

HEC MONTRÉAL

**Predicting Stock Returns :
Using PCA of IV Surfaces**

par

Parsa Moaven

**Vincent Grégoire
HEC Montréal
Directeur de recherche**

**Sciences de la gestion
(Spécialisation Finance)**

*Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)*

December 2025
© Parsa Moaven, 2025

Résumé

Cette thèse examine si les surfaces de volatilité implicite contiennent des informations utiles pour prédire les rendements hebdomadaires des actions. En utilisant des actions américaines à grande capitalisation sur une décennie, j'extrais les composantes principales des surfaces de volatilité implicite et j'emploie la régression elastic net dans un cadre de fenêtres glissantes pour prévoir les rendements.

Le résultat central est une divergence entre la précision statistique et la valeur économique. Tous les modèles présentent une précision prédictive hors échantillon négative, mais génèrent des portefeuilles long-short rentables. Le modèle avec la pire précision ponctuelle obtient la meilleure performance de portefeuille, démontrant que la capacité de classement importe plus que la précision ponctuelle. Elastic net surpasse nettement Random Forest, qui souffre de surapprentissage. Cependant, les coûts de transaction érodent les rendements, limitant la viabilité aux implémentations à faible coût.

Mots-clés

Volatilité implicite, prévisibilité des rendements, analyse en composantes principales, elastic net, régularisation, portefeuille long-short, coefficient d'information, évaluation transversale des actifs

Méthodes de recherche

Analyse en composantes principales, régression elastic net, estimation par fenêtres glissantes, construction de portefeuille

Abstract

This thesis examines whether implied volatility surfaces contain information useful for predicting weekly stock returns. Using large-cap U.S. stocks over a decade, I extract principal components from implied volatility surfaces and employ elastic net regression within a rolling window framework to forecast returns.

The central finding is a divergence between statistical accuracy and economic value. All models exhibit negative out-of-sample predictive accuracy, yet achieve positive information coefficients and generate profitable long-short portfolios. The model with the worst point prediction accuracy achieves the best portfolio performance, demonstrating that ranking ability matters more than point accuracy for security selection. Elastic net substantially outperforms Random Forest, which suffers severe overfitting. However, transaction costs substantially erode returns, limiting viability to low-cost implementations.

Keywords

Implied volatility, return predictability, principal component analysis, elastic net, regularization, long-short portfolio, information coefficient, cross-sectional asset pricing

Research Methods

Principal component analysis, elastic net regression, rolling window estimation, portfolio construction

Contents

Résumé	i
Abstract	iii
List of Tables	ix
List of Figures	xi
List of Acronyms	xiii
Acknowledgements	xv
Introduction	1
Literature Review	7
1 Data	17
1.1 Sample Selection and Data Sources	17
1.2 Notation	18
1.3 Equity Return Construction	19
1.4 Implied Volatility Surface Construction	19
1.4.1 Raw Option Quotes and Filters	19
1.4.2 Interpolation to a Standard Grid	20
1.4.3 Weekly Aggregation	21
1.4.4 De-meaning the Surface	22

1.5	Merged Panel and Coverage Filters	22
1.6	Discussion	23
2	Methodology	25
2.1	Overview of the Empirical Framework	25
2.2	Dimensionality Reduction: Rolling PCA	27
2.2.1	Handling Missing Values	28
2.2.2	Standardization and PCA Estimation	29
2.2.3	Selecting the Number of Components	30
2.3	Elastic Net Return Prediction	31
2.3.1	The Elastic Net Objective	31
2.3.2	Stock-by-Stock, Window-by-Window Estimation	32
2.3.3	Hyperparameter Selection	32
2.4	Evaluation Metrics	33
2.4.1	Out-of-Sample R^2	33
2.4.2	Information Coefficient	34
2.4.3	Directional Accuracy	34
2.4.4	Cumulative SSE Difference	34
2.5	Alternative Model: Random Forest	35
3	Results	37
3.1	Model Comparison	37
3.2	Feature Set Comparison	39
3.2.1	Candidate Model Specifications	40
3.2.2	Point Prediction versus Ranking Metrics	40
3.2.3	The Accuracy-Ranking Divergence	41
3.2.4	Interaction Terms	42
3.2.5	Model Selection for Portfolio Construction	43
3.3	Coefficient Distributions	44
3.3.1	Summary Statistics	44

3.3.2	Distributional Shape	45
3.3.3	Implications for Interpretation	48
3.4	Model Performance Across Market Regimes	48
3.4.1	Cumulative SSE Difference	48
3.4.2	Pre-COVID Performance	50
3.4.3	Reconciling Point Prediction and Ranking Performance	51
4	Portfolio Construction and Economic Value	53
4.1	Portfolio Construction Methodology	53
4.1.1	Long-Short Strategy Design	53
4.1.2	Performance Metrics	54
4.1.3	Transaction Cost Assumptions	55
4.2	Model Comparison and Selection	55
4.2.1	Candidate Models	55
4.2.2	Portfolio Performance Comparison	56
4.2.3	The Accuracy-Ranking Paradox	57
4.2.4	Model Selection	58
4.3	Main Results	59
4.3.1	Performance Summary	59
4.3.2	Cumulative Wealth	61
4.3.3	Transaction Cost Sensitivity	63
4.4	Performance Across Time	65
4.4.1	Market Regime Analysis	65
4.4.2	Annual Performance	66
4.4.3	Summary	68
	Conclusion	71
	Bibliography	73

Appendix A – Declaration of Generative AI Assistance	i
Appendix B – Additional Figures	iii

List of Tables

1.1	Sample Summary Statistics	23
2.1	Distribution of Cross-Validation Selected Hyperparameters	33
3.1	Out-of-Sample Performance: Model Comparison	38
3.2	Out-of-Sample Performance by Model and Feature Specification	39
3.3	Elastic Net Performance by Feature Specification	40
3.4	Long-Short Portfolio Performance by Model Specification	41
3.5	Impact of Interaction Terms on Return Prediction	42
3.6	Model Selection Summary	43
3.7	Distribution of Elastic Net Coefficients Across 319 Models	44
3.8	Distributional Shape of Elastic Net Coefficients	47
4.1	Long-Short Portfolio Performance by Model Specification	56
4.2	The Accuracy-Ranking Paradox	57
4.3	Selected Model Characteristics	59
4.4	Long-Short Portfolio Performance Summary	60
4.5	Transaction Cost Sensitivity Analysis	64
4.6	Long-Short Performance by Market Regime	65
4.7	Annual Long-Short Performance	67

List of Figures

0.1	Schematic overview of the estimation pipeline, showing the sequential stages from data processing through portfolio construction and evaluation.	3
2.1	Rolling window cross-validation with 3-year training and 1-year test windows.	27
2.2	Scree plot of variance explained by the first five principal components.	30
3.1	Kernel density estimates of elastic net coefficients for PC1, PC2, and PC3. . .	46
3.2	Q-Q plots comparing elastic net coefficient distributions to a standard normal.	47
3.3	Cumulative SSE difference between elastic net and prevailing mean benchmark.	49
3.4	Two-panel view of cumulative SSE difference with VIX overlay.	51
4.1	Growth of \$1 under different transaction cost assumptions.	62
4.2	Cumulative wealth for gross and net returns with drawdown panel.	63
4.3	Annual long-short returns (gross of transaction costs).	67
1	Variance explained in Window 1	iii

List of Acronyms

bps	Basis Points
CBOE	Chicago Board Options Exchange
COVID-19	Coronavirus Disease 2019
CV	Cross-Validation
IC	Information Coefficient
IV	Implied Volatility
KDE	Kernel Density Estimate
MAE	Mean Absolute Error
Max DD	Maximum Drawdown
MSc	Master of Science
OLS	Ordinary Least Squares
OOS	Out-of-Sample
PC	Principal Component
PCA	Principal Component Analysis
Q-Q	Quantile-Quantile

RF	Random Forest
RMSE	Root Mean Squared Error
SSE	Sum of Squared Errors
US	United States
VIX	Volatility Index

Acknowledgements

I would like to extend my heartfelt appreciation to my supervisor, Professor Vincent Grégoire. His guidance, deep expertise, and constant encouragement have played a crucial role in shaping this research and my academic growth. His mentorship has been genuinely invaluable, and I feel fortunate to have had the opportunity to work under his supervision.

I am deeply grateful to my dear family, my parents Maryam Reihani and Reza Moaven, and my brother Aria Moaven for their unwavering support and encouragement throughout my academic journey.

Finally, I would like to thank my friend and classmate Navid Namazi for the many thought-provoking discussions that helped shape this thesis.

Introduction

Predicting stock returns is a central problem in empirical finance. A large literature has proposed macroeconomic variables, firm characteristics, and statistical factors as predictors, yet most variables that look promising in-sample deliver weak or even negative performance out-of-sample. At the same time, listed options provide a rich set of forward-looking prices that reflect market expectations about volatility, downside risk, and short-selling constraints. Because option prices embed investors' views about the entire distribution of future returns, they are natural candidates for return prediction. The challenge is that the implied volatility surface is high-dimensional and highly correlated across maturities and moneyness, which makes it difficult to extract robust signals with standard linear regressions.

This thesis investigates whether option-implied information can be used to predict the cross-section of stock returns in the U.S. equity market. The analysis focuses on large-cap equities with actively traded options and asks whether a data-driven representation of the implied volatility surface can yield useful predictors for weekly returns. Rather than working with ad hoc summaries such as simple skews or call-put spreads, I construct the full implied volatility surface on a standardized grid and use principal component analysis to obtain a small number of orthogonal factors. These factors are then combined with regularized prediction models that are designed for low signal-to-noise environments.

To provide a clear overview, the empirical analysis follows a structured multi-stage estimation pipeline. First, daily option data are aggregated to a weekly frequency and used to construct implied volatility surfaces for each stock on a fixed maturity and money-

ness grid. Second, principal component analysis is applied to these surfaces using rolling training windows to extract a small number of stock-specific volatility factors. Third, the resulting factor scores are used as inputs to machine learning forecasting models to generate one-week-ahead return predictions. Finally, stocks are ranked based on predicted returns and long–short portfolios are formed and evaluated using both statistical and economic performance measures. Figure 0.1 provides a schematic representation of this full estimation pipeline.

The work builds on Muravyev, Vasquez, and Wang (2020), who show that implied volatility surfaces have a stable low-dimensional structure and that the first three principal components of the surface predict future stock returns. Their study uses a pooled sample of U.S. equities and applies principal component analysis to de-meaned implied volatility surfaces on a fixed maturity and moneyness grid. I adopt their construction of the surface but extend the methodology in several directions that are relevant for both empirical asset pricing and practical forecasting. First, I focus on the fifty largest U.S. firms by average market capitalization over 2014 to 2023 and construct weekly implied volatility surfaces for the subset of forty-six stocks with sufficient options coverage. Second, I estimate principal components separately for each stock using a rolling window structure with three-year training periods and one-year test periods. This design allows the volatility factors to adapt to local dynamics while avoiding look-ahead bias. Third, I use elastic net regression to forecast one-week-ahead stock returns from these volatility factors and lagged returns, which introduces shrinkage and variable selection into the predictive model.

A central theme of the thesis is the distinction between statistical accuracy and economic value. Out-of-sample prediction performance is evaluated using both regression-based measures, such as the predictive R^2 , and ranking-based measures, such as information coefficients and the performance of long-short portfolios constructed from model forecasts. In line with recent work on machine learning in asset pricing, the models often produce negative out-of-sample R^2 when compared to a historical mean benchmark, yet they retain the ability to rank stocks in a way that generates economically meaning-

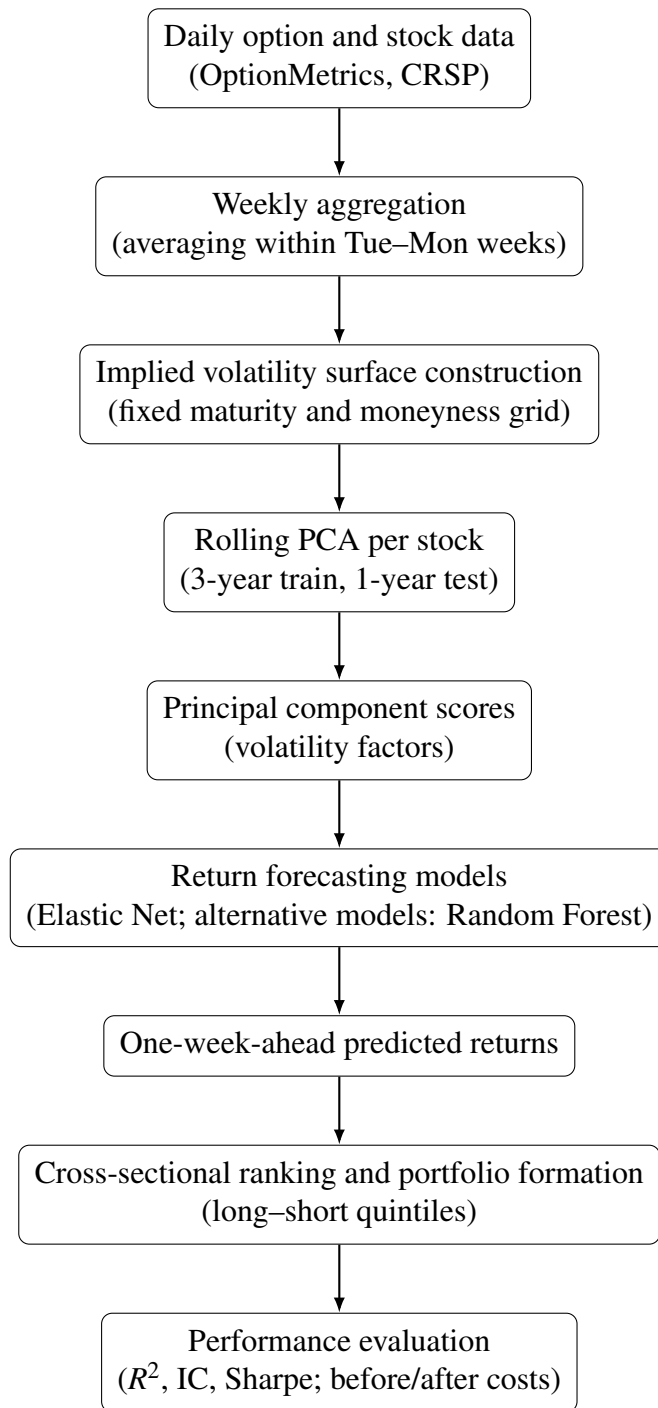


Figure 0.1: Schematic overview of the estimation pipeline, showing the sequential stages from data processing through portfolio construction and evaluation.

ful portfolio returns. The results highlight that low or negative predictive R^2 does not necessarily imply that a model is useless for security selection.

Transaction costs and implementation frictions are explicitly taken into account. The trading strategies considered in this thesis involve weekly rebalancing of long-short quintile portfolios based on forecasted returns. Such strategies are high-turnover by construction, so their profitability is very sensitive to transaction costs, bid-ask spreads, and shorting fees. I therefore combine the return forecasts with realistic assumptions about trading costs drawn from the recent literature on anomaly implementation. The net performance of the strategies is reported both before and after these costs, which provides a more realistic assessment of whether the inferred signals could be exploited by an institutional investor.

The remainder of the thesis is organized as follows. Chapter 1 describes the construction of the dataset. It explains how the sample of large-cap U.S. equities is selected, how daily OptionMetrics and CRSP data are merged, and how weekly stock returns and implied volatility surfaces are obtained. Particular attention is given to the construction of the maturity and moneyness grid, the interpolation of implied volatilities to that grid, and the filters used to ensure that each stock-week surface has sufficient coverage.

Chapter 2 presents the empirical methodology. It introduces the notation and describes the rolling principal component analysis that is applied to the de-meaned implied volatility surfaces. The chapter then details the elastic net regression framework, including the choice of tuning parameters, the cross-validation scheme, and the way in which principal component scores and lagged returns enter the prediction model. It also sets out the evaluation metrics used to assess forecasting performance, including predictive R^2 , information coefficients, and the returns and Sharpe ratios of long-short portfolios based on model forecasts. The treatment of transaction costs and the construction of net-of-cost performance measures are presented at the end of the chapter.

Chapter 3 reports the empirical results. It first documents the properties of the implied volatility factors, including their average variance explained and their interpretation as level, slope, and curvature components of the surface. It then presents the out-of-sample

prediction results for the elastic net models across rolling windows and stocks, comparing alternative predictor specifications. The chapter interprets the sign and magnitude of the estimated coefficients and examines how the contribution of each principal component varies over time. The performance of long-short portfolios formed on the basis of forecasted returns is analyzed both before and after transaction costs, with particular emphasis on how profitability varies across market conditions.

Finally, the conclusion summarizes the main findings and discusses their implications. It reflects on what the results say about the information contained in option-implied volatility factors, the usefulness of regularized models for return prediction, and the trade-off between statistical fit and economic value. The conclusion also outlines several directions for future research, such as extending the analysis to other markets, incorporating additional option-based variables, or combining implied volatility factors with macroeconomic information.

Literature Review

Return Predictability Puzzle

The empirical literature shows that forecasting stock returns is difficult, even when many variables appear predictive in historical samples. Welch and Goyal (2008) document that standard predictors of the U.S. equity premium, such as dividend yields and interest rate spreads, fail to generate positive out-of-sample performance. Their findings highlight that traditional predictors are unstable over time and often deliver negative out-of-sample statistics. This motivates the use of forward-looking market information rather than backward-looking macroeconomic or accounting series.

