ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

M. Sc. Volkswirtschaftslehre

Predict Financial Prices With Binary Response Models

Master Thesis

Examiner: Prof. Dr. Roxana Halbleib

Name: Stephan Berke

Student ID: 5548152

Start of the Thesis: 1st January 2025

End of the Thesis: 1st July 2025

universität freiburg

Abstract

This paper investigates the predictability of return directions in the context of volatility forecasting. The objective of this research is to identify confident upward predictions using interpretable and robust modeling approaches based on two specifications of binary response models. The first specification builds on the direction-of-change model proposed by Christoffersen and Diebold (2006). This work extends the framework by introducing probabilistic confidence intervals using the delta method, enabling statistical testing of the significance of the modeled directional return probabilities. The second specification applies a logistic regression model based on a set of statistical features to classify return directions. Cost-sensitive and conformal prediction models are applied to extend the classifications to control for confident upward predictions. The results of both specifications suggest that a certain degree of directional predictability exists, particularly under calm market regimes.

Zusammenfassung

Diese Arbeit untersucht die Vorhersagbarkeit finanzieller Renditen im Kontext von Volatilitätsprognosen. Ziel der Untersuchung ist es, mit Hilfe robuster Modellierungsansätze die auf zwei Spezifikationen binärer Regressionsmodelle basieren, verlässliche und interpretierbare Aufwärtsprognosen zu identifizieren. Die erste Spezifikation baut auf dem Richtungsmodell von Christoffersen and Diebold (2006) auf und wird in dieser Arbeit durch die Einführung probabilistischer Konfidenzintervalle mittels der Delta-Methode erweitert. Diese Erweiterung ermöglicht das statistische Testen der Signifikanz der modellierten Wahrscheinlichkeiten. Die zweite Spezifikation verwendet ein logistisches Regressionsmodell zur Vorhersage positiver Renditerichtungen. Dieser Klassifizierungsrahmen wird durch zwei Erweiterungen ergänzt: Ein kostensensitives sowie ein konfidenzbasiertes Vorhersagemodell. Die Ergebnisse beider Spezifikationen deuten auf ein nachweisbares Maß an Vorhersagbarkeit der täglichen Renditerichtungen in ruhigen Marktphasen hin.

Contents

List of Figures

List of Tables

Abbreviations

ACF	Autocorrelation Function
AR(1)	Auto Regressive Model of Order 1
ARCH	Auto Regressive Conditional Heteroskedasticity
CDF	Cumulative Distribution Function
CI	Confidence Interval
CP	Conformal Prediction
CS	Cost-Sensitive
CTS	Calendar Time Sampling
DNN	Deep Neural Network
EMH	Efficient Market Hypothesis
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
IBM	International Business Machines Corporation
JB	Jarque-Bera-Test
JPM	JP Morgan
Logit	Logistic Regression
ML	Machine Learning
OOS	Out-of-Sample
PACF	Partial Autocorrelation Function
PDF	Probability Density Function
PFE	Pfizer
PPR	Positive Prediction Rate
RV	Realize Variance
VIX	Volatility Index

1 Introduction

Predicting financial asset returns remains a central challenge in both academic and practical finance applications. In the literature, the return level is widely seen as unpredictable, which is consistent with the efficient market hypothesis (EMH) by Fama (1970). However, a growing number of studies highlight that directional predictability may exist and carries economic value. Directional forecasts are particularly relevant for market timing and risk-based decision-making, as they do not require the exact magnitude of returns to be known.

This thesis investigates the problem of return direction prediction on a daily horizon using the framework of binary response models. The methodological foundation for this work is the model proposed by Christoffersen and Diebold (2006), who developed a binary response framework to examine the interaction between volatility dynamics and return signs. They demonstrate that volatility dependence produces sign dependence, as long as return distributions have non-zero mean and are asymmetric. Accordingly, the central assumption of this work is therefore that periods of low predicted volatility are associated with increased predictability of returns. Specifically, when past returns display a clear trend, such as consecutive positive returns, and forecasted volatility is low, the conditional probability of observing a subsequent positive return is assumed to increase.

The objective of this thesis is to extend the directional forecasting framework by Christoffersen and Diebold (2006). A key contribution is the development of time-varying confidence intervals for binary return predictions. These intervals are derived using the delta method and allow inference on whether the predicted probabilities significantly exceed a benchmark of random guessing. The interval width adjusts with the forecasted volatility, widening during periods of high volatility and narrowing in calmer market regimes, thus providing a time-varying measure of predictive uncertainty.

In the second part of the thesis, a logistic regression model (Logit) is implemented using s set of statistical features to forecast the direction of daily returns. This model is extended in two ways: (i) by applying cost-sensitive learning to penalize false positive classifications in the training set, and (ii) by incorporating conformal prediction techniques to quantify the confidence level of each classification. All models are calibrated in-sample and evaluated out-of-sample. As a robustness check, a decision tree (DT) classifier is applied to benchmark the logit-based binary response models. Lastly, the economic significance of the directional forecasts is evaluated by simulating trading strategies that accumulate realized returns on days with positive return predictions.

The remainder of this thesis is structured as follows. Chapter 2 provides a brief review of the relevant literature on return predictability and directional forecasting. Chapter 3 presents the dataset and outlines key statistical characteristics. Chapter 4 presents the methodological framework, including

theoretical derivation and model specifications. Chapter 5 reports the empirical results, including model outputs, parameter calibration procedures, and an assessment of economic significance. A critical discussion of the findings and and methodological limitations is provided in Chapter 6. Chapter 7 concludes the thesis and summarizes the main contributions.

2 Literature review

The question of whether financial markets allow for predictable patterns has been a longstanding topic of debate in empirical finance. According to the efficient market hypothesis (EMH) stock returns are regarded as unpredictable, as prices are assumed to fully incorporate all publicly available information (Fama, 1970). In its semi-strong form, the EMH implies that price changes only occur in response to new, unanticipated information. Although some studies report evidence of return predictability in recession periods (Nyberg, 2011), several contributions argue that such predictability is time-inconsistent and negligible out-of-sample (Welch & Goyal, 2008) or not economically meaningful (Bajgrowicz & Scaillet, 2012).

In contrast to point forecasts, Christoffersen and Diebold (2006) focused on directional predictions, motivated by the following decomposition:

$$r_t = sign(r_t)|r_t|, \tag{1}$$

where $sign(r_t) = 1$ if $r_t > 0$ and $sign(r_t) = 0$ if $r_t \leq 0$.

The authors argue that directional predictability in $sign(r_t)$ is generated by persistence in absolute returns $|r_t|$, provided the conditional mean is non-zero. Extending this framework, Christoffersen et al. (2007) show that directional predictability can also emerge under zero conditional mean, as long as the return distribution is asymmetric, by incorporating skewness and kurtosis effects. Christoffersen and Diebold (2006) further argue that directional forecastability is most likely to be found on intermediate time horizons, such as monthly or quarterly, due to the characteristics of financial return means and volatilities. Complementary findings by Leung et al. (2000) suggest that the predictability of return signs can be stronger than that of return levels, highlighting the potential values of directional forecasts in practice. Additionally, Linton and Whang (2007) introduced quantilograms, the autocorrelation in quantiles, in a non-parametric framework for directional predictability at the daily frequency.

Recent studies have increasingly applied machine learning (ML) methods to (directional) return predictions. For instance, Zhong and Enke (2019) applied deep neural networks (DNNs) to capture nonlinear patterns in return signs and report higher classification accuracy relative to traditional models. In another approach, Campisi et al. (2024) compare several machine learning algorithms for predicting the direction of the US stock market using volatility indices at

a monthly frequency. Their findings indicate that ML methods outperform classical least-squares regressions.

Approaches using cost-sensitive learning have been applied in the context of price range forecasting using a deep forest framework (Ma et al., 2020). However, to the best of current knowledge, cost-sensitive models have not yet been applied to return direction prediction. Similarly, while conformal prediction provides a probabilistic framework for reliable classification, it has not been covered extensively in the financial literature yet.

Despite the potential of machine learning approaches, these models are often prone to overfitting, particularly when applied in out-of-sample settings. Arian et al. (2024) emphasize that characteristics of financial time series such as non-stationarity, structural breaks or regime shifts, can lead to unstable ML-based return predictions when applied out-of-sample. These findings highlight the need of robust model calibration and robustness in empirical forecasting approaches.

3 Data

3.1 Characterization

This study utilizes high-frequency data from the TAQ database on 5-minute frequency from three stocks: J.P. Morgan (JPM), International Machines Business Corporation (IBM) and Pfizer (PFE). The stocks represent large-cap U.S. equities across different sectors: financials, technology and healthcare. The 5-minute high-frequency data spans the period from 01/2001 to 03/2024, contains 79 observations per day and has been resampled using a calendar-time sampling (CTS) scheme. Prices are calculated as the average of the bid and ask quotes sampled at the particular second. Non-trading days and holidays are excluded from the dataset by default. Daily returns, as the primary objective, are calculated as close-to-close changes. Throughout the modeling process, logarithmic returns are used to facilitate computations and due to their desirable statistical properties such as time additivity and approximate normality:

$$r_t = log(\frac{P_t^{close}}{P_{t-1}^{close}}), \tag{2}$$

where r_t is the log return on day t, P_t^{close} the closing price in t and P_{t-1}^{close} at day $t-1$. When evaluating the economic significance of the modeling, the returns are expressed as discrete returns. Realized variance has been calculated in the standardized framework:

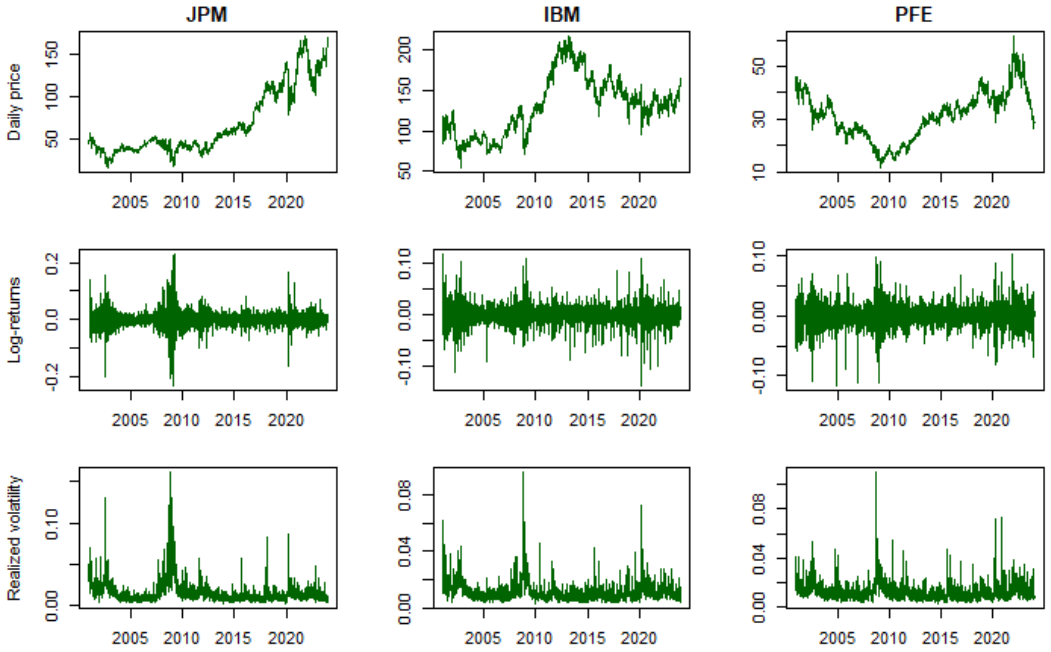
$$RV_t^{79} = \sum_{j=1}^{79} r_{j,t}^2. \tag{3}$$

Although studies such as Ahoniemi and Lanne (2013) find improved in-

sample fit for single-stock volatility models excluding overnight returns due to their noisy characteristics, they are retained in this setting to ensure consistency between realized volatility inputs and the close-to-close return dynamics. However, using open-to-close directions with omitted overnight returns may improve predictive performance due to reduced noise.

