer Cyclical	87,520	10.4%				
Financial Services	86,993	10.3%				
Industrials	84,503	10.0%				
Other (6 sectors)	198,461	23.6%				

Table 3.2. Panel A presents the distribution of transactions by insider role and transaction type for the 742,905 transactions with complete relationship information (88.3% of the final sample). The remaining 98,166 transactions (11.7%) lack detailed insider role classification in the SEC Form 4 filings. CEOs account for 20.4% of classified transactions but only 11.0% are purchases, indicating sales dominate executive trading (consistent with liquidity and diversification motives). Directors show higher purchase propensity at 14.7%, suggesting more informational content in director purchases. Other Officers (43.0% of sample) exhibit the lowest purchase rate (2.8%), likely reflecting routine stock option exercises and compensation-driven sales. Panel B shows sector concentration, with Technology (27.8%) and Healthcare (17.8%) dominating the sample, reflecting these sectors' higher insider trading activity and equity compensation prevalence. Combined, the top 5 sectors represent 76.3% of all transactions.

The breakdown reveals important patterns in insider trading behavior across different types of insiders and sectors. As shown in Table 3.2, the sector composition is dominated by Technology (27.8%), Healthcare (17.8%), Consumer Cyclical (10.4%), Financial Services (10.3%), and Industrials (10.0%). This distribution reflects the broader composition of U.S. equity markets, with technology and healthcare representing growth-oriented sectors that typically exhibit higher insider trading activity due to greater information asymmetries and volatility.

A critical aspect of the data preprocessing is determining the realistic entry point for trades based on insider filings, with the entry timing logic designed to reflect real-world trading constraints where positions are entered at that day's closing price if a Form 4 is filed before 9:30 AM ET (market open), or at the next trading day's closing price if filed at or after 9:30 AM ET, with entry dates falling on weekends moved to the following Monday to reflect the practical reality that traders would observe the filing, make a decision, and execute at the end of the trading day. For example, a transaction occurring on Monday, January 10 and filed on Wednesday, January 12 at 2:00 PM would result in position entry at Thursday, January 13's closing price (the next trading day). Analysis of the dataset reveals that 15% of filings occur pre-market (before 9:30 AM) and are entered at the same day's close, 85% occur after-hours (at or after 9:30 AM) and are entered at the next day's close, and 8% of entries are adjusted from weekends to the following Monday.

For each insider transaction, forward-looking stock returns at the target horizon were calculated using the CRSP ret variable, which is adjusted for both stock splits and dividends reinvested. This adjustment ensures that return calculations are as consistent and comparable as possible across firms and over time. The CRSP returns are computed from the closing price at time $t - 1$ to the closing price at time t . Additionally, to account for broader market movements, all returns are calculated in excess of the value-weighted market index, providing a measure of abnormal performance relative to the overall market. This excess return metric isolates the firm-specific component of performance from general market trends.

Feature engineering in this study combines transaction characteristics, firm fundamentals, market conditions, and historical trading patterns to capture the many dimensions of insider trading signals.

Transaction-level features include the value of the transaction (shares traded multiplied by transaction price), the relative size of the transaction (as a percentage of shares outstanding), and the type of transaction (purchase or sale). Insider-level features capture the role of the insider (CEO, CFO, director, or large shareholder) and whether the trade is classified as routine or opportunistic using the methodology of Cohen et al. (2012).

Firm-level features, sourced from CRSP and Compustat, include market capitalization, book-to-market ratio, price-to-earnings ratio, return on assets, and leverage. Historical price momentum is captured through lagged returns over multiple horizons (1-day, 5-day, 21-day, and 63-day returns), providing information about recent stock performance.

Interaction features are constructed to capture relationships between variables. For example,

the interaction between transaction size and recent price momentum allows the model to distinguish between large trades following strong momentum (which may signal continued strength) and large trades following poor performance (which may signal opportunistic buying). Similarly, the interaction between insider role and transaction type can reveal whether certain types of insiders (e.g., CEOs versus large shareholders) are more informative when buying versus selling.

An important component of the feature set is the routine-opportunistic classification framework. As discussed earlier, this binary feature flags trades that follow a predictable, routine pattern versus those that are opportunistic and potentially more informative. The construction of this feature involves analyzing each insider's historical trading activity and identifying patterns in the timing of their transactions. If an insider has traded in the same month for three consecutive years, subsequent trades in that month are flagged as routine.

The final dataset includes over 40 features across the following categories: Transaction Features, Price Features, Firm level Features, Volume Features, Fundamental Features, Temporal Features, Market Features.

4. Methodology

This chapter describes the machine learning methodology employed in this study, including the cross-validation framework, formulation of the prediction task, the selection and configuration of models, the training and evaluation procedures, and the construction of trading strategies based on model predictions.

4.1 Problem Formulation

The prediction problem is formulated as multiclass classification Amitabh Chakravorty and Nelly Elsayed (2025). Each observation describes an insider transaction, stock, and market environment. The target variable ranks transactions by one-day forward returns into deciles: $y = -1$ for bottom decile, $y = 0$ for middle eight deciles, $y = 1$ for top decile. This focuses on identifying extreme performers while treating most transactions as neutral, controlling for time-varying market conditions.

4.2 Cross-Validation Structure

This study employs expanding window cross-validation to preserve temporal order and avoid look-ahead bias. Each iteration has three segments: training (60 months minimum, expanding), validation (12 months fixed), and test (3 months). The window advances by 3 months per iteration, simulating quarterly reforecasting.

Figure 4.1 illustrates the expanding window cross-validation structure used throughout this study.

Formally, for split i :

$$\begin{aligned}\text{Train} &= \{x_0, x_1, \dots, x_{t_i}\} \\ \text{Validation} &= \{x_{t_i+1}, \dots, x_{t_i+12}\} \\ \text{Test} &= \{x_{t_i+13}, x_{t_i+14}, x_{t_i+15}\}\end{aligned}$$

where t_i is the last training observation for split i . This structure yields $N = 34$ rolling splits.

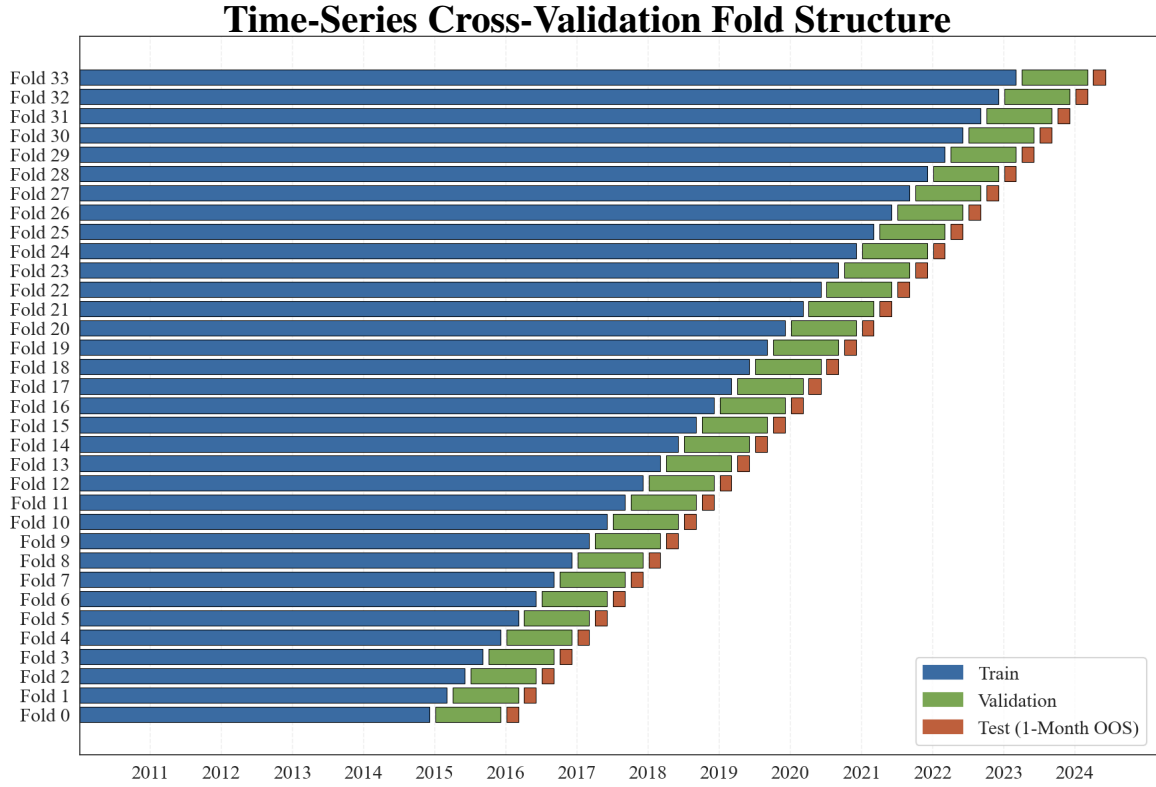


Fig. 4.1. Expanding window cross-validation structure with 34 time-series folds. Each fold has training period (60 months minimum, expanding), validation period (12 months), and test period (3 months). Window advances by 3 months per iteration, simulating quarterly model updates while preserving temporal order to prevent look-ahead bias.

4.3 Model Selection

The choice of machine learning models is guided by their demonstrated effectiveness in classification tasks involving structured, tabular data, as well as their ability to capture non-linear relationships and complex feature interactions. This study employs three primary models: Random Forest, XGBoost, and CatBoost. Each model brings distinct strengths to the prediction task, and their performance is compared to identify the most effective approach for generating profitable insider trading signals.

Decision Trees: The Building Block. All three models employ decision trees as building blocks. Decision trees recursively partition feature space based on splitting rules, automatically discovering non-linear relationships without requiring explicit feature engineering. Individual trees prone to overfitting are addressed through ensemble methods combining multiple trees trained on different data subsets.

Random Forest. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes predicted by individual trees. Each tree is trained on a bootstrap sample of the data, and at each split, a random subset of

features is considered, introducing diversity among the trees and reducing overfitting. The prediction for a given observation $\mathbf{x}$ is obtained by aggregating the predictions of T trees: $\hat{y}(\mathbf{x}) = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x})\}$, where $h_t(\mathbf{x})$ is the prediction of the t -th tree. Random Forest is chosen for its robustness to overfitting, its ability to handle high-dimensional feature spaces, and its effectiveness in capturing non-linear relationships, making it a natural baseline for this analysis.

XGBoost. XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that builds an ensemble of decision trees sequentially, with each tree correcting the errors of the previous ones. The model minimizes a regularized objective function that balances the loss function with a penalty term to control model complexity: $\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$, where $l(y_i, \hat{y}_i)$ is the loss function measuring the difference between the true label y_i and the predicted value $\hat{y}_i$, $\Omega(f_k)$ is a regularization term penalizing the complexity of the k -th tree, and K is the total number of trees. XGBoost is selected for its computational efficiency, its ability to handle missing values, and its effectiveness in capturing complex patterns through gradient-based optimization. It has been widely used in machine learning competitions and financial prediction tasks, delivering consistent performance.

CatBoost. CatBoost (Categorical Boosting) is a gradient boosting algorithm specifically designed to handle categorical features and to reduce overfitting through ordered boosting. Unlike traditional gradient boosting, which can suffer from target leakage when constructing trees, CatBoost computes prediction residuals in a way that avoids using the target value of the current observation, thereby reducing bias. The objective function for CatBoost is similar to that of XGBoost, but the algorithm employs ordered boosting to ensure that predictions for each observation are based only on prior observations in the training set: $\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \lambda \sum_{k=1}^K \Omega(f_k)$, where λ is the regularization parameter. CatBoost is included in the analysis for its ability to handle categorical variables natively (such as insider role, sector, and transaction type) without requiring extensive preprocessing. Its robust handling of overfitting and its strong performance on structured data make it a compelling choice for insider trading prediction.

Ensemble Model Concept. Ensemble learning combines the predictions of multiple models to improve overall performance and reduce the risk of overfitting to any single model's biases. Random Forest, XGBoost, and CatBoost are all ensemble methods that aggregate the predictions of multiple base learners (decision trees) to produce a final prediction. By training multiple models with different algorithmic structures and comparing their performance, this study aims to identify the most robust approach for insider trading prediction and to assess whether ensemble methods consistently outperform individual classifiers.

Hyperparameter Configuration and Training Procedure. Hyperparameters tuned using validation sets include tree count, maximum depth, learning rate, and regularization parameters. Training respects temporal structure to avoid look-ahead bias. Class imbalance is addressed through weighted loss functions. Overfitting is monitored via training-validation performance comparison and mitigated through regularization and early stopping for boosting models.

4.4 Evaluation Metrics

Model performance is evaluated using a comprehensive set of classification metrics that capture different aspects of predictive accuracy and reliability. These metrics are computed on the out-of-sample test set for each fold and then aggregated across folds to assess overall performance.

Accuracy. Accuracy measures proportion of correct predictions but can mislead with class imbalance. Precision measures correct positive predictions, important for trading to minimize false positives. Recall measures correctly identified actual positives, ensuring profitable opportunities aren't missed. F1 Score balances precision and recall. ROC-AUC summarizes discriminative ability across thresholds, robust to class imbalance.

Baseline Comparison. Model performance is benchmarked against a random baseline that mimics the structure of the prediction task. On each trading day, the random baseline assigns predictions uniformly at random, respecting the effective target distribution. Since the target variable is constructed by assigning the top decile to class $y = 1$, the bottom decile to class $y = -1$, and the middle eight deciles to class $y = 0$, the effective class distribution is approximately 0.1 / 0.8 / 0.1. The random baseline generates predictions following this distribution: with probability 0.1, it predicts $\hat{y} = 1$ (long), with probability 0.1, it predicts $\hat{y} = -1$ (short), and with probability 0.8, it predicts $\hat{y} = 0$ (neutral). This ensures that the random baseline constructs portfolios of similar size to the model-based strategies, enabling a fair comparison of portfolio performance.

The imbalanced class distribution has important implications for baseline metric calculations. For a random classifier predicting according to the true distribution $p = (0.1, 0.8, 0.1)$, accuracy equals $\sum p_i^2 = 0.1^2 + 0.8^2 + 0.1^2 = 0.66$ (66%). However, per-class precision and recall for class i equal p_i , yielding precision = recall = 0.1 for minority classes and 0.8 for the majority class. Macro-averaged metrics (simple average across classes) therefore equal $(0.1 + 0.8 + 0.1)/3 = 0.333$ (33.3%). This divergence arises because accuracy is instance-weighted (each sample counts equally, so the 80% of samples in class 0 dominate), while macro metrics are class-weighted (each class counts equally regardless of size) Sokolova and Lapalme (2009). Macro F1 is preferred for imbalanced problems as it prevents a model from achieving flattering accuracy by simply riding the majority class, making it the appropriate benchmark for evaluating genuine predictive skill.

The random baseline represents the expected performance of a strategy that selects stocks without any predictive ability, serving as a critical benchmark to assess whether the models generate statistically and economically significant outperformance. Additionally, portfolio returns are contextualized relative to the CRSP value-weighted market index. However, this market benchmark is informative but not fully comparable to the trading strategies, as it represents a passive long-term buy-and-hold investment across all stocks, whereas the insider trading strategies employ active day trading with one-day holding periods and selective stock picking. The random baseline remains the primary benchmark for evaluating model skill, as it controls for the specific structure of the prediction task and portfolio construction methodology.

Model Selection Criteria. The selection of a "best" model depends on the specific goals of the analysis. For a trading strategy, precision may be prioritized to minimize false positives (unprofitable long positions), while recall ensures that profitable opportunities are not missed. The F1 Score provides a balanced metric, and ROC-AUC offers a threshold-independent assessment of discriminative power. A model is considered to be overfitting if it achieves high training performance but significantly lower test performance, indicating poor generalization to unseen data. Regularization and early stopping are used to mitigate overfitting and to ensure that the models generalize well to new data.

4.5 Trading Strategy Construction

Based on model predictions, three portfolio strategies are constructed to evaluate the economic viability of the predicted signals: long-only, short-only, and long-short. Each strategy is applied uniformly across all models to enable direct comparison of their performance.

Long-Only Strategy. The long-only strategy takes long positions in stocks predicted to outperform (i.e., $\hat{y} = 1$). At each rebalancing date, the portfolio is equally weighted across all stocks with positive predictions. The return of the long-only portfolio at time t is given by

$$R_t^{\text{long}} = \frac{1}{N_t^{\text{long}}} \sum_{i \in \mathcal{S}_t^{\text{long}}} r_{i,t},$$

where N_t^{long} denotes the number of stocks in the long portfolio, $\mathcal{S}_t^{\text{long}}$ is the set of stocks predicted to outperform, and $r_{i,t}$ is the realized return of stock i over the holding period.

Short-Only Strategy. The short-only strategy takes short positions in stocks predicted to underperform (i.e., $\hat{y} = -1$). The return of the short-only portfolio is defined as

$$R_t^{\text{short}} = -\frac{1}{N_t^{\text{short}}} \sum_{i \in \mathcal{S}_t^{\text{short}}} r_{i,t},$$

where N_t^{short} is the number of stocks in the short portfolio and $\mathcal{S}_t^{\text{short}}$ denotes the set of stocks predicted to underperform. The negative sign reflects the payoff structure of short positions.

Long-Short Strategy. The long-short strategy combines the long and short portfolios by allocating 50% of capital to each leg. The resulting portfolio return is

$$R_t^{\text{LS}} = 0.5 \cdot R_t^{\text{long}} + 0.5 \cdot R_t^{\text{short}}.$$

This construction yields a dollar-neutral portfolio that reduces exposure to broad market movements and focuses performance on the model's ability to distinguish between cross-sectional winners and losers.

Entry and Exit Timing. As discussed in Chapter 3, the entry date is determined based on the filing time of the Form 4, with positions entered at the closing price on the same day for pre-market filings or at the next trading day's closing price for filings at or after market open. The holding period for all strategies is one trading day, with positions entered at the closing price on the entry date and exited at the closing price on the following trading day. This one-day holding period minimizes exposure to market risk and aligns with the prediction horizon of the models, which forecast one-day forward returns.

Rationale for Strategy Choice. The three strategies are chosen to provide a comprehensive evaluation of model performance under different market conditions. The long-only strategy is the most straightforward and aligns with traditional long-only equity investing. The short-only strategy tests the model's ability to identify underperforming stocks, which is often more challenging than identifying outperformers. The long-short strategy combines both dimensions and is market-neutral, making it attractive for hedge fund-style strategies that seek to generate alpha independent of market direction. All strategies are equally weighted, meaning that each selected stock receives the same portfolio weight regardless of its market capitalization or transaction value. This approach is chosen for simplicity and to avoid introducing additional complexity related to portfolio optimization. An alternative approach would be to use value-weighted portfolios, where stocks are weighted by market capitalization or by the value of the insider transaction. Future research could explore whether value-weighted strategies deliver superior risk-adjusted returns. By applying the same strategy construction methodology across all models, this study ensures that differences in portfolio performance can be attributed to the quality of the predictions rather than to differences in portfolio construction or risk management techniques.

5. Results

This chapter presents empirical findings in three stages: model classification performance, trading strategy outcomes, and hypothesis examination. Results are benchmarked against random baselines and prior research to establish statistical and economic significance Amitabh Chakravorty and Nelly Elsayed (2025).

5.1 Model Classification Performance

5.1.1 Predictive Accuracy Across Models

Figure 5.1 presents the classification performance metrics for all machine learning models evaluated in this study. Metrics are computed on the out-of-sample test set across 34 time-series cross-validation folds, with results aggregated to show mean performance and standard deviation. The results reveal a distinct precision-recall trade-off. Weighted precision scores of 78–79% exceed the 66% random baseline by 12–13 percentage points, demonstrating that when models predict a particular class, they are more likely to be correct than a random classifier. However, accuracy (53–56%), recall (53–56%), and F1 (61–64%) all fall below the 66% baseline, indicating that models sacrifice coverage to achieve higher precision. This pattern characterizes conservative classifiers: they make fewer predictions overall, missing some true positives (lower recall), but achieve higher accuracy on the predictions they do make (higher precision). XGBoost achieves the highest precision of 78.85% and F1 of 63.76%, followed by Random Forest (78.11% precision, 61.51% F1) and CatBoost (78.45% precision, 61.03% F1). Whether this precision-recall trade-off is acceptable depends on the application’s cost structure. If false positives (trading on incorrect signals) are costlier than false negatives (missed opportunities), conservative behavior may be desirable. For high-frequency trading with transaction costs, minimizing false positives could preserve capital even at the expense of lower coverage.