Options markets provide such information because option prices embed expectations about future volatility, downside risk, and short sale constraints. Several studies show that option-implied variables predict the cross-section of stock returns. Cremers and Weinbaum (2010) find that deviations from put-call parity forecast future returns, consistent with limits to arbitrage and informed trading in options. Xing, Zhang, and Zhao (2010) show that stocks with steeper volatility smirks earn lower subsequent returns, indicating that out-of-the-money puts contain information about negative news that equity markets incorporate slowly. Johnson and So (2012) show that a high option-to-stock volume ratio predicts lower near-term returns, which they interpret as informed traders preferring options when short selling the stock is costly. These findings indicate that option-implied information contains forward-looking signals that are not captured by traditional predictors.

Most existing studies focus on a single feature of the implied volatility surface. Because the surface contains implied volatilities across moneyness and maturities, reducing it to one skew or spread can miss important variation. Muravyev, Vasquez, and Wang (2020) address this issue by applying principal component analysis to de-meaned implied volatility surfaces. They show that a small number of components summarize most variation in surface shapes and predict stock returns both in-sample and out-of-sample, while subsuming the information in standard option-based predictors.

A number of studies further document that option-implied variables contain information about the cross-section of expected stock returns beyond traditional predictors. Bali and Hovakimian (2009) investigate whether realized and implied volatilities of individual stocks can predict future returns. They find that the levels of volatility from the physical and risk-neutral distributions do not forecast returns, but that volatility spreads do. In particular, there is a negative and statistically significant relation between expected returns and the spread between realized and implied volatility, which they interpret as a proxy for volatility risk. At the same time, they document a positive relation between expected returns and the spread between call and put implied volatilities, which they view as a proxy for jump risk. Estimates from a VAR–bivariate–GARCH model indicate significant information flow from equity options to the underlying stocks, consistent with informed trading in options by investors with private information.

An et al. (2014) study the joint cross-section of stocks and options and provide complementary evidence that option-implied volatility dynamics predict equity returns. They show that stocks with large increases in call implied volatility over the previous month tend to have high subsequent returns, whereas stocks with large increases in put implied volatility tend to have low subsequent returns. Sorting stocks into decile portfolios based on past call implied volatility changes yields spreads in average returns of about one percent per month, and these return differences persist for up to six months. They also find that the cross-section of stock returns predicts future changes in option implied volatilities, with stocks that performed well in the past subsequently exhibiting higher implied volatility but lower realized volatility. These patterns are consistent with rational models

of informed trading in which information is reflected jointly in stock and option markets.

Using a different option-implied measure, Guo and Qiu (2014) revisit the relation between idiosyncratic risk and expected stock returns. They construct an options-implied variance measure as a forward-looking proxy for conditional variance and study its link to future returns at both the individual stock and market levels. In cross-sectional and time-series regressions, they find a negative relation between options-implied variance and subsequent returns. They further show that the negative correlation of realized idiosyncratic variance and analyst forecast dispersion with future returns mainly reflects their close correlation with the conditional idiosyncratic variance extracted from options. Together, these additional studies reinforce the view that option-implied quantities such as volatility spreads, changes in implied volatility, and options-implied variances contain forward-looking information about risk and return. The present thesis contributes to this literature by using principal components of the full implied volatility surface rather than a small number of scalar option-implied measures.

The Informational Content of the Implied Volatility Surface

Option markets offer a rich, forward-looking source of information about future risk that can complement or even surpass traditional backward-looking indicators such as accounting ratios or past returns. A typical U.S. stock is associated with hundreds of listed options across different strikes and maturities, yielding a high-dimensional implied volatility surface. Researchers have long noted that certain ad hoc summaries of this surface carry predictive power for returns. For example, simple metrics like the implied volatility skew or a call–put IV spread are intuitive and have proven useful in forecasting stock returns. However, collapsing a complex surface into one or two variables inevitably discards a large share of the information. This limitation motivates a more data-driven approach to extract the informational content embedded in the options market.

Recent work by Muravyev, Vasquez, and Wang (2020) demonstrates the benefits of applying principal component analysis to the full implied volatility surface. PCA is a statistical technique that reduces dimensionality while preserving as much variance as possible, yielding a small set of orthogonal factors that summarize the surface’s shape. In their study, the first five principal components explain roughly 78 percent of the variation in de-measured volatility surfaces, indicating that a few factors capture most of the relevant information. Importantly, the first three components significantly forecast future returns both in-sample and out-of-sample and subsume conventional option-based predictors when included in joint regressions. These PCA factors encapsulate signals that simpler measures, such as volatility skew or slope, aim to capture but in a more robust and comprehensive way. Moreover, because the components are orthogonal by construction, they reduce multicollinearity and mitigate overfitting when used in predictive regressions. The leading factors are also economically interpretable. For example, the dominant principal component is closely related to option-implied shorting costs, suggesting that option prices reflect market frictions and risk premia relevant for future returns.

Other studies corroborate the predictive power of implied volatility surfaces. Doran and Krieger (2010) show that the implied volatility skew derived from call and put spreads is significantly related to future stock returns, even after controlling for other equity characteristics. Fu et al. (2016) evaluate multiple implied volatility metrics and find that the IV skew consistently exhibits strong predictive power in both time-series and cross-sectional regressions. Eksi and Roy (2025) document that SP 500 index inclusions induce shifts in implied volatility skew that anticipate reversals in the underlying stocks, highlighting the anticipatory nature of option prices around non-fundamental shocks. Açıkalın and Önder (2024) show that option-based variables improve stock return forecasts across market regimes, with call–put spreads contributing most in both normal and recessionary periods. Together, these findings reinforce that option-implied volatility surfaces, when properly extracted through techniques such as PCA, contain forward-looking signals relevant for return prediction.

Implied Volatility Surface Dynamics

It is well-established that equity implied volatility (IV) surfaces exhibit a low-dimensional factor structure. Principal component analysis (PCA) of daily IV changes typically reveals that only a few orthogonal modes drive most variations. In particular, the first principal component usually acts as a broad level shift in overall volatility, while the next components correspond to a tilt (skew) and a curvature change. This mirrors the classic level–slope–curvature decomposition known from yield curve analysis, and indeed the first three factors often explain over 95 percent of total IV movement.

Cont and Fonseca (2002) provided early empirical evidence of this structure by applying PCA to index option surfaces. They found that the dynamic IV surface of the S&P 500 can be effectively described by a three-factor model: a parallel volatility shift, a skew adjustment, and a smile curvature change. These three factors captured the vast majority of daily implied volatility variation, supporting the notion that the surface’s complex movements can be summarized in just a few dimensions.

Subsequent studies confirmed and extended these findings. Fengler, Härdle, and Villa (2003) introduced a Common PCA approach to analyze the IV surface across multiple maturities simultaneously. Their analysis demonstrated that different maturity slices of the surface share a common eigenstructure, meaning the same set of principal component shapes can describe IV dynamics at all maturities. This result further justified modeling IV surface dynamics with a low number of factors. In a related vein, Fengler, Härdle, and Mammen (2007) proposed a semiparametric factor model for IV surfaces, using DAX index options to show that the surface can be approximated in a finite-dimensional function space. They successfully extracted a small set of latent factors and modeled their time evolution, reinforcing the view that most IV surface complexity can be captured by a handful of components.

More recent research on individual equity options has also identified strong factor structures in implied volatilities. Christoffersen, Fournier, and Jacobs (2017) examine the cross-section of single-stock option returns using PCA on volatility metrics. They report

that a few principal components explain a sizable fraction of the variation in delta-hedged option returns, indicating the presence of common underlying risk factors in equity option markets. Likewise, Avellaneda et al. (2020) perform PCA on implied-volatility dynamics across a large cross-section of U.S. equities. They identify a dominant market-wide factor, closely related to an open-interest-weighted index of implied-volatility changes, alongside at least one additional factor, which together capture most of the co-movements in implied-volatility dynamics across firms. This evidence across studies consistently supports the use of PCA-based dimensionality reduction: the high-dimensional implied-volatility surface can be well-represented by a few interpretable factors, making PCA an academically justified and effective method for analyzing equity options data.

Cross-Sectional Ranking vs Point Prediction

Predicting the exact level or sign of stock returns out-of-sample is difficult because returns are noisy and the signal-to-noise ratio is low. Models evaluated purely on out-of-sample R^2 therefore often appear to perform poorly, even when they contain information that is useful for portfolio construction. In contrast, cross-sectional ranking models focus on ordering stocks by expected return rather than accurately matching the realized level. This distinction is central to the methodology of the present thesis, which evaluates predictive content using long–short decile strategies and the information coefficient (IC) instead of relying on regression-based R^2 alone.

A growing literature shows that low or even negative out-of-sample R^2 is compatible with economically meaningful cross-sectional predictability. Gu, Kelly, and Xiu (2020) train machine learning models on hundreds of firm-level characteristics and find that, although the models deliver out-of-sample R^2 below 1%, they generate long–short portfolio Sharpe ratios between 1.2 and 2.0 and ICs as high as 6–10%. Lewellen (2015) documents that standard characteristic-based return models typically explain less than 0.5% of the variance in monthly returns out-of-sample, yet the same signals often produce sizable spreads between high- and low-characteristic portfolios. Cenesizoglu and Timmermann

(2012) directly study the link between statistical fit and investment value and show that many return prediction models with negative out-of-sample R^2 still yield positive Sharpe ratios and utility gains for investors. Together, these studies demonstrate that relative ordering of assets can be informative even when the model explains very little of the level of returns.

The implication for empirical design is that models should be judged by their ability to rank assets profitably rather than by mean squared error alone. This thesis follows that approach by focusing on cross-sectional ranking metrics such as IC and long–short portfolio performance. The PCA-derived implied volatility features and elastic net regressions are therefore evaluated primarily on their ability to sort stocks into deciles with systematically different future returns, rather than on maximizing R^2 at the individual-stock level.

Transaction Costs & Implementation

Accurately accounting for transaction costs is essential when evaluating the real-world profitability of high-turnover long–short strategies. Several components contribute to implementation frictions, including bid–ask spreads, market impact, and short-selling fees.

Novy-Marx and Velikov (2016) show that most equity anomalies experience a significant drop in profitability after adjusting for trading costs. They document that high-turnover strategies often fail to survive bid–ask spread and price impact adjustments, particularly when implemented using equal-weighted portfolios. Strategies with monthly one-way turnover above 50% are especially vulnerable. In a related context, Kociński (2014) decomposes commissions, bid–ask spreads, and market impact on the Warsaw Stock Exchange and concludes that bid–ask spreads are a dominant component of trading costs, with official average spread statistics often understating the effective costs faced by active investors.

Muravyev, Vasquez, and Wang (2020) analyze short-sale constraints and show that shorting costs are a critical friction for long–short anomaly portfolios. Their findings indicate that excluding only the most expensive-to-borrow stocks is often sufficient to

eliminate the statistical significance of long–short returns.

Muravyev and Pearson (2019) further emphasize this point. They document that stock borrow fees can entirely eliminate anomaly profits, reducing average monthly long–short returns from 15 basis points to nearly zero. This underscores the importance of incorporating shorting costs into performance evaluations.

Beyond these key studies, several papers provide important estimates and methodologies for transaction cost measurement. Abdi and Ranaldo (2017) develop a robust low-frequency estimator for effective spreads that closely tracks intra-day TAQ benchmarks using only daily high–low–close data. Jahan-Parvar and Zikes (2023) warn that volatility can distort commonly used spread proxies, especially in highly liquid stocks. Their work suggests caution when using standard spread measures to approximate costs. Finally, Bikker, Spierdijk, and Sluis (2007) estimate market impact costs for institutional equity trades, finding they average 20–30 basis points depending on trade direction and size.

Together, these studies provide a rigorous framework for understanding and quantifying the key frictions that affect the implementability of equity return signals.

Gap in Literature & Positioning

A large body of literature has studied the link between option-implied variables and future stock returns. Early work shows that volatility-based measures such as realized–implied volatility spreads and call–put implied volatility spreads contain information about expected returns and volatility risk (Bali and Hovakimian 2009). Subsequent studies document that changes in individual-stock implied volatilities and options-implied variances forecast the cross-section of stock returns and reflect informed trading and short-sale constraints (An et al. 2014; Guo and Qiu 2014). However, most of this literature compresses the full implied volatility (IV) surface into a single feature such as a skew, a slope, or a spread. This simplification risks discarding much of the information embedded in the surface. Muravyev, Vasquez, and Wang (2020) address this limitation by applying principal

component analysis (PCA) to the de-meaned IV surface. They show that the first three principal components explain most of the variation in the surface and significantly predict stock returns both in-sample and out-of-sample. Their work demonstrates that the full surface contains richer predictive content than ad hoc summaries and that PCA provides a robust method to extract this information.

This thesis extends Muravyev, Vasquez, and Wang (2020) in several important ways. First, the implied volatility surface is constructed using 0.2 to 0.8 delta options across standardized maturities from 30 to 360 calendar days, with data from 2013 to 2023 covering the 46 largest U.S. stocks. Second, after demeaning the surface, PCA is used to extract orthogonal factors. These principal components are then used as predictors in elastic net regressions. The elastic net estimator is well suited for this task because it applies both ℓ_1 and ℓ_2 regularization, allowing for shrinkage and variable selection in the presence of correlated predictors.

Unlike earlier work that focuses on point prediction and regression R^2 , this thesis emphasizes cross-sectional predictability through long–short decile portfolios and information coefficients. Even though the regression R^2 is often negative out-of-sample, the predictive signal is strong enough to produce Sharpe ratios exceeding 2 and directional accuracy above 50 percent. This is consistent with recent literature showing that economically useful ranking-based forecasts can arise even when standard predictive metrics like R^2 are low or negative.

By combining PCA-based factor extraction with modern regularized regression and a focus on cross-sectional ranking performance, this thesis contributes to the growing literature on machine learning and asset pricing. It provides evidence that option-implied information, when properly extracted and applied, continues to offer valuable insights into the cross-section of equity returns.

Chapter 1

Data

1.1 Sample Selection and Data Sources

The empirical analysis is based on U.S. large-cap equities for which both option and underlying stock data are available from Wharton Research Data Services (WRDS). The sample construction follows Muravyev, Vasquez, and Wang (2020) in spirit but is tailored to the present thesis.

In a first step, I use CRSP monthly data to identify the largest U.S. firms by market capitalization. For each CRSP common share, I compute the market value of equity as the absolute stock price times shares outstanding and average this market capitalization over the period January 2014 to December 2023. I then rank firms by their average size and retain the top fifty names. This step ensures that the sample focuses on liquid and mature firms with actively traded options, which is important for obtaining a dense implied volatility surface and reliable price quotes.

In a second step, I map these CRSP identifiers to OptionMetrics security identifiers (`secid`) and download the corresponding daily option and underlying data from the OptionMetrics Ivy DB U.S. database. The option data provide, for each security and date, the contract type (call or put), time to maturity in calendar days, delta, and implied volatility. The underlying data contain the daily closing price, return, and shares outstanding. I

restrict attention to individual equities by excluding any observations where the Option-Metrics index flag is nonzero.

The sample period runs from 31 August 2013 to 31 August 2023. All dates outside this window are dropped before any further processing. After filtering for equities and the sample period, the raw implied volatility file contains approximately 31.4 million contract-day observations, while the underlying price file contains about 118,000 daily stock-level observations across the (up to) fifty firms.

Because some firms enter or leave the large-cap universe over time and because option coverage is incomplete for a few names, not all of the initial fifty firms have usable data over the entire period. After merging options and returns and imposing the implied volatility coverage requirements described below, the final panel used in the predictive regressions contains $N = 46$ stocks.

The start date of August 2013 is chosen to allow the first three-year training window (September 2013 to August 2016) required for the rolling principal component analysis and forecasting models. Starting earlier would either reduce the number of complete rolling windows available for out-of-sample evaluation or include periods with less consistent option coverage for some firms. The chosen window ensures a balanced trade-off between sample length and the need for sufficiently long estimation periods within each rolling window.

1.2 Notation

Throughout this chapter and the remainder of the thesis, I use the following notation. Let $i = 1, \dots, N$ index stocks and let $t = 1, \dots, T$ index weeks. Let d index trading days within a week, and let $m = 1, \dots, M$ index grid points on the implied volatility surface. The weekly return of stock i in week t is denoted $r_{i,t}$, and the one-week-ahead return (the prediction target) is $r_{i,t+1}$. Implied volatility at grid point m for stock i in week t is denoted $\sigma_{i,t,m}$, and its de-meaned counterpart is $\tilde{\sigma}_{i,t,m}$. Week t is identified by its ending date, which I refer to as the “week-end” date.

1.3 Equity Return Construction

Weekly equity returns are constructed from the OptionMetrics security price file in a manner consistent with Muravyev, Vasquez, and Wang (2020). For each stock, I sort daily observations by date and compute the daily market capitalization as the closing price times shares outstanding (shares outstanding are reported in thousands in the raw data and scaled accordingly).