The following Figure 1 displays the daily share prices, log-returns and realized volatilities of the analyzed stocks:

Figure 1: Asset prices, log-returns and realized volatility plots



The JPM stock showed mostly positive growth trend over the sample period, with spikes in 2020 until 2022. In contrast, IBM increased until 2012 followed by a continuous decline characterized by volatile fluctuations. PFE experienced an initial decline until 2009, after which it entered a growth phase lasting until 2021, before experiencing a downturn. Volatility levels peaked during the financial crisis and again during the COVID-19 pandemic across stocks, reflecting periods of heightened market uncertainty.

For modeling purposes, each stock is split into a training set, a calibration set and a test set. The training set has been particularly used to train the Logit classifiers in the second approach, while the calibration set has been used to calibrate the parameters. All methods are evaluated on the test sets. To account for different market regimes, the analysis is conducted on a calm and a volatile period.

Table 1: Temporal split of calm and volatile periods

Regime	Train	Calibration	Test
Calm (P1)	2002-2011	2012-2015	2016–2019
Volatile (P2)	2005-2014	2015-2018	2019–2022

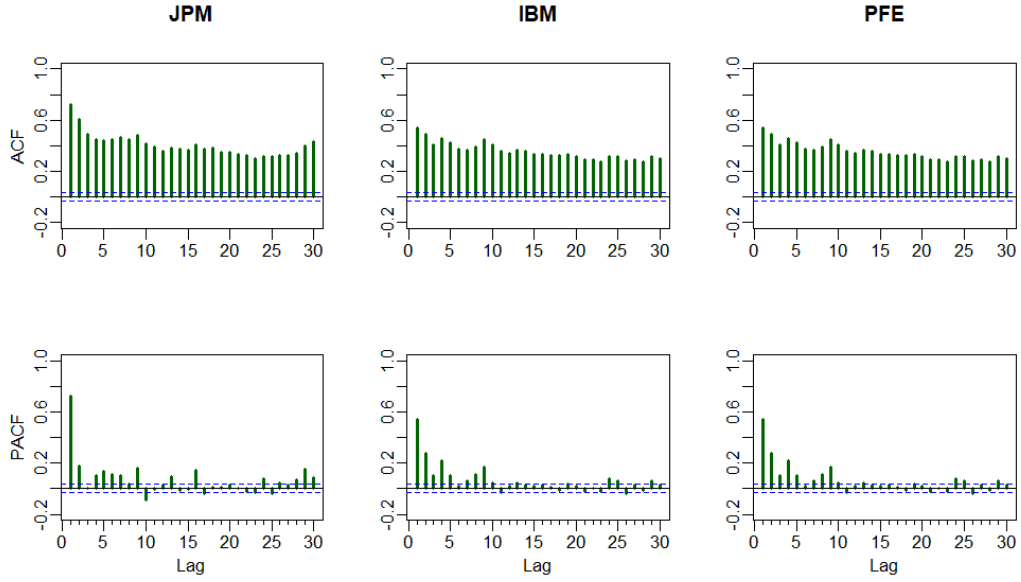
Note: Train: 56%, Calibration: 22%, Test: 22% of total sample

The calm period from 2016 to 2019 is marked by relatively low volatility (see Table 1). In contrast, the volatile period from 2019 to 2022 includes the COVID-19 pandemic, which led to strong volatility spikes and more frequent return outliers. Figure 1 shows that all three stocks experienced higher volatility and a larger number of extreme return observations during the volatile period.

To incorporate volatility dynamics into the modeling framework, a suitable model for forecasting realized variance (RV) is required. Financial return series typically exhibit volatility clustering, where periods of high volatility are followed by high volatility and vice versa. Consequently, modeling the conditional variance is essential for capturing temporal structures.

Figure 2 presents the autocorrelation and partial autocorrelation functions of the RV series, based on the training and calibration data of the calm periods.

Figure 2: ACF and PACF plots of realized variances



The ACF plots of RV decay slowly across all assets, indicating persistence and long-memory behaviour in volatility. The PACF plots show significant first-order autocorrelation and additional significance at higher lags. The patterns suggest that lagged volatility contains predictive information, supporting the use of autoregressive structures for volatility forecasting. The baseline

specification is an AR(1) on the square root of realized variance:

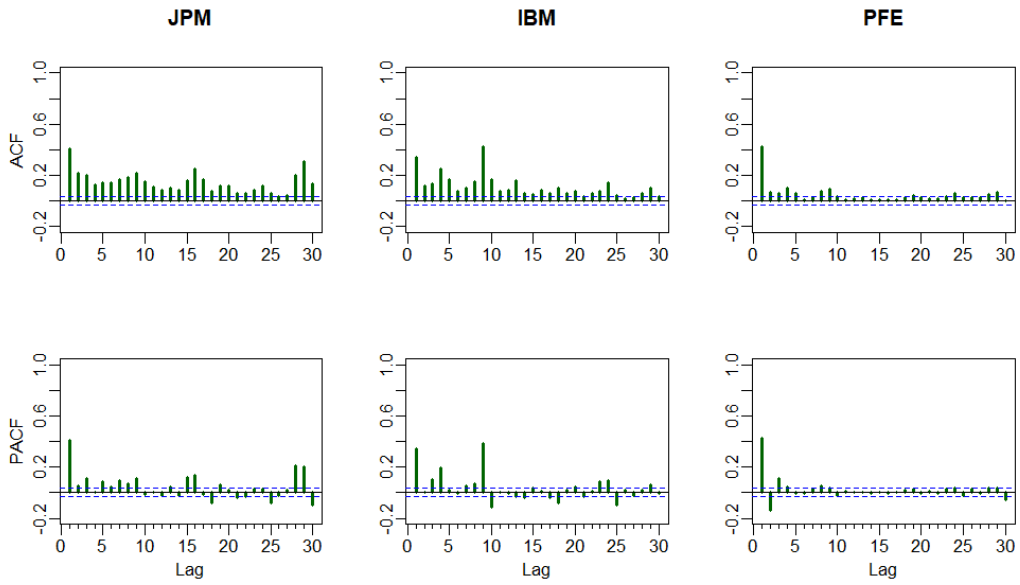
$$X_{t+1} = \alpha + \beta X_t + \varepsilon_t, \tag{4}$$

where $X_t = \sqrt{RV_t}$.

However, the multi-lag structure observed in the PACF motivates the consideration of multi-lag models such as the HAR specification proposed by Corsi et al. (2008), which captures volatility dynamics across daily, weekly and monthly horizons. While HAR models might produce better in-sample fit, this does not guarantee better out-of-sample performance.

To evaluate the AR(1) specification, the autocorrelation structure of its squared residuals is further analyzed.

Figure 3: ACF and PACF plots of squared residuals from AR(1) model



The ACF and PACF of the squared residuals show significant autocorrelation at multiple lags, indicating remaining conditional heteroskedasticity. An ARCH-LM test rejects the null hypothesis of no auto-regressive conditional heteroskedasticity (ARCH) effects for all assets:

Table 2: ARCH-LM test results on squared residuals

Asset	JPM	IBM	PFE
p-value	0^{***}	0^{***}	0^{***}

Note: p-values from the ARCH-LM test applied to the squared residuals of AR(1) models fitted to realized variance series. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

The results indicate that a GARCH-type component could improve the modeling of the innovation variance. If the HAR model also produces residual heteroskedasticity, a combined HAR-GARCH specification (Corsi, 2009) may,

in theory, better capture both the persistence in realized volatility and the time-varying nature of the error variance.

Nevertheless, due to the increased model complexity of other specifications and the focus on predicting return directions, the AR(1) on $\sqrt{RV}$ is retained as the main volatility model in the forecasting frameworks.

3.2 Descriptive statistics

For better data understanding, the descriptive statistics of daily log-returns of each asset are analyzed for both the calm and the volatile period. The tables report the summary measures and indicate the p-values of the test statistics of normality tests. For the Jarque-Bera test of non-normality, the test statistic is directly reported.

Table 3: Descriptive statistics of log returns in the calm period

	JPM			**IBM**			**PFE**		
Moment	Train	Calib	Test	Train	Calib	Test	Train	Calib	Test
Mean	0.0000	0.0007	0.0007*	0.0002	-0.0003	0.0000	-0.0002	0.0004	0.0002
SD	0.029	0.014	0.013	0.016	0.012	0.013	0.017	0.010	0.011
Skewness	0.263***	-0.296***	-0.006	0.075	-1.053***	-0.438***	-0.294***	-0.006	-0.129*
Kurtosis	15.227***	6.684***	6.790***	9.023***	10.541***	11.016***	8.619***	4.637***	7.300***
JB	15707***	583***	601***	3806***	2569***	2725***	3347 ***	112***	777***

Note: Table reports statistics of log return series. Significance levels are based on tests of: Mean = 0, Kurtosis = 3, and the Jarque-Bera (JB) test for normality. ***p $<$ 0.01, **p $<$ 0.05, *p $<$ 0.10

In the calm period, the Jarque-Bera test of normality rejects the null hypothesis of normality for all assets, indicating that the return distributions deviate significantly from the normal distribution. All series exhibit significant excess kurtosis, suggesting the presence of fat tails caused by outliers. For JPM, both the training and calibration sets show statistically significant skewness, while the test set appears approximately symmetric. In the case of IBM, only the training set is not skewed, whereas both calibration and test sets are significantly skewed. PFE displays skewness only in the training data, with the calibration and the test sets showing more asymmetric behaviour.