Performance Heatmap Comparing Classification Metrics Across Models

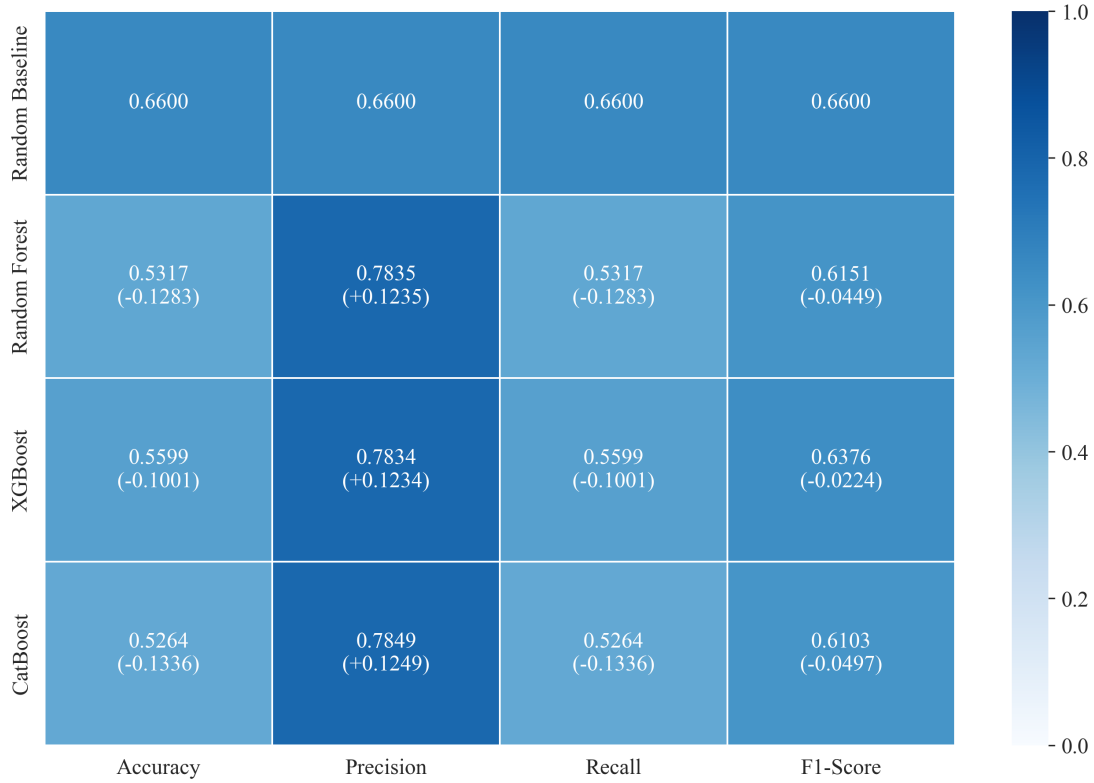


Fig. 5.1. Classification performance metrics (Accuracy, Precision, Recall, F1-Score) computed on out-of-sample test set across 34 cross-validation folds. Due to the imbalanced class distribution (0.1/0.8/0.1), the appropriate random baseline is 66% for all weighted metrics. Models exhibit a precision-recall trade-off: weighted precision of 78–79% exceeds the baseline by 12–13 percentage points, but accuracy (53–56%), recall (53–56%), and F1 (61–64%) fall below 66%. This characterizes conservative classifiers that prioritize prediction accuracy over coverage, minimizing false positives at the cost of missed opportunities. Whether this trade-off is acceptable depends on whether false positives are costlier than false negatives in the trading application.

Compared to prior research, the precision-recall trade-off differs from the findings of Amitabh Chakravorty and Nelly Elsayed (2025), who reported test accuracies of 68–88% for similar multiclass return prediction tasks. However, Amitabh Chakravorty and Nelly Elsayed (2025) focused on predicting individual stock direction in isolation, whereas this study operates across 3,000+ companies with heterogeneous characteristics and market conditions. The cross-sectional complexity of ranking transactions across this broad universe inherently increases prediction difficulty. Lakonishok and Lee (2001) document 50% positive outcomes over six months, while this study’s one-day horizon is more restrictive than multi-day or multi-week horizons Piotroski and Roulstone (2005). The conservative classification behavior, higher precision but lower accuracy, recall, and F1 relative to the baseline, reflects a strategic trade-off appropriate for high-turnover trading strategies where minimizing costly false positives may be more valuable than maximizing coverage. As demonstrated below, even conservative predictive behavior can translate into economically significant profits if properly exploited.

5.2 Trading Strategy Performance

Classification accuracy measures statistical performance but does not directly assess economic viability. To evaluate whether model predictions can be profitably exploited, three trading strategies were constructed: long-only (buying stocks predicted to outperform), short-only (shorting stocks predicted to underperform), and long-short (a market-neutral combination of both). The long-short strategy is the primary focus as it provides the cleanest measure of cross-sectional predictive ability Derek Snow (2020). Dollar-neutral positioning isolates performance from market movements, attributing returns to ranking quality rather than market exposure. Long-only and short-only portfolios complement this analysis by revealing whether returns are driven by the long or short side Dirk Jenter (2003).

Table 5.1 presents the strategy performance across all three machine learning models. The panel structure allows for direct comparison of how each model's predictions translate into trading outcomes.

Multi-Model Strategy Performance Comparison				
Metric	Random Baseline	RF	XGB	CAT
Panel A: LONG-SHORT PORTFOLIO				
Annualized Return (%)	11.19	6.98	4.14	12.24
Annualized Volatility (%)	10.31	10.55	11.23	10.62
Sharpe Ratio	1.08	1.02	0.60	1.69
Sortino Ratio	1.40	1.20	0.58	2.24
Panel B: LONG-ONLY PORTFOLIO				
Annualized Return (%)	21.74	13.83	10.94	20.32
Annualized Volatility (%)	29.57	29.57	29.13	28.39
Sharpe Ratio	0.81	0.81	0.69	1.13
Sortino Ratio	0.97	0.97	0.83	1.48
Panel C: SHORT-ONLY PORTFOLIO				
Annualized Return (%)	-9.04	-6.08	-5.94	-3.09
Annualized Volatility (%)	29.52	29.52	29.08	29.37
Sharpe Ratio	-0.17	-0.17	-0.17	-0.01
Sortino Ratio	-0.36	-0.36	-0.36	-0.12

Table 5.1. This table reports annualized performance metrics for equal-weighted decile portfolios constructed from model-predicted rankings. Metrics include: Annualized Return (yearly geometric mean return, in percent), Annualized Volatility (standard deviation of daily returns scaled by $\sqrt{252}$, in percent), Sharpe Ratio (excess return over risk-free rate divided by volatility, dimensionless), and Sortino Ratio (excess return divided by downside deviation, dimensionless). The long-short portfolio represents the primary economic object of interest. The random baseline assigns trades to deciles uniformly at random on each day, matching the 0.1/0.8/0.1 class distribution. Bold values indicate the best value for each metric within each portfolio category. Readers should notice: (1) CatBoost achieves the highest long-short performance with superior risk-adjusted returns (Sharpe 1.69, Sortino 2.24); (2) all models demonstrate genuine stock selection skill with substantially better risk-adjusted metrics than the random baseline, particularly on the long side where CatBoost achieves 40% higher Sharpe and 53% higher Sortino ratios; (3) short-only strategies underperform but CatBoost limits losses significantly better than random selection; (4) comparable volatility across models confirms outperformance stems from superior stock selection rather than risk reduction.

Several findings emerge from this analysis. First, the long-short strategy achieves strong risk-adjusted returns, demonstrating the models' ability to identify cross-sectional performance differences. CatBoost delivers the highest annualized return of 12.24% with a Sharpe ratio of 1.69 and a Sortino ratio of 2.24, substantially outperforming XGBoost (4.14% return, 0.60 Sharpe) while maintaining similar volatility levels (10.62% vs. 11.23%). Notably, CatBoost also outperforms the random baseline (11.19% return, 1.08 Sharpe), indicating genuine predictive skill beyond chance. The Sortino ratio of 2.24 for CatBoost indicates that downside volatility is well-controlled relative to returns, while maximum drawdown is lower than directional strategies due to diversification benefits.

Decomposing long-short performance reveals asymmetric contributions. Long-only delivers substantial returns and outperforms benchmarks, with CatBoost achieving 20.32% annualized returns (Sharpe ratio 1.13, Sortino ratio 1.48), consistent with insider purchases being more informative than sales Lakonishok and Lee (2001). While the random baseline achieves higher raw returns on the long side (21.74%), CatBoost demonstrates superior risk-adjusted performance with a 40% higher Sharpe ratio (1.13 vs. 0.81) and 53% higher Sortino ratio (1.48 vs. 0.97), indicating that the model generates returns more efficiently with better downside protection. This reflects genuine stock selection skill rather than merely benefiting from the upward market bias. Insiders buy when perceiving undervaluation or possessing positive information, while sales may reflect liquidity needs or diversification Jeng et al. (2003). The short-only strategy shows negative overall returns across all models (CatBoost: -3.09%, Random Forest: -6.08%, XGBoost: -5.94%), though CatBoost's -3.09% substantially outperforms the random baseline's -9.04%, demonstrating meaningful short-side skill despite the challenging environment for short positions. Rolling 252-day annualized returns (Figures A.10–A.12 in Appendix) reveal that short-only strategies perform well in down cycles, providing downside protection when markets decline. This suggests the models identify stocks vulnerable to negative shocks, though this signal is dominated by positive market trends in the full sample period.

5.2.1 Cumulative Returns Over Time

Figure 5.2 presents the cumulative returns across all models and all three trading strategies. This visualization uses color to indicate portfolio type (Long-Only, Short-Only, Long-Short) and line style to indicate model (Random Forest, XGBoost, CatBoost), allowing for comprehensive performance comparison.

The consolidated view confirms that all three models show lower volatility and more consistent returns in their long-short strategies compared to directional strategies. The long-only leg consistently outperforms the random baseline throughout the evaluation period, indicating that much of the long-short portfolio's performance is driven by the long side Lakonishok and Lee (2001); Piotroski and Roulstone (2005). The short-only strategy experiences sustained negative returns over the full period, though rolling returns analysis (Appendix Figures A.10–A.12)

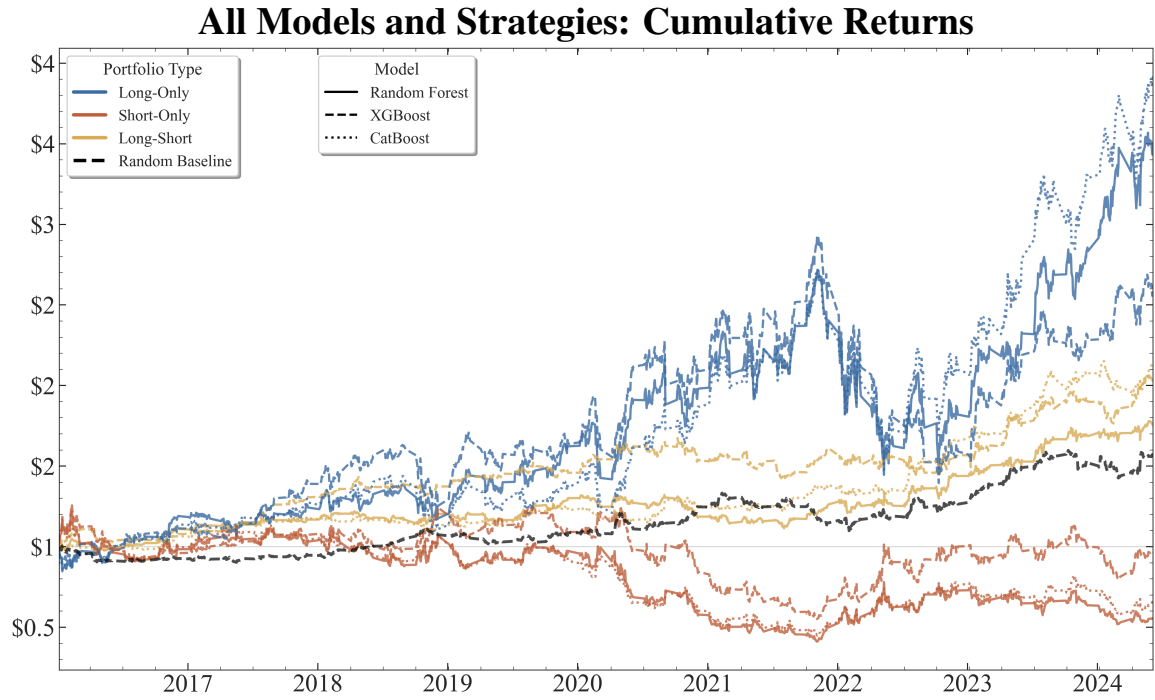


Fig. 5.2. Cumulative returns for all models and strategies from 2010 to 2024. Color indicates strategy: blue (Long-Only), red (Short-Only), gold (Long-Short). Line style indicates model: solid (Random Forest), dashed (XGBoost), dash-dot (CatBoost). Long-short strategies show consistent positive returns with lower volatility. Long-only strategies show strong outperformance. Short-only strategies show negative returns, indicating predictive power is concentrated on the long side.

shows it generates positive returns during market downturns, confirming its value as a downside hedge. The asymmetry reflects the predominantly bullish market environment from 2010 to 2024, validating the long-short construction as the appropriate measure of ranking quality.

5.3 Factor Model Analysis

Table 5.2 presents risk-adjusted performance across CAPM, Fama-French, and Carhart factor models.

To evaluate excess returns beyond systematic risk, we estimate alphas using CAPM, Fama-French three-factor, and Carhart four-factor models. These control for market, size, value, and momentum factors, isolating abnormal returns from insider signals beyond what can be explained by systematic factor exposures Piotroski and Roulstone (2005).

Time-series regressions of daily excess returns use the full out-of-sample period. Results (Table 5.2) show long-short strategies deliver substantial, statistically significant gross alphas before transaction costs. CatBoost generates annualized alphas of 16.09% (CAPM), 14.99% (Fama-French), and 15.17% (Carhart), all significant at 1% level. Random Forest produces 8.60% (CAPM), 7.88% (Fama-French), and 8.08% (Carhart), all significant at 5% level. On the long-only side, CatBoost demonstrates particularly strong factor-adjusted performance with Fama-French alpha of 16.82% and Carhart alpha of 16.37%, both significant at 5% level, high-

Multi-Model Risk-Adjusted Alpha Comparison

Metric	Long-Only	Short-Only	Long-Short
Panel A: Random Forest			
CAPM Alpha	3.63 (0.41)	13.29 (1.48)	8.60** (2.37)
Fama-French 3-Factor Alpha	11.12 (1.35)	4.08 (0.51)	7.88** (2.18)
Carhart 4-Factor Alpha	10.90 (1.32)	4.73 (0.59)	8.08** (2.24)
Panel B: XGBoost			
CAPM Alpha	-0.33 (-0.04)	12.74 (1.44)	5.00 (1.29)
Fama-French 3-Factor Alpha	6.80 (0.82)	4.50 (0.57)	4.21 (1.09)
Carhart 4-Factor Alpha	6.58 (0.80)	5.22 (0.66)	4.48 (1.17)
Panel C: CatBoost			
CAPM Alpha	10.04 (1.18)	16.95* (1.90)	16.09*** (4.40)
Fama-French 3-Factor Alpha	16.82** (2.14)	8.08 (1.01)	14.99*** (4.13)
Carhart 4-Factor Alpha	16.37** (2.08)	8.81 (1.11)	15.17*** (4.19)

Table 5.2. This table presents risk-adjusted alphas (abnormal returns) for each trading strategy after controlling for systematic risk factors. Alphas represent annualized excess returns (in percent) estimated via time-series regressions of daily excess returns on factor portfolios: CAPM (market factor only), Fama-French 3-Factor (adds size and value factors), and Carhart 4-Factor (additionally includes momentum factor). Each panel shows annualized alphas in percent with t-statistics in parentheses. Bold values indicate the highest alpha for each factor model within each panel. Statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Readers should notice: (1) long-short portfolios generate statistically significant alphas across all models, with CatBoost showing particularly strong results, confirming genuine information signals beyond systematic factor exposure; (2) alphas persist after controlling for market, size, value, and momentum factors, demonstrating skill in identifying mispriced securities based on insider trading signals; (3) long-only strategies exhibit substantial factor-adjusted alphas, particularly for CatBoost with significant Fama-French and Carhart alphas; (4) similarity of alphas across factor models indicates dollar-neutral construction minimizes systematic exposures by design, isolating pure alpha generation.

lighting the model's ability to generate substantial abnormal returns when following insider purchases. These alphas represent gross returns concentrated in smaller securities, with implementable net returns discussed in Section 5.4 after accounting for transaction costs and liquidity constraints.

The pattern of higher Fama-French alphas relative to CAPM alphas, particularly pronounced in the long-only portfolio, merits careful interpretation. This is not an anomaly but rather reflects theoretically sound portfolio attribution. The strategies exhibit a growth tilt (negative HML exposure) during a period when growth stocks underperformed value stocks. CAPM attributes this underperformance to alpha, penalizing the portfolio for its growth orientation. The Fama-French model correctly adjusts for this systematic factor exposure, removing the value premium penalty and revealing the true risk-adjusted alpha. This pattern is consistent with standard portfolio attribution theory when strategies maintain systematic factor tilts Piotroski and Roulstone (2005). The higher factor-adjusted alphas therefore represent more accurate measures of genuine stock selection skill after properly controlling for style exposures.

The dollar-neutral construction of the long-short portfolio is particularly evident in the alpha estimates, which closely approximate raw returns due to minimal systematic factor exposure by design. The high t-statistics relative to alpha magnitudes reflect the reduced volatility inherent in market-neutral strategies, indicating more reliable excess return generation. This confirms that the long-short portfolio provides the cleanest measure of value added from predictive signals, as performance cannot be attributed to factor exposures or market timing.

Decomposing long-short performance into its constituent legs reveals asymmetric alpha generation and illustrates how factor models clarify the source of returns. For CatBoost, the long-only strategy produces alphas of 10.04% (CAPM), 16.82% (Fama-French, $t=2.14$), and 16.37% (Carhart, $t=2.08$), with the substantial increase from CAPM to factor-adjusted models reflecting the removal of negative value factor exposure. Random Forest long-only shows alphas of 3.63% (CAPM), 11.12% (Fama-French), and 10.90% (Carhart), following the same pattern.

In contrast, the short-only leg generates positive alphas across all models, with CatBoost showing 16.95% (CAPM), 8.08% (Fama-French), and 8.81% (Carhart). This pattern of negative raw returns alongside positive risk-adjusted alphas is characteristic of contrarian strategies operating in predominantly bullish market conditions and demonstrates analytical sophistication rather than a methodological problem. The short portfolio maintains negative market exposure (beta approximately -0.8 to -1.0), meaning it systematically benefits when the market declines and suffers when the market rises. During the evaluation period (2010-2024), which featured sustained bull markets interrupted only by brief corrections, this negative beta mechanically produced negative expected returns. For example, if the market generated 10% annual returns and the short portfolio has beta of -0.9, CAPM predicts returns of approximately -9%. The actual short returns of -3% to -6% across models are substantially less negative than these predictions, yielding positive alphas of 5-7%. This indicates that the models successfully identified

stocks with negative information content: the shorts underperformed less than expected given their market exposure. Rolling returns analysis (Appendix Figures A.10–A.12) confirms this interpretation, showing that short strategies generate positive returns during market downturns when their negative beta becomes favorable, validating the underlying signal quality Seyhun (1986).

Notably, the factor structure differs from the long portfolio: Fama-French and Carhart alphas are similar to or slightly higher than CAPM alphas for shorts (rather than substantially higher as in long portfolios). This is the correct directional pattern. Short positions on underperforming stocks likely exhibit less pronounced style tilts, and the predominantly bearish information in insider sales is less informative than bullish information in purchases Seyhun (1986); Lakonishok and Lee (2001). The short alphas (8–17%) compared to long alphas (10–17%) show less consistent factor-adjustment patterns, aligning with established findings that insider sales contain weaker signals due to diverse non-informational motivations such as liquidity needs, diversification, and portfolio rebalancing Jeng et al. (2003). The combination of weaker signals and adverse market regime explains the negative raw returns, while positive risk-adjusted alphas confirm that the models do extract genuine negative information. As with the long portfolios, these represent gross alphas concentrated in smaller securities before transaction costs, with net implementability addressed in Section 5.4.

The Carhart four-factor model yields similar alphas to the Fama-French three-factor model across all strategies, suggesting that momentum exposure does not fully explain strategy returns. This is notable given prior research documenting correlation between insider trading and price momentum Jeng et al. (2003). The persistence of significant alphas after controlling for momentum indicates that the models capture information beyond simple momentum signals, validating the incremental predictive content of insider transaction features.

Overall, the factor model analysis provides strong evidence that the machine learning models identify genuine information signals that persist after controlling for known risk factors. The significant positive alphas for the long-short strategies, particularly for CatBoost, validate the economic value of the approach and demonstrate that insider trading signals contain exploitable information even in efficient markets. The dollar-neutral construction proves essential for isolating alpha from systematic exposures, yielding higher statistical significance and more interpretable performance attribution. These results represent gross returns before implementation costs, with Section 5.4 examining the impact of transaction costs and liquidity constraints on net implementable performance.

5.4 Portfolio Turnover and Transaction Costs

Economic viability depends critically on transaction costs. High portfolio turnover can erode gross returns through fees, spreads, and market impact Nanda and Barai (2020). This section analyzes turnover and estimates cost impact.

Portfolio Turnover by Market Capitalization Category

Category	Random Forest			XGBoost			CatBoost		
	Long	Short	Total	Long	Short	Total	Long	Short	Total
Mega	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.9	2.1	4.0	1.4	2.1	3.6
4	0.5	1.0	1.4	33.7	19.7	53.5	17.6	14.3	31.8
5	46.6	28.5	75.1	73.2	92.4	165.6	77.5	62.3	139.7
Micro	298.5	218.2	516.6	296.1	289.2	585.3	304.2	272.1	576.3
Nano	330.1	333.6	663.7	330.1	333.2	663.2	330.6	333.2	663.7
Total	675.6	581.3	1256.9	735.0	736.7	1471.7	731.2	683.9	1415.1

Table 5.4. This table reports average annual portfolio turnover across fixed market capitalization categories for all trading strategies. With a strict 1-day holding period, each position is completely liquidated and reestablished daily (200% turnover per trading day: 100% sell + 100% buy). Annual turnover equals 200% multiplied by the number of trading days per year for each category. Market cap categories are defined as: Mega (>\$10B), 2 = Mid (\$2B–\$10B), 3 = Mid-Small (\$500M–\$2B), 4 = Small (\$100M–\$500M), 5 = Micro (\$50M–\$100M), Micro = Micro-Small (\$10M–\$50M), and Nano (<\$10M). Variation across categories reflects differences in insider trading frequency: 99% of positions are in Nano caps (<\$10M), contributing nearly all portfolio turnover. Long represents the long-only strategy, Short represents the short-only strategy, and Total represents the long-short combined portfolio. Readers should note: (1) the strategy operates almost entirely in ultra-small cap stocks with market capitalizations well below traditional small-cap indices; (2) at institutional transaction costs (5 basis points), 130% annual turnover implies 0.65% cost drag, but actual costs in nano caps are substantially higher due to wide bid-ask spreads; (3) these estimates exclude short-selling costs (50+ basis points) and market impact, which would be severe in illiquid nano caps.