Following standard practice in the options literature, I define trading weeks on a Tuesday-to-Monday basis. This convention aligns with the timing of option settlements and ensures that the implied volatility surface observed at the end of week t does not overlap with the return period being predicted. Specifically, I assign each trading day to the week ending on the following Monday. The weekly stock return $r_{i,t}$ for stock i and week t is computed as the compounded daily return:

$$r_{i,t} = \prod_{d \in \mathcal{D}_t} (1 + r_{i,d}) - 1, \quad (1.1)$$

where $\mathcal{D}_t$ denotes the set of trading days in week t and $r_{i,d}$ is the daily return on day d .

The prediction target in all regressions is the one-week-ahead return. For each stock, I shift the weekly return series one period forward:

$$y_{i,t} \equiv r_{i,t+1}, \quad (1.2)$$

which represents the return from the end of week t to the end of week $t + 1$. The last week for each stock is dropped because the forward return is undefined. After this construction, the weekly stock-side panel contains 24,496 stock-week observations before merging with option data.

1.4 Implied Volatility Surface Construction

1.4.1 Raw Option Quotes and Filters

The option data provide, for each stock, date, and contract, the time to maturity in calendar days, the Black–Scholes delta (expressed in percentage points), the contract type

(call or put), and the quoted implied volatility. I apply several filters to ensure data quality.

First, I convert the quoted deltas, which in the raw data lie between -90 and $+90$ (representing -0.90 to $+0.90$ in standard notation), to absolute values in the unit interval:

$$|\Delta| = \frac{|\text{delta}_{\text{raw}}|}{100}. \quad (1.3)$$

I then retain only contracts with maturities τ between 30 and 365 calendar days and absolute deltas between 0.20 and 0.80:

$$30 \leq \tau \leq 365, \quad 0.20 \leq |\Delta| \leq 0.80. \quad (1.4)$$

This region covers moderately out-of-the-money to in-the-money options and avoids extremely short-dated or deep out-of-the-money contracts that are often illiquid or poorly priced.

1.4.2 Interpolation to a Standard Grid

To compare implied volatility surfaces across stocks and over time, I project the raw implied volatilities onto a fixed maturity–moneyness grid following Muravyev, Vasquez, and Wang (2020). The grid consists of eight maturities and seven delta levels:

$$\tau \in \{30, 60, 91, 122, 152, 182, 273, 365\} \text{ days}, \quad (1.5)$$

$$|\Delta| \in \{0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80\}. \quad (1.6)$$

This yields $8 \times 7 = 56$ grid points per option type (call or put) and 112 grid points when calls and puts are considered jointly.

For each stock i , date d , and option type $s \in \{\text{C}, \text{P}\}$, I treat the available implied volatility quotes as observations on a two-dimensional surface over $(\tau, |\Delta|)$. I then use bilinear interpolation to estimate the implied volatility $\hat{\sigma}_{i,d,s}(\tau_k, |\Delta|_j)$ at each grid point $(\tau_k, |\Delta|_j)$ from the set of $L_{i,d,s}$ available raw quotes. Interpolation is attempted only when at least three distinct raw quotes are available (i.e., $L_{i,d,s} \geq 3$); otherwise, that stock-date surface is discarded.¹

¹The interpolation is implemented using `scipy.interpolate.griddata` with the linear method in Python.

The resulting variables are named following the pattern used by Muravyev, Vasquez, and Wang (2020). For example, C_T91_D40 denotes the call implied volatility for a maturity of 91 days and absolute delta $|\Delta| = 0.40$, while P_T365_D50 denotes the put implied volatility for a maturity of 365 days and absolute delta $|\Delta| = 0.50$. The letter indicates call (C) or put (P), the number after T is maturity in days, and the number after D is $100 \times |\Delta|$.

1.4.3 Weekly Aggregation

Because the prediction horizon is weekly, I aggregate the daily implied volatility surfaces to weekly frequency. I assign each daily observation to the week ending on the following Monday, consistent with the return construction. For each stock i , week t , and grid point m , I compute the simple average of the interpolated implied volatilities across all trading days within that week:

$$\sigma_{i,t,m} = \frac{1}{|\mathcal{D}_t|} \sum_{d \in \mathcal{D}_t} \sigma_{i,d,m}, \quad (1.7)$$

where $\mathcal{D}_t$ denotes the set of trading days in week t with available data and $|\mathcal{D}_t|$ is its cardinality. Each grid point m corresponds to a unique combination of option type $s \in \{C, P\}$, maturity τ_k , and delta $|\Delta|_j$.

The data are then pivoted so that each stock-week observation forms one row, with 112 implied volatility columns (plus identifiers). In practice, some grid points at the boundary of the moneyness range (particularly $|\Delta| = 0.20$) have insufficient raw data for reliable interpolation and are missing for all observations. After dropping these 16 structurally missing columns, $M = 96$ implied volatility series remain in the final dataset.

The choice of weekly frequency reflects both econometric and practical considerations. First, the prediction target is the one-week-ahead return, so aligning implied volatility measures to a weekly horizon ensures consistency between the information set and the forecasted outcome. Second, weekly aggregation reduces high-frequency microstructure noise and bid–ask bounce present in daily option quotes, which can distort principal component estimates in high-dimensional settings. Third, weekly rebalancing aligns naturally

with the portfolio construction exercise in Chapter 4, where trading costs and turnover are evaluated at a weekly frequency. Finally, aggregating to weekly frequency keeps the dimensionality of the rolling PCA manageable while preserving sufficient time-series variation for factor estimation.

1.4.4 De-meaning the Surface

Following Muravyev, Vasquez, and Wang (2020), I focus on the *shape* of the implied volatility surface rather than its overall level. For each stock-week observation, I de-mean the surface by subtracting the cross-sectional average across all observed grid points. Let $\mathcal{M}_{i,t} \subseteq \{1, \dots, M\}$ denote the set of grid points with non-missing values for stock i in week t . I compute the row mean as

$$\bar{\sigma}_{i,t} = \frac{1}{|\mathcal{M}_{i,t}|} \sum_{m \in \mathcal{M}_{i,t}} \sigma_{i,t,m} \quad (1.8)$$

and define the de-meaned implied volatility as

$$\tilde{\sigma}_{i,t,m} = \sigma_{i,t,m} - \bar{\sigma}_{i,t}. \quad (1.9)$$

By construction, the mean of $\tilde{\sigma}_{i,t,m}$ across $m \in \mathcal{M}_{i,t}$ is zero for each stock-week. This transformation removes variation in the overall level of implied volatility (which largely reflects the stock’s unconditional volatility) and isolates relative differences across moneyness and maturity, the “shape” of the surface. These de-meaned surfaces $\{\tilde{\sigma}_{i,t,m}\}$ serve as inputs to the principal component analysis described in Chapter 2.

1.5 Merged Panel and Coverage Filters

To construct the final panel for the empirical analysis, I merge the weekly return data with the implied volatility surfaces using the stock identifier and week-end date as keys. The merge is performed as an inner join, so only weeks with both a valid one-week-ahead return and an implied volatility surface are retained. After dropping the final observation

for each stock (for which the forward return is undefined), the merged panel contains 23,881 stock-week observations.

Implied volatility data remain missing at some grid points for certain weeks due to sparse trading or interpolation failures at the surface boundaries. To ensure that the surfaces used in the principal component analysis are well populated, I impose a row-level coverage filter: a stock-week observation is retained only if at least 90 percent of the grid points are non-missing. Formally, the observation (i, t) is kept if

$$\frac{|\mathcal{M}_{i,t}|}{M} \geq 0.90. \quad (1.10)$$

In practice, this filter removes very few observations, and the final analysis-ready panel continues to contain 23,881 stock-week pairs across 46 stocks.

Table 1.1 summarizes the key characteristics of the final sample.

Table 1.1: Sample Summary Statistics

Characteristic	Value
Sample period	August 2013 – August 2023
Number of stocks (N)	46
Number of weeks (T)	521
Total stock-week observations	23,881
Average observations per stock	519
Number of IV grid points (M)	96
<i>Weekly Return Statistics</i>	
Mean	0.31%
Standard deviation	3.61%
Minimum	−33.3%
Maximum	56.5%

1.6 Discussion

Several aspects of the sample construction merit discussion.

The sample selection based on average market capitalization over 2014–2023 introduces survivorship bias, as firms must have existed and remained large throughout (or for

most of) the sample period to be included. This design choice follows from the research objective of studying liquid, optionable stocks with dense implied volatility surfaces but may limit the generalizability of the findings to smaller or younger firms.

The choice of week boundaries (Tuesday to Monday) follows the convention in the options literature and ensures a clean separation between the information set (the implied volatility surface observed through Monday) and the prediction target (the return from Tuesday through the following Monday). This timing avoids look-ahead bias in the predictive regressions.

The structural missingness of certain grid points (particularly at $|\Delta| = 0.20$) reflects the limited liquidity of deep out-of-the-money options. The 90 percent coverage filter ensures that the surfaces used in the analysis are well populated while retaining essentially all observations. The principal component analysis in the next chapter handles the remaining missing values through appropriate imputation within each rolling window.

Chapter 2

Methodology

2.1 Overview of the Empirical Framework

This chapter describes the empirical framework used to test whether the shape of the option-implied volatility surface predicts the cross-section of weekly stock returns for the largest U.S. firms. The starting point is the merged panel constructed in Chapter 1, which contains, for each stock i and week t , the one-week-ahead return $y_{i,t} = r_{i,t+1}$ and a de-meaned implied volatility surface represented by M grid points $\{\tilde{\sigma}_{i,t,m}\}_{m=1}^M$. The methodology transforms this high-dimensional options information into a small set of factors and then uses these factors as predictors in regularized return forecasting models.

Following Muravyev, Vasquez, and Wang (2020), I first apply principal component analysis (PCA) to the de-meaned implied volatility surfaces. For each stock and each rolling estimation window, the M grid points are stacked into a vector and PCA is performed on the covariance matrix of these vectors. The first five principal components are retained, with the leading three interpreted as level, slope, and curvature shifts in the implied volatility surface. These components summarize most of the cross-sectional variation in surface shapes and provide a compact representation of the option-implied information that can be used in subsequent regressions.

Importantly, principal component analysis is conducted separately for each stock rather

than on a pooled panel of surfaces across firms. For a given stock i and rolling estimation window, I construct the time-series matrix of de-meaned implied volatility vectors for that stock alone and compute the eigenvectors of its sample covariance matrix. No cross-sectional pooling across different stocks is used in the PCA stage. As a result, the extracted principal components are stock-specific and time-local.

This approach differs from Muravyev, Vasquez, and Wang (2020), who estimate principal components on a pooled panel of implied volatility surfaces across many firms, thereby extracting a common set of factors. In contrast, this thesis allows the factor structure of the implied volatility surface to vary across firms and over time by estimating PCA separately for each stock within rolling windows.

The second step specifies and estimates elastic net return prediction models that use these PCA factors as explanatory variables. For each stock i and week t , I construct a feature vector $x_{i,t}$ that contains the contemporaneous principal components and, in some specifications, lagged returns and simple polynomial interactions of the components. I then estimate elastic net regressions of the form

$$y_{i,t} = \beta_{0,i} + x_{i,t}^\top \beta_i + \varepsilon_{i,t},$$

where the coefficient vector β_i is obtained by minimizing a penalized least-squares criterion with both ℓ_1 and ℓ_2 penalties as in (Zou and Hastie 2005). The models are estimated in a rolling 3-year training window and evaluated in the subsequent 1-year test period, and this procedure is repeated over the full sample. I consider both stock-specific models, where a separate elastic net is estimated for each stock, and pooled models that use the entire panel of stocks with appropriate fixed effects. Tuning parameters are chosen by time-series cross-validation within each training window, following the approach advocated by Gu, Kelly, and Xiu (2020) for return prediction with many correlated predictors.

The final step, developed in Chapter 4, uses the out-of-sample fitted values from these elastic net models to form cross-sectional rankings and long-short portfolios. Model performance is evaluated using forecast error measures, information coefficients, and portfolio-level statistics such as Sharpe ratios. The present chapter focuses on the con-

struction of the PCA factors and the specification and estimation of the elastic net models that generate these forecasts.

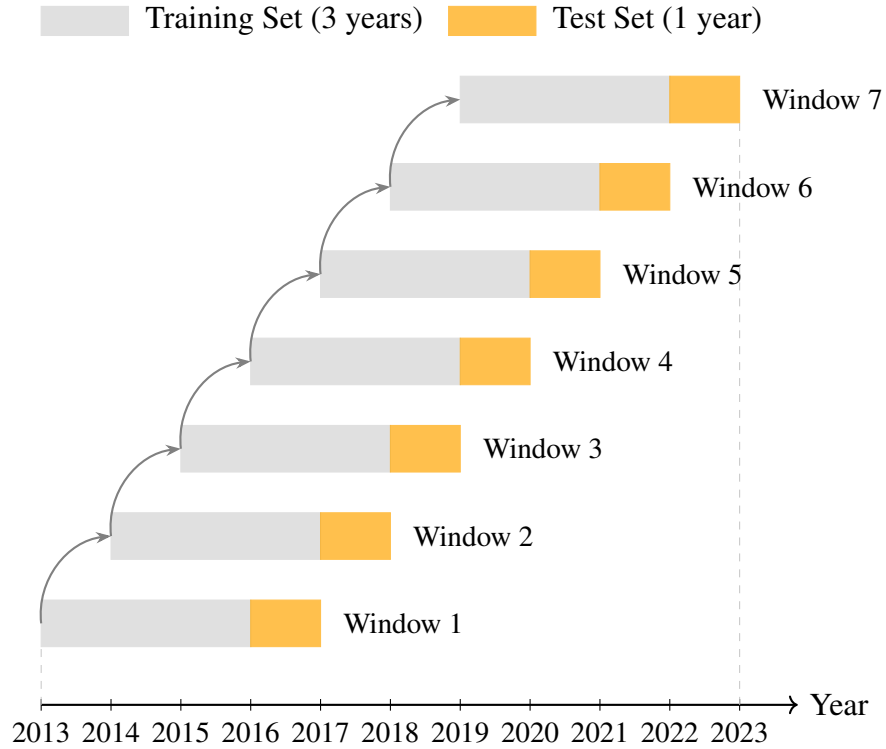


Figure 2.1: Rolling window cross-validation with 3-year training windows and 1-year test windows over the period 2013–2023.

2.2 Dimensionality Reduction: Rolling PCA

Figure 2.1 illustrates the rolling window scheme used to extract option-implied factors from the volatility surface. Rather than estimating a single static PCA on the full sample, I apply principal component analysis in a rolling manner in order to respect the time-series structure of the data and avoid look-ahead bias.

For each stock i , the weekly panel of de-meaned implied volatilities $\{\tilde{\sigma}_{i,t,m}\}$ is ordered by the week-end date t . I then construct a sequence of non-overlapping test windows. Window 1 uses the first three years of data (September 2013 to August 2016, approximately 156 weeks) as the training period and the following year (September 2016 to

August 2017, approximately 52 weeks) as the test period. Window 2 shifts forward by one year, so that the training period runs from September 2014 to August 2017 and the test period from September 2017 to August 2018. This procedure continues until Window 7, which trains on the three years from September 2019 to August 2022 and tests on the final year from September 2022 to August 2023. Each stock-week observation appears in exactly one test window, and only past data are used to estimate the PCA for that window.

The choice of a three-year training window reflects a bias–variance trade-off in estimating both the principal components and the forecasting models. A shorter window would reduce the number of observations available for estimating the covariance matrix underlying the PCA, potentially leading to unstable factor loadings and noisy coefficient estimates in the elastic net regressions. Conversely, a much longer window would increase estimation precision but at the cost of assuming a stable factor structure and return–predictor relationship over long horizons, which may be unrealistic in the presence of structural breaks or changing market regimes. With 156 weekly observations per training window, the three-year length provides sufficient time-series variation for reliable covariance estimation while allowing the factor structure and predictive relationships to adapt gradually over time.

Additional robustness checks using two-year and four-year training windows yield qualitatively similar patterns in explained variance and forecasting performance, suggesting that the results are not driven mechanically by the specific window length.

2.2.1 Handling Missing Values

Before applying PCA within each window, I address missing implied volatility values that arise from sparse option trading or interpolation failures. For each training window, I first identify grid-point columns where more than 50 percent of observations are missing and exclude these columns from the analysis for that window. This ensures that the PCA is not distorted by columns with predominantly imputed values. Additionally, I require that each training window contains at least 20 weeks with complete observations across

the retained columns; windows failing this criterion are skipped.

For test-period observations, any remaining missing values in the retained columns are imputed using the corresponding column mean from the training period. This approach ensures that the test data can be projected onto the principal component space estimated from the training data without introducing forward-looking information.

2.2.2 Standardization and PCA Estimation

Within a given window w , I stack all training observations for stock i into a matrix $X_i^{(w)}$ of dimension $T_{i,w}^{\text{train}} \times M_w$, where M_w denotes the number of implied volatility grid points retained after the missing-value filters for that window. Before applying PCA, I standardize each column of $X_i^{(w)}$ to have zero mean and unit variance. Specifically, for each grid point m , I compute the column mean $\mu_{i,m}^{(w)}$ and standard deviation $\sigma_{i,m}^{(w)}$ from the training data and transform each entry as

$$\tilde{x}_{i,t,m}^{(w)} = \frac{X_i^{(w)}(t, m) - \mu_{i,m}^{(w)}}{\sigma_{i,m}^{(w)}}.$$

The same transformation is applied to the test-period observations using the training-period means and standard deviations. This step ensures that all implied volatility grid points enter the PCA on a comparable scale.