Table 4: Descriptive statistics of log returns in the volatile period

	JPM			IBM			PFE		
Moment	Train	Calib	Test	Train	Calib	Test	Train	Calib	Test
Mean	0.0002	0.0004	0.0003	0.0002	-0.0003	0.0002	0.0001	0.0003	0.0002
SD	0.027	0.014	0.021	0.014	0.013	0.018	0.015	0.011	0.017
Skewness	0.312***	-0.089	-0.045	-0.179***	-0.684***	-0.731***	-0.149***	0.208***	0.166**
Kurtosis	18.038***	6.319***	14.955***	9.642***	9.432***	12.433***	10.286***	6.395***	7.158***
JB	23749***	463***	5991***	4639***	1813***	3819***	5575***	490***	729***

Note: Significance levels are based on tests of: Mean = 0, Kurtosis = 3, and the Jarque-Bera (JB) test for normality. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

In the volatile period shown in Table 4, the tests for normality and the absence of excess kurtosis are rejected as well. All series show significant skewness of log returns distributions, except for JPM in the calibration and test sets. These results are further illustrated by the standardized log return densities, where standardized returns are defined as $r_t^{\text{std}} = \frac{r_t}{\hat{\sigma}_t}$.

Figure 4: Normalized log returns of train, calibration and test sets

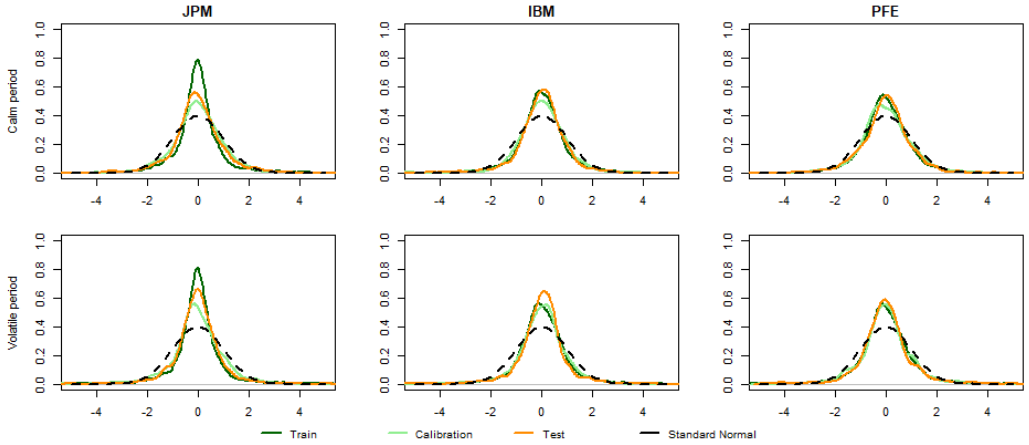


Figure 4 presents the kernel densities of standardized log returns for each stock and set across the calm and volatile periods, overlaid with the standard normal distribution (dotted line). In both regimes, the return distributions are more peaked and exhibit heavier tails compared to the normal distribution, indicating leptokurtosis. This effect is more pronounced for JPM, whose distribution shows the narrowest spread around the mean.

3.3 Explanatory variables in the logistic regression approach

Since directional predictability of daily stock returns has been largely ruled out in the literature, there is no established feature set for explanatory variables. While many other studies rely mainly on macroeconomic or market-wide indicators that might be related to the individual stock, this approach focuses on

features derived directly from each assets own intraday price data. In addition, three general indicators, the VIX, the economic policy uncertainty index and 10-year U.S. treasury yield are included as macroeconomic controls to capture market uncertainties. This feature design improves external validity, as it allows the model to be applied across different stocks without requiring prior stock-specific correlation analysis. The feature set is structured into four categories:

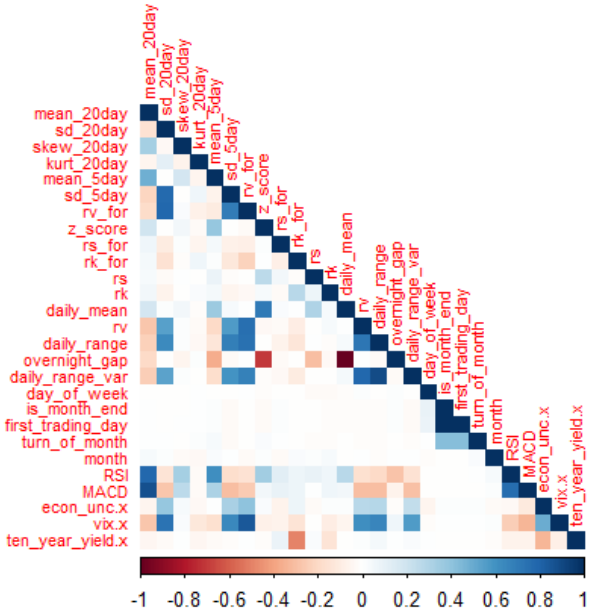
Table 5: Overview of technical, statistical, calendaric and macroeconomic indicators

Indicator	Feature name	Brief description	Formula / Logic
Statistical indicators			
Realized Volatility	rv	Return dispersion over the day[a]	$RV_t = \sum r_{t,i}^2$
Realized Skewness	rs	Asymmetry in intraday returns	$RS_t = \frac{\sum r_{t,i}^3}{(\sum r_{t,i}^2)^{3/2}}$
Realized Kurtosis	rk	Tail risk measure	$RK_t = \frac{\sum r_{t,i}^4}{(\sum r_{t,i}^2)^2}$
RV Forecast	rv_for	AR(1) forecast of volatility	$RV_{t+1} = \alpha + \beta RV_t + \epsilon_t$
RS Forecast	rs_for	AR(1) forecast of skewness	$RS_{t+1} = \alpha + \beta RS_t + \epsilon_t$
RK Forecast	rk_for	AR(1) forecast of kurtosis	$RK_{t+1} = \alpha + \beta RK_t + \epsilon_t$
20-Day Mean	mean_20day	Rolling mean of returns	$\bar{r}_t^{(20)} = \frac{1}{20}\sum_{i=0}^{19} r_{t-i}$
20-Day SD	sd_20day	Rolling standard deviation	$SD_t^{(20)} = \sqrt{\frac{1}{19}\sum(r_{t-i} - \bar{r})^2}$
20-Day Skewness	skew_20day	Rolling skewness	$Skew_t^{(20)} = \frac{\sum(r_{t-i}-\bar{r})^3}{SD^3}$
20-Day Kurtosis	kurt_20day	Rolling kurtosis	$Kurt_t^{(20)} = \frac{\sum(r_{t-i}-\bar{r})^4}{SD^4}$
5-Day Mean	mean_5day	Rolling mean	$\bar{r}_t^{(5)} = \frac{1}{5}\sum_{i=0}^{5} r_{t-i}$
5-Day SD	sd_5day	Rolling standard deviation	$SD_t^{(5)} = \sqrt{\frac{1}{4}\sum(r_{t-i} - \bar{r})^2}$
Technical indicators			
MACD	MACD	Momentum/trend signal	$MACD = EMA_{12} - EMA_{26}$
PROC	PROC	Rate of price change	$PROC = \frac{X_t - X_{t-1}}{\sum(X_i - X_{i-1})^+}$
Overnight Gap	overnight_gap	Return from prev. close to curr. open	$OG_t = \log(Open_t) - \log(Close_{t-1})$
Daily Range	daily_range	High–low price difference	$Range_t = High_t - Low_t$
Daily Range Variance	daily_range_var	Volatility from high-low spread	$HLVar_t = \frac{1}{4\log 2}\left(\log\frac{High_t}{Low_t}\right)^2$
Calendaric indicators			
Day of Week	day_of_week	Encodes weekday (1–7)	$d_t = \text{weekday}(date)$
Month	month	Encodes calendar month of date	$m_t = \texttt{month(date)}$
Is Month End	is_month_end	End of month flag	1 if $\Delta m_t \neq 0$ or $t = T$
First Trading Day	first_trading_day	Start of new month	1 if $\Delta m_t \neq 0$
Turn of Month	turn_of_month	Turn-of-month effect	1 if $day_t \in \{1, 2, 3, 28, 29, 30, 31\}$
Macroeconomic indicators			
Economic Uncertainty	econ_uncertainty	Economic Policy Uncertainty Index	Daily value of U.S. EPU index
10-Year Bond Yield	bond_yield_10y	Long-term interest rate indicator	Daily U.S. 10-year treasury yield
VIX	vix	Implied volatility index	Daily closing value of the VIX index

[a] Including overnight returns.

Descriptive statistics for all features are reported in Appendix, Table 13. Figure 5 shows the average correlation structure of the explanatory variables. Variance-based features display moderate intercorrelations, while several technical indicators are linked to price-based statistics. The macroeconomic controls exhibit only weak correlations with other features. Most calendaric variables show near-zero correlation with return-related metrics, confirming their role as independent controls in the modeling framework.

Figure 5: Correlation plot as average of all three series across all period



Note: Created by the author.

4 Model framework

This section introduces the theoretical model frameworks used to predict the direction of daily stock returns. The analysis builds on the concept of binary response models, which provide the foundation for the directional forecasting tasks.

The first specification follows the approach of Christoffersen and Diebold (2006), who propose a binary response model estimated in a rolling-window forecasting framework using information on the conditional mean and the volatility of returns. This baseline model is extended by introducing confidence intervals around the predicted probabilities, constructed via the delta method. This allows for formal inference against the null hypothesis of random guessing.

The second specification applies a supervised classification model using a logistic regression. To improve the precision of predicted upward movements, the model is extended in two ways:

1. **Cost-sensitive learning:** Penalizing false positive classification during training

2. **Conformal prediction:** Modeling confidence-calibrated prediction sets based on probability thresholds.

To evaluate the robustness, of the second specification, a decision tree model is applied using the same methodology.

4.1 Binary response models

Binary response models are parametric models used when the dependent variable y takes on two possible outcomes:

$$y = \begin{cases} 1 & \text{if the event occurs,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

These models estimate the conditional probability of y = 1 given a set of explanatory variables X:

$$P(y = 1|X) = F(X\beta), \tag{6}$$

where $F()$ is the cumulative distribution function (CDF).

In the probit model, the error term is assumed to follow a normal, while in the Logit case, it follows a logistic distribution :

$$\begin{aligned} F_{Probit}(X\beta) &= \Phi(X\beta) \\ F_{Logit}(X\beta) &= \Lambda(X\beta), \end{aligned} \tag{7}$$

where $\Phi()$ denotes the standard normal CDF and $\Lambda()$ the logistic CDF.

4.2 Approach by Christoffersen

Christoffersen and Diebold (2006) propose a binary response framework models for forecasting the direction of returns. The model assumes the following structure of returns:

$$R_{t+1} = \mu_t + \sigma_{t+1}\epsilon_{t+1}, \tag{8}$$

where μ is the conditional mean, σ_{t+1} the conditional standard deviation and $\varepsilon \sim N(0, 1)$ is a standard normal innovation.

The binary response variable is defined as the sign of the next-day return:

$$y_t = \begin{cases} 1 & \text{if } R_{t+1} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The goal is to estimate the probability of observing a positive return:

$$\begin{aligned} P(R_{t+1} > 0) &= 1 - P(R_{t+1} \leq 0) \\ &= 1 - P(\epsilon_{t+1} \leq \frac{-\mu_t}{\sigma_{t+1}}) \\ &= 1 - F_\epsilon(\frac{-\mu_t}{\sigma_{t+1}}) \\ &= \Phi(\frac{\mu_t}{\sigma_{t+1}}). \end{aligned} \tag{10}$$

This serves as the baseline equation of the Christoffersen and Diebold

(2006) model. In this framework, the predictive probability depends on the conditional mean and volatility. Forecasting volatility is a central component of this framework. In this approach, a standard $AR(1)$ realized variance forecast is used (see Chapter 3.1). Depending on modeling preferences and data characteristics, other forecasting routines, such as HAR models, risk metrics, or GARCH-type models can be applied as well.