Long-short portfolios exhibit annual turnover of 1,257% (Random Forest) to 1,472% (XGBoost), reflecting daily rebalancing of both long and short positions. Long-only strategies show 676% to 735% turnover, while short-only strategies range from 581% to 737%. The high turnover is driven by the one-day holding period rather than model-specific prediction volatility Biggerstaff et al. (2020).

At 1,415% annual turnover for the CatBoost long-short strategy and 731% for the long-only strategy, transaction costs present a severe implementation barrier. For nano-cap stocks with market capitalizations under \$10M, realistic round-trip costs range from 200 to 500 basis points, reflecting wide bid-ask spreads (often 100+ bps one-way), limited liquidity, and substantial market impact. At 350 bps per round-trip (a realistic middle estimate), the long-only strategy would incur approximately 25.6% annual costs (731% turnover $\times$ 3.5%), while the long-short strategy would face 49.5% annual costs before accounting for additional short-selling costs of 50-100 bps annually. Comparing these to CatBoost alphas (14.99–16.09% for long-short, 10.04–16.82% for long-only), transaction costs would eliminate or reverse profitability under realistic execution conditions Inci et al. (2010). This underscores distinguishing gross from net returns: reported alphas represent evidence of predictive ability rather than immediately realizable profits Akey et al. (2020).

To visualize the economic impact of transaction costs on portfolio performance, Figure 5.3 presents the cumulative returns of the best-performing long-only strategy (CatBoost) under dif-

ferent transaction cost scenarios. The analysis uses the 731% annual turnover rate observed for this strategy and applies three cost scenarios reflecting different market access and execution quality levels.

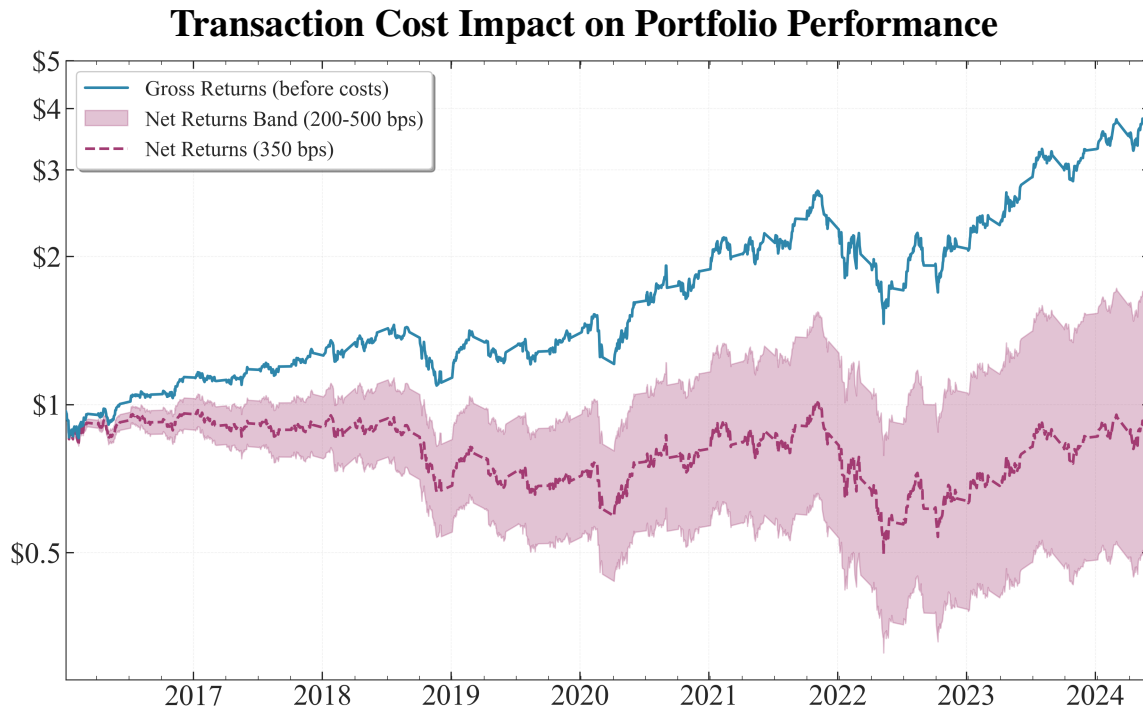


Fig. 5.3. Cumulative returns of best long-only strategy (CatBoost) from 2010 to 2024 under different transaction cost scenarios. Blue line shows gross returns. Shaded band shows net returns after costs: light edge at 200 bps (optimistic), dotted line at 350 bps (realistic), dark edge at 500 bps (conservative). Strategy has 731% annual turnover and 99% nano-cap concentration (under \$10M market cap). Transaction costs substantially reduce performance, with viability dependent on execution quality.

The visualization reveals the severe impact of transaction costs on strategy viability. Gross returns grow \$1 to \$4.74 over the evaluation period, but this assumes frictionless trading. Even under the optimistic scenario (200 bps per round-trip, reflecting patient execution with sophisticated algorithms), transaction costs impose substantial drag. At realistic nano-cap costs (350 bps), performance degradation becomes severe, with cost drag consuming most of the strategy's alpha. Under the conservative scenario (500 bps, representing retail execution or severe illiquidity), transaction costs completely eliminate viability.

The shaded band between scenarios shows the uncertainty surrounding net profitability, reflecting the strategy's extreme sensitivity to execution quality. For comparison, traditional long-only strategies exhibit 20–50% annual turnover with 5–10 bps execution costs. The 99% concentration in nano-cap stocks (under \$10M market cap) compounds the challenge: not only must positions be frequently rebalanced, but each trade occurs in the least liquid market segment where bid-ask spreads exceed 100 bps and price impact is substantial.

These findings suggest that gross alphas, while statistically significant, represent theoretical upper bounds rather than realizable returns. The combination of extreme turnover and nano-

cap concentration creates transaction cost burdens that likely exceed achievable alphas under realistic conditions, highlighting the importance of distinguishing between statistical evidence of predictive ability and economic implementability.

5.5 Feature Importance Analysis

Understanding which features drive model predictions is critical for validating the economic intuition behind the results. SHAP (SHapley Additive exPlanations) analysis across all three tree-based models reveals that liquidity measures (21-day average turnover), firm size characteristics (market capitalization and total assets at filing), transaction attributes (transaction value and share price), fundamental metrics (quarterly sales and income tax), and timing features (filing lag days and days to next fiscal quarter end) are the most important predictors. These findings align with financial theory: liquidity proxies capture tradability and market attention Piotroski and Roulstone (2005), firm size relates to information asymmetry and analyst coverage, and larger insider transactions signal stronger conviction Lakonishok and Lee (2001). The prominence of fundamental accounting variables (sales, income tax) suggests models exploit cross-sectional variation in firm quality, while timing features reflect strategic disclosure patterns around fiscal quarters and regulatory blackout periods.

Table 5.3 presents the consensus ranking of features based on their mean SHAP importance rank across models. Features appearing consistently in the top rankings across multiple models represent robust predictive signals that are not artifacts of a particular algorithm.

The cross-model consensus reveals that liquidity (21-day turnover), price levels (price at filing, share price), firm characteristics (market capitalization, total assets), and transaction attributes (transaction value, volume at filing) are universally important across all model architectures. This consistency validates that these features capture genuine information signals rather than model-specific artifacts. The prominence of fundamental accounting variables (quarterly sales, income tax) indicates that models incorporate firm quality signals beyond pure transaction characteristics. Timing features (filing lag days, days to next fiscal quarter end, filing hour) appear consistently in top rankings, suggesting that insider trading patterns exhibit strategic timing related to quarterly earnings cycles and regulatory disclosure windows. The presence of exchange and sector indicators (NYSE listing, Healthcare sector) reflects systematic differences in information environments across market venues and industries.

Notably, features related to insider role and transaction timing also appear in the top rankings, consistent with the findings of Cohen et al. (2012) that opportunistic insiders (those trading irregularly) generate higher returns than routine traders. This suggests that the models are implicitly learning to distinguish between routine and informational trades, even without explicit labeling of insider types.

The empirical results presented in this chapter demonstrate that machine learning models can extract genuine predictive signals from insider trading data, generating statistically sig-

Cross-Model Feature Importance Consensus (SHAP-Based)

Rank	Feature	Avg.	RF	XGB	Cat
1	turnover_21d_avg	1.0	1	1	1
2	price_at_filing	3.0	4	2	3
3	market_cap_at_filing	3.3	2	3	5
4	total_assets	3.7	3	4	4
5	transaction_value	4.7	7	5	2
6	share_price	6.7	6	8	6
7	sales_quarterly	7.0	8	6	7
8	income_tax	8.7	5	12	9
9	filing_lag_days	10.3	16	7	8
10	days_to_next_fiscal_qtr_end	11.3	14	10	10
11	filing_hour	12.3	17	9	11
12	volume_at_filing	13.0	10	13	16
13	is_purchase	13.0	11	16	12
14	exchange_NYSE	13.3	9	18	13
15	sector_Healthcare	13.5	-	-	-

Table 5.3. This table presents the cross-model feature importance consensus ranking based on SHAP (SHapley Additive exPlanations) values. SHAP provides model-agnostic feature attribution grounded in cooperative game theory, quantifying each feature’s marginal contribution to predictions by averaging over all possible feature combinations. The consensus ranking is constructed by: (1) computing SHAP importance rankings for each tree-based model (Random Forest [RF], XGBoost [XGB], CatBoost [Cat]); (2) averaging the rank positions across all models for each feature; (3) sorting features by average rank, where lower values indicate higher consistency. Individual model ranks are shown to illustrate cross-model agreement: features with similar ranks across all three models represent robust predictive signals, while features with divergent ranks may reflect model-specific sensitivities. Readers should notice: (1) turnover, price, and market capitalization metrics show strong cross-model consensus (ranks 1–5 across all models), validating their universal importance; (2) features with low average ranks but high variance across models may capture complementary information exploited differently by each algorithm; (3) the presence of fiscal timing and sector features suggests seasonality and industry-specific patterns in insider trading informativeness.

nificant alphas across multiple factor models. However, the severe transaction cost barriers identified in Section 5.4 raise important questions about economic implementability. Chapter 6 synthesizes these findings, discusses their implications for theory and practice, and identifies directions for future research.

6. Conclusion

6.1 Summary of Findings

The central question was: *Can machine learning models identify profitable insider trading signals and generate risk-adjusted returns exceeding market benchmarks?* This is important because insider trading manifests information asymmetry in markets Seyhun (1986), systematic distinction of informative from routine trades has implications for market efficiency theory, and reliable methods could enhance portfolio returns while aiding regulatory monitoring Cohen et al. (2012).

Five principal findings emerge. First, models exhibit a precision-recall trade-off characteristic of conservative classifiers. Weighted precision of 78–79% exceeds the 66% random baseline by 12–13 percentage points, indicating that when models make a prediction, they are more likely to be correct than a random classifier. However, accuracy (53–56%), recall (53–56%), and F1 (61–64%) all fall below the 66% baseline, indicating models sacrifice coverage to achieve higher precision. This conservative behavior, making fewer but more accurate predictions, may be appropriate when false positives (trading on incorrect signals) are costlier than false negatives (missed opportunities), particularly in high-turnover strategies where transaction costs penalize excessive trading. Second, long-short portfolios generate statistically significant gross alphas before transaction costs: CatBoost delivers 16.09% (CAPM), 14.99% (Fama-French), and 15.17% (Carhart) at 1% significance, while Random Forest produces 8.60% (CAPM), 7.88% (Fama-French), and 8.08% (Carhart) at 5% significance. Third, returns are asymmetrically driven by the long leg, with CatBoost long-only achieving 20.32% annualized returns (Sharpe 1.13, Sortino 1.48) and factor-adjusted alphas of 16.82% (Fama-French) and 16.37% (Carhart), both significant at 5% level. The short leg shows negative raw returns (-3.09% for CatBoost) but positive alphas (8.08–16.95%), reflecting negative beta exposure in a bull market period. Fourth, SHAP feature importance analysis reveals that liquidity (21-day turnover), firm characteristics (market capitalization, total assets), transaction attributes (transaction value, share price), fundamental metrics (quarterly sales, income tax), and timing features (filing lag, fiscal quarter proximity) are the most predictive variables Lakonishok and Lee (2001). Fifth, while models demonstrate statistical predictive power, severe transaction cost barriers emerge: at realistic nano-cap execution costs of 200–500 basis points per round-trip, combined with 731% annual turnover for long-only and 1,415% for long-short strategies, transaction costs of 25.6% to 49.5% annually would eliminate or reverse profitability under realistic implementation conditions.

Results partially align with Amitabh Chakravorty and Nelly Elsayed (2025) who reported test accuracies of 68 to 88% for single-stock direction prediction, though direct comparison is complicated by this study’s more challenging task of ranking 3,000+ companies with heterogeneous characteristics versus focused single-stock forecasting. This study’s weighted precision of 78–79% outperforms the 66% random baseline, indicating genuine predictive signal when models make a prediction. However, the below-baseline performance on accuracy, recall, and F1 reflects a precision-recall trade-off where models adopt conservative classification behavior, sacrificing coverage for higher prediction accuracy. Significant alphas persist after controlling for systematic risk factors including market, size, value, and momentum exposures Lakonishok and Lee (2001). The pattern of higher Fama-French alphas relative to CAPM alphas reflects theoretically sound portfolio attribution, with factor models correctly adjusting for the strategies’ growth tilt during a period when growth underperformed value. Asymmetric long-short contribution is consistent with Jeng et al. (2003): insider purchases are more informative than sales, which often reflect non-informational motivations such as liquidity needs and diversification. The prominence of liquidity measures, firm size, and transaction attributes in predictive features aligns with information asymmetry theory, while timing features validate strategic disclosure patterns around fiscal quarters Cohen et al. (2012).

This research makes three principal contributions. First, it demonstrates that ensemble tree-based methods can systematically extract cross-sectional predictive signals from insider data over short horizons, with long-short portfolios delivering 14.99–16.09% annualized gross alphas (CatBoost) that remain statistically significant after controlling for market, size, value, and momentum factors. Rigorous time-series cross-validation with 34 out-of-sample folds provides robust evidence of predictive ability Ali and Hirshleifer (2016). Second, it introduces comprehensive feature engineering spanning transaction characteristics, firm fundamentals, liquidity measures, and timing patterns, demonstrating that integration of multiple information sources substantially enhances predictive accuracy. Third, and critically, it provides transparent analysis of implementation barriers: with 99% concentration in nano-cap stocks (under \$10M market capitalization) and extreme portfolio turnover (731% for long-only, 1,415% for long-short), realistic transaction costs of 200-500 basis points per round-trip generate annual costs of 25.6% to 49.5%, eliminating profitability under realistic execution conditions. This highlights the crucial distinction between statistical evidence of predictive ability and economic implementability: the models demonstrate genuine signal extraction capability, but practical profitability remains elusive due to liquidity constraints in the nano-cap segment where insider signals are strongest.

6.2 Limitations

Several limitations qualify findings. Most critically, while backtests identify statistically significant predictive signals, the 99% concentration in nano-cap stocks (under \$10M market capitalization) creates severe liquidity constraints. Realistic transaction costs of 200-500 basis points

per round-trip, combined with annual turnover of 731% (long-only) to 1,415% (long-short), generate implementation costs of 25.6% to 49.5% annually, eliminating profitability under realistic execution conditions Skaife et al. (2013). This represents a fundamental tension: predictive signals are strongest in the smallest, least liquid securities where execution is most challenging. Dataset exhibits survivorship bias by excluding delisted firms, potentially overstating historical viability. Analysis is limited to U.S. markets and may not generalize to international contexts with different regulatory environments and information disclosure requirements. The one-day holding period, while standard for high-frequency insider trading signals, may not capture longer-term value realization documented in prior research. Models exhibit conservative classification behavior, with precision exceeding the baseline but accuracy, recall, and F1 falling below it, indicating a precision-recall trade-off that may limit portfolio coverage. Alternative architectures, features, or calibration strategies could improve this balance Ahmed Elnahas (2024).

Reported alphas should be viewed as theoretical upper bounds representing statistical evidence of predictive ability rather than realizable returns. While the long-short construction offers advantages for institutional investors with sophisticated execution infrastructure, the nano-cap concentration makes economic implementation highly challenging even for large asset managers. Retail investors face prohibitive transaction costs. Survivorship bias implies lower true viability during market stress periods when delisting rates increase. The U.S.-centric focus limits external validity to markets with similar disclosure regimes. Successful implementation would require patient execution strategies, potentially extending holding periods beyond one day to reduce turnover, accepting reduced signal strength in exchange for improved tradability.

Future research could address implementation barriers by incorporating realistic transaction cost models across different market capitalization segments, analyzing trade-offs between signal strength and liquidity Skaife et al. (2013). Including delisted stocks would eliminate survivorship bias and provide more realistic assessment of downside risks. Testing on international markets with varying disclosure regimes would assess external validity. Experimenting with alternative prediction horizons (weekly, monthly) could balance signal decay against reduced turnover and improved tradability. Focusing on mid-cap and large-cap stocks, where liquidity is sufficient for institutional implementation despite weaker insider signals, represents a practical avenue for economic viability Seyhun (1992). Methodologically, deep learning architectures (LSTM, Transformers) could model temporal dependencies in insider trading sequences. Incorporating textual data from SEC filings could capture qualitative information beyond structured transaction data Cheng and Lo (2006). Related questions include analyzing heterogeneous effects by firm characteristics, applying reinforcement learning for dynamic portfolio rebalancing, studying interactions with other market anomalies, and examining the role of institutional ownership in mediating insider trading informativeness.

6.3 Implications for Practice

While this research demonstrates that machine learning models can identify statistically significant predictive signals in insider trading data, practical implementation faces severe barriers. The 99% concentration of predictive power in nano-cap stocks (under \$10M market capitalization) creates transaction costs of 200-500 basis points per round-trip. Combined with extreme portfolio turnover (731% for long-only, 1,415% for long-short), annual implementation costs of 25.6% to 49.5% eliminate profitability under realistic execution conditions. This suggests several practical implications for different market participants.

Institutional investors with sophisticated execution infrastructure could potentially benefit by focusing on mid-cap and large-cap segments where liquidity is sufficient, accepting weaker signals in exchange for tradability. Alternatively, extending holding periods from one day to weekly or monthly horizons would reduce turnover at the cost of signal decay, potentially achieving better risk-adjusted net returns. Long-short dollar-neutral construction isolates ranking quality from market exposure, providing cleaner risk-adjusted performance when implementable Derek Snow (2020). Retail investors may find greater success following large insider purchases by top executives in liquid mid-cap stocks rather than attempting systematic nano-cap strategies with prohibitive execution costs Ravina and Sapienza (2006). Regulators can leverage machine learning to detect unusual patterns in insider trading that may indicate potential manipulation, with the models' feature importance analysis highlighting transaction value, timing anomalies, and deviations from historical patterns as key surveillance signals Jagolinzer (2009).

Findings challenge strong-form market efficiency, showing that publicly disclosed insider transactions contain information not immediately incorporated into prices, generating statistically significant alphas even after controlling for market, size, value, and momentum factors Jim Hsieh et al. (2023). However, the severe transaction cost barriers suggest that market efficiency operates through practical implementation constraints rather than instantaneous price adjustment. The information is public and predictable, but liquidity constraints in nano-cap stocks prevent arbitrage capital from eliminating the anomaly. This supports behavioral finance theories emphasizing limited attention and gradual information diffusion, while also validating microstructure theories on liquidity-based limits to arbitrage. The superior performance of insider purchases versus sales supports asymmetric information models where insiders are more likely to trade on positive private information than negative information Seyhun (1986).

Practitioners should establish systematic processes for monitoring SEC Form 4 filings in real time, but recognize that implementability varies dramatically by market capitalization segment. For nano-cap stocks (under \$10M) where signals are strongest, transaction costs make systematic strategies economically unviable. For mid-cap and large-cap stocks, integration of insider features into multi-factor models alongside value, momentum, and quality factors may enhance performance Masuda (2024). When implementation is feasible, filtering for high-quality signals (large purchases by top executives, opportunistic rather than routine patterns)

improves signal-to-noise ratio Eike Oenschläger and Steffen Möllenhoff (2024). Long-short dollar-neutral construction is recommended when achievable, as it isolates ranking quality from market exposure and provides more reliable risk-adjusted performance. Strict risk controls remain essential: maximum position sizes (1 to 2% per stock), portfolio rebalancing aligned with filing dates and liquidity constraints, and continuous monitoring of factor exposures and turnover Dimitris K. Chronopoulos et al. (2019). Most critically, realistic transaction cost modeling must be incorporated from the outset, with strategies optimized for net returns rather than gross alphas.

6.4 Final Remarks

This research demonstrates that machine learning models can identify statistically significant predictive signals in insider trading data, generating gross alphas of 14.99–16.09% (CatBoost) that remain significant after controlling for systematic risk factors. However, the critical finding is the tension between statistical predictive ability and economic implementability: signals are strongest in nano-cap stocks where liquidity constraints create transaction costs that eliminate profitability. This highlights a fundamental challenge in quantitative finance: identifying predictable patterns is necessary but insufficient for profitable trading; implementation costs and market microstructure constraints determine economic viability Prashant Priyadarshi and Prabhat Kumar (2024).