I then apply PCA to the standardized training matrix and extract the first five principal components. Let $P_i^{(w)} \in R^{M_w \times 5}$ denote the matrix whose columns are the eigenvectors associated with the five largest eigenvalues of the sample covariance matrix of the standardized training data. For each week t in the test period of window w , I compute the principal component scores as

$$\text{PC}_{i,t} = \tilde{x}_{i,t}^\top P_i^{(w)},$$

where $\tilde{x}_{i,t} \in R^{M_w}$ is the standardized implied volatility vector for stock i in week t . The PCA loadings $P_i^{(w)}$ are estimated using only the training segment and held fixed when projecting test observations.

The procedure produces a panel of rolling principal component scores for each stock and week. For every test year, I obtain five standardized factors that summarize the shape of the implied volatility surface at the end of week t . These PC scores are then used as predictors in the elastic net return forecasting models described in the following section, with the first three components playing the central role in the portfolio construction analysis.

2.2.3 Selecting the Number of Components

A key decision in the methodology is the number of principal components to retain. Because PCA is estimated separately for each stock and rolling window, there is no single scree plot for the full sample. Instead, I compute the explained variance ratio for each stock-window PCA and then average these ratios across all stocks and all rolling windows (a total of $46 \times 7 = 322$ PCA estimations). Figure 2.2 therefore presents the average explained variance profile across stock-window PCAs rather than a scree plot from a single covariance matrix. The first principal component (Level) explains approximately 58% of

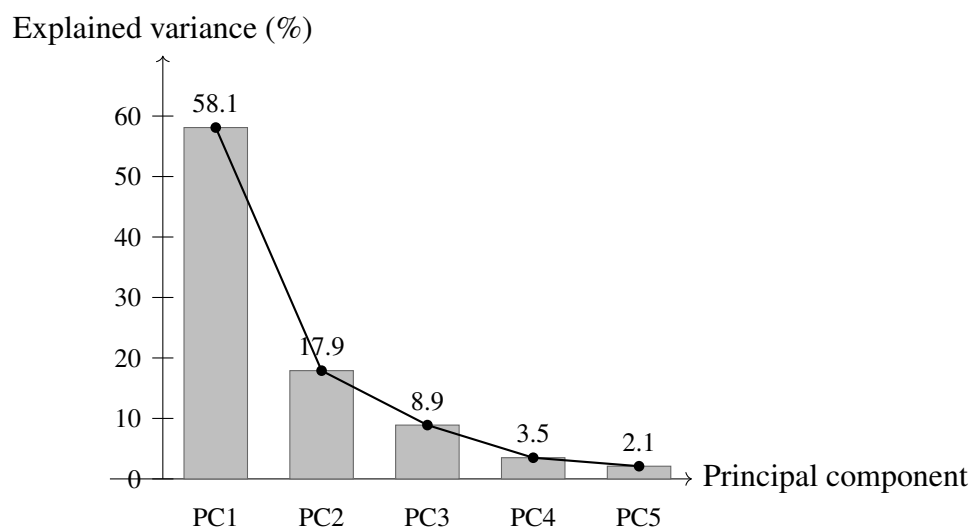


Figure 2.2: Scree plot showing the average percentage of variance explained by the first five principal components across all rolling windows.

the variation in the implied volatility surface. The second (Slope) and third (Curvature)

components explain about 18% and 9%, respectively. Together, the first three components account for roughly 85% of the total variance, confirming that the implied volatility surface admits a low-dimensional factor structure. While I retain five components to capture subtle variations in the wings of the surface, the steep drop-off after the third component suggests that most of the signal is concentrated in the leading factors. Importantly, the first rolling window alone yields similar proportions (PC1: 53.9%, PC2: 18.3%, PC3: 9.6%), confirming that the decision to retain five principal components does not rely on information from later periods (see Figure 1 in Appendix B).

2.3 Elastic Net Return Prediction

Having extracted the principal component factors from the implied volatility surface, I now describe the regularized regression framework used to forecast weekly stock returns.

2.3.1 The Elastic Net Objective

The elastic net estimator of Zou and Hastie (2005) combines the variable selection properties of the lasso with the grouping behavior of ridge regression. For each stock i and rolling window w , I estimate coefficients by minimizing the penalized least-squares objective:

$$\hat{\beta}_i^{(w)} = \arg \min_{\beta} \left\{ \frac{1}{2n} \sum_{t=1}^n \left(y_{i,t} - \beta_0 - x_{i,t}^\top \beta \right)^2 + \lambda \left[\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right] \right\}, \quad (2.1)$$

where $y_{i,t} = r_{i,t+1}$ denotes the one-week-ahead return, $x_{i,t}$ is the vector of predictors observed at week t , and $\lambda > 0$ controls the overall regularization strength. The mixing parameter $\alpha \in [0, 1]$ determines the balance between the ℓ_1 penalty (which induces sparsity by shrinking some coefficients exactly to zero) and the ℓ_2 penalty (which shrinks correlated predictors together). When $\alpha = 1$, the estimator reduces to the lasso; when $\alpha = 0$, it reduces to ridge regression.

The elastic net is particularly well-suited for return prediction with implied volatility factors for two reasons. First, the principal components may be correlated across adja-

cent maturities or strikes, and the ℓ_2 penalty encourages similar coefficients on correlated predictors rather than arbitrarily selecting one. Second, in the low signal-to-noise environment typical of return forecasting, the ℓ_1 penalty can shrink irrelevant predictors to zero, reducing overfitting and improving out-of-sample performance.

2.3.2 Stock-by-Stock, Window-by-Window Estimation

I estimate separate elastic net models for each stock and each rolling window, yielding $46 \times 7 = 322$ distinct models. This approach allows coefficient heterogeneity across stocks, reflecting differences in how option-implied information relates to returns for different firms, and permits the relationships to evolve over time as market conditions change.

For each stock-window combination, the model is trained on approximately 150 weekly observations (3 years) and evaluated on approximately 52 observations (1 year). This separation ensures that all performance metrics reported in Chapter 3 are genuinely out-of-sample and free from look-ahead bias.

2.3.3 Hyperparameter Selection

The regularization parameters λ and α are selected via time-series cross-validation within each training window. I employ `TimeSeriesSplit` with 5 folds, which respects the temporal ordering of observations by always using earlier data for training and later data for validation within each fold. This avoids the data leakage that would arise from standard k -fold cross-validation in a time-series context.

The grid search considers 60 values of λ log-spaced between 10^{-6} and 10^2 , and five values of $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The combination minimizing mean squared error on the validation folds is selected, and the model is refit on the entire training window before generating out-of-sample predictions for the test period.

Table 2.1 summarizes the distribution of selected hyperparameters across all 322 models.

Table 2.1: Distribution of Cross-Validation Selected Hyperparameters

Parameter	Mean	Median	Std. Dev.	Range
λ (penalty strength)	34.79	0.026	48.15	$[10^{-6}, 100]$
α (L1 ratio)	0.28	0.10	0.28	$[0.1, 0.9]$

Notes: Statistics computed across 322 stock-window models.

The median α of 0.10 indicates that most models favor a ridge-like penalty with only mild sparsity, consistent with the view that multiple principal components contain complementary information. The wide range of selected λ values reflects heterogeneity in signal strength across stocks and time periods.

2.4 Evaluation Metrics

I evaluate model performance using four complementary metrics that capture different aspects of forecast quality.

2.4.1 Out-of-Sample R^2

The out-of-sample coefficient of determination measures the proportion of return variance explained by the model relative to a naive benchmark. Following Campbell and Thompson (2008), I define:

$$R_{\text{OOS}}^2 = 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y})^2}, \quad (2.2)$$

where y_t is the realized return, $\hat{y}_t$ is the model’s prediction, and $\bar{y}$ is the historical mean return computed from the training period (the “prevailing mean” benchmark). A positive R_{OOS}^2 indicates that the model outperforms the historical mean in terms of squared forecast errors; a negative value indicates underperformance.

2.4.2 Information Coefficient

The information coefficient (IC) measures the rank correlation between predicted and realized returns:

$$IC_t = \text{Spearman}(\hat{y}_{i,t}, y_{i,t}), \quad (2.3)$$

where the correlation is computed across all stocks i within each week t . The IC captures the model's ability to rank stocks correctly, a property that is directly relevant for portfolio construction, where relative performance matters more than absolute forecast accuracy. I report the time-series average of weekly ICs.

2.4.3 Directional Accuracy

Directional accuracy measures the fraction of predictions where the model correctly forecasts the sign of the return:

$$\text{Direction} = \frac{1}{N} \sum_{i,t} \mathbf{1}[\text{sign}(\hat{y}_{i,t}) = \text{sign}(y_{i,t})]. \quad (2.4)$$

A directional accuracy above 50% indicates that the model predicts the direction of returns better than chance, even if the magnitude of predictions is imprecise.

2.4.4 Cumulative SSE Difference

Following Welch and Goyal (2008), I track model performance over time using the cumulative sum of squared error differences:

$$\Delta_T = \sum_{t=1}^T d_t, \quad \text{where} \quad d_t = (y_t - \bar{y})^2 - (y_t - \hat{y}_t)^2. \quad (2.5)$$

A positive d_t indicates that the model produces a smaller squared error than the benchmark in week t . The cumulative sum Δ_T provides a visual representation of how forecast accuracy evolves over time and whether performance gains or losses are concentrated in particular periods. I construct 95% confidence intervals using an expanding-window standard error with a lag-1 autocorrelation adjustment.

2.5 Alternative Model: Random Forest

To assess whether nonlinear relationships between implied volatility factors and returns improve predictive performance, I estimate Random Forest regressions as an alternative to the linear elastic net framework. Random Forests are ensemble methods that aggregate predictions from many decision trees, each trained on a bootstrap sample of the data with random feature subsets at each split (Breiman 2001). This approach can capture complex interactions and nonlinearities without explicit specification. I estimate separate Random Forest models for each stock and window using the same training/test split as the elastic net models. Hyperparameters, including the number of trees, maximum depth, minimum samples per split, and maximum features per split, are selected via `TimeSeriesSplit` cross-validation. Comparing Random Forest and elastic net performance reveals whether the additional flexibility of nonlinear methods improves out-of-sample forecasts or leads to overfitting in the low signal-to-noise environment of weekly return prediction.

Chapter 3

Results

This chapter presents the empirical findings from the forecasting models developed in Chapter 2. Section 3.1 compares the performance of elastic net, Fama-MacBeth OLS, and Random Forest models. Section 3.2 examines which feature specifications yield the best forecasts and reveals a striking divergence between point prediction accuracy and portfolio performance. Section 3.3 analyzes the distribution of elastic net coefficients. Section 3.4 tracks model performance over time relative to market volatility conditions.

3.1 Model Comparison

Table 3.1 presents out-of-sample performance metrics for the three estimation approaches: cross-sectional Fama-MacBeth OLS, stock-by-stock elastic net, and stock-by-stock Random Forest. All models use the same rolling window structure and feature set (five principal components plus four lagged returns) to ensure comparability.

Three key findings emerge from this comparison. First, all models exhibit negative out-of-sample R^2 . The elastic net achieves $R^2_{\text{OOS}} = -0.023$, indicating that its squared forecast errors exceed those of the simple prevailing mean benchmark by approximately 2.3%. This finding is consistent with the broader literature on return predictability, which documents that beating the historical mean is extraordinarily difficult at short horizons (Welch and Goyal 2008). The Fama-MacBeth OLS performs slightly worse ($R^2_{\text{OOS}} =$

Table 3.1: Out-of-Sample Performance: Model Comparison

Model	R_{OOS}^2	IC	Direction	Models
Fama-MacBeth OLS	−0.030	0.035	50.2%	7
Elastic Net	−0.023	0.010	55.1%	319
Random Forest	−0.808	−0.003	52.3%	319

Notes: All models use 5 PCs + 4 lagged returns as features. R_{OOS}^2 is the out-of-sample coefficient of determination relative to the prevailing mean benchmark. IC is the average weekly Spearman correlation between predicted and realized returns. Direction is the percentage of correct sign predictions. Fama-MacBeth estimates weekly cross-sectional regressions and averages coefficients; Elastic Net and Random Forest estimate separate models for each stock-window combination. Three stock-window combinations are excluded due to insufficient observations, yielding 319 models rather than $46 \times 7 = 322$.

−0.030), while Random Forest performs dramatically worse ($R_{\text{OOS}}^2 = -0.808$), suggesting severe overfitting.

Second, elastic net substantially improves directional accuracy. Despite failing to outperform the benchmark in terms of squared errors, elastic net correctly predicts the sign of returns 55.1% of the time, compared to 50.2% for Fama-MacBeth OLS and 52.3% for Random Forest. This 5 percentage point improvement over a coin flip, compounded across many stocks and weeks, can translate into economically meaningful portfolio returns.

Third, Random Forest severely overfits. The Random Forest models achieve $R_{\text{OOS}}^2 = -0.808$, meaning their squared errors are nearly double those of the prevailing mean. This poor performance arises because the flexible tree-based models fit noise in the training data that does not persist out-of-sample. The negative information coefficient (−0.003) indicates that Random Forest predictions are essentially uncorrelated with or even slightly inversely related to realized returns. This finding underscores the importance of regularization in the low signal-to-noise environment of return forecasting, where model complexity is penalized rather than rewarded.

Table 3.2 provides a more detailed comparison across feature specifications.

Panel A reveals a striking pattern: the interaction specification achieves the *worst* R_{OOS}^2 (−0.030) yet the *best* information coefficient (0.036) and directional accuracy (55.5%).

Table 3.2: Out-of-Sample Performance by Model and Feature Specification

Model	Feature Set	R^2_{OOS}	IC	Direction
<i>Panel A: Elastic Net</i>				
Elastic Net	5 PCs only	−0.024	0.024	55.4%
Elastic Net	5 PCs + Lag1	−0.024	0.034	55.4%
Elastic Net	All features (9)	−0.023	0.010	55.1%
Elastic Net	With interactions (45)	−0.030	0.036	55.5%
<i>Panel B: Random Forest</i>				
Random Forest	5 PCs only	−0.079	0.002	52.0%
Random Forest	5 PCs + Lag1	−0.809	−0.002	52.3%
Random Forest	All features (9)	−0.808	−0.003	52.3%

Notes: “5 PCs only” includes PC1–PC5. “5 PCs + Lag1” adds the most recent weekly return. “All features” includes 5 PCs plus 4 lagged returns. “With interactions” adds 36 pairwise interaction terms to the 9 base features. Bold indicates best performance within each metric for Elastic Net. Statistics are averages across stock-window models.

This divergence between point prediction and ranking metrics, explored in detail in Section 3.2 which represents a central finding of this thesis.

Panel B confirms that Random Forest severely overfits regardless of feature specification, with performance degrading catastrophically when lagged returns are included. The regularization imposed by elastic net prevents this overfitting, demonstrating the value of penalized regression methods for return prediction.

3.2 Feature Set Comparison

This section examines the marginal contribution of different predictor variables and tests whether nonlinear feature transformations improve forecast performance. A central finding emerges: the model with the worst point prediction accuracy achieves the best portfolio performance, revealing a fundamental divergence between statistical and economic measures of forecast quality.

3.2.1 Candidate Model Specifications

I evaluate four feature specifications of increasing complexity. The first specification uses only the five principal components extracted from the implied volatility surface ($K = 5$ features). The second adds the most recent weekly return to these five components ($K = 6$ features). The third includes the five principal components plus four lagged weekly returns ($K = 9$ features). The fourth augments the nine base features with all 36 pairwise interaction terms ($K = 45$ features).

The interaction terms capture potential nonlinear relationships between implied volatility factors and past returns. For features x_j and x_k , the interaction term is simply their product $x_j \times x_k$. With 9 base features, there are $\binom{9}{2} = 36$ unique pairwise interactions, yielding 45 total predictors.

3.2.2 Point Prediction versus Ranking Metrics

Table 3.3 presents out-of-sample performance for each specification. Critically, I report both point prediction metrics (R^2_{OOS} , RMSE) and ranking metrics (IC, directional accuracy), as these can diverge substantially.

Table 3.3: Elastic Net Performance by Feature Specification

Feature Set	Features	R^2_{OOS}	RMSE	IC	Direction
5 PCs only	5	−0.024	0.0346	0.024	55.4%
5 PCs + Lag1	6	−0.024	0.0345	0.034	55.4%
All features	9	−0.023	0.0348	0.010	55.1%
With interactions	45	−0.030	0.0349	0.036	55.5%

Notes: R^2_{OOS} is the out-of-sample coefficient of determination relative to the prevailing mean. IC is the average Spearman correlation between predicted and realized returns. Direction is the percentage of correct sign predictions. Bold indicates best (highest IC, best direction) or worst (most negative R^2) within each column. Statistics are averages across 319 stock-window models.

3.2.3 The Accuracy-Ranking Divergence

A striking pattern emerges from Table 3.3: the model with the *worst* point prediction accuracy achieves the *best* ranking metrics. The interaction model has $R^2_{\text{OOS}} = -0.030$, the most negative among all specifications, indicating that its squared forecast errors exceed those of the simple historical mean by 3%. Yet this same model achieves the highest information coefficient ($\text{IC} = 0.036$) and best directional accuracy (55.5%).

This divergence arises because R^2 penalizes the magnitude of forecast errors, while IC rewards correct relative rankings regardless of magnitude. The interaction terms may introduce noise that inflates point prediction errors while simultaneously improving the model’s ability to distinguish which stocks will outperform others cross-sectionally. For portfolio construction, where the goal is to identify relative winners and losers rather than predict exact returns, ranking ability matters more than point accuracy.