The conditional mean is estimated using averages of the previous T days log returns in a rolling window:

$$\mu_t = \frac{1}{T} \sum_{i=1}^{T} r_i. \tag{11}$$

The choice of the window length T introduces a key modeling parameter, as both the estimated mean and volatility depend on it. It is therefore recommended to evaluate model performance across different window sizes to assess the sensitivity of results.

4.2.1 Derivation of confidence intervals

Confidence intervals complement point predictions by quantifying the uncertainty associated with forecasted probabilities. In statistical forecasting, it is standard practice to reflect the inherent uncertainty of model estimates. The strength of this approach is its ability to account for time-varying uncertainty, driven primarily by changes in volatility. This enables a formal testing framework based on the null hypothesis that the predicted probability corresponds to a random guess, defined by a probability of 0.5:

$$H_0 : Pr(R_{t+1} > 0) = 0.5. \tag{12}$$

Under this framework, only predictions whose confidence intervals exclude a probability the value of 0.5 are considered statistically significant and therefore indicative of meaningful directional predictability.

There exist multiple approaches in the literature for constructing confidence intervals for functions of maximum likelihood estimators. One of the most commonly used techniques is the delta method, which employs a first-order Taylor expansion to approximate the variance of nonlinear functions (see, e.g., Casella and Berger (2002)). The following derivation follows the work of Xu and Long (2005), who proposed a routine for constructing confidence intervals for predicted probabilities. A step-by-step derivation is presented in the following.

First, let the previously defined prediction function of interest be:

$$G(\theta) = \Phi(\frac{\mu_t}{\sigma_{t+1}}), \tag{13}$$

where $\theta \in (\mu_t, \sigma_{t+1})$.

The function $G(\hat{\theta})$ can be approximated by a first-order Taylor series expansion to derive its variance:

$$G(\hat{\theta}) \approx G(\theta) + (\hat{\theta} - \theta)' G'(\theta), \tag{14}$$

where $G'(\theta)$ denotes the gradient with respect to θ. Subtracting $G(\theta)$ yields:

$$G(\hat{\theta}) - G(\theta) \approx (\hat{\theta} - \theta)' G'(\theta). \tag{15}$$

Under the standard regularity conditions and assuming asymptotic normality:

$$\sqrt{n}[G(\hat{\theta}) - G(\theta)] \xrightarrow{d} N(0, Var(G(\hat{\theta}))), \tag{16}$$

where the variance of the function is:

$$\begin{aligned} Var(G(\hat{\theta})) &= Var(G(\theta) + (\hat{\theta} - \theta)' G'(\theta)) \\ &= Var((\hat{\theta} - \theta)' G'(\theta)) \qquad \text{(since } G(\theta) \text{ is a constant)} \\ &= Var(\hat{\theta}\, G'(\theta)) \\ &= G'(\theta)^T Var(\hat{\theta}) G'(\theta). \end{aligned} \tag{17}$$

Hence, the gradient of the objective functions needs to be derived, as well as the variances of the estimates. The gradient can be defined as:

$$G'(\theta) = \frac{\partial G(\theta)}{\partial \theta} = [\frac{\partial \Phi(\mu_t/\sigma_{t+1})}{\partial \mu_t} \quad \frac{\partial \Phi(\mu_t/\sigma_{t+1})}{\partial \sigma_{t+1}}]^T, \tag{18}$$

where

$$\frac{\partial \Phi(\mu_t/\sigma_{t+1})}{\partial \mu_t} = \phi(\frac{\mu_t}{\sigma_{t+1}}) \frac{1}{\sigma_{t+1}}, \tag{19}$$

and

$$\frac{\partial \Phi(\mu_t/\sigma_{t+1})}{\partial \sigma_{t+1}} = -\phi(\frac{\mu_t}{\sigma_{t+1}}) \frac{\mu_t}{\sigma_{t+1}^2}. \tag{20}$$

Here, ϕ denotes the normal probability density function (PDF) as the derivative of the cumulative distribution function.

The variance of $G(\hat{\theta})$ is stated as the variance-covariance matrix:

$$Var(G(\hat{\theta})) = \begin{bmatrix} \text{Var}(\mu_t) & \text{Cov}(\mu_t, \sigma_{t+1}) \\ \text{Cov}(\mu_t, \sigma_{t+1}) & \text{Var}(\sigma_{t+1}) \end{bmatrix}, \tag{21}$$

and requires deriving the sample variances and covariances.

1. **Variance of the sample mean:**

   Let $\mu_t = \frac{1}{T} \sum_{i=1}^{T} r_i$ denote the mean of the past T log returns. Assuming the returns r_i are i.i.d. with variance σ_r^2, the variance of the sample mean is:

   $$\text{Var}(\mu_t) = \frac{\sigma_r^2}{T}. \tag{22}$$

In empirical applications, σ_r^2 is estimated using the sample variance of the returns in the rolling window.

2. **Variance of the volatility forecast:**

   Recall the AR(1) specification for the square root of realized volatility:

$$X_{t+1} = \alpha + \beta X_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2), \tag{23}$$

   where $X_t = \sqrt{RV_t}$. After estimating $\hat{\alpha}$ and $\hat{\beta}$, the residuals are computed as:

$$\hat{\varepsilon}_{t+1} = X_{t+1} - (\hat{\alpha} + \hat{\beta} X_t). \tag{24}$$

   The variance of the one-step ahead forecast is given by the residual variance:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{T-1} \sum_{t=1}^{T-1} \hat{\varepsilon}_{t+1}^2. \tag{25}$$

   Under the model assumptions, this represents the conditional variance of the forecast:

$$\mathrm{Var}(X_{t+1} \mid \mathcal{F}_t) = \hat{\sigma}_\varepsilon^2. \tag{26}$$

3. **Covariance of the sample mean and volatility forecast**

   The covariance between μ_t and σ_{t+1} is assumed to be zero:

$$Cov(\mu_t, \sigma_{t+1}) = 0. \tag{27}$$

   This assumption simplifies the application of the delta method. In future, the model can be extended by incorporating return-volatility interactions.

The resulting confidence intervals take the form:

$$CI = G(\theta) \pm z_{\alpha/2} \sqrt{G'(\theta)^T V(\hat{\theta}) G'(\theta)}. \tag{28}$$

This allows for statistical inference on the predicted probability. Specifically, the null hypothesis that the estimated probability corresponds to a random guess of 0.5 can be tested using the test statistic:

$$TS = \frac{G(\theta) - 0.5}{\sqrt{\mathrm{Var}(G(\hat{\theta}))}}. \tag{29}$$

For completion, the p-value can be stated as:

$$p = 2(1 - \Phi(|TS|)). \tag{30}$$

This framework enables the differentiation between statistically significant and non-significant forecasts, allowing for a statistical evaluation framework.

4.3 Logistic regression approach

The second specification is a logistic regression model. It provides a statistically interpretable computationally robust approach to predicting returns directions.

The binary outcome is defined as:

$$y_{t+1} = \mathbf{1}(R_{t+1} > 0) \tag{31}$$

where $y_{t+1} = 1$ if the return in period $t+1$ is positive and 0 otherwise.

The conditional probability of a positive return is modeled as:

$$P(y_{t+1} = 1 \mid X_t) \ = \Lambda(X_t\beta) = \frac{1}{1 + e^{-X_t\beta}} \tag{32}$$

with X_t denoting the vector of explanatory variables, β the vector of model coefficients and $\Lambda()$ denotes the logistic function.

Predictions are classified as positive if the forecasted probability exceeds a threshold of 0.5:

$$\hat{p}_t > 0.5.$$

The baseline logistic model showed relatively balanced predictions in terms of accuracy. To further improve predictive precision, two extensions are introduced. Each model is trained and evaluated in a supervised learning set-up to ensure generalizability and avoid overfitting. The procedure follows three steps:

1. Train the model on the training set Z^{train}
2. Tune the hyperparameters based using a separate calibration set Z^{calib}
3. Evaluate predictive performance on the out-of-sample test set Z^{test}

4.3.1 Cost-sensitive learning

Standard logistic regression assumes equal misclassification costs for false positive and false negative predictions. However, in financial applications, the cost of false positives can be particularly relevant. In the context of this thesis, focusing on modeling confident upward return predictions, the cost of false positives is explicitly increased to improve directional certainty.

The literature on cost-sensitive classification (see, e.g. Elkan (2001)) provides various frameworks for incorporating asymmetric costs. The implementation used in this approach represents a simplified version, where only the cost of false positives is modeled via scalar weighting in the log-likelihood function. While this is only a partial formulation of the cost minimization framework, it specifically contributes to improving the precision of the predicted return directions.

In the baseline model, the parameters are estimated by minimizing the standard log-likelihood:

$$\mathcal{L}(\beta) = -\frac{1}{n}\sum_{t=1}^{n}\left[y_{t+1}\log(p_t) + (1-y_{t+1})\log(1-p_t)\right], \tag{33}$$

where $p_t = \Lambda(X_t\beta)$ is the predicted probability of a positive return.

To reflect asymmetric costs, observation weights are introduced during model training, following the structure of Bahnsen et al. (2014). The weight w_t is defined as:

$$w_t = \begin{cases} \lambda & \text{if } y_{t+1} = 0, \\ 1 & \text{if } y_{t+1} = 1, \end{cases} \tag{34}$$

where $\lambda > 1$ increases the cost of false positives.

The weighted log-likelihood function becomes:

$$\mathcal{L}_{\text{weighted}}(\beta) = -\frac{1}{n}\sum_{t=1}^{n} w_t\left[y_{t+1}\log(p_t) + (1-y_{t+1})\log(1-p_t)\right]. \tag{35}$$

This adjustment is only applied during training. Model evaluation on calibration and test sets remains unbiased. Increasing λ makes the model more conservative, issuing positive forecasts only when highly confident. This typically improves precision at the expense of recall.

In the Logit model, the weighted log-likelihood was implemented using the glm function in R. The decision tree model was fitted using the rpart package with cost-sensitivity specified via the loss matrix in the model control settings.

4.3.2 Conformal prediction

Conformal prediction provides a statistically grounded framework to quantify prediction uncertainty by associating each forecast with a confidence level based on past model performance (Shafer and Vovk, 2008). In binary classification settings, it enables the construction of predictions set that are guaranteed to contain the true label with a specified probability $1-\alpha$, assuming exchangeability.

In the context of this research, conformal predictions is applied to identify days with high confidence upward return predictions. This is achieved by filtering for test-set instances where the model is confident to include only in the positive class ($\Gamma^{\alpha}(X_t = \{1\}$ in the prediction set.