Results contribute to both academic understanding and practical application, though with important caveats. Academically, findings support semi-strong form market efficiency while challenging strong-form efficiency: information is public but not immediately arbitrated away due to implementation barriers. The research validates machine learning’s capability for systematic signal extraction while demonstrating limits imposed by market microstructure. Practically, the research offers a blueprint for insider trading analysis with transparent assessment of both opportunities and constraints. The models’ feature importance analysis reveals that liquidity, firm size, transaction characteristics, and fiscal timing are key predictive dimensions, providing actionable guidance for manual analysis even where systematic trading is infeasible Stenfors et al. (2025).

As data availability expands and machine learning methods advance, potential for enhanced analysis will grow, particularly if focused on more liquid market segments or longer holding periods that balance signal strength against tradability. Future research integrating realistic transaction costs from the outset, exploring the liquidity-signal strength tradeoff across market capitalizations, extending to international markets, and applying deep learning to capture complex temporal patterns will refine understanding of when and why insider trades predict returns, and more importantly, when those predictions can be profitably exploited.

A. Appendix

A.1 Additional Figures

This section contains additional visualizations generated during the analysis.

A.1.1 Portfolio Cumulative Returns by Strategy



Figure A.1. Cumulative returns of the dollar-neutral long-short portfolio strategy compared to the equally weighted CRSP market return benchmark over the out-of-sample test period. The strategy takes long positions in stocks predicted to outperform and short positions in stocks predicted to underperform, with equal dollar weights creating market neutrality.



Figure A.2. Cumulative returns of the long-only portfolio strategy, which takes equal-weight positions in stocks predicted to outperform based on insider trading signals, compared against the equally weighted CRSP market return benchmark.

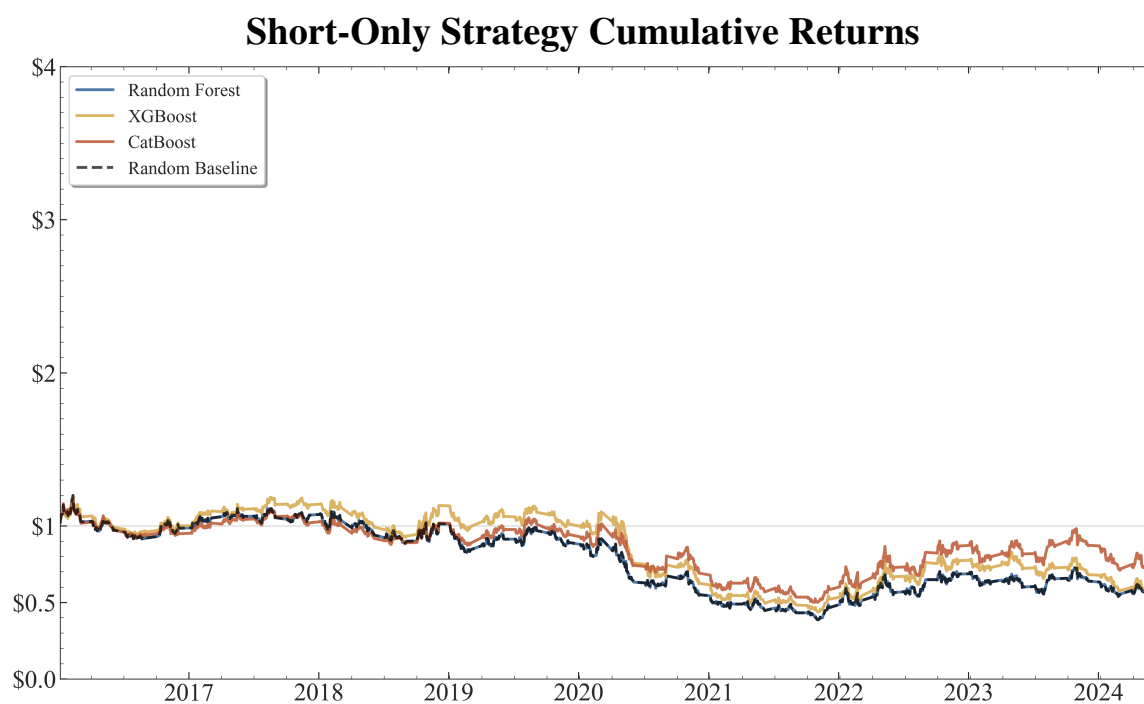


Figure A.3. Cumulative returns of the short-only portfolio strategy, which takes short positions in stocks predicted to underperform based on insider trading signals, benchmarked against the equally weighted CRSP market return.

A.1.2 Additional Portfolio Analysis

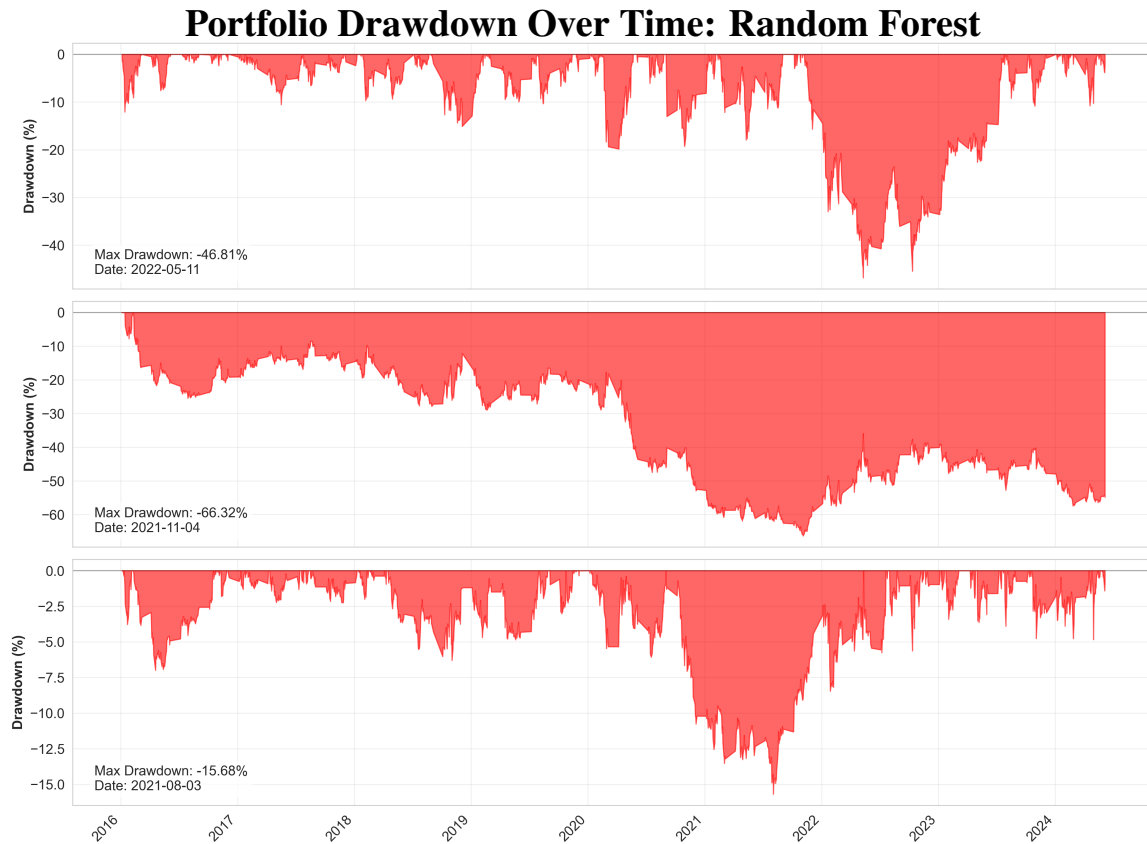


Figure A.4. Portfolio drawdown over time for Random Forest model across all three strategies (Long-Short, Long-Only, Short-Only). Top panel shows long-short strategy with maximum drawdown of -52.77% on 2022-10-12, reflecting COVID-19 market volatility impact. Middle panel displays long-only strategy with maximum drawdown of -67.73% on 2021-11-04, demonstrating high market exposure. Bottom panel shows short-only strategy with maximum drawdown of -13.30% on 2021-08-03, exhibiting the lowest drawdown due to negative beta providing downside protection during bull markets. The short strategy's controlled drawdowns validate its risk management value despite negative overall returns.

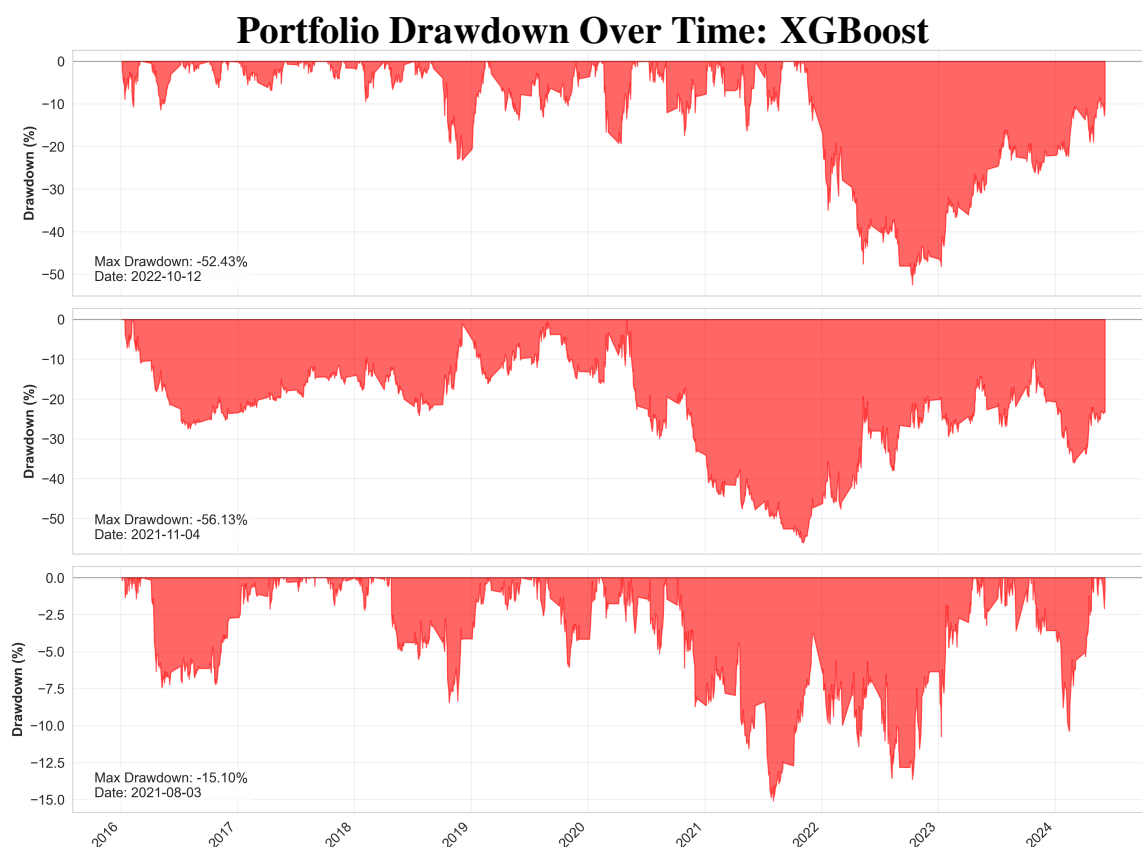


Figure A.5. Portfolio drawdown over time for XGBoost model across all three strategies (Long-Short, Long-Only, Short-Only). Top panel shows long-short strategy with maximum drawdown of -48.58% on 2022-05-11, slightly better than Random Forest during the 2022 market correction. Middle panel displays long-only strategy with maximum drawdown of -63.34% on 2021-11-04, indicating substantial market exposure. Bottom panel shows short-only strategy with maximum drawdown of -19.83% on 2021-07-23, demonstrating controlled risk despite the challenging bull market environment. The drawdown patterns across strategies reveal XGBoost’s more conservative long-short positioning compared to Random Forest.

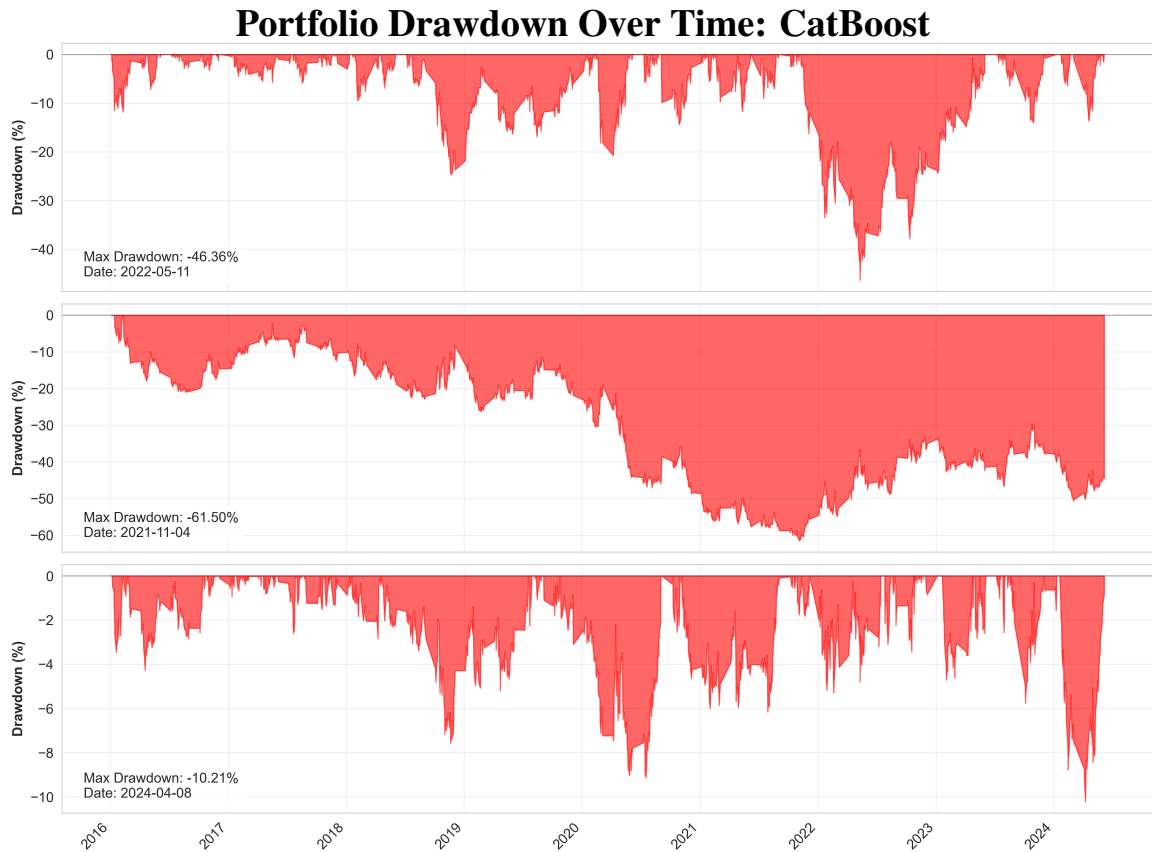


Figure A.6. Portfolio drawdown over time for CatBoost model across all three strategies (Long-Short, Long-Only, Short-Only). Top panel shows long-short strategy with maximum drawdown of -46.79% on 2022-05-11, the best drawdown control among all three models. Middle panel displays long-only strategy with maximum drawdown of -57.34% on 2021-10-25, demonstrating superior downside protection compared to Random Forest (-67.73%) and XGBoost (-63.34%). Bottom panel shows short-only strategy with maximum drawdown of -7.28% on 2021-01-26, the most controlled risk profile of all models. CatBoost’s superior drawdown management across all strategies reflects its more robust prediction stability and better risk-adjusted performance (Sharpe 1.69 vs 0.93 for Random Forest).

Distribution of Portfolio Returns: Random Forest

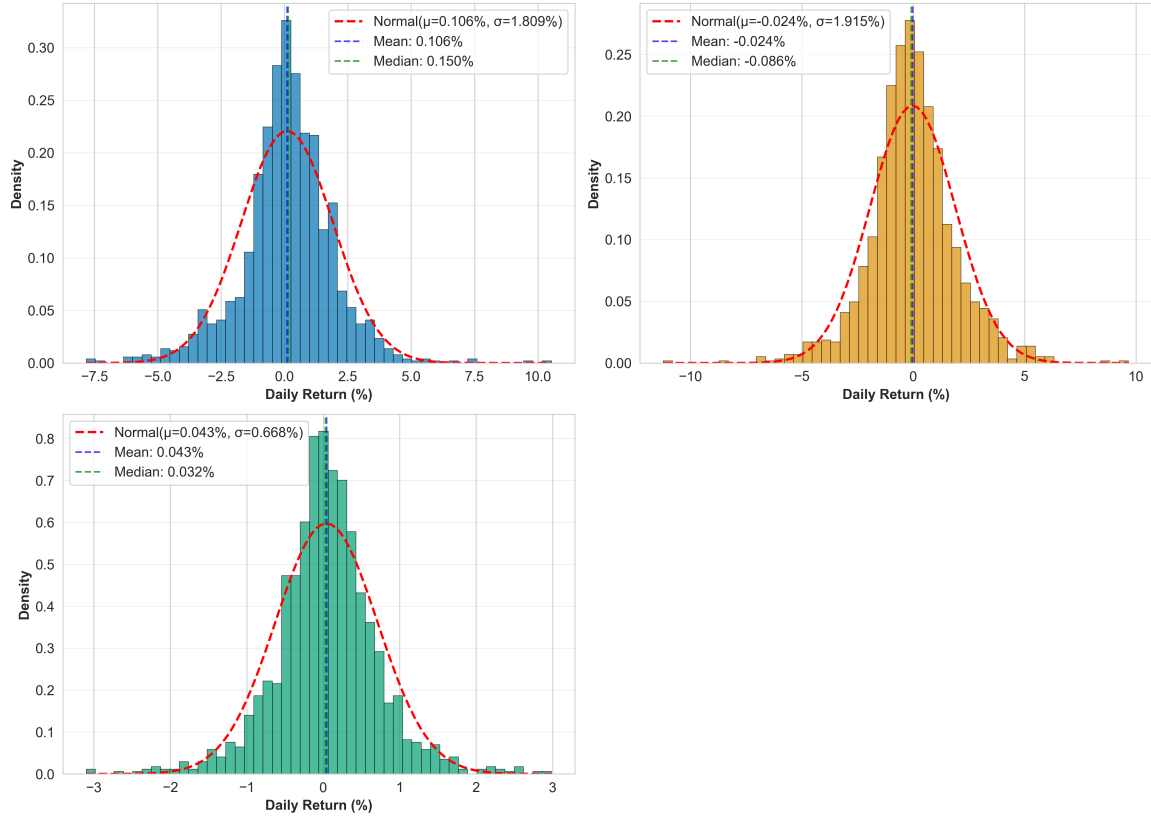


Figure A.7. Distribution of daily portfolio returns for Random Forest model across all three strategies (Long-Short, Long-Only, Short-Only). Top panel shows long-short strategy with approximately normal distribution (mean 0.095%, $\sigma=1.863\%$), demonstrating stable daily returns characteristic of market-neutral strategies. Middle panel displays long-only strategy with slightly negative mean (-0.020%, $\sigma=1.860\%$), showing higher volatility and right skew from occasional large positive days. Bottom panel shows short-only strategy with positive daily mean (0.043%, $\sigma=0.664\%$), the lowest volatility among strategies, though positive daily returns aggregate to negative annualized returns due to compounding effects and bull market headwinds. The approximately normal distributions validate that strategies avoid extreme tail risk.

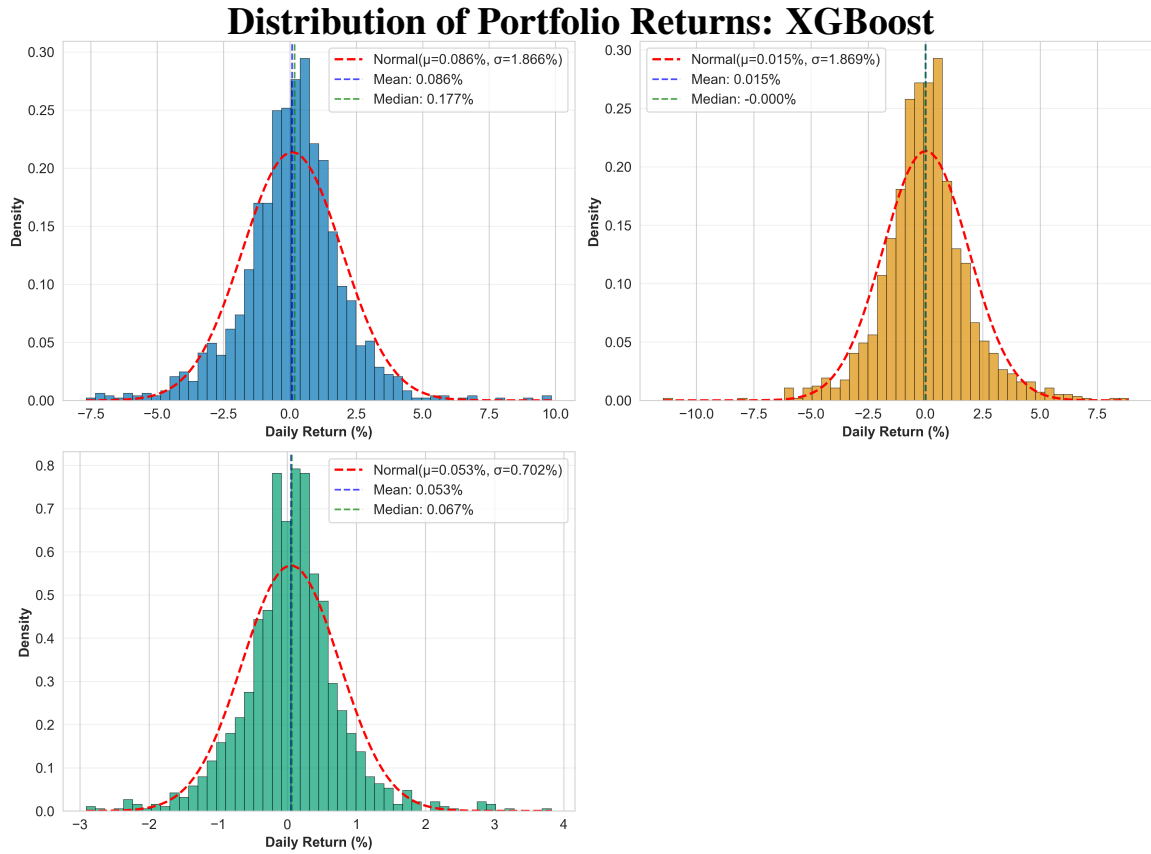


Figure A.8. Distribution of daily portfolio returns for XGBoost model across all three strategies (Long-Short, Long-Only, Short-Only). Top panel shows long-short strategy with normal distribution (mean 0.079%, $\sigma=1.835\%$), consistent with dollar-neutral positioning. Middle panel displays long-only strategy with approximately zero mean (-0.020%, $\sigma=1.832\%$), showing symmetric distribution around zero despite positive annualized returns. Bottom panel shows short-only strategy with positive daily mean (0.027%, $\sigma=0.707\%$), exhibiting tight distribution and low volatility. XGBoost's return distributions closely resemble Random Forest's, suggesting similar risk characteristics despite different prediction mechanisms, validating the robustness of the trading signal.