To verify this insight, Table 3.4 compares the long-short portfolio performance of the three most promising specifications.

Table 3.4: Long-Short Portfolio Performance by Model Specification

Model	R^2_{OOS}	IC	Sharpe	Ann. Return	Max DD
5 PCs + Lag1	−0.024	0.034	0.615	11.98%	−39.7%
All features (no interactions)	−0.023	0.010	0.562	10.76%	−39.6%
With interactions	−0.030	0.036	0.618	12.33%	−40.4%

Notes: Long-short portfolios go long the top 20% and short the bottom 20% of stocks ranked by predicted returns each week. Sharpe ratio and annualized return assume zero transaction costs. Max DD is maximum drawdown.

The results confirm the accuracy-ranking divergence. The interaction model, despite having the most negative R^2_{OOS} , generates the highest Sharpe ratio (0.618) and annualized return (12.33%). This finding has important implications for forecast evaluation in finance: traditional accuracy metrics like R^2 may be misleading indicators of economic value when the ultimate goal is portfolio construction rather than point prediction.

3.2.4 Interaction Terms

To test whether nonlinear relationships between predictors improve forecasts, I estimate models that augment the nine base features with pairwise interaction terms. Interaction terms capture the possibility that the effect of one predictor on returns depends on the value of another predictor. For example, the predictive content of the slope factor (PC2) may vary with the overall level of implied volatility (PC1), or the momentum signal from lagged returns may be stronger or weaker depending on the shape of the volatility surface.

Formally, for base features $x_1, x_2, \dots, x_9$, the interaction specification includes all pairwise products:

$$x_j \times x_k \quad \text{for all } j < k, \quad (3.1)$$

yielding $\binom{9}{2} = 36$ interaction terms. Combined with the 9 original features, this produces 45 total predictors. The elastic net's ℓ_1 penalty can shrink irrelevant interactions to zero, allowing the model to select which (if any) interactions contain predictive information.

Table 3.5 presents a detailed comparison of the baseline and interaction specifications.

Table 3.5: Impact of Interaction Terms on Return Prediction

Metric	Without Interactions	With Interactions
Number of features	9	45
R_{OOS}^2	−0.023	−0.030
RMSE	0.0348	0.0349
Information coefficient	0.010	0.036
Directional accuracy	55.1%	55.5%
Portfolio Sharpe ratio	0.562	0.618

Notes: Both specifications use 5 PCs + 4 lagged returns as base features. The interaction specification adds all 36 pairwise products. Bold indicates better performance.

The results present a nuanced picture. Adding interaction terms *worsens* point prediction accuracy: R_{OOS}^2 declines from −0.023 to −0.030, indicating larger squared forecast errors relative to the prevailing mean benchmark. However, interaction terms *improve* cross-sectional ranking metrics: the information coefficient increases from 0.010 to 0.036,

directional accuracy rises from 55.1% to 55.5%, and the portfolio Sharpe ratio improves from 0.562 to 0.618.

This divergence between point prediction and ranking performance reinforces a central theme of this thesis: in return forecasting, the ability to rank stocks correctly matters more for portfolio construction than the ability to predict return magnitudes accurately. The interaction terms appear to capture nonlinear relationships that help distinguish relative winners from losers, even though they introduce additional noise that degrades absolute forecast accuracy.

3.2.5 Model Selection for Portfolio Construction

Based on these findings, I select the **elastic net model with interaction terms** as the primary specification for portfolio construction in Chapter 4. This choice prioritizes economic value over statistical accuracy, reflecting the ultimate goal of translating forecasts into profitable trading strategies.

Table 3.6 summarizes the selection criteria.

Table 3.6: Model Selection Summary

Criterion	Best Model	Value
Information Coefficient	With interactions	0.036
Directional Accuracy	With interactions	55.5%
Portfolio Sharpe Ratio	With interactions	0.618
Sample Coverage	With interactions	364 weeks
Out-of-sample R^2	All features (no int.)	−0.023

Notes: The interaction model ranks first on all portfolio-relevant metrics despite ranking last on R^2_{OOS} . This divergence illustrates the importance of aligning evaluation metrics with investment objectives.

The selection of the interaction model, despite its inferior point prediction accuracy, reflects a fundamental insight: for long-short equity strategies, the ability to correctly rank stocks cross-sectionally matters more than the ability to predict exact return magnitudes. The information coefficient directly measures this ranking ability, and the inter-

action model’s IC of 0.036 while modest in absolute terms translates into economically meaningful portfolio returns before transaction costs.

This choice also highlights a broader methodological point for applied financial forecasting. Researchers and practitioners should align their evaluation metrics with their ultimate objective. When the goal is portfolio construction, ranking metrics such as IC should take precedence over traditional accuracy metrics such as R^2 . The interaction model’s superior portfolio performance, despite its worst-in-class R^2 , provides compelling evidence for this perspective.

3.3 Coefficient Distributions

This section analyzes the distribution of elastic net coefficient estimates across the 319 stock-window models. Understanding these distributions provides insight into which predictors consistently contribute to forecasts and whether the regularization is achieving its intended effect of inducing sparsity.

3.3.1 Summary Statistics

Table 3.7 reports summary statistics for the estimated coefficients on the first three principal components, which together explain approximately 85% of the variation in the implied volatility surface.

Table 3.7: Distribution of Elastic Net Coefficients Across 319 Models

Coefficient	Mean	Median	Std. Dev.	% Positive	% Negative	% Zero
β_{PC1}	−0.00027	0.000	0.00143	5.3%	15.7%	79.0%
β_{PC2}	0.00019	0.000	0.00114	11.6%	3.1%	85.3%
β_{PC3}	−0.00000	0.000	0.00107	5.3%	5.0%	89.7%

Notes: Statistics computed across 319 stock-window elastic net models using the “All features” specification (5 PCs + 4 lagged returns). Coefficients are classified as zero if $|\beta| < 10^{-10}$.

The most striking feature of the coefficient distributions is their extreme sparsity. Between 79% and 90% of all coefficient estimates are exactly zero, a direct consequence of the elastic net's ℓ_1 penalty, which shrinks small coefficients entirely to zero rather than merely reducing their magnitude. The cross-validation procedure consistently selects regularization parameters that produce this high degree of sparsity, suggesting that for most stock-window combinations, the principal components do not contain sufficient predictive signal to survive penalization.

Among the non-zero coefficients, meaningful patterns emerge that align with economic intuition. The coefficient on PC1, the level factor capturing overall implied volatility, exhibits a negative bias, with 15.7% of estimates being negative compared to only 5.3% positive. This asymmetry suggests that elevated levels of implied volatility tend to precede lower subsequent returns. One possible interpretation is that high implied volatility reflects temporary investor fear that dissipates without fully materializing, leading to mean reversion in prices. The coefficient on PC2, the slope factor capturing the steepness of the volatility smile, displays the opposite pattern, with 11.6% of estimates positive and only 3.1% negative. A steeper smile, indicating relatively higher implied volatility for out-of-the-money puts, may signal demand for downside protection that does not fully materialize, leading to higher average returns. The coefficient on PC3, the curvature factor, shows no directional bias, with roughly equal proportions of positive and negative estimates among the non-zero values.

3.3.2 Distributional Shape

Figure 3.1 displays kernel density estimates of the coefficient distributions for the first three principal components. Each panel reveals the same characteristic pattern: a massive spike at zero representing the sparse solutions selected by cross-validation, flanked by thin tails of non-zero estimates extending in both directions. The bottom-right panel overlays all three distributions, confirming their similar shapes while highlighting subtle differences in dispersion.

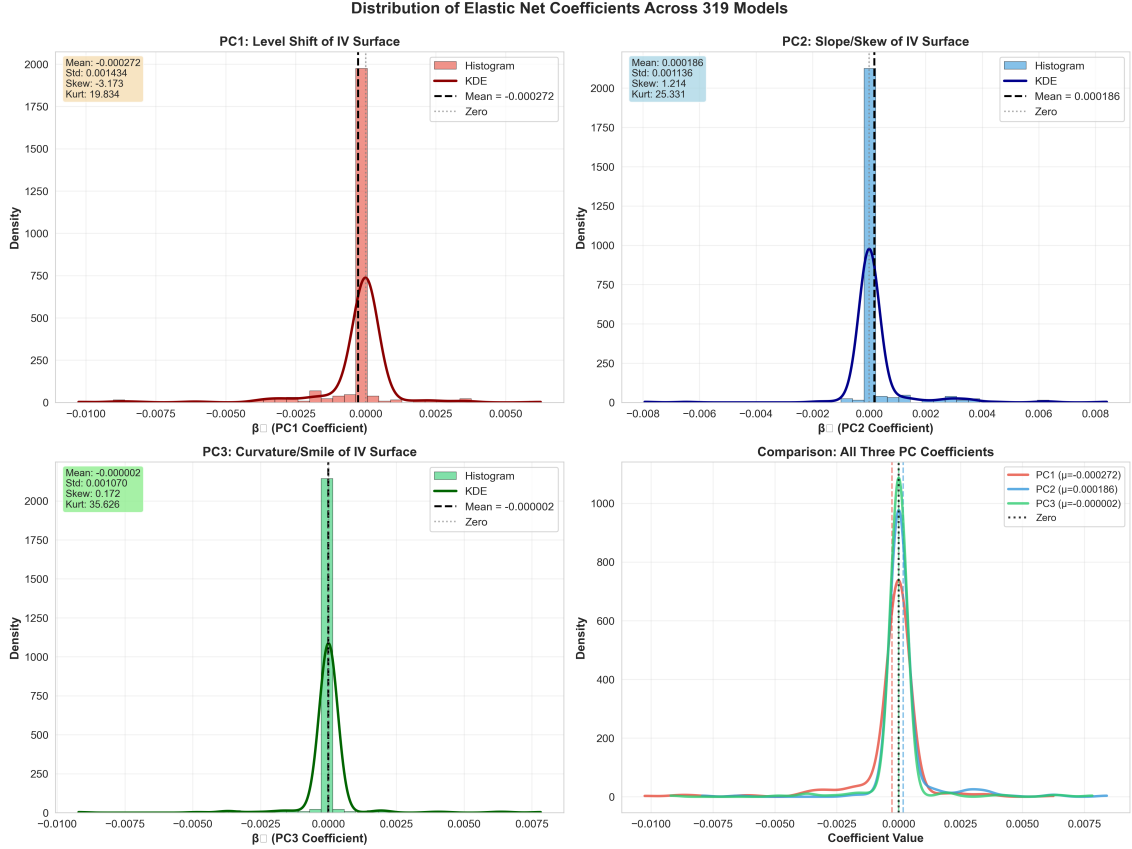


Figure 3.1: Kernel density estimates of elastic net coefficients for PC1, PC2, and PC3 across 319 stock-window models. The bottom-right panel overlays all three distributions.

The formal measures of distributional shape reported in Table 3.8 confirm what the visual inspection suggests. All three coefficient distributions exhibit extreme leptokurtosis, with kurtosis values ranging from 19.8 for PC1 to 35.6 for PC3, far exceeding the value of 3 that characterizes a normal distribution. This excess kurtosis arises from the combination of two features: the concentration of probability mass at zero, which creates a sharp peak, and the presence of occasional large outliers, which extend the tails. The Shapiro-Wilk test decisively rejects the null hypothesis of normality for all three coefficients, with p -values below 0.001 in each case.

The skewness statistics reveal additional structure. The PC1 coefficient distribution is strongly negatively skewed (-3.17), reflecting the preponderance of negative estimates documented in Table 3.7. The PC2 distribution exhibits positive skewness (1.21), con-

Table 3.8: Distributional Shape of Elastic Net Coefficients

Coefficient	Skewness	Kurtosis	Shapiro-Wilk W	p -value
β_{PC1}	-3.17	19.8	0.412	< 0.001
β_{PC2}	1.21	25.3	0.387	< 0.001
β_{PC3}	0.17	35.6	0.298	< 0.001

Notes: Skewness and kurtosis are sample moments. The Shapiro-Wilk test evaluates the null hypothesis that the coefficients are drawn from a normal distribution. A normal distribution has skewness = 0 and kurtosis = 3.

sistent with its tendency toward positive values. The PC3 distribution is approximately symmetric (0.17), aligning with the balanced proportions of positive and negative estimates observed for this factor.

Figure 3.2 presents quantile-quantile plots that compare the empirical coefficient distributions to a theoretical standard normal distribution. The departures from the diagonal reference line are severe and systematic. In each panel, the S-shaped pattern reveals that the empirical distribution has far more mass concentrated near zero than a normal distribution would predict, while simultaneously exhibiting heavier tails with more extreme outliers. This pattern is characteristic of ℓ_1 -regularized estimators and provides visual confirmation that the elastic net is functioning as intended, aggressively shrinking weak signals to zero while allowing genuinely informative coefficients to take non-zero values.

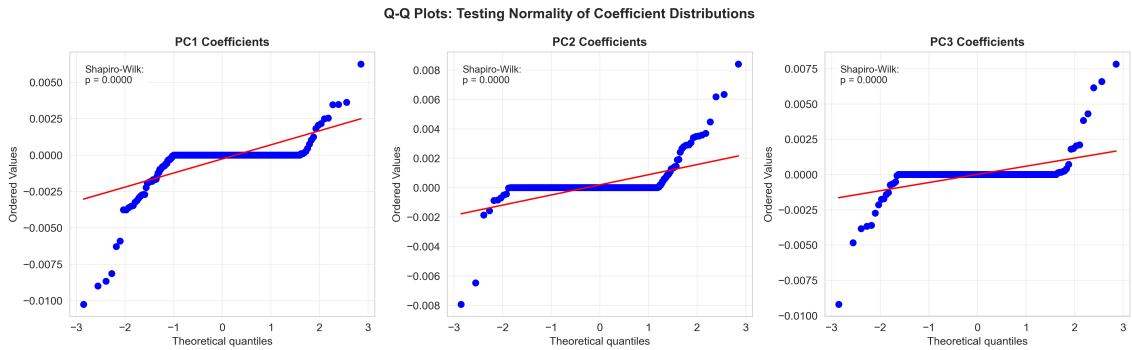


Figure 3.2: Quantile-quantile plots comparing elastic net coefficient distributions to a standard normal. Departures from the diagonal indicate non-normality.

3.3.3 Implications for Interpretation

The coefficient analysis reveals that implied volatility surface factors provide sporadic rather than consistent predictive power. For most stock-window combinations, the elastic net selects a completely sparse solution, implying that the principal components do not improve upon a simple constant forecast. However, for a meaningful minority of cases (roughly 10% to 20% depending on the component), the model identifies non-zero relationships that contribute to the aggregate forecasting performance documented in Section 3.1. This heterogeneity underscores the value of stock-specific estimation, as a pooled model imposing common coefficients across all stocks would miss the idiosyncratic patterns that drive the positive information coefficient. At the same time, the high proportion of zero coefficients suggests that the implied volatility surface is not a universally reliable predictor; its informativeness varies across firms and time periods in ways that the elastic net can adaptively detect.

3.4 Model Performance Across Market Regimes

This section examines how the elastic net model’s forecasting performance varies over time and across different market conditions. Following the methodology of Welch and Goyal (2008), I track the cumulative sum of squared error differences between the model and the prevailing mean benchmark, overlaying this measure with the CBOE Volatility Index (VIX) to assess whether performance deteriorates during periods of market stress.

3.4.1 Cumulative SSE Difference

Figure 3.3 plots the cumulative SSE difference over the full out-of-sample period from September 2016 to August 2023. The black line tracks the running sum of $d_t = (\text{SSE}_{\text{benchmark},t} - \text{SSE}_{\text{model},t})$, where positive values indicate weeks in which the model produced smaller squared errors than the prevailing mean. The blue line displays the

VIX index on the right axis, and the red shaded regions highlight periods when the VIX exceeded 25, a commonly used threshold for elevated market fear.

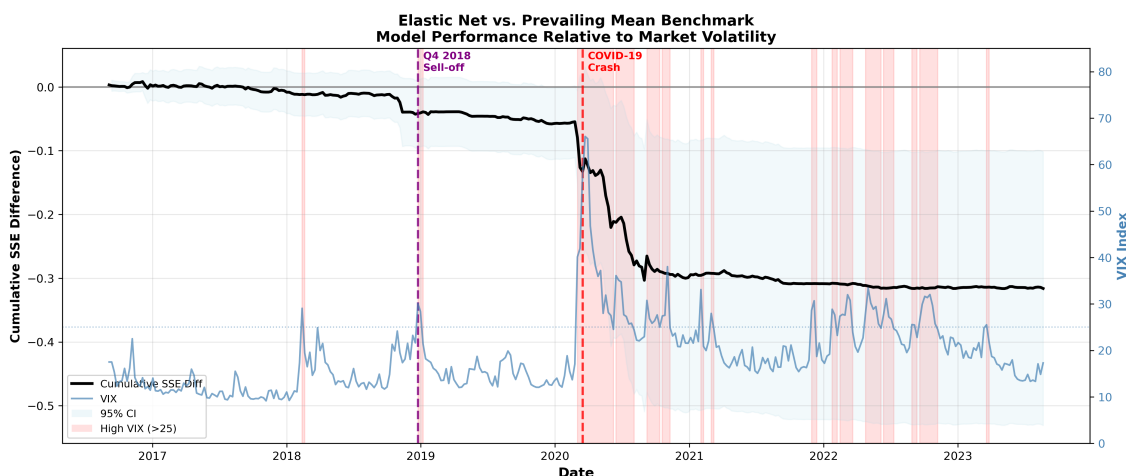


Figure 3.3: Cumulative SSE difference between the elastic net model and the prevailing mean benchmark, plotted alongside the VIX index. Red shaded regions indicate high-volatility periods ($VIX > 25$). Vertical dashed lines mark the Q4 2018 sell-off and the COVID-19 crash. The light blue band represents a 95% confidence interval.