Let the labeled dataset be denoted by $Z = \{(X_t, y_{t+1})\}_{t=1}^{n}$, with feature vectors $X_t \in \mathbb{R}^d$ and binary labels $y_{t+1} \in \{0, 1\}$. The conformal prediction procedure follows these steps:

1. **Dataset partitioning:** Split the data into a training set Z^{train}, a calibration set Z^{calib} and a test set Z^{test}.

2. **Model fitting:** Fit a probabilistic classifier (logistic regression) to Z^{train}. For each $(X_t, y_{t+1}) \in Z^{\text{calib}}$, compute the predicted class probabilities:

$$\hat{p}_t = \hat{p}(y_{t+1} = 1 \mid X_t). \tag{36}$$

3. **Non-conformity score:** For each calibration sample, define the non-conformity score as:

$$s_t = 1 - \hat{p}(y_{t+1} \mid X_t). \tag{37}$$

4. **Quantile threshold:** Compute the conformal threshold as the $(1-\alpha)$-quantile of the calibration scores:

$$q_\alpha = \text{Quantile}_{1-\alpha}\left(\{s_t : (X_t, y_{t+1}) \in Z^{\text{calib}}\}\right). \tag{38}$$

5. **Prediction set construction:** For each test sample $X_t \in Z^{\text{test}}$, construct the prediction set

$$\Gamma^\alpha(X_t) = \{y \in \{0,1\} : 1 - \hat{p}(y \mid X_t) \leq q_\alpha\}. \tag{39}$$

The prediction set $\Gamma^\alpha(X_t)$ satisfies the marginal coverage guarantee:

$$P\left(y_{t+1} \in \Gamma^\alpha(X_t)\right) \geq 1 - \alpha.$$

Conformal prediction provides a non-parametric alternative for controlling prediction risk. However, its effectiveness depends strongly on the calibration quality. Since financial applications are typically non-i.i.d and non-stationary, the calibration becomes more challenging, as regime shifts may effect the validity of the coverage guarantee.

4.4 Robustness check: Decision tree approach

As a robustness check, a decision tree classifier is estimated on the same feature set as the baseline Logit model. Decision trees are non-parametric models that recursively partition the feature space via binary splits, aiming to minimize node-impurity (James et al., 2013). The model defines a classification function f_{tree} such that:

$$\hat{y}_{t+1} = f_{\text{tree}}(X_t), \quad \hat{y}_{t+1} \in \{0,1\}. \tag{40}$$

Model training and evaluation follow the known train-calibration-test structure.

4.5 Evaluation

The objective of this study is to identify high-confidence predictions of positive daily return directions. Hence, the evaluation focuses on the out-of-sample

precision of the forecasts:

$$\text{Precision} = \frac{\sum_{t \in Z^{\text{test}}} \mathbf{1}\left[\hat{y}_{t+1} = 1 \ \wedge \ \hat{y}_{t+1} = y_{t+1}\right]}{\sum_{t \in Z^{\text{test}}} \mathbf{1}\left[\hat{y}_{t+1} = 1\right]} = \frac{TP}{TP + FP} \tag{41}$$

where $\hat{y}_{t+1} = \mathbf{1}[\hat{p}_t > 0.5]$ denotes the predicted return direction, y_{t+1} the actual return direction and Z^{test} denotes the out-of-sample evaluation set.

Model parameters are selected to maximize precision on the calibration set. To ensure a sufficient number of positive predictions, the positive prediction rate (PPR) is monitored:

$$\text{Positive prediction rate} = \frac{1}{|Z^{\text{test}}|} \sum_{t \in Z^{\text{test}}} \mathbf{1}[\hat{y}_{t+1} = 1] = \frac{TP + FP}{TP + FP + TN + FN}. \tag{42}$$

To avoid solutions with near-zero prediction activity, a lower bound of 2.5% is imposed on the PPR. The use of PPR in place of recall ensures a consistent tracking of model performance across all approaches. Since some of the methods in this research emphasize selective high-confidence predictions, traditional recall can be misleading. Instead, the PPR captures the models selectiveness. The overall evaluation is thus based on the trade-off between maximizing precision and maintaining a minimum prediction level.

Future research may formalize this evaluation approach through custom loss functions to model the precision-coverage trade-off. Alternatively, a tailored metric such a precision weighted F1-score can be applied.

5 Empirical analysis

This section presents the results of the empirical analysis into the directional predictability of financial asset prices using binary response models. The entire modeling is evaluated on a separate test set to ensure out-of-sample validity and avoid look-ahead bias. The forecast performances are assessed using precision and the positive prediction rate as described in Chapter 4.5. To examine the economic relevance, the predictions will be used to construct trading strategies.

5.1 Confidence interval approach

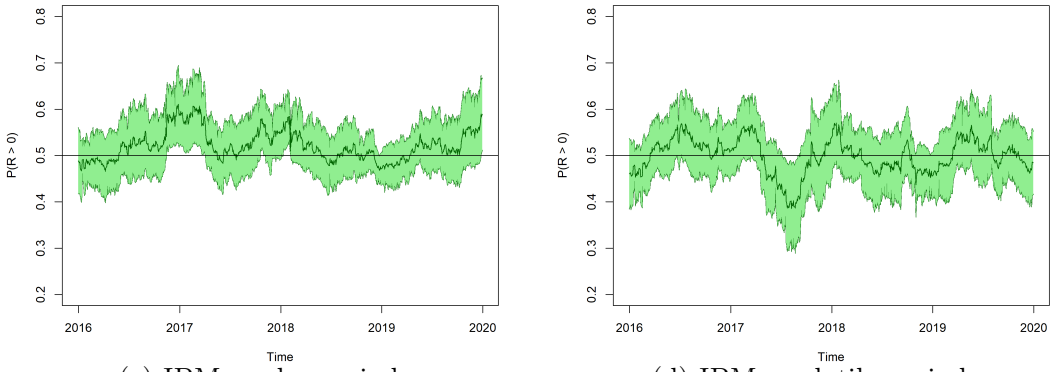
In this rolling window forecasting setup, the window length determines the number of past log returns and realized volatilities used to compute both the mean and the volatility forecasts. Given that return dynamics differ across assets, the optimal window length is selected individually for each series. Short windows (e.g. 20 or 30 days) may respond more quickly to structural breaks but are sensitive to noise, while very long windows (e.g. 250 days) may oversmooth meaningful shifts.

The significance level α controls the width of the predictive confidence intervals. Lower values of α lead to wider intervals and more conservative predictions, while higher values increase prediction coverage but may not filter a sufficient number of days. An α-value of 0.1, corresponding to 90% confidence intervals, has been selected.

Figure 6: Predicted probabilities and confidence intervals of the Christoffersen model

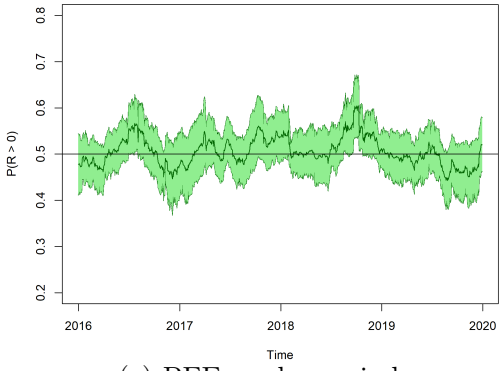
(a) JPM - calm period

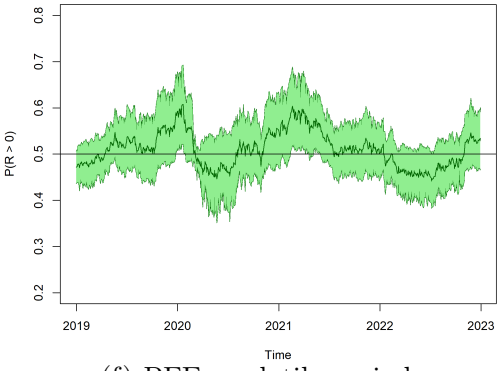
(b) JPM - volatile period



(c) IBM - calm period

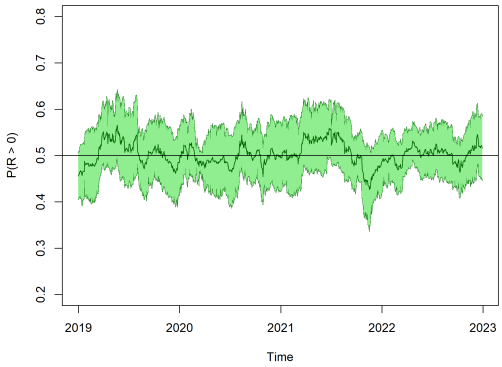
(d) IBM - volatile period

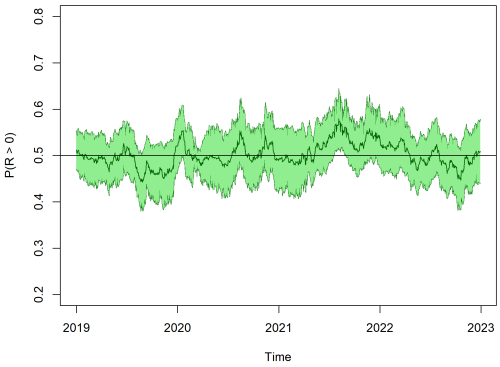




(e) PFE - calm period

(f) PFE - volatile period





Note: P1, P2 refer to the calm and volatile periods, respectively. The thick lines indicate the predicted probabilities.

The probability prediction plots in Figure 6 display the time-varying confidence intervals with individual confidence levels. A probability level of 0.5 corresponds to a random probability in this framework. Higher predictive probabilities occur for instances where the conditional mean over the most recent T observations is elevated and the associated forecasted volatility is low,

indicating stronger directional signals. If the lower bound of the constructed confidence interval exceeds the 0.5 threshold, the prediction is interpreted as statistically significant in favor of a positive return at the specified significance level (e.g. 5%).

Table 6: Performance metrics of the Christoffersen model and the CI extension on the test sets

Asset	**Precision**		**PPR**	
	Baseline	**Confident**	**Baseline**	**Confident**
JPM P1	0.5224	0.5682	0.7404	0.1312
IBM P1	0.5218	1.0000	0.5208	0.0030
PFE P1	0.5276	0.5556	0.5286	0.0626
JPM P2	0.5123	0.5172	0.6149	0.0865
IBM P2	0.5293	0.0000	0.5412	0.0010
PFE P2	0.4705	0.5455	0.4335	0.0328

Note: P1, P2 refer to the calm and volatile periods, respectively. Window lengths: 100, significance level: 10%.