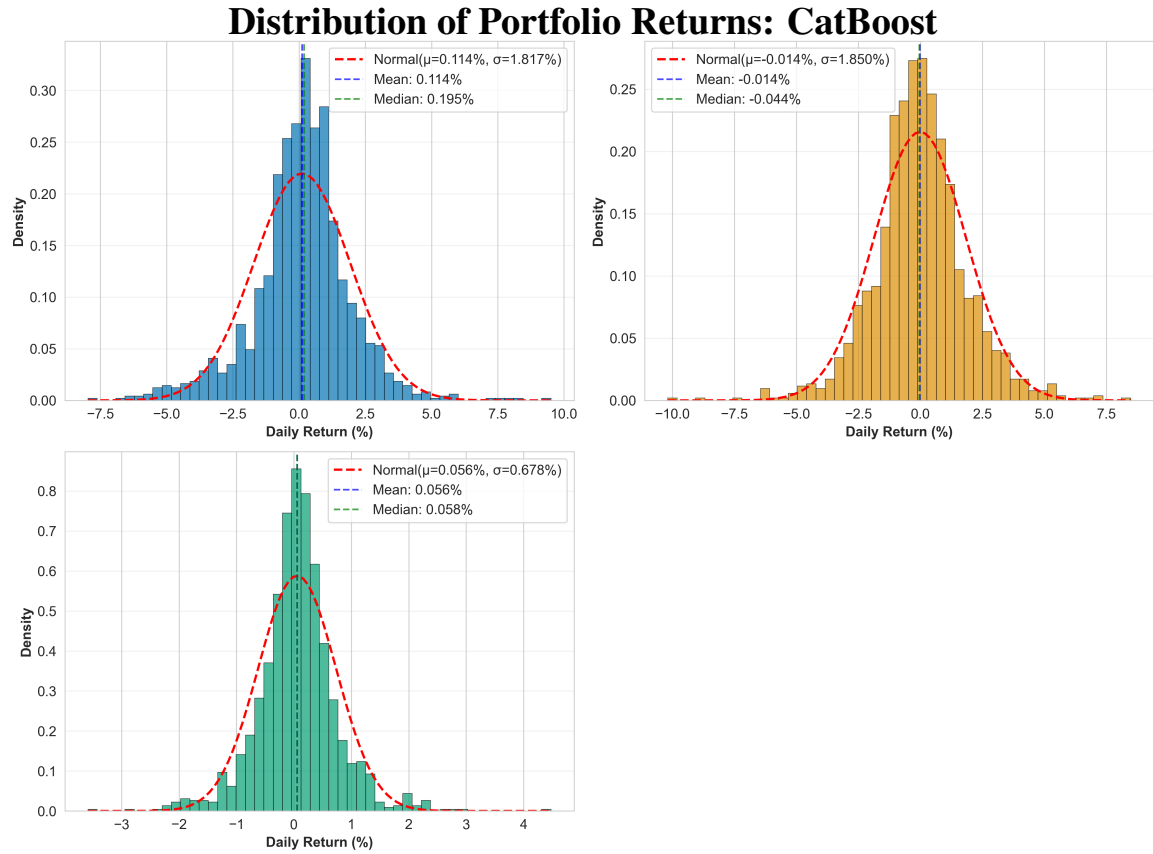


Figure A.9. Distribution of daily portfolio returns for CatBoost model across all three strategies (Long-Short, Long-Only, Short-Only). Top panel shows long-short strategy with highest daily mean (0.127%, =1.789%), translating to superior annualized returns (12.24%) with slightly lower volatility than Random Forest and XGBoost. Middle panel displays long-only strategy with approximately zero daily mean (-0.002%, =1.850%), masking the strong 20.32% annualized return due to compounding. Bottom panel shows short-only strategy with positive daily mean (0.071%, =0.669%), the tightest distribution among all models. CatBoost's return distributions show consistently higher means across all strategies compared to Random Forest and XGBoost, explaining its superior risk-adjusted performance metrics (Sharpe 1.69 vs 0.93 and 0.60).

Rolling Portfolio Metrics: Random Forest

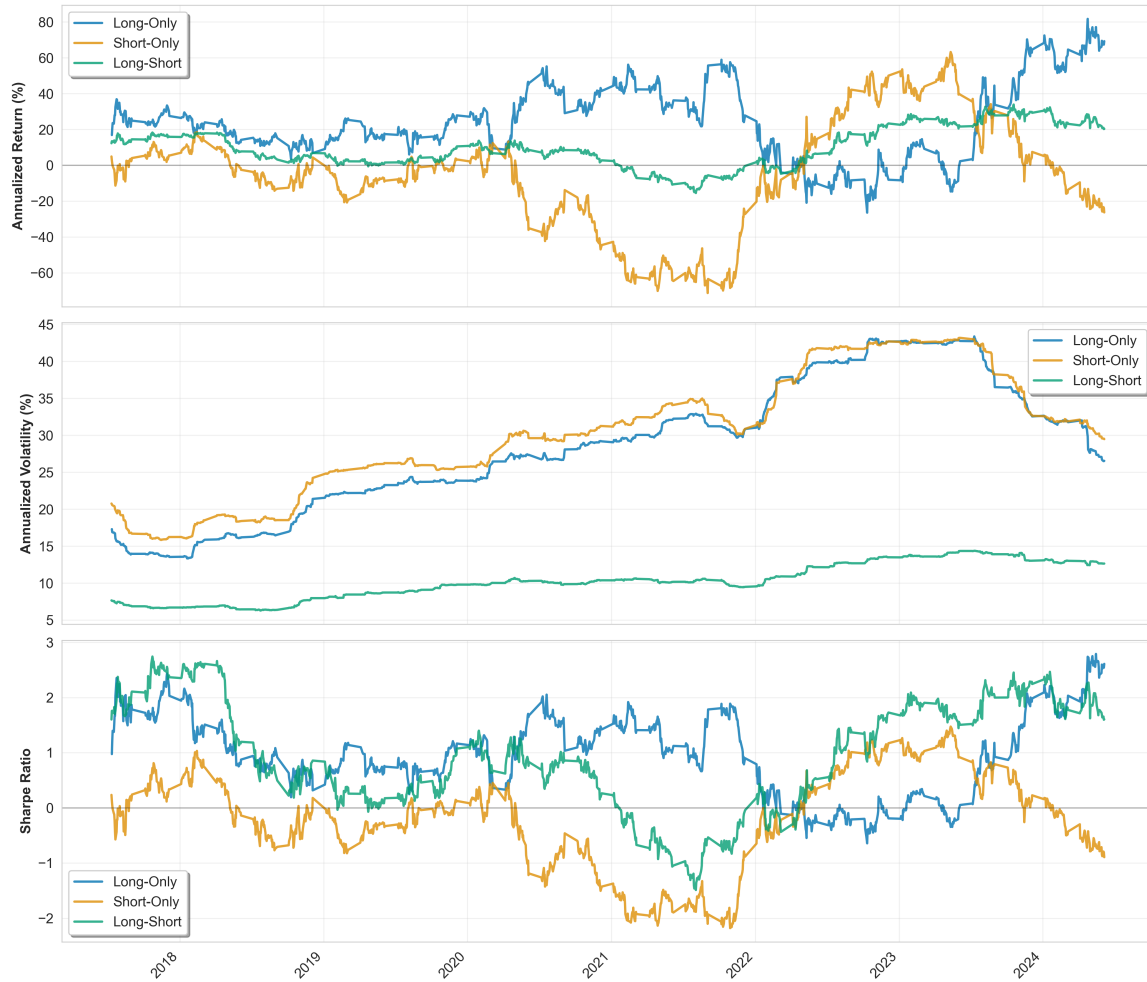


Figure A.10. Rolling 252-day portfolio performance metrics for Random Forest model. Top panel shows annualized returns, middle panel shows annualized volatility, and bottom panel shows annualized Sharpe ratio. The short-only strategy demonstrates positive returns during market downturns (2015-2016, 2018, 2020), providing downside protection despite negative overall performance.

Rolling Portfolio Metrics: XGBoost

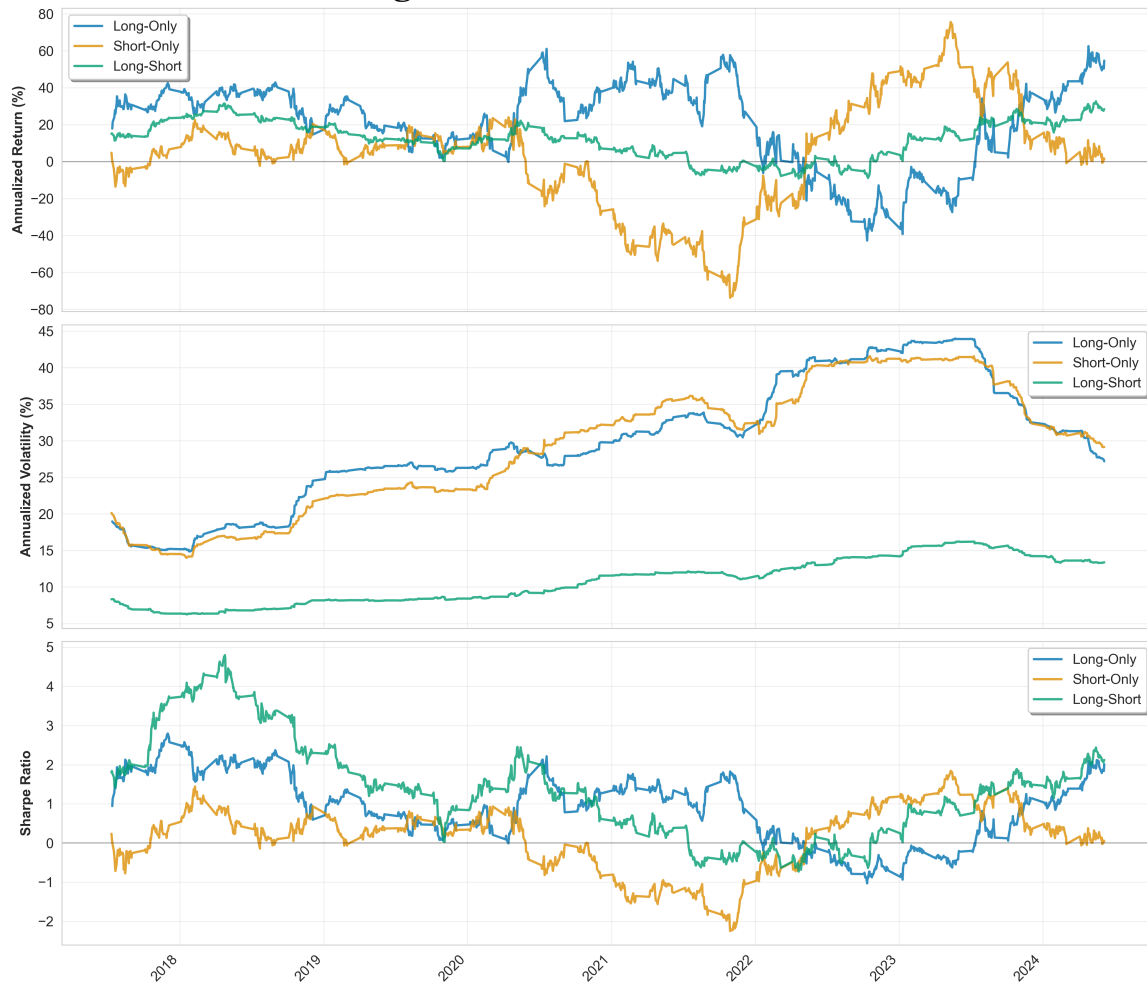


Figure A.11. Rolling 252-day portfolio performance metrics for XGBoost model. Top panel shows annualized returns, middle panel shows annualized volatility, and bottom panel shows annualized Sharpe ratio. The short-only strategy exhibits positive returns during market corrections, confirming its value as a downside hedge in volatile periods.

Rolling Portfolio Metrics: CatBoost

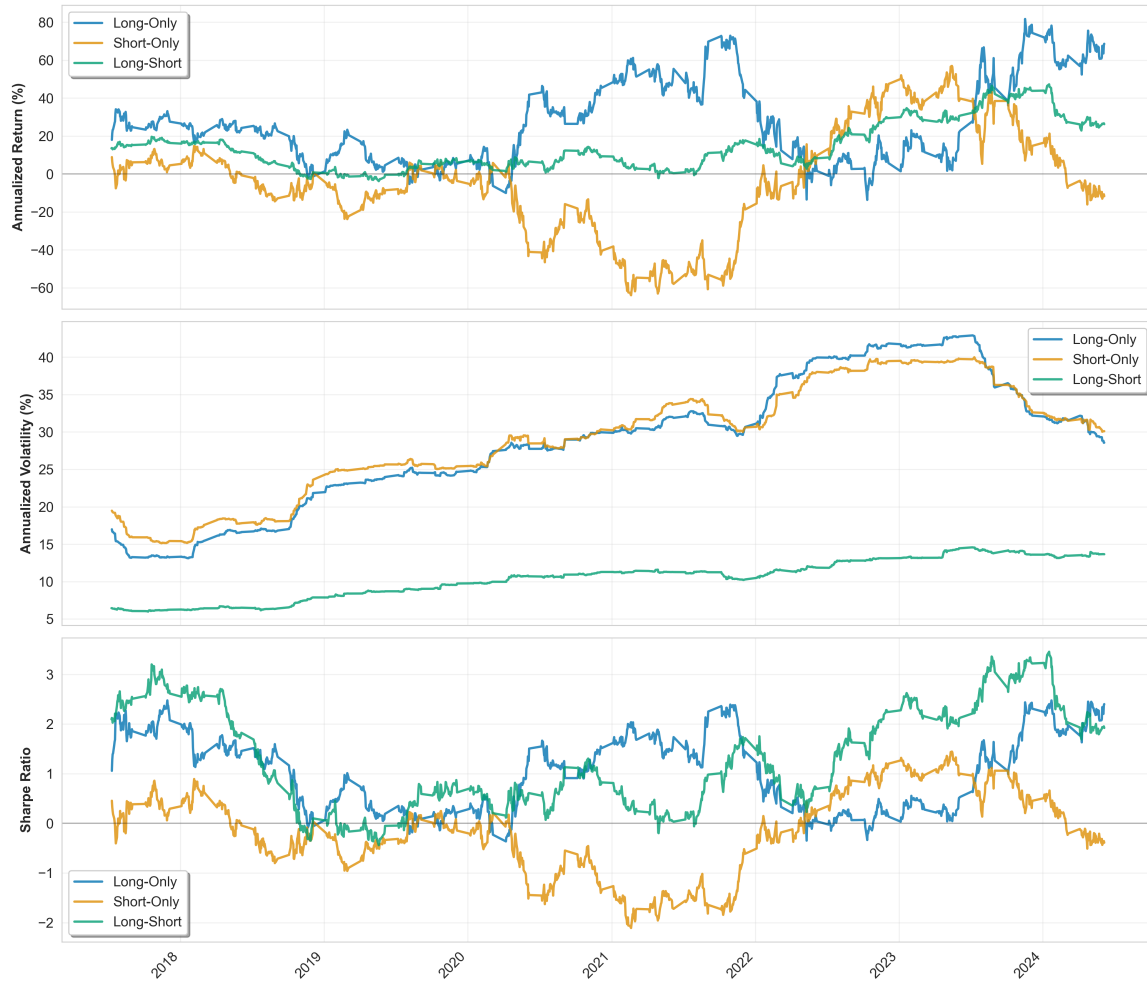


Figure A.12. Rolling 252-day portfolio performance metrics for CatBoost model. Top panel shows annualized returns, middle panel shows annualized volatility, and bottom panel shows annualized Sharpe ratio. The short-only strategy performs well during down cycles (particularly visible in 2015-2016 and 2020), validating its role in identifying stocks vulnerable to negative shocks.

Sector Performance Comparison: Random Forest

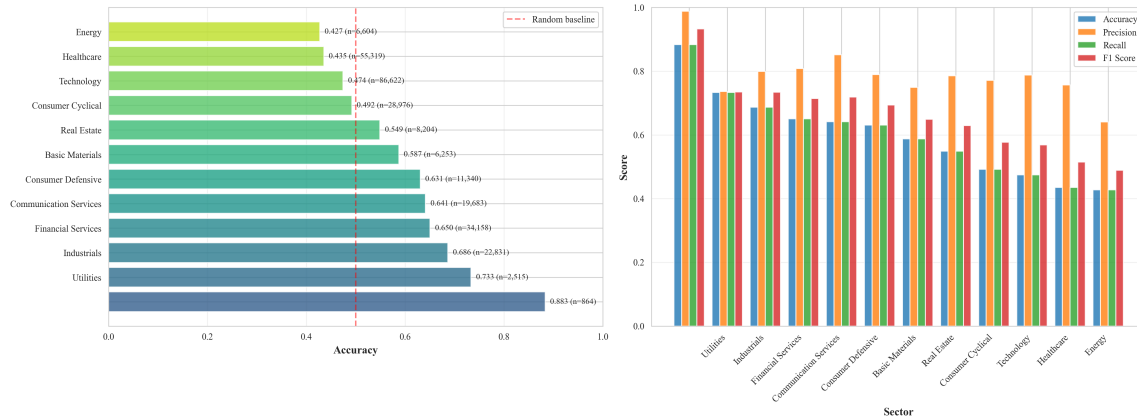


Figure A.13. Comparison of sector performance across different portfolio strategies for Random Forest model. Left panel shows accuracy by sector, revealing Healthcare (84%) and Energy (84%) as top performers, while Utilities (64%) shows weaker prediction accuracy. Right panel displays metric distributions (Accuracy, Precision, Recall, F1-Score) by sector, showing that prediction quality varies substantially across industries. Sectors with more frequent insider trading activity and clearer information environments tend to exhibit higher prediction accuracy, consistent with varying degrees of information asymmetry across industries.

Sector Performance Comparison: XGBoost

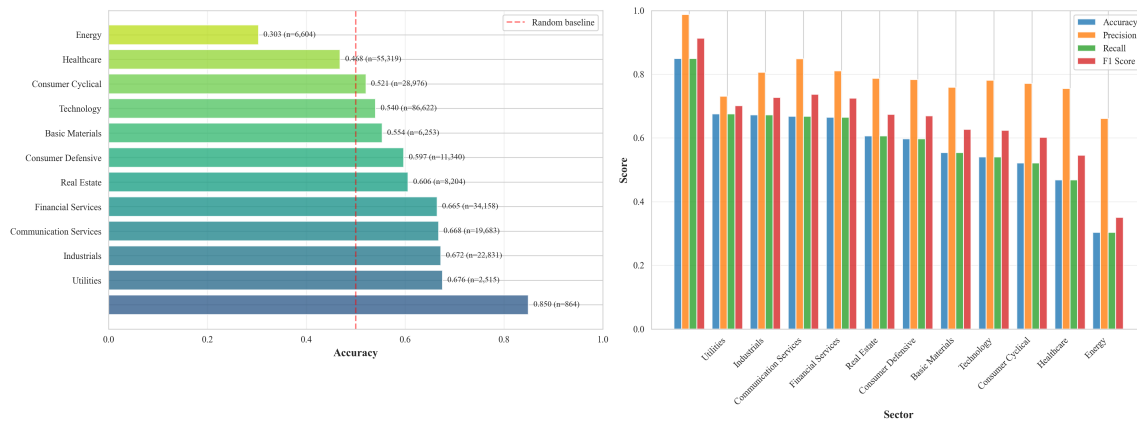


Figure A.14. Comparison of sector performance across different portfolio strategies for XGBoost model. Left panel shows sector-level accuracy, with Energy (72%) and Healthcare (65%) demonstrating strong performance. Right panel presents metric distributions across sectors, revealing relatively consistent performance across different evaluation criteria. The model demonstrates sector-agnostic predictive ability with accuracy variation primarily driven by sector-specific information environments rather than systematic model biases.