Several important patterns emerge from this visualization. During the first two years of the out-of-sample period (late 2016 through late 2018), the cumulative SSE difference hovers near zero, indicating that the model and the benchmark perform comparably in terms of point prediction accuracy. The model neither consistently outperforms nor underperforms the simple historical mean during this relatively calm market environment, when the VIX generally remained below 20.

The Q4 2018 market sell-off, marked by the purple dashed line, coincides with a modest deterioration in model performance. As volatility spiked in December 2018, the cumulative SSE difference drifts into negative territory, suggesting that the model's predictions became less accurate relative to the benchmark during this period of market stress. However, the magnitude of this deterioration is modest, with the cumulative difference declining by approximately 0.04 over several weeks.

The COVID-19 crash in March 2020 represents a watershed moment for model performance. As the VIX spiked to unprecedented levels above 80, the cumulative SSE

difference plunges sharply, falling from approximately -0.05 to below -0.25 within a matter of weeks. This dramatic deterioration indicates that the model's squared forecast errors substantially exceeded those of the prevailing mean during the most volatile period in recent market history. The extreme returns observed during the COVID crash (both negative during the initial sell-off and positive during the subsequent recovery) were poorly predicted by the implied volatility surface factors, which were themselves undergoing extraordinary dislocations.

Following the COVID crash, the cumulative SSE difference stabilizes at approximately -0.30 and remains relatively flat through the end of the sample period. Notably, the subsequent high-VIX episodes in 2021 and 2022 (including the inflation-driven volatility of early 2022) do not produce comparable deterioration. This pattern suggests that the COVID crash represented a structural break rather than a general pattern of underperformance during volatile markets; the model's incremental performance returned to roughly break-even after the initial shock, even as volatility remained elevated, though the cumulative losses from the COVID period were not recovered.

3.4.2 Pre-COVID Performance

To better visualize the model's performance before it was overwhelmed by the COVID shock, Figure 3.4 presents a two-panel display with the full sample in the top panel and a zoomed view of the pre-COVID period (September 2016 through January 2020) in the bottom panel.

The zoomed panel reveals nuances that are obscured in the full-sample view. During 2017, the cumulative SSE difference fluctuated in a narrow band around zero, with the model occasionally outperforming the benchmark (positive drift) and occasionally underperforming (negative drift). Beginning in mid-2018, however, a gradual negative trend emerges. The cumulative difference drifts steadily downward from approximately zero to roughly -0.06 by the end of 2019, even before the COVID shock. This pre-existing trend suggests that the model's point prediction accuracy was already deteriorating modestly

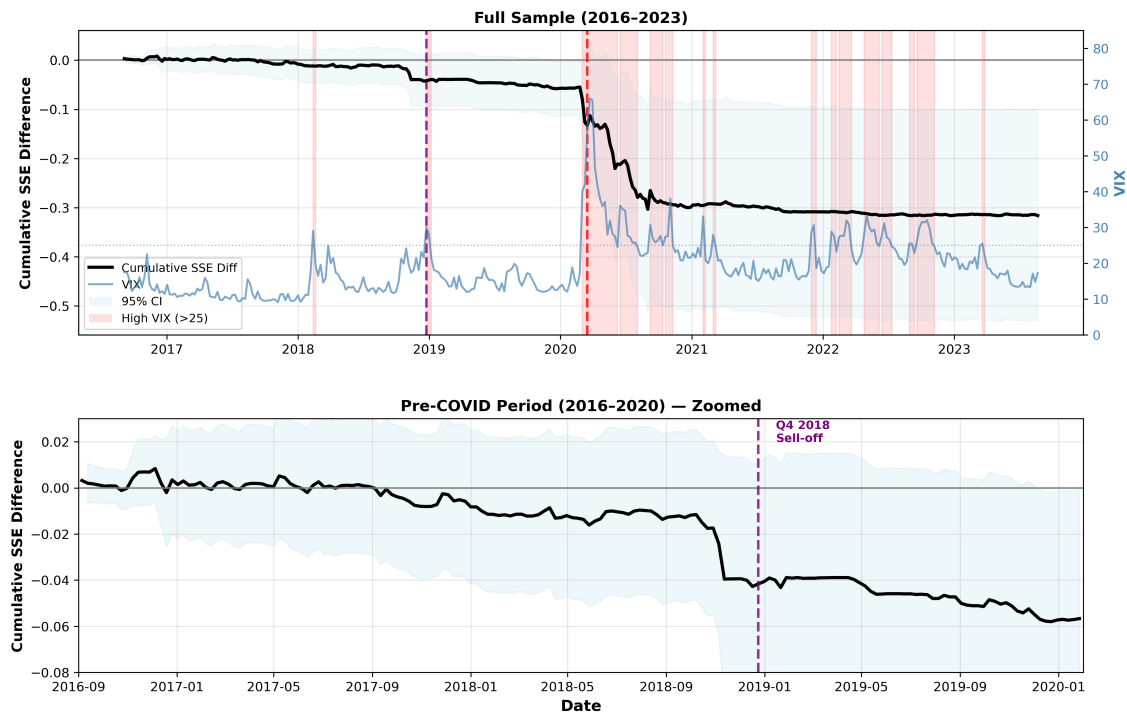


Figure 3.4: Two-panel view of cumulative SSE difference. Top panel: Full sample (2016–2023) with VIX overlay. Bottom panel: Zoomed view of the pre-COVID period (2016–2020).

before the pandemic, though the magnitude of this deterioration was small compared to what followed.

The 95% confidence interval, shown as the light blue band, provides a measure of statistical uncertainty around the cumulative SSE difference. Throughout the pre-COVID period, the confidence interval comfortably includes zero, indicating that the observed underperformance is not statistically distinguishable from noise. Only after the COVID crash does the confidence interval shift entirely below zero, confirming that the post-crash underperformance represents a statistically significant departure from the benchmark.

3.4.3 Reconciling Point Prediction and Ranking Performance

The cumulative SSE analysis presents an apparent puzzle: the elastic net model consistently underperforms the prevailing mean in terms of squared forecast errors, yet as

documented in Section 3.2, the interaction specification achieves a positive information coefficient (0.036), directional accuracy above 55%, and a portfolio Sharpe ratio of 0.618. How can a model that loses to the benchmark in point prediction simultaneously provide useful cross-sectional rankings and generate positive risk-adjusted returns?

The resolution lies in recognizing that the SSE metric and the information coefficient measure fundamentally different aspects of forecast quality. The SSE penalizes large errors heavily due to the squaring operation, making it sensitive to the extreme returns that occur during market crashes. The COVID shock generated weekly returns of $\pm 10\%$ or more for individual stocks, magnitudes that are virtually impossible to predict accurately and that dominate the SSE calculation. In contrast, the information coefficient measures rank correlation, asking only whether stocks with higher predicted returns tend to have higher realized returns, regardless of the magnitude of either.

A model can excel at ranking while failing at point prediction if it correctly identifies relative winners and losers without accurately forecasting return magnitudes. During the COVID crash, for example, the model may have correctly predicted that defensive stocks would outperform cyclicals (generating a positive IC) while simultaneously underestimating the magnitude of returns for both groups by a wide margin (producing large squared errors). This interpretation is consistent with the portfolio results presented in Chapter 4, where the positive IC translates into economically meaningful long-short returns despite the negative out-of-sample R^2 .

The VIX overlay in Figure 3.3 reinforces this interpretation. The model's SSE performance deteriorates most dramatically during periods of extreme volatility, when return magnitudes are most difficult to predict. However, the information coefficient (which depends only on rankings) may remain positive even during these turbulent periods, provided that the model's relative predictions retain validity. This divergence between point prediction accuracy and ranking ability represents a central finding of this thesis and motivates the portfolio-centric evaluation approach developed in the following chapter.

Chapter 4

Portfolio Construction and Economic Value

4.1 Portfolio Construction Methodology

4.1.1 Long-Short Strategy Design

To evaluate the economic value of the return forecasts developed in Chapter 3, I construct quintile-sorted long-short portfolios using the predictions from each model specification. At the end of each week t , I obtain predicted returns $\hat{r}_{i,t+1}$ for all stocks i in the universe from the elastic net model estimated on data through week t . Stocks are then ranked by their predicted returns and sorted into quintiles, with Quintile 5 containing the top 20% of stocks by predicted return and Quintile 1 containing the bottom 20%.

The long-short portfolio takes a long position in Quintile 5 stocks and a short position in Quintile 1 stocks. Both legs are equal-weighted, meaning each stock within a quintile receives identical portfolio weight regardless of market capitalization or other characteristics. The long-short return for week $t + 1$ is computed as

$$r_{t+1}^{LS} = \bar{r}_{t+1}^{Q5} - \bar{r}_{t+1}^{Q1}, \quad (4.1)$$

where $\bar{r}_{t+1}^{Q5}$ denotes the equal-weighted average return of stocks in the top quintile and

$\bar{r}_{t+1}^{Q1}$ denotes the equal-weighted average return of stocks in the bottom quintile. This construction isolates the model’s ability to distinguish relative winners from losers. If the predicted returns contain information about the cross-section, stocks in Quintile 5 should outperform those in Quintile 1 on average, generating positive long-short returns.

The portfolio is rebalanced weekly, with positions established at the Monday close and held for one week. With 46 stocks in the universe, each quintile contains approximately 9 stocks, providing reasonable diversification within each leg. Weeks with fewer than 10 stocks available are excluded to ensure meaningful quintile formation. The long-short portfolio is dollar-neutral by construction, with equal dollar amounts allocated to the long and short legs.

4.1.2 Performance Metrics

I evaluate portfolio performance using several standard metrics. The annualized return is computed as $(1 + \bar{r}^{LS})^{52} - 1$, where $\bar{r}^{LS}$ is the mean weekly long-short return. This compounding approach accounts for the reinvestment of weekly profits into subsequent positions. The annualized volatility is calculated as $\sigma^{LS} \times \sqrt{52}$, where σ^{LS} is the standard deviation of weekly returns. The Sharpe ratio, defined as annualized return divided by annualized volatility, provides a risk-adjusted measure of performance. Following convention in the academic literature, I assume a zero risk-free rate when computing the Sharpe ratio.

Beyond the Sharpe ratio, I report several additional metrics to characterize the strategy’s risk profile. The maximum drawdown measures the largest peak-to-trough decline in cumulative portfolio value, capturing the worst-case loss an investor would have experienced. The win rate reports the proportion of weeks with positive long-short returns, indicating the consistency of the strategy’s performance. I also compute separate hit rates for the long and short legs, measuring the proportion of weeks in which each leg’s return has the correct sign (positive for the long leg, negative for the short leg).

To assess ranking ability directly, I calculate the weekly information coefficient, de-

defined as the Spearman rank correlation between predicted and realized returns across all stocks in each week. A positive information coefficient indicates that the model's cross-sectional rankings align with subsequent realized performance, even if the predicted magnitudes are inaccurate.

4.1.3 Transaction Cost Assumptions

The baseline results assume zero transaction costs to isolate the gross economic value of the forecasts. However, implementing the strategy requires weekly rebalancing of both long and short positions, incurring trading costs that can substantially erode returns. To assess practical viability, I conduct sensitivity analysis under varying transaction cost assumptions.

I assume 100% weekly turnover in each leg, meaning all positions are closed and reopened at each rebalancing date. This assumption represents an upper bound on actual turnover, as some positions may persist across consecutive weeks when rankings remain stable. Under this assumption, total weekly transaction costs equal twice the one-way cost, accounting for trading in both the long and short legs. I examine one-way transaction costs ranging from 5 to 30 basis points, spanning the range from low-cost institutional execution to retail trading. The breakeven transaction cost, at which the strategy's net Sharpe ratio equals zero, provides a useful benchmark for assessing implementation feasibility.

4.2 Model Comparison and Selection

4.2.1 Candidate Models

Chapter 3 identified three elastic net specifications as the most promising candidates for portfolio construction based on their information coefficients and directional accuracy. The first specification, denoted "5 PCs + Lag1," includes the five principal components extracted from the implied volatility surface plus a single lagged return, yielding six fea-

tures. This parsimonious model achieved the highest information coefficient among simple specifications ($IC = 0.034$) in the forecast evaluation. The second specification, “All features,” augments the five principal components with four lagged returns, yielding nine features. The third specification, “With interactions,” further expands the feature set by adding all 36 pairwise interaction terms between the nine base features, resulting in 45 total predictors.

These three specifications represent different points on the complexity spectrum. The 5 PCs + Lag1 model prioritizes parsimony and interpretability, using only six predictors with clear economic meaning. The All features model adds additional momentum controls through multiple lagged returns. The interaction model allows for nonlinear relationships between predictors, capturing the possibility that the effect of one variable on returns depends on the value of another. The elastic net’s regularization penalty can shrink irrelevant interactions to zero, allowing the model to adaptively select which nonlinearities contain predictive information.

4.2.2 Portfolio Performance Comparison

Table 4.1 presents the long-short portfolio performance for each of the three candidate specifications. All portfolios follow the methodology described in Section 4.1, going long the top quintile and short the bottom quintile of stocks ranked by predicted returns.

Table 4.1: Long-Short Portfolio Performance by Model Specification

Model	R^2_{OOS}	IC	Sharpe	Ann. Return	Max DD
5 PCs + Lag1	−0.024	0.034	0.615	11.98%	−39.7%
All features (no interactions)	−0.023	0.010	0.562	10.76%	−39.6%
With interactions	−0.030	0.036	0.618	12.33%	−40.4%

Notes: All portfolios use the long-short quintile strategy described in Section 4.1. Sharpe ratio and annualized return assume zero transaction costs. Max DD denotes maximum drawdown.

The interaction model achieves the highest Sharpe ratio (0.618) and annualized return (12.33%), followed by the 5 PCs + Lag1 model (Sharpe 0.615, return 11.98%) and the All

features model (Sharpe 0.562, return 10.76%). Maximum drawdowns are similar across all three specifications, ranging from 39.6% to 40.4%, indicating comparable downside risk despite the differences in return generation.

4.2.3 The Accuracy-Ranking Paradox

A striking pattern emerges from Table 4.1 that constitutes a central finding of this thesis. The model with the worst point prediction accuracy achieves the best portfolio performance. Table 4.2 summarizes this paradox by ranking the three specifications along each performance dimension.

Table 4.2: The Accuracy-Ranking Paradox

Model	R^2_{OOS}	IC	Sharpe
With interactions	−0.030 (worst)	0.036 (best)	0.618 (best)
5 PCs + Lag1	−0.024	0.034	0.615
All features	−0.023 (best)	0.010 (worst)	0.562 (worst)

Notes: Parenthetical labels indicate best and worst performance within each column.

The interaction model has $R^2_{\text{OOS}} = -0.030$, the most negative among all specifications, indicating that its squared forecast errors exceed those of the prevailing mean benchmark by 3%. By traditional forecast evaluation standards, this model performs worst. Yet the same model achieves the highest information coefficient (0.036) and generates the highest Sharpe ratio (0.618). Conversely, the All features model achieves the best R^2_{OOS} (−0.023) but the lowest information coefficient (0.010) and only the second-highest Sharpe ratio.

This divergence between point prediction accuracy and portfolio performance arises because R^2 and the Sharpe ratio measure fundamentally different aspects of forecast quality. The R^2 metric penalizes errors in predicted magnitudes, with larger errors receiving quadratically greater penalties. A model that correctly predicts the direction of returns but overestimates their magnitude will be penalized heavily. In contrast, the Sharpe ratio of a long-short portfolio depends only on whether the model correctly ranks stocks. A model predicting returns of 1%, 2%, and 3% for three stocks generates identical portfolio

weights to a model predicting 10%, 20%, and 30%, because both models rank the stocks in the same order.

The interaction terms appear to capture nonlinear relationships that improve cross-sectional rankings while introducing noise that degrades point predictions. For example, the predictive content of implied volatility level may depend on the volatility smile slope, or momentum signals from lagged returns may vary with the shape of the volatility surface. These conditional relationships help distinguish relative winners from losers even though they increase the variance of point forecasts. The elastic net's regularization ensures that only interactions with genuine predictive content survive, as irrelevant terms are shrunk to zero.

This finding has important implications for applied financial forecasting. Researchers and practitioners commonly evaluate return predictions using R^2 or mean squared error, metrics that emphasize point accuracy. However, when the ultimate goal is portfolio construction, these metrics may be misleading. A model that fails by traditional standards can still generate substantial economic value if it correctly ranks securities. The results suggest that forecast evaluation should be aligned with the intended application, with ranking metrics such as information coefficient and portfolio Sharpe ratio taking precedence when the goal is security selection.

4.2.4 Model Selection

Based on the portfolio performance comparison, I select the elastic net model with interaction terms as the primary specification for the detailed analysis that follows. Table 4.3 summarizes the key characteristics of the selected model.