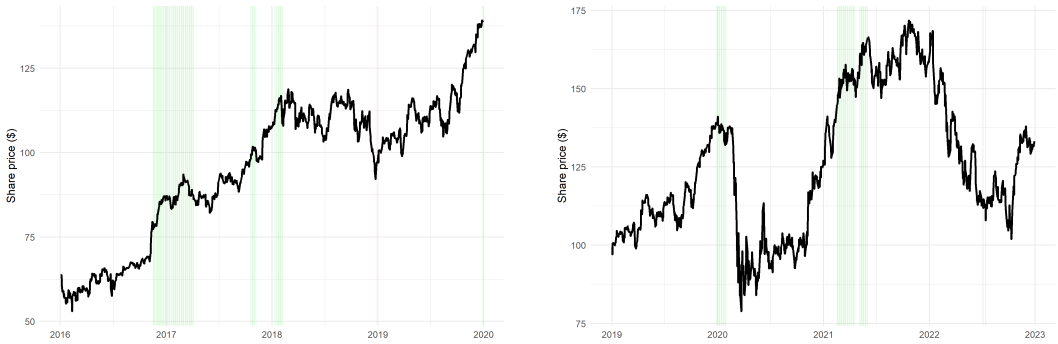
Table 6 summarizes the performances of the standard Christoffersen model and the confidence interval (CI) extension. During the calm market period, the baseline specification consistently achieved precision scores above 52% accompanied by moderately to heavily weighted positive prediction rates. When applying the confidence interval model, the precision improved to levels of 55.6% and 56.8% for JPM and PFE in the calm period. In the volatile period, the precision levels of JPM and PFE improved less pronounced to levels of 51.7% and 54.6%. For IBM, the model did not select a significant amount of days with positive predicted return directions. Overall, the results indicate robust improvements in predictive accuracy when applying the confidence interval extension.

To evaluate the modeled statistically significant prediction on realized prices, the corresponding days are highlighted on the asset's share price plot.

Figure 7: Share price plots with highlighted significant periods during calm and volatile periods

(a) JPM - calm period

(b) JPM - volatile period

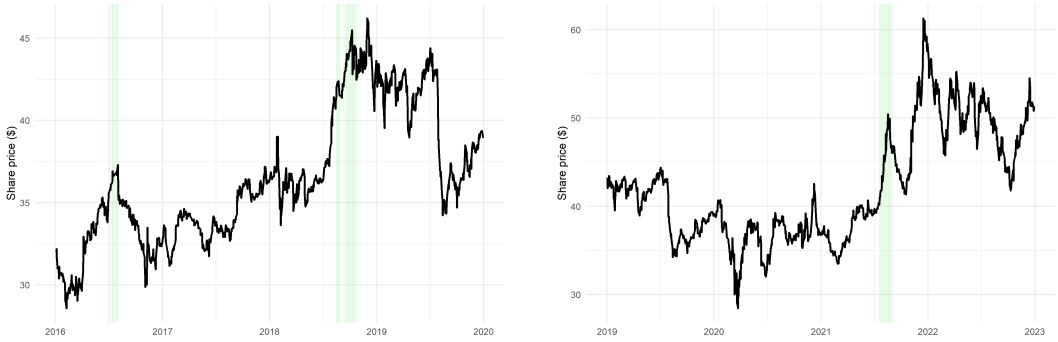


(c) IBM - calm period

(d) IBM - volatile period

(e) PFE - calm period

(f) PFE - volatile period



Note: Green highlighted periods indicate significant positive, red symbolizes significant negative predicted days. Significance is determined by confidence intervals above/below the 0.5 threshold.

Figure 7 shows that the model captures key movements in the stock price dynamics. The highlighted days often align with actual market trends, suggesting that the model detects relevant patterns. However, not all directional changes have been detected correctly. The model tends to issue predictions around saddle points after spikes, due to their characteristics of lower volatility. Selecting shorter windows for computing rolling return means increases the models responsiveness to return dynamics, but also increases the risk of capturing more noise rather than systematic patterns.

5.2 Logistic regression approach with supervised learning

This section analyzes the supervised classification setup using a logistic regression (Logit) model. As outlined in Chapter 3.3, the feature set includes statistical, technical, calendaric and macroeconomic indicators. The estimated parameters of the Logit baseline models are reported in Table 14 of the Appendix. Due to the rather low correlations among the features, a feature selection approach has not been applied in this analysis. However, future work may investigate model-specific or asset-specific feature selection procedures to improve predictive efficiency.

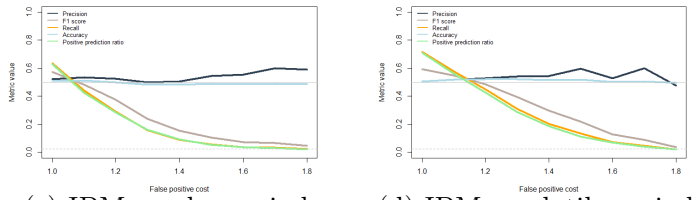
5.2.1 Cost-sensitive learning

In this supervised framework, a Logit model is trained under a cost-sensitive specification that penalizes false positive predictions. The corresponding weighting parameter is selected on a hold-out calibration set and subsequently applied to the test set. Due to the non-stationarity, noise and potential of structural breaks in financial return series, backtest overfitting remained a major concern.

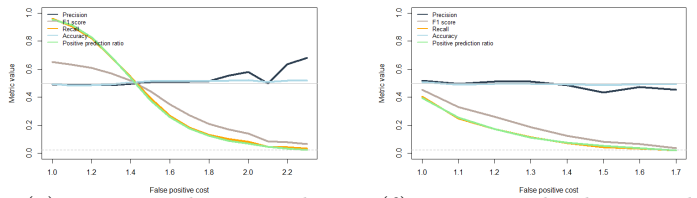
The following figure displays calibration set metrics across different cost weights:

Figure 8: Cost-sensitive calibration results of the Logit model

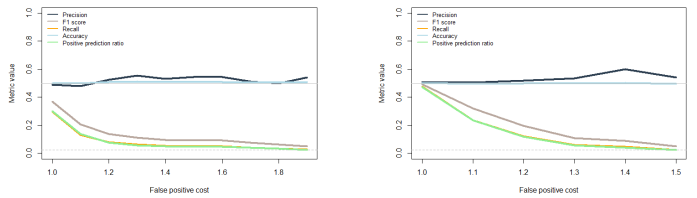
(a) JPM – calm period (b) JPM – volatile period



(c) IBM – calm period (d) IBM – volatile period



(e) PFE – calm period (f) PFE – volatile period



Note: Precision, F1 score, accuracy, recall, and positive prediction rate across varying false positive cost weights in the Logit model. Metrics are computed on the calibration set.

The cost parameter on false positives is iteratively increased in steps of 0.1 starting from the baseline of 1, until the rate of positive predictions falls

below a threshold of 2.5%. The calibration plots in Figure 8 indicate that increasing the cost parameter improves precision in several cases. During the calm periods, all three assets show higher precision under penalized estimation, suggesting that false positive control helps filter noisy signals. In volatile periods, precision also tends to improve, but the effect is less consistent. A reason might be a greater uncertainty of separating true positives from false positives under unstable market conditions. The estimated parameters of the Logit cost-sensitive model are reported in Table 15 of the Appendix.

Compared to the baseline Logit model, the cost-sensitive model extension selects only a small subset of instances for positive prediction.

Table 7: Conformal prediction metrics of Logit models evaluated on calibration and test sets

Asset	**Cost**	**Precision**		**PPR**	
		Calib	**Test**	**Calib**	**Test**
JPM P1	1.7	0.600	0.714	0.030	0.035
IBM P1	2.2	0.636	0.455	0.033	0.033
PFE P1	1.3	0.554	0.591	0.056	0.066
JPM P2	1.7	0.600	0.520	0.040	0.025
IBM P2	1.3	0.513	0.546	0.112	0.182
PFE P2	1.4	0.600	0.507	0.040	0.147

Note: P1, P2 refer to the calm and volatile periods, respectively. Table displays metrics of in-sample calibration set and out-of-sample test set for the Logit conformal prediction model.

The performance metrics in Table 7 reported on the calibration set show a divergence between results obtained on the calibration and test sets, particularly during volatile periods. A careful selection of tuning parameters is therefore substantial. The effect of cost-sensitive tuning on precision is not uniformly positive across all stocks. For JPM and PFE stocks in the calm period the precision scores remain stable or improve out-of-sample. In contrast, other cases show reduced precision in the test set despite improved calibration results. This highlights that the benefits of asymmetric penalization are stock- and regime-dependent. Aggressive tuning may reduce robustness if not adequately regularized.

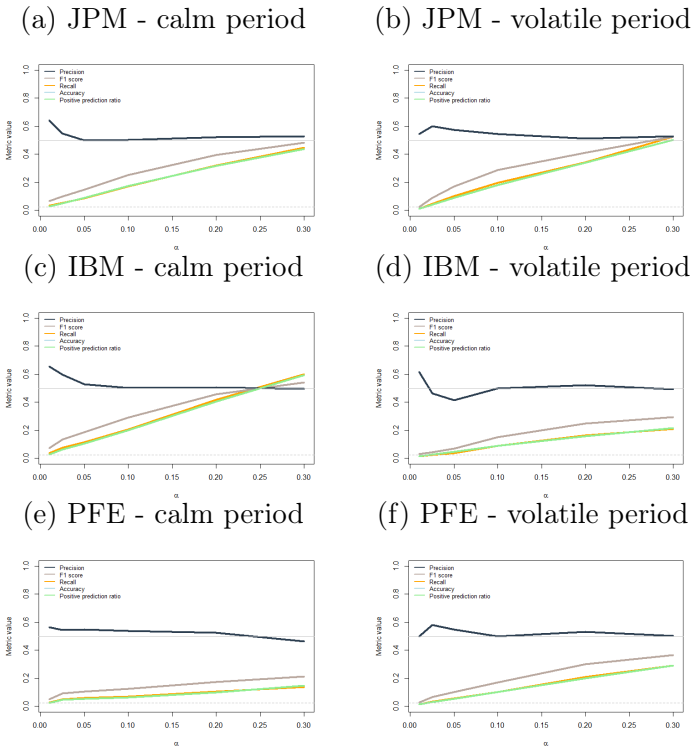
5.2.2 Conformal prediction

Similar to the cost-sensitive learning approach, the performances of the conformal prediction model are first evaluated on the calibration set to select the significance level α. This parameter controls the confidence level of the predictive sets, where lower values of α correspond to higher confidence and narrower prediction sets. The optimal α is selected to maximize predictive precision

while maintaining a sufficient rate of positive predictions to ensure economic relevance. The selected parameter is then applied to the test set to assess out-of-sample performance. This procedure allows the model to adapt the level of predictive confidence to the data structure. The grid of α-candidates is:

$$\alpha \in \{0.01, 0.025, 0.05, 0.10, 0.20, 0.3\}.$$

Figure 9: Conformal prediction calibration results of the Logit model



Note: Precision, F1 score, accuracy, recall, and positive prediction ratio for iteratively increased confidence levels. Metrics are computed on the calibration set.

Precision rates tend to increase at lower significance levels under the conformal prediction model, as shown in Figure 9. Pronounced spikes in precision are observed up to $\alpha = 5\%$, suggesting that stricter thresholds can improve predictive precision. However, in volatile periods, the relationship between α and the realized precision becomes less stable, indicated by irregular spikes of α.