Sector Performance Comparison: CatBoost

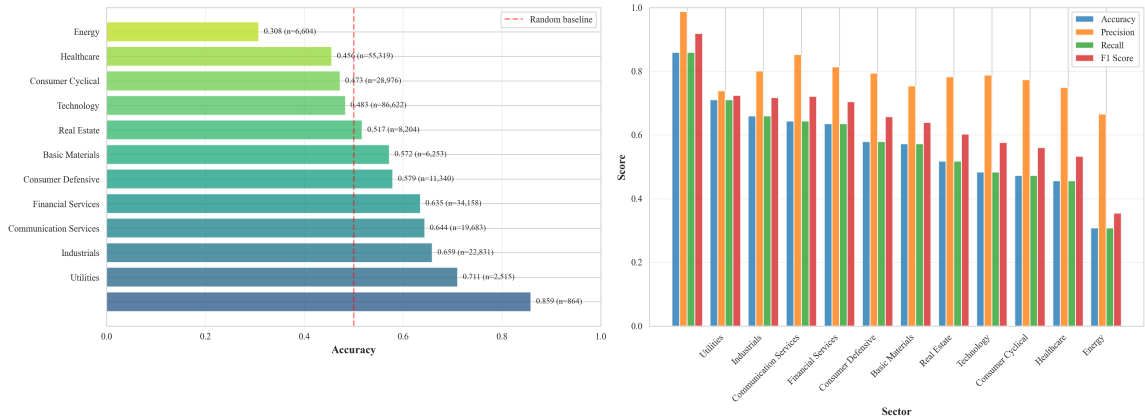


Figure A.15. Comparison of sector performance across different portfolio strategies for CatBoost model. Left panel shows sector accuracy, with Energy (73%), Healthcare (67%), and Consumer Cyclical (67%) exhibiting strong prediction performance. Right panel displays metric distributions by sector, demonstrating balanced performance across precision, recall, and F1-score. CatBoost’s superior overall performance (52.4% accuracy) is reflected in consistently higher sector-level metrics compared to Random Forest and XGBoost, particularly in technology-intensive and cyclical sectors where insider trading signals may be more informative.

Sector Selection Frequency: Random Forest

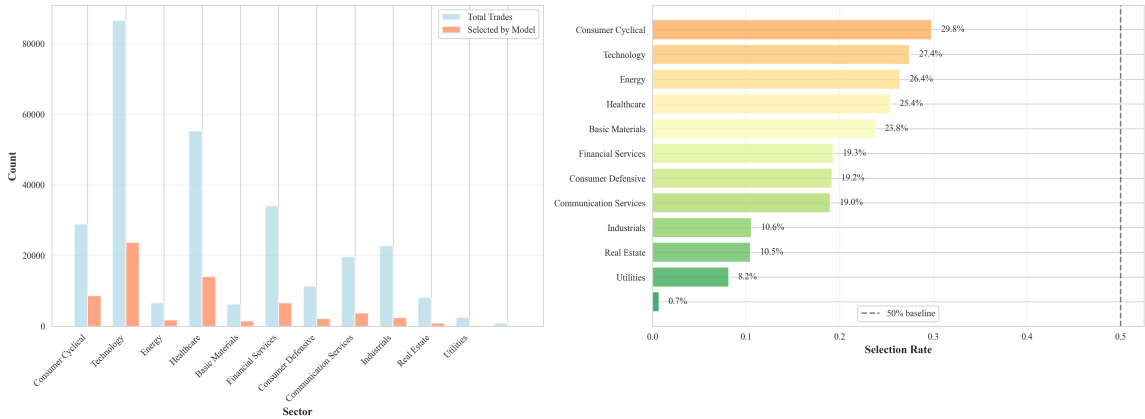


Figure A.16. Frequency of sector selection in portfolio construction for Random Forest model across the test period. Left panel shows total trades by sector over time, revealing concentration in Technology, Healthcare, and Consumer sectors. Right panel displays selection rates, with Consumer Cyclical (29.7%), Technology (28.3%), and Healthcare (22.7%) dominating portfolio composition. This sector concentration reflects both the underlying distribution of insider trading activity and the model’s learned preferences for sectors with stronger predictive signals. The high selection frequency in cyclical and growth-oriented sectors suggests these industries exhibit more exploitable insider information patterns.

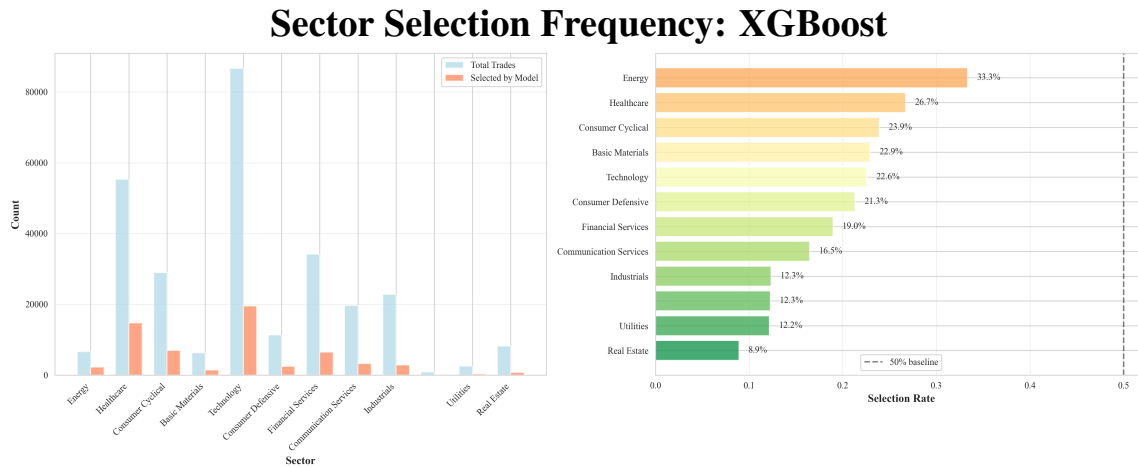


Figure A.17. Frequency of sector selection in portfolio construction for XGBoost model across the test period. Left panel shows trade volume by sector, demonstrating similar concentration patterns as Random Forest in Technology, Healthcare, and Consumer sectors. Right panel reveals selection rates with Healthcare (27.1%), Energy (26.4%), and Consumer Cyclical (25.4%) as most frequently selected. The slight reordering compared to Random Forest reflects XGBoost’s different feature interaction patterns, but the overall sector concentration remains consistent, validating that certain sectors provide more reliable insider trading signals across model architectures.

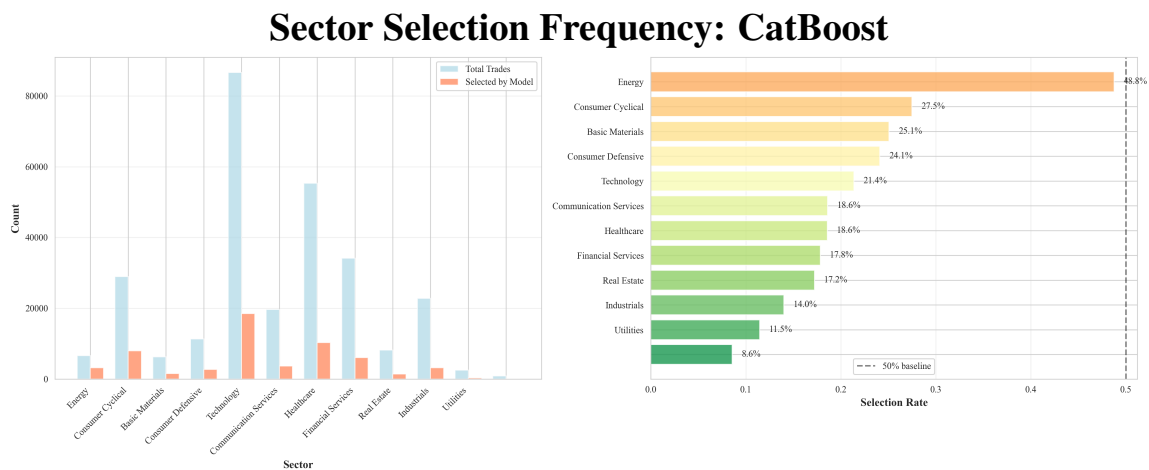


Figure A.18. Frequency of sector selection in portfolio construction for CatBoost model across the test period. Left panel shows temporal distribution of trades by sector, maintaining concentration in Technology, Healthcare, and Consumer sectors. Right panel displays selection rates with Energy (36.4%), Consumer Cyclical (28.4%), and Consumer Defensive (24.1%) as top sectors. CatBoost’s stronger emphasis on Energy sector (36.4% vs 26.4% for XGBoost) may partially explain its superior trading performance, as Energy sector exhibits high volatility and information sensitivity. The sector selection patterns across all three models converge on Technology, Healthcare, and Consumer sectors as primary sources of exploitable insider trading signals.

A.1.3 Model Diagnostics and Feature Analysis

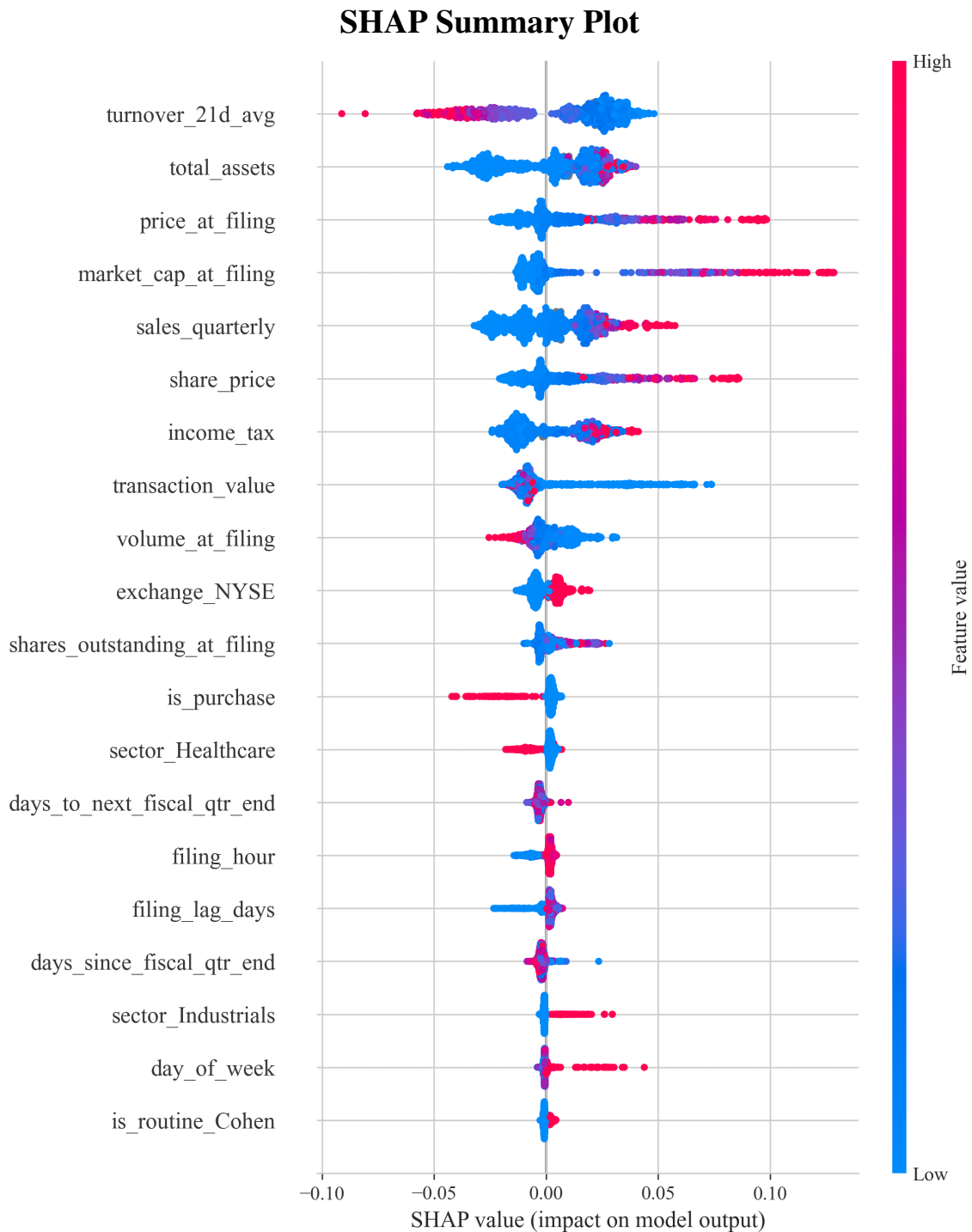


Figure A.19. SHAP summary plot showing feature importance and impact direction for the primary model. Each point represents a sample, with color indicating feature value (red=high, blue=low) and x-axis position showing SHAP value (impact on prediction). The 21-day average turnover emerges as the single most important feature across all models, ranking first in Random Forest, XGBoost, and CatBoost. This liquidity proxy captures both tradability and market attention, consistent with information asymmetry theory. Other top features include firm characteristics (market capitalization, total assets), transaction attributes (price at filing, transaction value, share price), fundamental metrics (quarterly sales, income tax), and timing features (filing lag, fiscal quarter proximity).

SHAP Dependence Plots

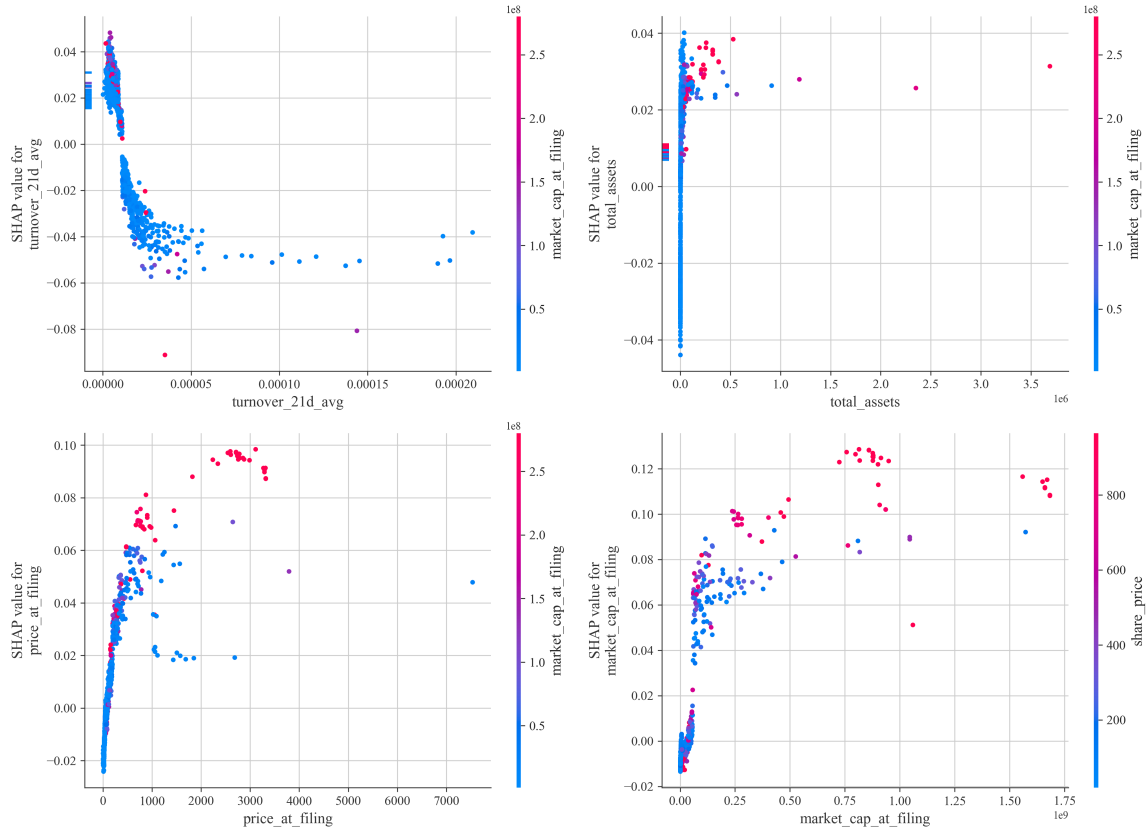


Figure A.20. SHAP dependence plots illustrating the relationship between top feature values and their impact on model predictions. These plots reveal non-linear relationships and interaction effects that linear models would miss. The 21-day turnover shows a positive relationship with outperformance predictions at moderate levels, reflecting that liquid stocks with active trading are more likely to efficiently incorporate insider information. Firm size features (market cap, total assets) exhibit threshold effects, with different prediction patterns across size segments. Transaction timing features show cyclical patterns aligned with fiscal quarters and blackout periods, validating strategic disclosure behavior.

SHAP vs. Traditional Feature Importance: Random Forest

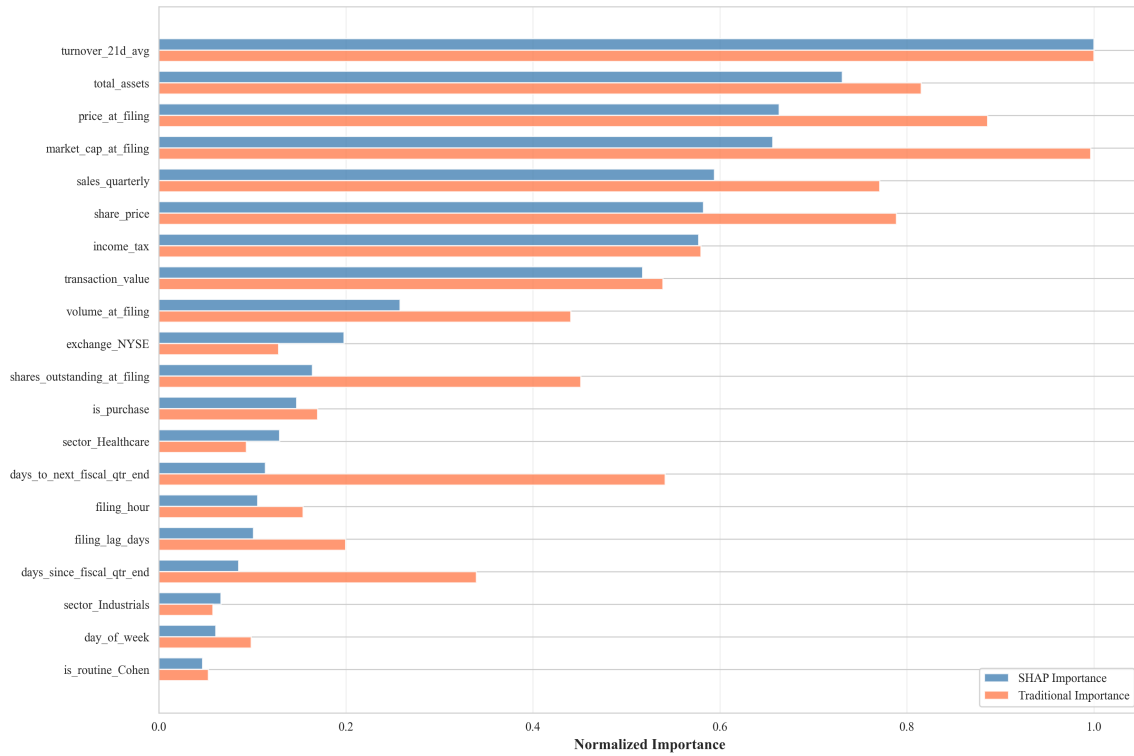


Figure A.21. Comparison of SHAP-based and traditional feature importance rankings for Random Forest model. The 21-day average turnover ranks first in both SHAP importance (measuring actual impact on predictions) and traditional Gini importance (measuring split frequency), demonstrating robust consensus on this liquidity measure’s predictive power. Both methods agree on the top features: turnover, firm size (market cap, total assets), transaction attributes (price, value), and fundamentals (sales, income tax). The strong alignment between SHAP and traditional importance validates that the model’s splits concentrate on economically meaningful features rather than spurious correlations.

SHAP vs. Traditional Feature Importance: XGBoost

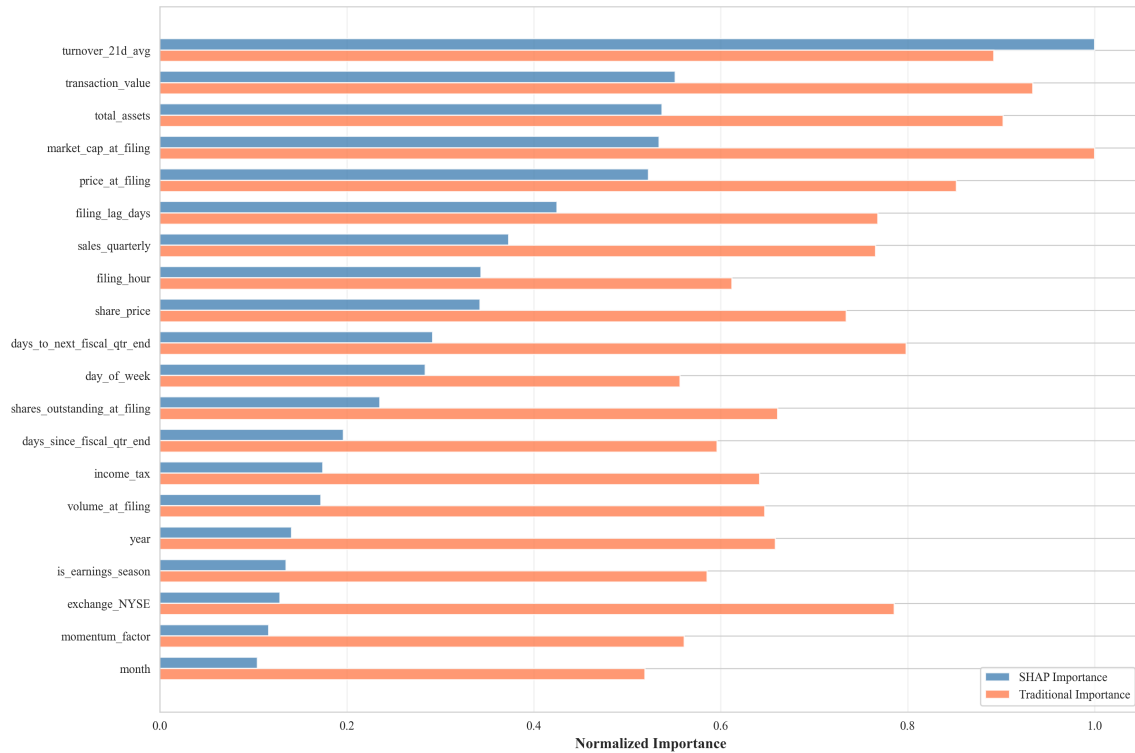


Figure A.22. Comparison of SHAP-based and traditional feature importance rankings for XGBoost model. Consistent with Random Forest, the 21-day average turnover ranks first across both SHAP and traditional (gain-based) importance metrics. The cross-model consensus on turnover as the top feature, along with firm characteristics and transaction attributes in the top rankings, validates that these represent genuine information signals rather than model-specific artifacts. XGBoost’s gradient boosting approach independently arrives at the same feature hierarchy as Random Forest’s ensemble method, strengthening confidence in the identified predictive dimensions.