The selection of the interaction model, despite its inferior point prediction accuracy, reflects the primacy of economic value over statistical accuracy in this analysis. The goal is to translate forecasts into profitable trading strategies, and the interaction model achieves this goal most effectively. The 45-feature specification may raise concerns about overfitting, but the elastic net's regularization addresses this issue by shrinking coeffi-

Table 4.3: Selected Model Characteristics

Characteristic	Value
Feature specification	5 PCs + 4 lags + 36 interactions
Number of features	45
Number of stock-window models	319
Out-of-sample R^2	−0.030
Information coefficient	0.036
Directional accuracy	55.5%
Portfolio Sharpe ratio	0.618

cients toward zero. The positive out-of-sample Sharpe ratio provides direct evidence that the model captures genuine predictive patterns rather than in-sample noise.

The remainder of this chapter presents detailed results for the selected model, examining its performance characteristics, sensitivity to transaction costs, and behavior across different market regimes.

4.3 Main Results

4.3.1 Performance Summary

Table 4.4 presents comprehensive performance statistics for the long-short portfolio constructed using the elastic net model with interaction terms. The sample spans 364 weeks from September 2016 through August 2023, with an average of 45.6 stocks available each week and approximately 9.7 positions in each leg.

The portfolio generates an annualized return of 12.33% with volatility of 19.96%, yielding a Sharpe ratio of 0.618. Over the full sample period, an initial investment of \$1 grows to \$1.96, representing a cumulative return of 96.35%. The return distribution exhibits mild negative skewness (−0.245) and near-normal kurtosis (0.529), suggesting that extreme negative returns are slightly more common than extreme positive returns but without the heavy tails often observed in individual stock returns. The median weekly return (0.329%) exceeds the mean (0.224%), consistent with the negative skewness.

Table 4.4: Long-Short Portfolio Performance Summary

<i>Panel A: Sample Characteristics</i>	
Sample period	Sep 2016 to Aug 2023
Number of weeks	364
Average stocks per week	45.6
Average positions per leg	9.7
<i>Panel B: Return Statistics</i>	
Mean weekly return	0.224%
Median weekly return	0.329%
Standard deviation (weekly)	2.768%
Skewness	−0.245
Kurtosis	0.529
<i>Panel C: Annualized Performance</i>	
Annualized return	12.33%
Annualized volatility	19.96%
Sharpe ratio	0.618
<i>Panel D: Risk Metrics</i>	
Win rate	53.6%
Long leg hit rate	61.8%
Short leg hit rate	43.4%
Maximum drawdown	−40.41%
Final cumulative return	96.35%
<i>Panel E: Information Coefficient</i>	
Mean weekly IC	0.045
Median weekly IC	0.061
Proportion of weeks IC > 0	57.4%

Notes: Results for the elastic net model with interaction terms (45 features). All returns are gross of transaction costs. Hit rate measures the proportion of weeks in which each leg's return has the correct sign (positive for long, negative for short). The weekly IC reported in Panel E is the average cross-sectional Spearman correlation computed each week across all stocks, which differs from the model-averaged IC of 0.036 reported in Chapter 3.

The win rate of 53.6% indicates that the strategy generates positive returns in slightly more than half of all weeks. While this may appear modest, the asymmetry between winning and losing weeks drives the positive overall performance. The mean information coefficient of 0.045 confirms that the model's predictions are positively correlated with realized returns, with 57.4% of weeks exhibiting positive IC.

Decomposing performance by leg reveals an important asymmetry. The long leg hit rate of 61.8% substantially exceeds the short leg hit rate of 43.4%, indicating that the model identifies future winners more reliably than future losers. This pattern is common in long-short equity strategies and may reflect several factors, including the general upward drift of equity markets, the difficulty of profiting from short positions due to borrowing costs and short-sale constraints, and the possibility that implied volatility signals are more informative about relative outperformance than relative underperformance.

4.3.2 Cumulative Wealth

Figure 4.1 displays the growth of \$1 invested in the long-short strategy under different transaction cost assumptions. The gross performance line (0 bps) shows substantial wealth accumulation over the sample period, nearly doubling from \$1.00 to \$1.96. The trajectory reveals several distinct phases. From late 2016 through early 2020, the portfolio exhibits steady growth with moderate volatility. The COVID-19 crash in March 2020 creates a sharp drawdown followed by a rapid recovery, with the portfolio reaching new highs by mid-2020. Performance remains strong through early 2021 before a prolonged drawdown period extends through 2022. The strategy recovers sharply in 2023, ending the sample near its all-time high.

The figure also illustrates the dramatic impact of transaction costs on cumulative wealth. At 5 basis points one-way, the final portfolio value falls from \$1.96 to \$1.36, representing a cumulative return of 36.49% rather than 96.35%. At 10 basis points, the portfolio ends the sample slightly below its starting value at \$0.95, erasing all gross profits. At 15 and 20 basis points, the strategy generates substantial losses, with final values

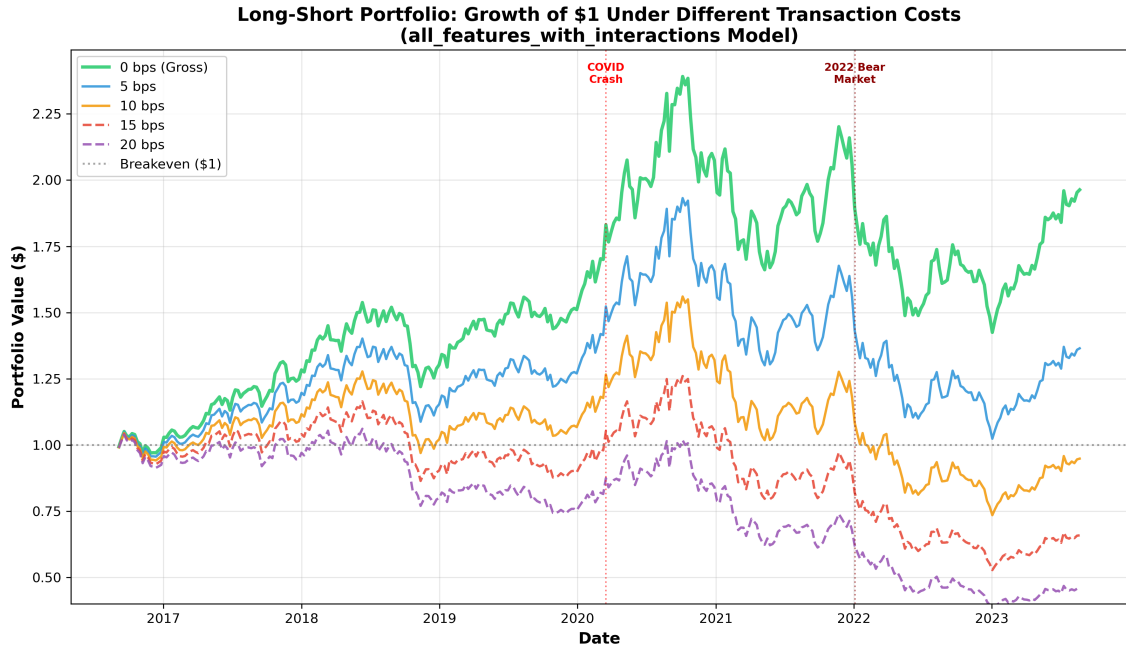


Figure 4.1: Growth of \$1 invested in the long-short portfolio under different transaction cost assumptions. Lines correspond to one-way transaction costs of 0, 5, 10, 15, and 20 basis points. Vertical dashed lines mark the COVID-19 crash (March 2020) and the onset of the 2022 bear market.

of \$0.66 and \$0.46 respectively. These results highlight the critical importance of implementation efficiency for strategies with high turnover.

Figure 4.2 presents the cumulative wealth alongside a drawdown panel, providing additional insight into the strategy’s risk characteristics. The top panel compares gross returns to net returns at 10 basis points, illustrating how transaction costs gradually erode performance over time. The divergence between the two lines widens as the sample progresses, reflecting the cumulative impact of weekly trading costs. The bottom panel displays the drawdown from peak cumulative wealth. The maximum drawdown of 40.41% occurs in early January 2023, following the difficult 2022 bear market during which the strategy lost 27.8%. This substantial drawdown represents a significant risk that investors must be willing to bear in exchange for the strategy’s long-run expected returns.

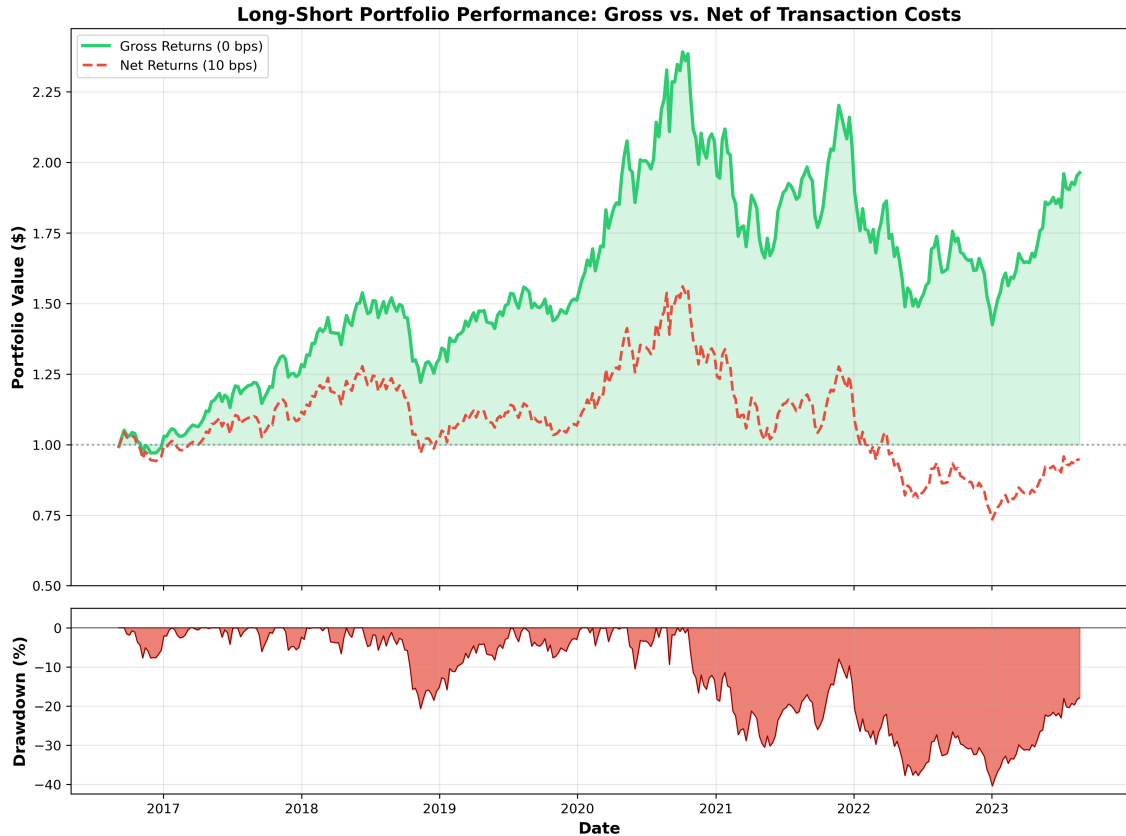


Figure 4.2: Top panel: Cumulative wealth for gross returns and net of 10 bps transaction costs. Bottom panel: Drawdown from peak cumulative wealth.

4.3.3 Transaction Cost Sensitivity

Table 4.5 presents a detailed sensitivity analysis of portfolio performance to transaction costs. The analysis assumes 100% weekly turnover in each leg, representing an upper bound on actual trading costs since some positions may persist across consecutive weeks when rankings remain stable.

The strategy breaks even at approximately 11.2 basis points one-way, the cost level at which the Sharpe ratio equals zero. This breakeven threshold has important implications for practical implementation. At 5 basis points, achievable for large institutional investors with efficient execution infrastructure and direct market access, the strategy retains a Sharpe ratio of 0.333 and annualized return of 6.65%. While substantially reduced from the gross performance, these figures remain economically meaningful. At 10 ba-

Table 4.5: Transaction Cost Sensitivity Analysis

TC (one-way)	Weekly Cost	Net Ann. Return	Net Sharpe	Final Wealth
0 bps	0.000%	12.33%	0.618	\$1.96
5 bps	0.100%	6.65%	0.333	\$1.36
10 bps	0.200%	1.25%	0.063	\$0.95
11.2 bps	0.224%	0.00%	0.000	\$1.00
15 bps	0.300%	−3.88%	−0.195	\$0.66
20 bps	0.400%	−8.76%	−0.439	\$0.46

Notes: Assumes 100% weekly turnover in each leg. Weekly cost equals twice the one-way cost to account for trading both legs. Final wealth shows the terminal value of \$1 invested at the start of the sample. The breakeven transaction cost (bold) is calculated by linear interpolation.

sis points, more typical of institutional equity trading, the strategy generates a marginal Sharpe ratio of 0.063 and annualized return of just 1.25%, barely compensating investors for the associated risks. At 15 basis points or higher, the strategy generates negative expected returns and should not be implemented.

These results suggest that the strategy’s economic viability depends critically on implementation efficiency. Low-cost traders such as quantitative hedge funds and proprietary trading firms with transaction costs below 10 basis points may find the strategy profitable. Institutional investors facing moderate trading costs between 10 and 15 basis points face marginal or negative expected returns that likely do not justify the operational complexity and risk of implementation. Retail investors with high trading costs exceeding 20 basis points cannot profitably implement the strategy under any reasonable assumptions.

The sensitivity analysis also suggests potential improvements through reduced turnover. A modified strategy that rebalances less frequently or trades only when predicted rankings change substantially could lower transaction costs while preserving much of the gross return. Such modifications would reduce the strategy’s responsiveness to new information but might improve net performance by economizing on trading costs.

4.4 Performance Across Time

4.4.1 Market Regime Analysis

The long-short strategy's performance varies substantially across different market environments. Table 4.6 presents performance statistics for six distinct market regimes spanning the sample period.

Table 4.6: Long-Short Performance by Market Regime

Regime	Period	Weeks	Mean Ret	Std Ret	Sharpe	Win Rate
Pre-COVID Bull	Sep 2016 to Feb 2020	182	0.300%	2.100%	1.03	54.9%
COVID Crash	Mar 2020	5	1.783%	4.144%	3.10	60.0%
COVID Recovery	Apr to Dec 2020	39	0.430%	3.821%	0.81	51.3%
2021 Bull	Jan to Dec 2021	52	0.041%	3.391%	0.09	50.0%
2022 Bear	Jan to Dec 2022	52	−0.572%	3.277%	−1.26	48.1%
2023 Recovery	Jan to Aug 2023	34	0.849%	2.222%	2.75	61.8%

Notes: Sharpe ratios computed as mean weekly return divided by weekly standard deviation, multiplied by $\sqrt{52}$. The COVID Crash period includes weeks when the VIX exceeded 50.

Several patterns emerge from the regime analysis. The strategy performs best during periods of elevated volatility and market stress. The COVID crash period achieves the highest Sharpe ratio (3.10) despite containing only five weeks of data. During this period of extreme market dislocation, the implied volatility surface factors appear to have been particularly informative about cross-sectional return differences. The model correctly identified which stocks would decline most severely and which would prove more resilient, generating an average weekly return of 1.783%. The 2023 recovery period exhibits similarly strong performance with a Sharpe ratio of 2.75, suggesting that the strategy capitalizes on return dispersion during market transitions.

Conversely, the strategy struggles during trending markets with low cross-sectional dispersion. The 2021 bull market represents a period of near-zero performance, with a Sharpe ratio of just 0.09 and a win rate of exactly 50%. During this period of broad-based equity gains driven by accommodative monetary policy and economic reopening,

the implied volatility surface provided little guidance about which stocks would outperform. The 2022 bear market represents the worst performance period, with a Sharpe ratio of -1.26 and an average weekly loss of 0.572% . The sustained selling pressure driven by inflation concerns and aggressive Federal Reserve rate hikes created a challenging environment in which the cross-sectional signals from implied volatility failed to identify relative winners and losers.

The pre-COVID bull market from September 2016 through February 2020 represents the longest continuous regime and exhibits solid performance with a Sharpe ratio of 1.03 . This extended period of relatively stable market conditions allowed the strategy to generate consistent, if unspectacular, returns. The COVID recovery period from April through December 2020 shows moderate performance with a Sharpe ratio of 0.81 , as the initial volatility of the crash subsided but cross-sectional dispersion remained elevated.

These regime-dependent results have important implications for investors considering the strategy. The strong performance during high-volatility periods suggests that the implied volatility surface is most informative when option markets are actively repricing risk across individual stocks. During calmer periods or sustained directional moves, the volatility surface may contain less cross-sectional information, reducing the strategy's effectiveness. Investors should expect extended periods of underperformance, particularly during trending markets, even if the long-run expected returns are positive.

4.4.2 Annual Performance

Table 4.7 presents year-by-year performance statistics for the long-short strategy. The decomposition reveals substantial variation in annual returns, ranging from -27.8% in 2022 to $+37.3\%$ in 2020.