Table 8: Conformal prediction metrics of Logit models evaluated on calibration and test sets

Stock	α	Precision		PPR	
		Calib	Test	Calib	Test
JPM P1	0.010	0.643	0.774	0.028	0.031
IBM P1	0.010	0.655	0.419	0.029	0.031
PFE P1	0.050	0.547	0.746	0.053	0.067
JPM P2	0.025	0.600	0.519	0.040	0.027
IBM P2	0.200	0.522	0.512	0.158	0.241
PFE P2	0.025	0.581	0.520	0.031	0.122

Note: P1, P2 refer to the calm and volatile periods, respectively. Table displays metrics of in-sample calibration set and out-of-sample test set for the Logit conformal prediction model.

A comparison between the in-sample calibrated parameters in the calibration set with the out-of-sample test set in Table 8 shows similar performance metrics for JPM and PFE in the calm period. For IBM, performance drops in the test set, consistent with other model variants. In the volatile period, test set precision stabilizes around 52%. The number of upward predictions remains similar between calibration and test sets in the calm period, indicating consistent model behavior across samples.

5.2.3 Evaluation of performance metrics metrics across Logit models

To assess the effect of cost-sensitive learning and conformal prediction on classification performance, Table 9 compares the precision scores across all Logit-based model variants.

Table 9: Evaluation of precision and PPR metrics across Logit models

Asset	Cost	α	Precision			PPR		
			Normal	CS	CP	Normal	CS	CP
JPM P1	1.7	0.010	0.522	0.714	0.774	0.690	0.035	0.031
IBM P1	2.2	0.010	0.533	0.455	0.419	0.871	0.033	0.031
PFE P1	1.3	0.050	0.547	0.591	0.746	0.338	0.066	0.067
JPM P2	1.7	0.025	0.517	0.520	0.519	0.607	0.025	0.027
IBM P2	1.3	0.200	0.503	0.546	0.512	0.452	0.182	0.241
PFE P2	1.4	0.025	0.492	0.507	0.520	0.614	0.147	0.122

Note: P1, P2 refer to the calm and volatile periods, respectively. Table displays out-of-sample test set metrics across all Logit model variants.

Both extensions improve precision in nearly all cases compared to the baseline Logit classifier, with the exception of IBM in the calm period. In volatile periods, only marginal gains in precision are observed across both extensions, highlighting the challenge of achieving consistent improvements under market stress. In contrast, JPM and PFE show notable precision gains during calm periods.

Overall, the cost-sensitive and conformal prediction models yield similar patterns in terms of both precision and PPR, with particularly close alignment in the calm period. On average, the conformal prediction approach performs slightly better out-of-sample. While the baseline model shows higher variability in rates of positive predictions, both extensions converged mostly to values around 3-4% in the calm period and expand to a broader range of 7-45% in the volatile period.

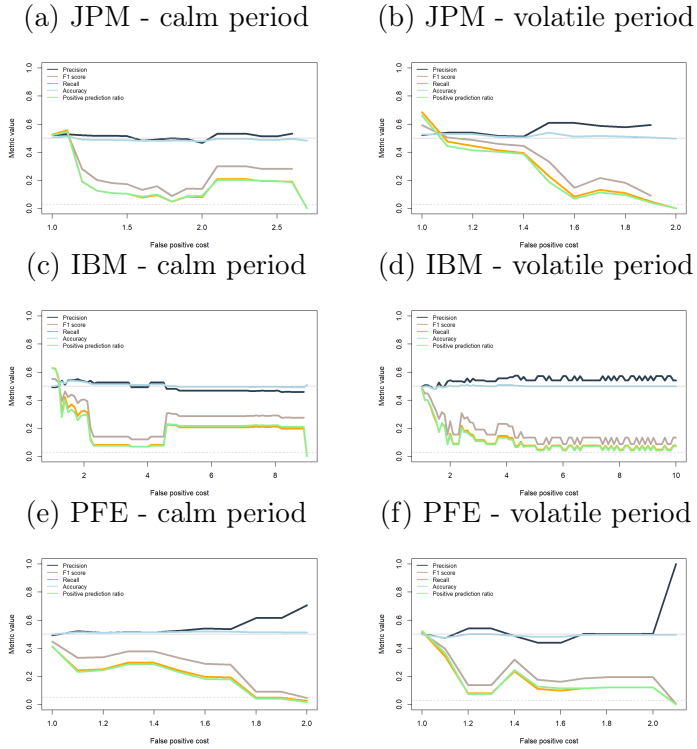
5.3 Robustness check: Decision Tree approach

The same methodology used for the Logit models has been subsequently applied to a Decision Tree classifier. It was initially believed that Decision Trees could effectively capture nonlinear interactions in volatility indicators and past return dynamics in line with the Christoffersen framework. However, the decision tree model showed strong sensitivity to overfitting, particularly during hyperparameter tuning (e.g. maximum depth and complexity) and parameters applied in the extensions. As a result, the out-of-sample performance was unstable, with inconsistent precision gains across market regimes. A visualization of the fitted Decision Tree for JPM during the calm period is reported in Figure 15 of the Appendix.

5.3.1 Cost-sensitive learning

The following chart represents the metric performances as the cost of false increases:

Figure 10: Cost-sensitive calibration results of the Decision Tree model



Precision, F1 score, accuracy, recall, and positive prediction rate across varying false positive cost weights in the Logit model. Metrics are computed on the calibration set.

Compared to the Logit model, the metrics of the Decision Tree model in Figure 10 exhibit substantially higher variability. Noticeable and consistent improvements are observed only for PFE in the calm period and JPM in the volatile period. Overall, the evaluation metrics indicate limited robustness of the decision tree model under cost-sensitive tuning.

Table 10: Conformal prediction metrics of decision tree models evaluated on calibration and test sets

Asset	**Cost**	**Precision**		**PPR**	
		Calib	**Test**	**Calib**	**Test**
JPM P1	2.6	0.532	0.486	0.185	0.176
IBM P1	1.8	0.550	0.533	0.256	0.209
PFE P1	1.6	0.541	0.523	0.180	0.173
JPM P2	1.5	0.611	0.506	0.189	0.495
IBM P2	4.3	0.576	0.535	0.066	0.128
PFE P2	1.2	0.541	0.519	0.074	0.289

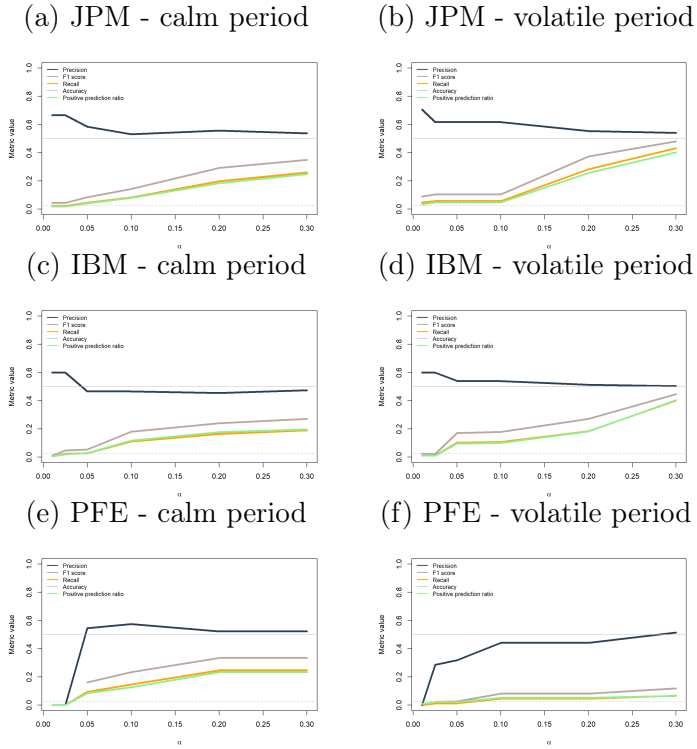
Note: P1, P2 refer to the calm and volatile periods, respectively. Alpha parameters were tuned on calibration set and then applied on OOS test set.

Table 10 reports precision and PPR metrics on the calibration and test set. While precision are high in the calibration set across all stocks and regimes, test set precision declines notably. This is consistent with the tendency of

tree-based models to overfit on financial data. In addition, PPR rates vary substantially between calibration and test sets in the volatile period, particularly for JPM and PFE. Overall, despite strong in-sample performance, the instability of out-of-sample test results limits the applicability in this setting.

5.3.2 Conformal prediction

Figure 11: Conformal prediction calibration results of the Decision Tree model



Note: Performance metrics for iteratively increased significance levels. Applied decision tree controls: minsplit = 10, maxdepth = 10, cp = 0.004.

Figure 11 shows the performance metrics of conformal prediction extension applied to the Decision Tree model. For JPM and IBM, the model achieves high precision at lower significance levels. In contrast, the performance for PFE is less stable, with precision falling below the 0.5 at higher α levels. This suggests that further tuning of model complexity may be necessary.

Table 11: Conformal prediction metrics of Logit models evaluated on calibration and test sets

Asset	α	Precision		PPR	
		Calib	Test	Calib	Test
JPM P1	0.050	0.585	0.491	0.041	0.057
IBM P1	0.300	0.475	0.524	0.197	0.232
PFE P1	0.100	0.575	0.510	0.126	0.148
JPM P2	0.010	0.706	0.500	0.034	0.016
IBM P2	0.050	0.542	0.490	0.095	0.154
PFE P2	0.300	0.515	0.526	0.066	0.289

Note: P1, P2 refer to the calm and volatile periods, respectively. Table displays metrics of in-sample calibration set and out-of-sample test set for the Logit conformal prediction model. Precision and PPR values are rounded to three decimal places.

Table 11 shows substantial differences between calibration and test set performance, consistent with the pattern observed in the cost-sensitive Decision Tree model.

Table 12: Evaluation of precision and PPR metrics across Decision Tree models

Asset	Cost	α	Precision			PPR		
			Base	CS	CP	Base	CS	CP
JPM P1	2.6	0.05	0.502	0.486	0.491	0.577	0.176	0.057
IBM P1	1.8	0.30	0.524	0.533	0.524	0.581	0.209	0.232
PFE P1	1.6	0.10	0.529	0.523	0.510	0.412	0.173	0.148
JPM P2	1.5	0.01	0.517	0.506	0.500	0.535	0.495	0.016
IBM P2	4.3	0.05	0.515	0.535	0.490	0.562	0.128	0.154
PFE P2	1.2	0.30	0.505	0.519	0.526	0.687	0.289	0.289

Note: P1, P2 refer to the calm and volatile periods, respectively. Table displays out-of-sample test set metrics across all decision tree model extensions.

Overall, the model extensions of the Decision Tree model did not lead to substantial improvements in precision, as shown in Table 12. In some cases, predictive performances declined relative to the baseline. This indicates that added model complexity does not reliably translate into better out-of-sample predictive performance under the current setup.