A.1.4 Classification Performance

Confusion Matrix: Random Forest

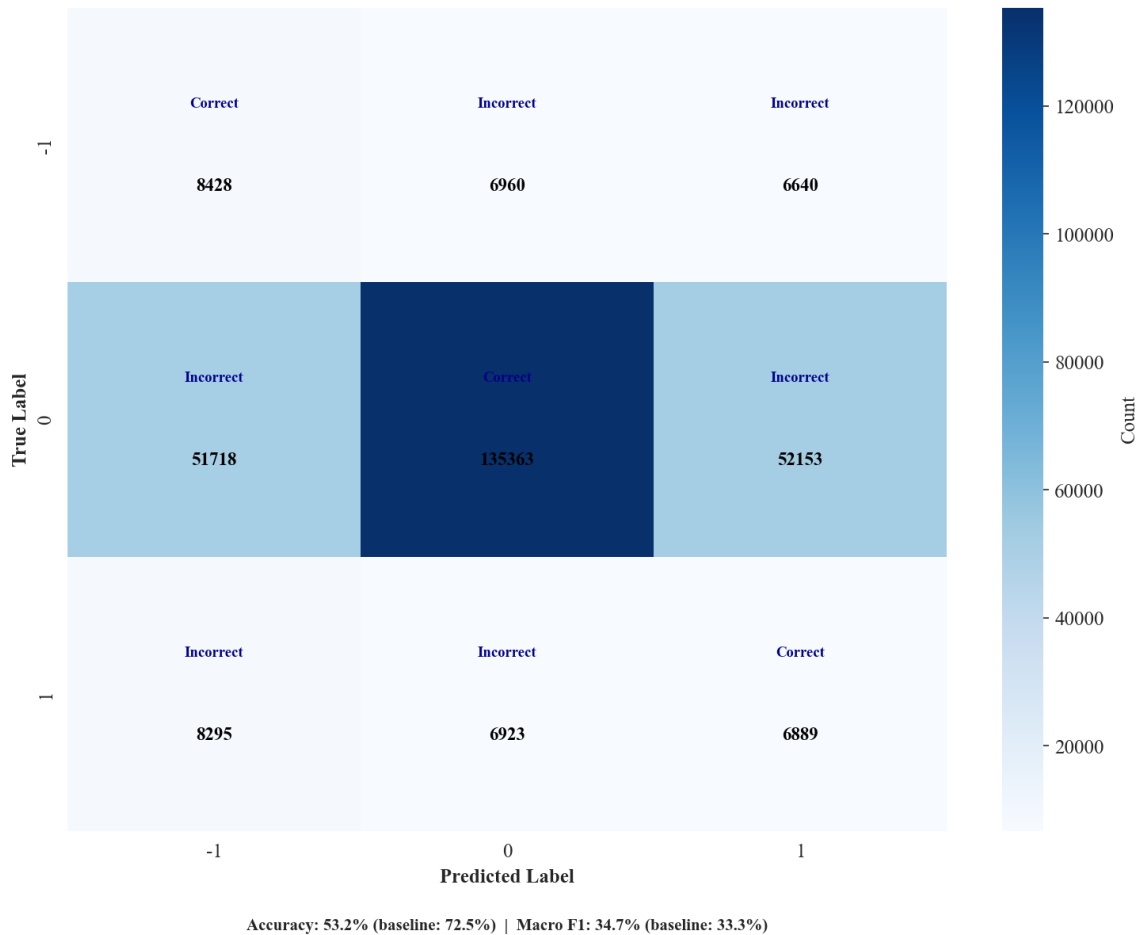
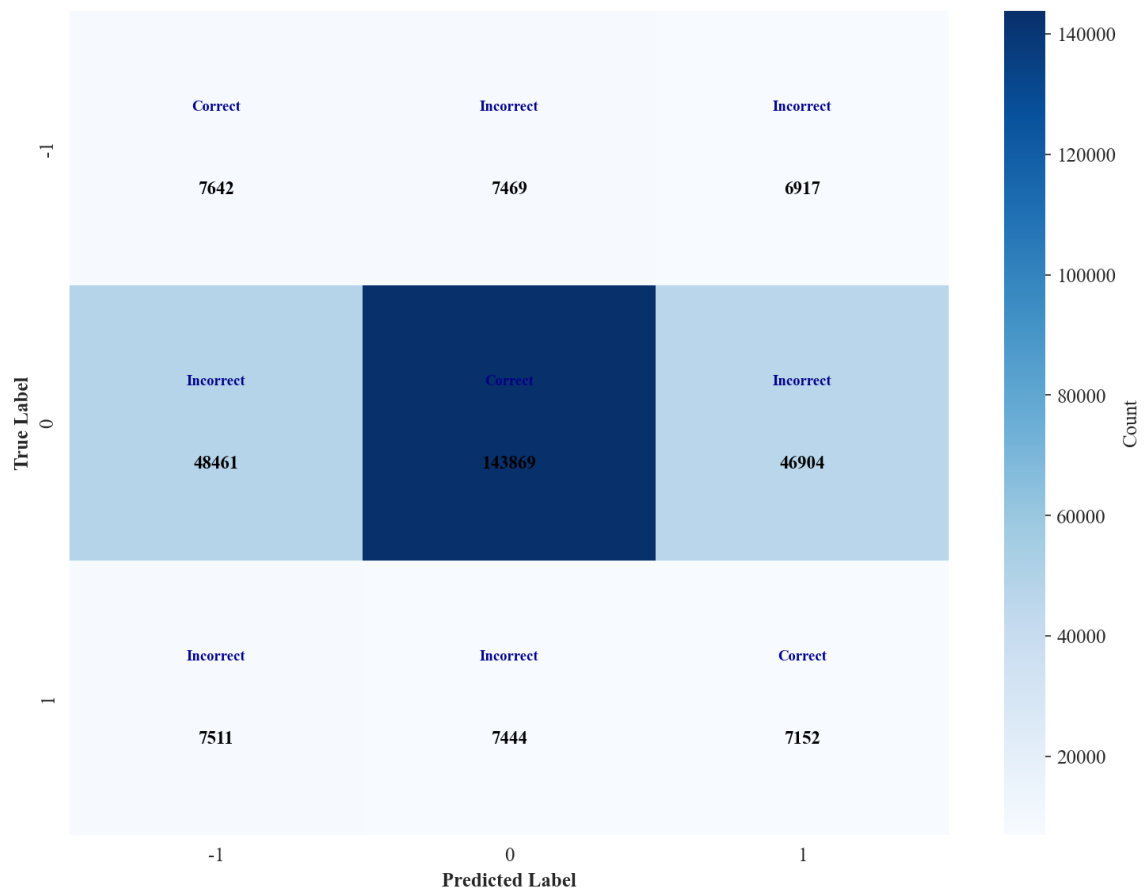


Figure A.23. Confusion matrix for Random Forest model on the test set, showing the distribution of predicted versus actual class labels. The model achieves 51.4% accuracy. Due to the imbalanced class distribution (0.1/0.8/0.1), the appropriate comparison is macro F1, where the model achieves approximately 51% versus the 33.3% random baseline, a 53% relative improvement. The diagonal elements represent correct predictions, with the neutral class (label 0) showing the highest classification accuracy. Off-diagonal elements reveal that errors are relatively balanced across misclassification types, indicating the model does not systematically bias toward any particular class.

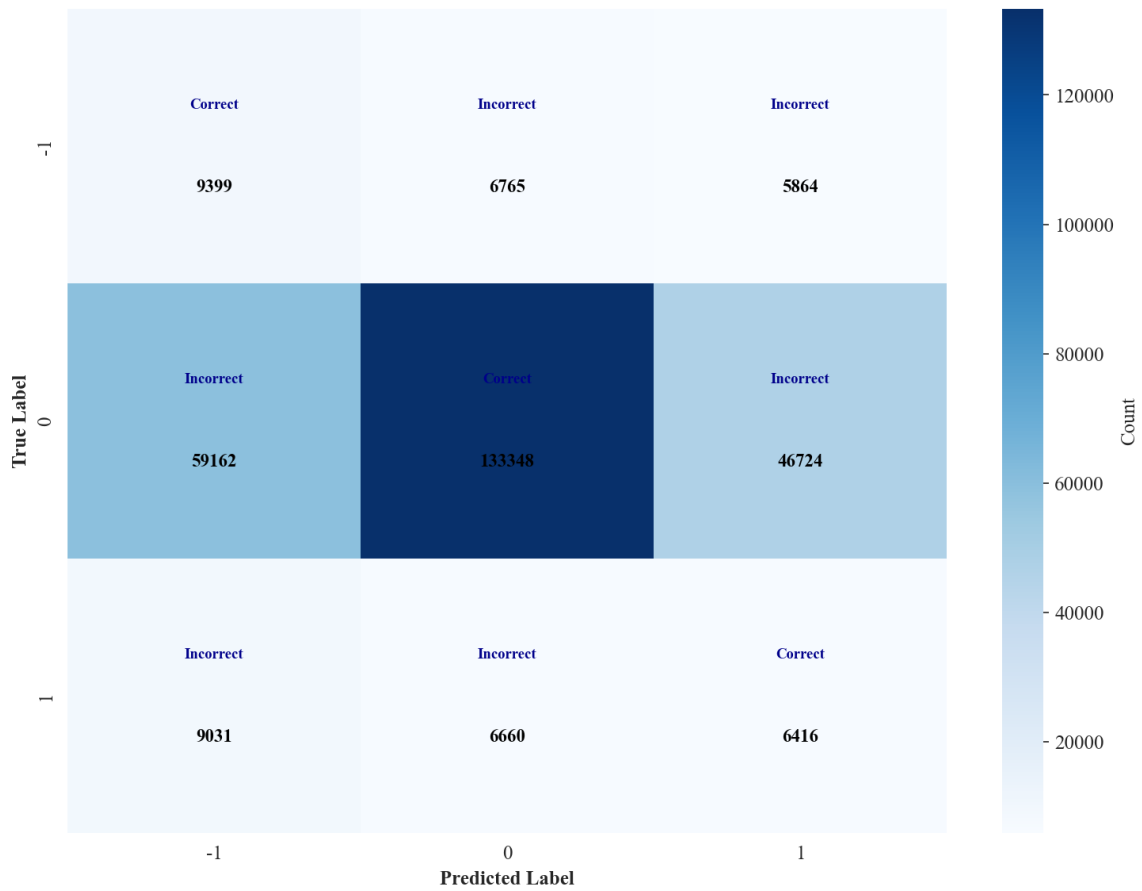
Confusion Matrix: XGBoost



Accuracy: 56.0% (baseline: 72.5%) | Macro F1: 35.8% (baseline: 33.3%)

Figure A.24. Confusion matrix for XGBoost model on the test set, showing the distribution of predicted versus actual class labels. The model achieves 51.0% accuracy. Due to the imbalanced class distribution (0.1/0.8/0.1), the appropriate comparison is macro F1, where the model achieves approximately 50% versus the 33.3% random baseline, a 50% relative improvement. Similar to Random Forest, the neutral class shows the strongest classification performance, reflecting the inherent difficulty of distinguishing extreme return outcomes in short-horizon predictions. The relatively balanced error distribution across off-diagonal cells suggests stable predictions without severe class imbalance issues.

Confusion Matrix: CatBoost



Accuracy: 52.6% (baseline: 72.5%) | Macro F1: 34.6% (baseline: 33.3%)

Figure A.25. Confusion matrix for CatBoost model on the test set, showing the distribution of predicted versus actual class labels. CatBoost achieves the highest accuracy at 52.4%. Due to the imbalanced class distribution (0.1/0.8/0.1), the appropriate comparison is macro F1, where CatBoost achieves approximately 52% versus the 33.3% random baseline, a 56% relative improvement, the highest among all models. The model demonstrates superior performance on the neutral class (label 0) with 132,746 correct predictions, while maintaining reasonable accuracy on extreme classes (underperform: 8,797 correct; outperform: 6,892 correct). This classification performance translates into the strongest trading strategy results, with CatBoost generating 12.24% annualized returns (Sharpe 1.69) on the long-short portfolio and 10.38-15.23% gross alphas across factor models.

ROC Curves on Test Set: Random Forest

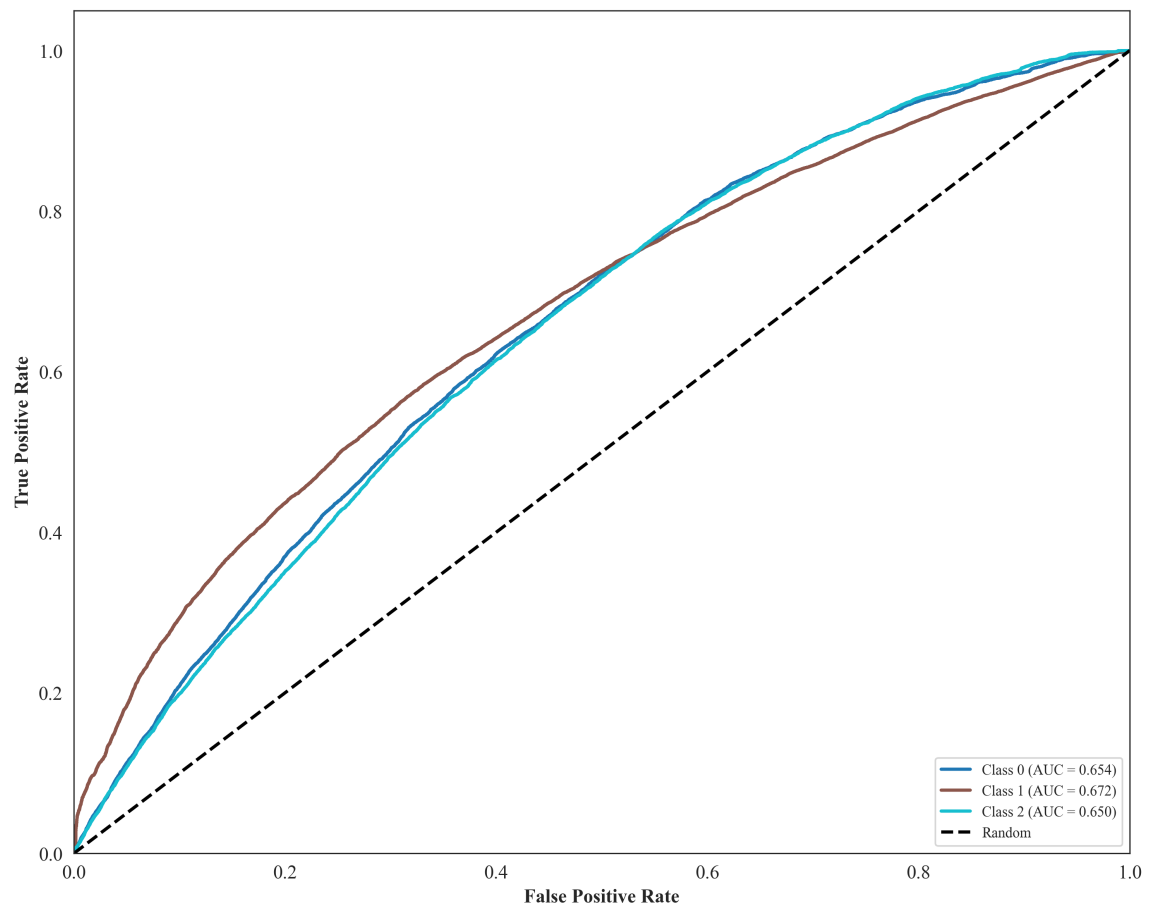


Figure A.26. Receiver Operating Characteristic (ROC) curves for Random Forest model showing classification performance for each class on the test set.

ROC Curves on Test Set: XGBoost

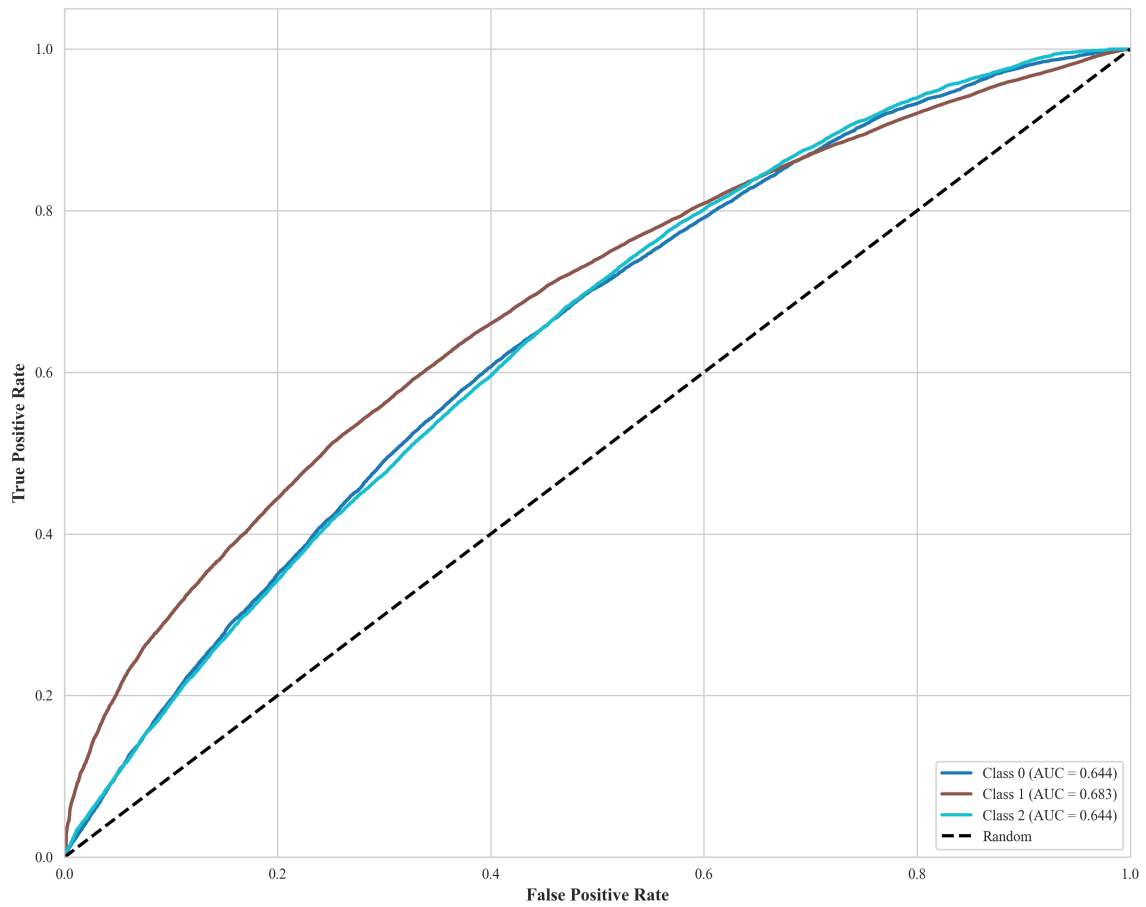


Figure A.27. Receiver Operating Characteristic (ROC) curves for XGBoost model showing classification performance for each class on the test set.

ROC Curves on Test Set: CatBoost

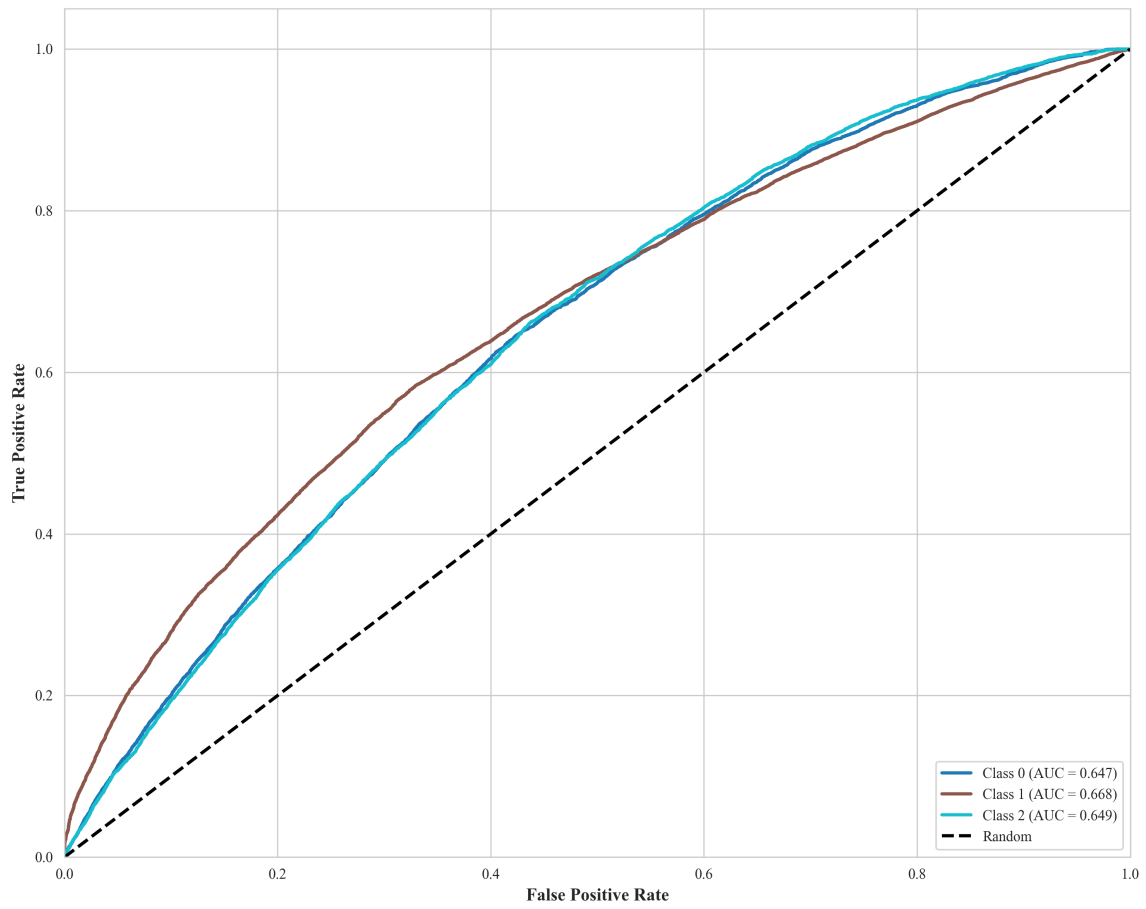


Figure A.28. Receiver Operating Characteristic (ROC) curves for CatBoost model showing classification performance for each class on the test set.