The strategy generates positive returns in five of eight calendar years, with three losing years in 2016, 2021, and 2022. The partial year of 2016 shows a modest loss of 1.0% , though this figure is based on only 17 weeks of data and may not be representative. The 2022 loss of 27.8% represents the strategy's worst annual performance and coincides with

Table 4.7: Annual Long-Short Performance

Year	Weeks	Mean Weekly	Std Weekly	Annual Return	Sharpe	Win Rate
2016	17	−0.042%	1.858%	−1.0%	−0.16	41.2%
2017	52	0.464%	1.884%	26.1%	1.78	63.5%
2018	53	0.110%	2.431%	4.4%	0.33	50.9%
2019	52	0.304%	1.925%	16.0%	1.14	51.9%
2020	52	0.678%	3.673%	37.3%	1.33	55.8%
2021	52	0.041%	3.391%	−0.8%	0.09	50.0%
2022	52	−0.572%	3.277%	−27.8%	−1.26	48.1%
2023	34	0.849%	2.222%	32.2%	2.75	61.8%

Notes: 2016 includes only September through December (17 weeks). 2023 includes only January through August (34 weeks). Annual return computed as the compounded product of weekly returns.

the difficult bear market regime discussed in the preceding subsection.

Figure 4.3 visualizes the annual return pattern, highlighting the alternation between strong positive years and the two losing periods.

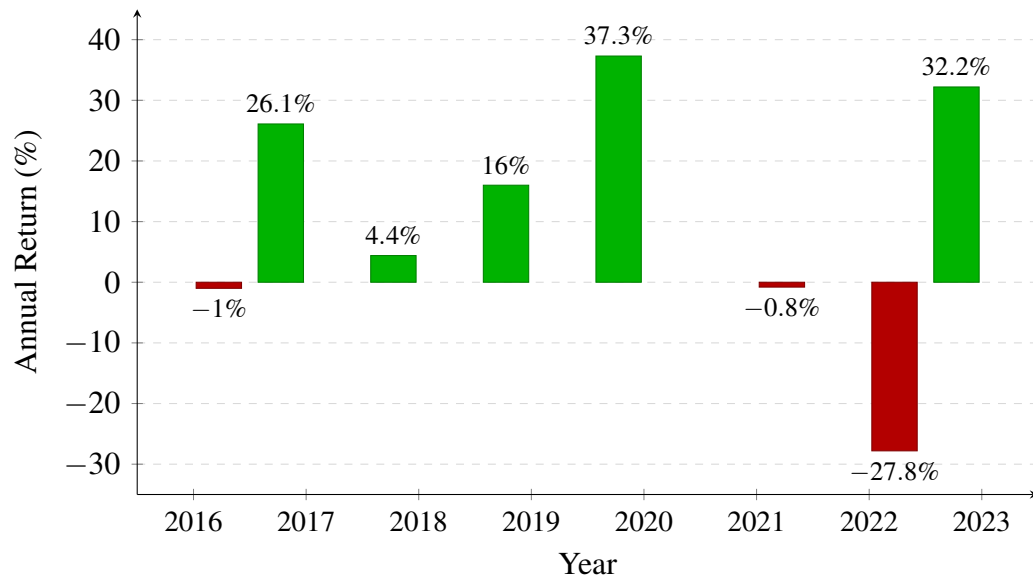


Figure 4.3: Annual long-short returns (gross of transaction costs). Green bars indicate positive years and red bars indicate negative years.

The best performing year is 2020 with a return of 37.3% and Sharpe ratio of 1.33. This strong performance reflects the strategy's ability to capitalize on the extreme volatility

and cross-sectional dispersion created by the COVID-19 pandemic. The second-best year is 2023 with a return of 32.2% and Sharpe ratio of 2.75, driven by the market recovery following the 2022 bear market. The third-best year is 2017 with a return of 26.1% and the highest full-year Sharpe ratio of 1.78, representing a period of steady gains with relatively low volatility.

The year-to-year variation in performance underscores the importance of a long investment horizon. An investor who began the strategy at the start of 2022 would have experienced a loss of 27.8% before the strong 2023 recovery. Similarly, the near-zero return in 2021 following the exceptional 2020 performance might have led some investors to abandon the strategy just before its 2023 rebound. The five profitable years out of eight suggest that the strategy has positive expected returns over longer horizons, but investors must be prepared for individual years of substantial underperformance.

4.4.3 Summary

The portfolio analysis reveals that the elastic net model with interaction terms generates economically meaningful long-short returns before transaction costs. The gross Sharpe ratio of 0.618, while not exceptional by hedge fund standards, represents a substantial improvement over the negative R^2 that characterizes the model's point prediction accuracy. This divergence between statistical and economic performance measures constitutes a central finding of this thesis. The model with the worst point prediction accuracy among the candidate specifications achieves the best portfolio Sharpe ratio, demonstrating that cross-sectional ranking ability matters more than point accuracy for long-short equity strategies.

Several important caveats temper these results. Transaction costs substantially erode returns, with the strategy breaking even at approximately 11.2 basis points one-way. At realistic institutional trading costs of 10 to 15 basis points, the strategy generates marginal or negative expected returns. Performance varies dramatically across market regimes, with the strategy performing best during high-volatility periods such as the COVID crash

and struggling during trending markets such as the 2022 bear market. The maximum drawdown of 40.4% and three losing calendar years highlight the substantial risks that accompany the strategy's positive expected returns.

These results suggest that implied volatility surfaces contain information useful for cross-sectional stock selection, but extracting economic value from this information requires low-cost implementation and tolerance for extended periods of underperformance. The strategy may be viable for quantitative hedge funds and proprietary trading firms with efficient execution infrastructure, but is unlikely to be profitable for institutional or retail investors facing higher transaction costs.

Conclusion

This thesis examined whether options implied volatility surfaces contain information useful for predicting weekly stock returns. Using principal components extracted from implied volatility surfaces of large-cap US stocks, I employed elastic net regression within a rolling window framework to forecast one-week-ahead returns and evaluated the economic value of these forecasts through long-short portfolio construction.

The central finding of this thesis is a striking divergence between statistical accuracy and economic value. The model with the worst point prediction accuracy, as measured by out-of-sample R^2 , achieves the best portfolio performance, as measured by Sharpe ratio. This paradox arises because point prediction metrics penalize errors in forecasted magnitudes, while long-short portfolio returns depend only on correct cross-sectional rankings. For portfolio construction, the ability to identify relative winners and losers matters more than predicting exact return magnitudes. This finding cautions researchers and practitioners against relying solely on traditional forecast accuracy metrics when the ultimate objective is security selection.

The analysis also demonstrates that regularization is essential in the low signal-to-noise environment of return prediction. Elastic net substantially outperforms both ordinary least squares and Random Forest, with the latter suffering severe overfitting when flexible tree-based models exploit spurious patterns in lagged returns. The interaction terms between implied volatility factors and past returns improve cross-sectional rankings despite degrading point predictions, suggesting that nonlinear relationships between predictors contain information about relative stock performance.

However, the economic viability of the strategy depends critically on implementation costs. Transaction costs substantially erode gross returns, and the strategy breaks even at a relatively low cost threshold. Performance also varies across market regimes, with the strategy performing best during high-volatility periods and struggling during trending markets. These results suggest that implied volatility surfaces contain useful cross-sectional information, but extracting economic value requires low-cost implementation and tolerance for extended periods of underperformance. Several limitations qualify these findings, including survivorship bias from a fixed stock universe, a limited sample period, and sensitivity to transaction cost assumptions.

Future research could extend this analysis in several directions. Testing the methodology on mid-cap and small-cap stocks would assess whether the cross-sectional signals from implied volatility are stronger or weaker outside the large-cap universe. Implementing longer holding periods or conditional rebalancing rules could reduce turnover and improve net returns. Combining implied volatility factors with established return predictors such as value, momentum, and quality might yield complementary signals that enhance portfolio performance. Finally, alternative machine learning methods with appropriate regularization, such as neural networks with dropout or gradient boosting with early stopping, could potentially capture nonlinear relationships more effectively than the interaction terms employed in this thesis.

Bibliography

- Abdi, Farshid and Angelo Rinaldo (2017). “A Simple Estimation of Bid-Ask Spreads from Daily Close, High, and Low Prices”. In: *The Review of Financial Studies* 30.12, pp. 4437–4473.
- Açıklan, Özgür Şafak and Zeynep Önder (2024). “Option-based Variables and Future Stock Returns in Normal Times and Recessions”. In: *Borsa Istanbul Review* 24.Supplement 1, pp. 80–87.
- An, Byeong-Je et al. (2014). “The Joint Cross Section of Stocks and Options”. In: *Journal of Finance* 69.5, pp. 2279–2337. DOI: 10.1111/jofi.12181.
- Avellaneda, Marco et al. (2020). “PCA for Implied Volatility Surfaces”. In: *Journal of Financial Data Science* 2.2, pp. 85–109.
- Bali, Turan G. and Armen Hovakimian (2009). “Volatility Spreads and Expected Stock Returns”. In: *Management Science* 55.11, pp. 1797–1812. DOI: 10.1287/mnsc.1090.1063.
- Bikker, Jacob A., Laura Spierdijk, and Pieter J. van der Sluis (2007). “Market Impact Costs of Institutional Equity Trades”. In: *Journal of International Money and Finance* 26.6, pp. 974–1000.
- Breiman, Leo (2001). “Random forests”. In: *Machine Learning* 45.1, pp. 5–32.
- Campbell, John Y and Samuel B Thompson (2008). “Predicting excess stock returns out of sample: Can anything beat the historical average?” In: *The Review of Financial Studies* 21.4, pp. 1509–1531.

- Cenesizoglu, Tolga and Allan Timmermann (2012). “Do Return Prediction Models Add Economic Value?” In: *Journal of Banking & Finance* 36.11, pp. 2974–2987.
- Christoffersen, Peter, Mathieu Fournier, and Kris Jacobs (2017). “The Factor Structure in Equity Options”. In: *The Review of Financial Studies* 31.2, pp. 595–637. DOI: 10.1093/rfs/hhx089.
- Cont, Rama and José da Fonseca (2002). “Dynamics of Implied Volatility Surfaces”. In: *Quantitative Finance* 2.1, pp. 45–60. DOI: 10.1088/1469-7688/2/1/304.
- Cremers, Martijn and David Weinbaum (2010). “Deviations from Put-Call Parity and Stock Return Predictability”. In: *Journal of Financial and Quantitative Analysis* 45.2, pp. 335–367.
- Doran, James and Kevin Krieger (2010). “Implications for Asset Returns in the Implied Volatility Skew”. In: *Financial Analysts Journal* 66.1, pp. 65–76.
- Eksi, Asli and Saurabh Roy (2025). *Stock Return Predictability of Realized-Implied Volatility Spread and Abnormal Turnover*. Working Paper 5234112. SSRN. URL: <https://ssrn.com/abstract=5234112> (visited on 01/02/2026).
- Fengler, Matthias R., Wolfgang K. Härdle, and Enno Mammen (2007). “A Semiparametric Factor Model for Implied Volatility Surface Dynamics”. In: *Journal of Financial Econometrics* 5.2, pp. 189–218. DOI: 10.1093/jjfinec/nbm005.
- Fengler, Matthias R., Wolfgang K. Härdle, and Christophe Villa (2003). “The Dynamics of Implied Volatilities: A Common Principal Components Approach”. In: *Review of Derivatives Research* 6.3, pp. 179–202. DOI: 10.1023/B:REDR.0000004823.77464.2d.
- Fu, Xi et al. (2016). “Option-Implied Volatility Measures and Stock Return Predictability”. In: *Journal of Derivatives* 24.1, pp. 58–78.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). “Empirical Asset Pricing via Machine Learning”. In: *The Review of Financial Studies* 33.5, pp. 2223–2273.
- Guo, Hui and Buhui Qiu (2014). “Options-Implied Variance and Future Stock Returns”. In: *Journal of Banking and Finance* 44, pp. 93–113. DOI: 10.1016/j.jbankfin.2014.04.002.

- Jahan-Parvar, Mohammad R. and Filip Zikes (2023). “When Do Low-Frequency Measures Really Measure Effective Spreads? Evidence From Equity and Foreign Exchange Markets”. In: *The Review of Financial Studies* 36.10, pp. 4190–4232.
- Johnson, Travis L. and Eric C. So (2012). “The Option to Stock Volume Ratio and Future Returns”. In: *Journal of Financial Economics* 106.2, pp. 262–286.
- Kociński, Marek (2014). “Transaction Costs and Market Impact in Investment Management”. In: *Financial Internet Quarterly* 10.4, pp. 28–35.
- Lewellen, Jonathan (2015). “The Cross-section of Expected Stock Returns”. In: *Critical Finance Review* 4.1, pp. 1–44.
- Muravyev, Dmitriy and Neil D. Pearson (2019). *Option Trading Costs Are Lower than You Think*. Working Paper. Available at SSRN: <https://ssrn.com/abstract=2580548>. SSRN.
- Muravyev, Dmitriy, Aurelio Vasquez, and Wenzhi Wang (2020). *Making Better Use of Option Prices to Predict Stock Returns*. Working Paper. Research Document DR 20-03, Montréal, Québec, Canada. Canadian Derivatives Institute.
- Novy-Marx, Robert and Mihail Velikov (2016). “A Taxonomy of Anomalies and Their Trading Costs”. In: *The Review of Financial Studies* 29.1, pp. 104–147.
- Welch, Ivo and Amit Goyal (2008). “A comprehensive look at the empirical performance of equity premium prediction”. In: *The Review of Financial Studies* 21.4, pp. 1455–1508.
- Xing, Yuhang, Xiaoyan Zhang, and Rui Zhao (2010). “What Does the Individual Option Volatility Smirk Tell Us About Future Equity Returns?” In: *Journal of Financial and Quantitative Analysis* 45.3, pp. 641–662.
- Zou, Hui and Trevor Hastie (2005). “Regularization and Variable Selection Via the Elastic Net”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320.

Appendix A – Declaration of Generative AI Assistance

I declare that I used Anthropic Claude (Version: Claude Sonnet, <https://claude.ai/>) and OpenAI ChatGPT (Version: GPT-5.2 and 5.1, <https://chat.openai.com/>) between August 2025 and January 2026 to support various stages of my thesis work. These tools were used strictly as assistive resources to enhance my understanding, analysis, and writing process, in accordance with HEC Montréal’s responsible AI usage guidelines.

Specifically, I used generative AI tools for the following purposes. In code development and debugging, Claude and ChatGPT assisted in writing, debugging, and optimizing Python scripts for data preprocessing, rolling window PCA, elastic net regression, Random Forest models, and portfolio construction, helping identify coding errors and improve computational efficiency. For statistical output verification, Claude assisted in cross-checking numerical results across multiple code outputs to ensure consistency between reported statistics in tables and text throughout the thesis. In writing assistance, Claude helped draft and refine thesis chapters, improving sentence structure, ensuring academic tone appropriate for an MSc thesis, and maintaining consistency in terminology and notation across chapters. For LaTeX formatting, Claude provided guidance on syntax for tables, figures, equations, and proper formatting according to HEC Montréal’s thesis template requirements. In literature review support, ChatGPT and Claude helped identify relevant academic papers and assisted in understanding key concepts from the return

predictability and machine learning literature. Finally, for consistency checking, Claude systematically reviewed chapters to identify inconsistencies in reported numbers, missing figure labels, grammatical errors, and formatting issues.

All AI-generated content was carefully reviewed, critically evaluated, and revised to ensure alignment with my own understanding, analytical reasoning, and academic integrity. The empirical analysis, including all data processing, model estimation, and portfolio construction, was executed independently using my own Python code. No confidential data or unpublished research was shared with these tools at any time. The final content of this thesis represents my original work, with generative AI tools serving solely as assistants to facilitate my learning and expression.

Appendix B – Additional Figures

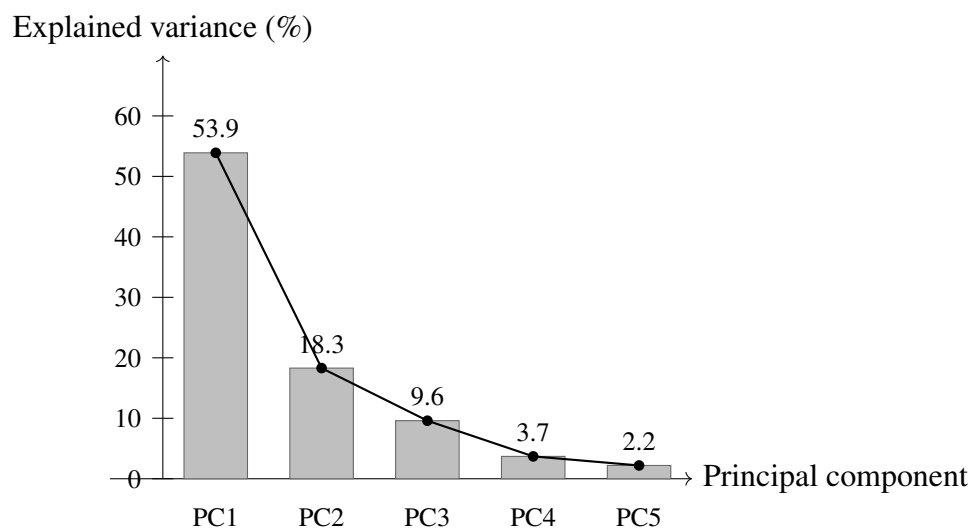


Figure 1: Variance explained by the first five principal components in Window 1 (training period September 2013 to August 2016). The first three components account for 81.8% and all five components account for 87.7% of total variation.