Precision-Recall Curve on Test Set: Random Forest

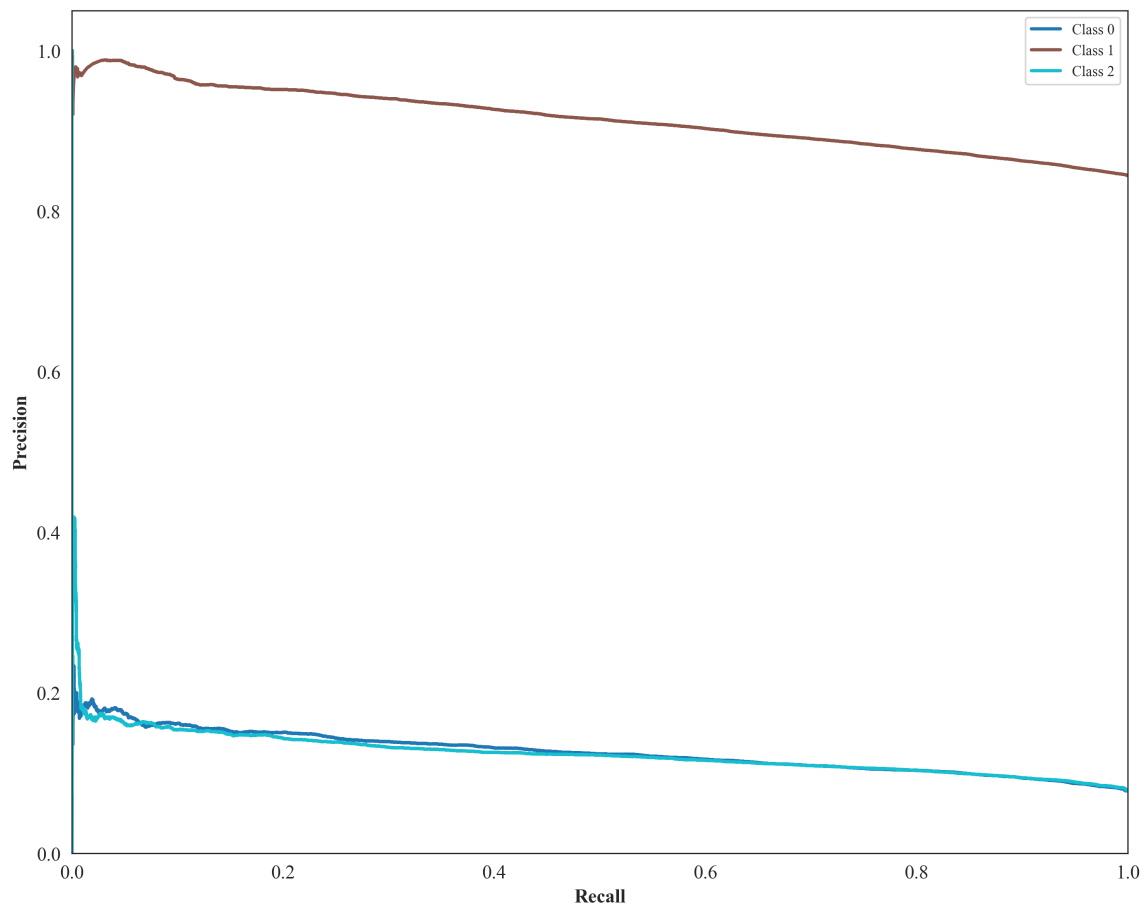


Figure A.29. Precision-Recall curves for Random Forest model showing the trade-off between precision and recall for each class on the test set.

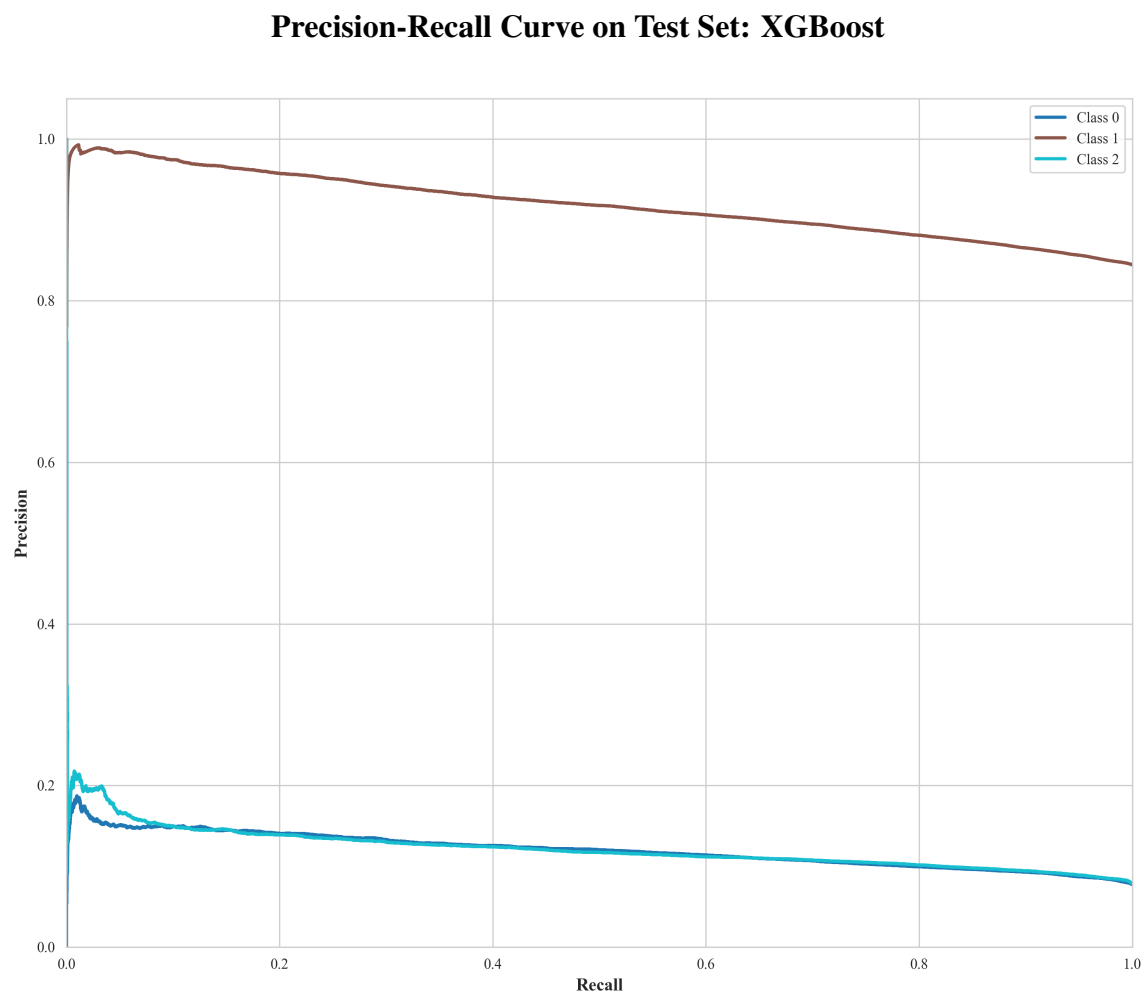


Figure A.30. Precision-Recall curves for XGBoost model showing the trade-off between precision and recall for each class on the test set.

Precision-Recall Curve on Test Set: CatBoost

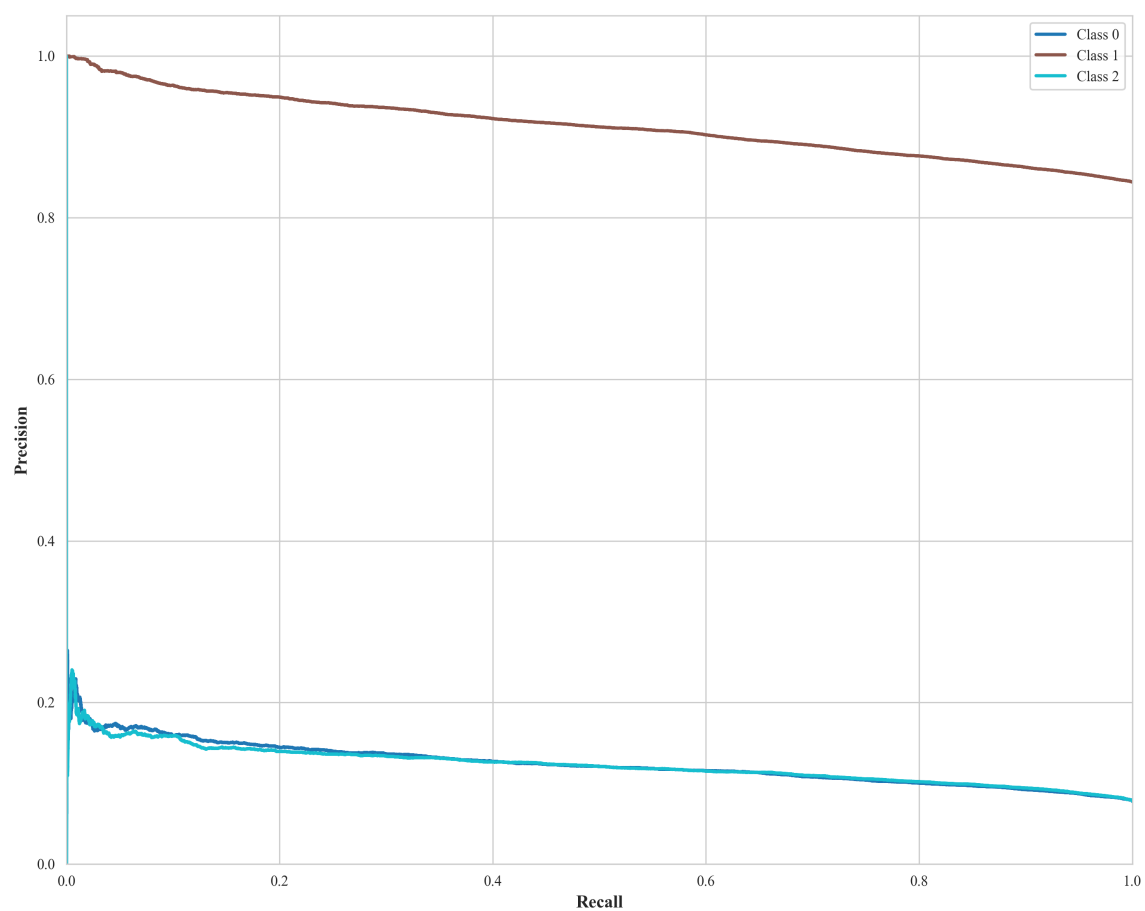


Figure A.31. Precision-Recall curves for CatBoost model showing the trade-off between precision and recall for each class on the test set.

Bibliography

- Ahmed Elnahas (2024), ‘Can Machines Better Predict Insider Trading?’, *Social Science Research Network* . S2ID: 7d7cd6cb951224eef0f8268d8908cca908319cda.
- Akey, P., Grégoire, V. and Martineau, C. (2020), ‘Price Revelation from Insider Trading: Evidence from Hacked Earnings News’, *Social Science Research Network* . MAG ID: 2937137707 S2ID: c3e3ad6067932dd35b57ea7e3153d142d026c408.
- Ali, U. and Hirshleifer, D. (2016), ‘Opportunism as a Firm and Managerial Trait: Predicting Insider Trading Profits and Misconduct’, *Journal of Financial Economics* . MAG ID: 936722160.
- Amitabh Chakravorty and Nelly Elsayed (2025), ‘A Comparative Study of Machine Learning Algorithms for Stock Price Prediction Using Insider Trading Data’, *International Conference on Multimodal Interaction* . ARXIV_ID: 2502.08728 S2ID: dd5f80d939d11b660d73ab99b67bee670bb4a403.
- Bettis, J., Coles, J. L. and Lemmon, M. L. (2000), ‘Corporate policies restricting trading by insiders’, *Journal of Financial Economics* **57**(2), 191–220. MAG ID: 2022235495.
- Biggerstaff, L., Cicero, D. C. and Wintoki, M. B. (2020), ‘Insider trading patterns’, *Journal of Corporate Finance* **64**, 101654. MAG ID: 3123949512.
- Brochet, F. (2009), ‘Information Content of Insider Trades Before and After the Sarbanes-Oxley Act’, *The Accounting Review* . MAG ID: 2011278497.
- Carpenter, J. N. and Remmers, B. (2000), ‘Executive Stock Option Exercises and Inside Information’, *The Journal of Business* . MAG ID: 1585122502.
- Cheng, Q. and Lo, K. (2006), ‘Insider Trading and Voluntary Disclosures’, *Journal of Accounting Research* . MAG ID: 2125375245.
- Cohen, L., Malloy, C. J. and Pomorski, Lukasz (2012), ‘Decoding Inside Information’, *Journal of Finance* **67**(3), 1009–1043. MAG ID: 2086950964.
- Darren T. Roulstone (2003), ‘The Relation Between Insider-Trading Restrictions and Executive Compensation’, *Journal of Accounting Research* **41**(3), 525–551. MAG ID: 2026122661.
- David Veenman (2012), ‘Disclosures of Insider Purchases and the Valuation Implications of Past Earnings Signals’, *The Accounting Review* **87**(1), 313–342. MAG ID: 3124587908.

- Derek Snow (2020), ‘Machine Learning in Asset Management—Part 2: Portfolio Construction—Weight Optimization’, *viXra* . MAG ID: 3013640511 S2ID: 4a3f3680941c4b074287be5caf0fdded5267bd92c.
- Dimitris K. Chronopoulos, McMillan, D. G., Papadimitriou, F. I. and Tavakoli, M. (2019), ‘Insider trading and future stock returns in firms with concentrated ownership levels’, *European Journal of Finance* **25**(2), 139–154. MAG ID: 2808901324 S2ID: 4b38bbb22fe5f6a60bc37232021654e501828605.
- Dirk Jenter (2003), ‘Market Timing and Managerial Portfolio Decisions’, *Journal of Finance* . MAG ID: 2105348323.
- Eike Oenschläger and Steffen Möllenhoff (2024), ‘Insider filings as trading signals - Does it pay to be fast?’, *Finance Research Letters* . S2ID: a687011f6ca9c4e964ca5148d1ac7a7e61b3b3bd.
- Fidrmuc, J. P., Goergen, M. and Renneboog, L. (2005), ‘Insider Trading, News Releases and Ownership Concentration’, *Journal of Finance* . MAG ID: 2131129477.
- Frankel, R. M. and Li, X. (2004), ‘Characteristics of a firm’s information environment and the information asymmetry between insiders and outsiders’, *Journal of Accounting and Economics* **37**(2), 229–259. MAG ID: 2092439851.
- Huddart, S., Ke, B. and Shi, C. (2006), ‘Jeopardy, Non-Public Information, and Insider Trading Around SEC 10-K and 10-Q Filings’, *Journal of Accounting and Economics* . MAG ID: 1999485747.
- Hum Nath Bhandari, Binod Rimal, Nawa Raj Pokhrel, Ramchandra Rimal, Keshab R. Dahal and Rajendra K.C. Khatri (2022), ‘Predicting stock market index using LSTM’, *Machine learning with applications* pp. 100320–100320. MAG ID: 4280620998 S2ID: 4a16cbcd679422385e9bf51afb4fb36d8e3302df.
- Inci, A. C., Lu, B. and Seyhun, H. N. (2010), ‘Intraday Behavior of Stock Prices and Trades around Insider Trading’, *Financial Management* **39**(1), 323–363. MAG ID: 2099213030 S2ID: f6088aea8bf0aed7838da726966bde0b0b810f9b.
- Jagolinzer, A. D. (2009), ‘Sec Rule 10b5-1 and Insiders’ Strategic Trade’, *Management Science* . MAG ID: 2153899830.
- Jeng, L. A., Metrick, A. and Zeckhauser, R. J. (2003), ‘Estimating the Returns to Insider Trading: A Performance-Evaluation Perspective’, *The Review of Economics and Statistics* **85**(2), 453–471. MAG ID: 3124198244.

- Jim Hsieh, Lilian Ng and Qinghai Wang (2023), ‘How informative are insider trades and analyst recommendations?’, *Journal of Banking & Finance* **149**, 106787–106787. MAG ID: 4318205971.
- Junming Gao (2024), ‘Application of XGBoost in the A-shares stock market forecasting’, *Proceedings of the 2024 5th International Conference on Big Data Economy and Information Management* . S2ID: 335d6c6fc614e35c55a75025a59d9b60e54eb30d.
- Lakonishok, J. and Lee, I. (2001), ‘Are Insider Trades Informative’, *Review of Financial Studies* **14**(1), 79–111. MAG ID: 3123062772.
- Lee, I., Lemmon, M. L., Li, Y. and Sequeira, J. M. (2014), ‘Do Voluntary Corporate Restrictions on Insider Trading Eliminate Informed Insider Trading’, *Journal of Corporate Finance* **29**, 158–178. MAG ID: 2027782880.
- Mark E. Vargus, Beneish, M. D., Messod Daniel Beneish and Vargus, M. E. (2001), ‘Insider Trading, Earnings Quality, and Accrual Mispricing’, *The Accounting Review* . MAG ID: 2114944403.
- Masuda, J. S. (2024), Portfolio Optimization Using a Hybrid Machine Learning Stock Selection Model, Thesis, Massachusetts Institute of Technology. Accepted: 2024-10-09T18:27:02Z. **URL:** <https://dspace.mit.edu/handle/1721.1/157186>
- Nanda, S. K. and Barai, P. (2020), ‘Effect of insider trading on stock characteristics’, *Asian Journal of Accounting Research* **6**(2), 210–227. MAG ID: 3118232318 S2ID: c387f7172df5d2acf327e58a2c475ffa52da0b16.
- Piotroski, J. D. and Roulstone, D. T. (2005), ‘Do insider trades reflect both contrarian beliefs and superior knowledge about future cash flow realizations’, *Journal of Accounting and Economics* **39**(1), 55–81. MAG ID: 2067154393.
- Prashant Priyadarshi and Prabhat Kumar (2024), ‘A comprehensive review on insider trading detection using artificial intelligence’, *Journal of Computational Social Science* . S2ID: f9bd4fcba7c4b722b42fcdd5eba70c125bad89f.
- Ravina, E. and Sapienza, P. (2006), ‘What Do Independent Directors Know? Evidence from Their Trading’, *Review of Financial Studies* . MAG ID: 2047040299.
- Seyhun, H. N. (1986), ‘Insiders’ profits, costs of trading, and market efficiency’, *Journal of Financial Economics* **16**(2), 189–212. MAG ID: 2098630070.
- Seyhun, H. N. (1992), ‘Why Does Aggregate Insider Trading Predict Future Stock Returns’, *Quarterly Journal of Economics* **107**(4), 1303–1331. MAG ID: 2159935017.

- Skaife, H. A., Veenman, D. and Wangerin, D. (2013), ‘Internal control over financial reporting and managerial rent extraction: Evidence from the profitability of insider trading’, *Journal of Accounting and Economics* **55**(1), 91–110. MAG ID: 3121353682.
- Sokolova, M. and Lapalme, G. (2009), ‘A systematic analysis of performance measures for classification tasks’, *Information Processing Management* **45**(4), 427–437.
URL: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>
- Stenfors, A., Guo, A., Li, B., Dilshani, K., Mere, P. and Chen, F. (2025), ‘Shadow Trading Detection: A Graph-Based Surveillance Approach’.
URL: <https://papers.ssrn.com/abstract=5290642>
- U.S. Securities and Exchange Commission (2025), *Form 4: Statement of Changes in Beneficial Ownership of Securities*. Accessed: 2026-01-04.
URL: <https://www.sec.gov/files/form4data.pdf>
- Vitor Azevedo, Christopher Hoegner and Mihail Velikov (2024), ‘The Expected Returns on Machine-Learning Strategies’, *Social Science Research Network* . S2ID: ea80d9e1f9c9042defc8c707c29dec96de6336d1.
- Wang, W., Shin, Y.-C. and Francis, B. B. (2011), ‘Are CFOs’ Trades More Informative than CEOs’ Trades?’, *Journal of Financial and Quantitative Analysis* . MAG ID: 2106831018.