5.4 Economic significance

The economic significance of the models can be assessed through several approaches. A particularly intuitive and interpretable method involves linking

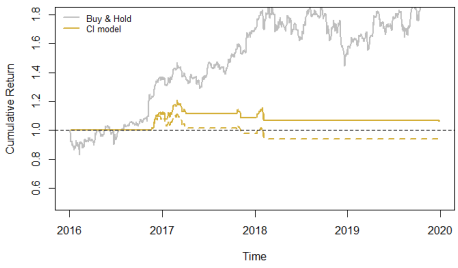
model-generated predictions to a simple trading strategy. Specifically, predicted upward movements are compared to actual realized returns of the following day to determine the profitability of the forecasting models. The corresponding investment strategy scheme is defined as follows:

1. At the end of day t, enter a long position if the model predicts a positive return for day $t+1$.
2. Hold the position overnight and sell at the closing price on day $t+1$.
3. Repeat this process over the test period and report the cumulative returns.

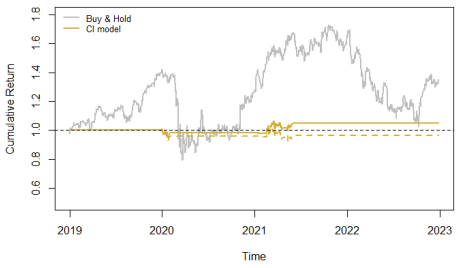
The following output plots display the cumulative returns of the buy & hold benchmark and the respective model approach. Dotted lines indicate the cumulative returns of the models with transaction costs. The transaction costs are set to 0.1% round-trip (i.e. 0.05% per transaction for entering and exiting a position).

Figure 12: Cumulative return plots of Confidence Interval model

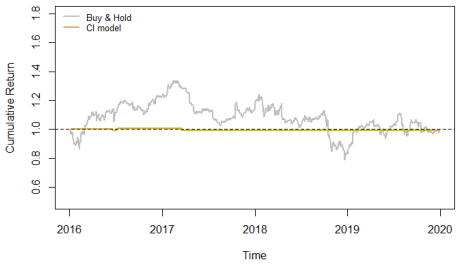
(a) JPM – Calm period



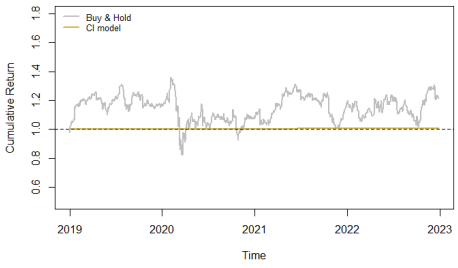
(b) JPM – Volatile period



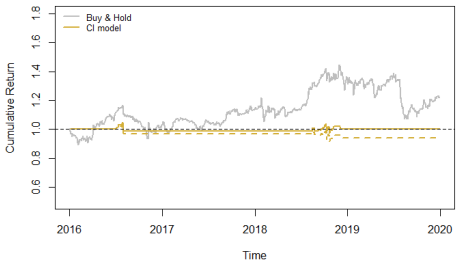
(c) IBM – Calm period



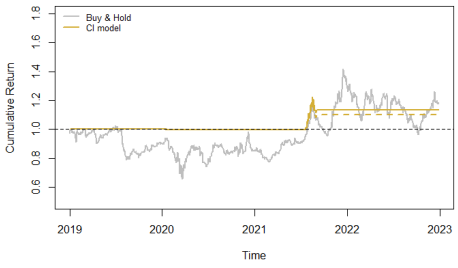
(d) IBM – Volatile period



(e) PFE – Calm period



(f) PFE – Volatile period



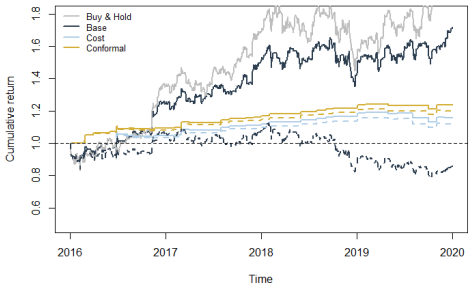
Note: Dotted lines indicate returns with transaction costs. Yellow line indicates cumulative returns where Christoffersen's CI model made upward predictions on 90% significance level.

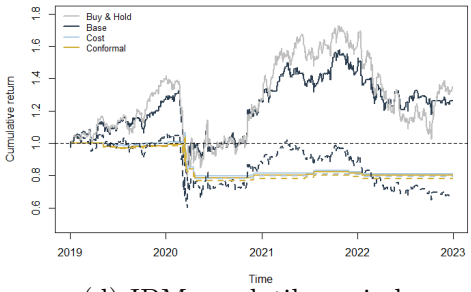
The cumulative return plots of the confidence interval model in Figure 12 report increases in cumulative returns for JPM in the calm and volatile period and PFE in the volatile period. With the current model settings of 90% confidence intervals, the model does not yield confident predictions for IBM, which is consistent with poor model performance of the Logit classifiers for IBM. Despite the confidence interval methods simplicity, major losses in returns are not reported. Further exploration of this method, such as increasing responsiveness or different methods for mean or volatility modeling, might improve the economic value of this method. Table 18 in the Appendix reports the return statistics including average returns, standard deviations and sharpe ratios.

Figure 13: Cumulative returns of logistic regressions by model and market regime

(a) JPM – calm period

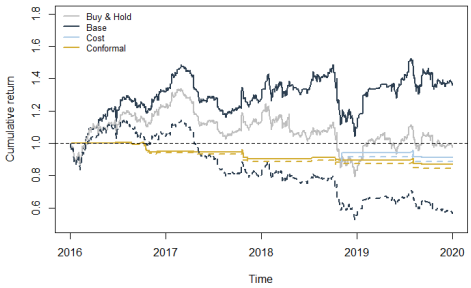
(b) JPM – volatile period

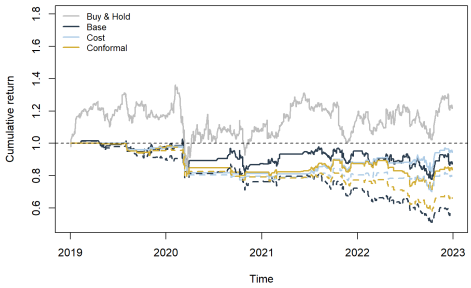




(c) IBM – calm period

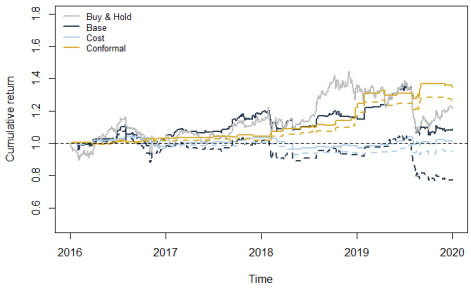
(d) IBM – volatile period

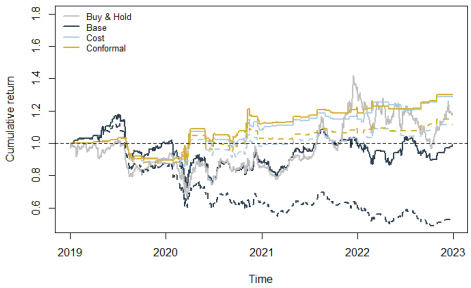




(e) PFE – calm period

(f) PFE – volatile period





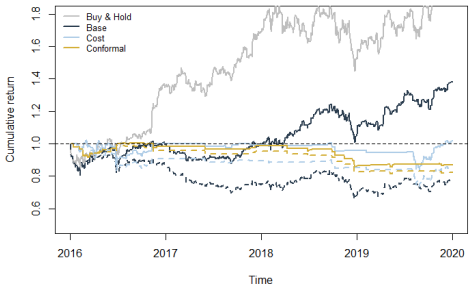
Note: Dotted lines indicate returns with transaction costs. Plots show cumulative returns of baseline, cost-sensitive, and conformal prediction specifications of Logit model.

The assessment of cumulative returns of the Logit models in Figure 13 indicates that the application of extensions resulted in positive and stable cumulative returns only for JPM and PFE, both with the cost-sensitive and conformal prediction model. During volatile periods, only the conformal prediction strategy for PFE achieved positive returns. It can be observed that

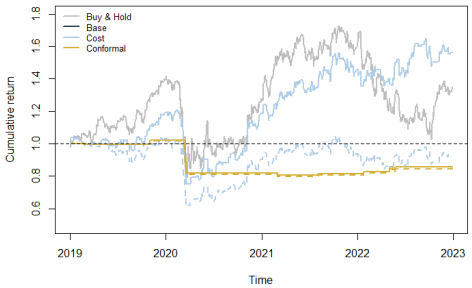
transaction costs had a relatively minor impact on the cumulative returns of the extended models, while their effect had a higher share in the baseline models with more trading instances. Complete return statistics are reported in Table 17 of Appendix.

Figure 14: Cumulative return plots for buy & hold, baseline, cost-sensitive, and conformal prediction decision tree models

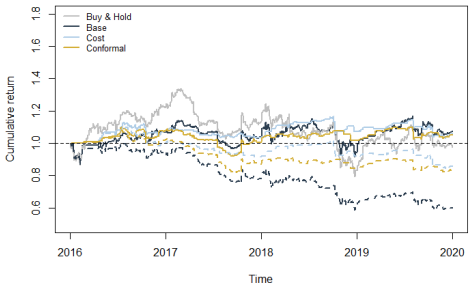
(a) JPM – Calm period



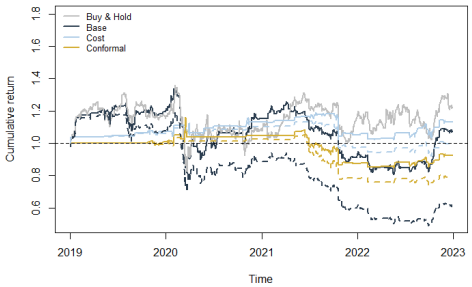
(b) JPM – Volatile period



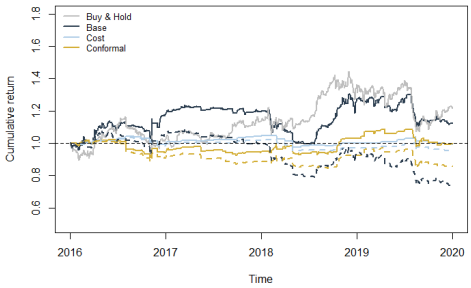
(c) IBM – Calm period



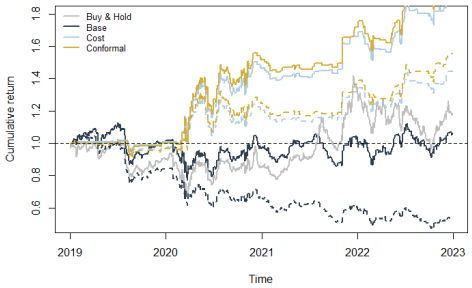
(d) IBM – Volatile period



(e) PFE – Calm period



(f) PFE – Volatile period



Note: Dotted lines indicate returns with transaction costs. Plots show cumulative returns of baseline, cost-sensitive, and conformal prediction specifications of Decision Tree models.

The decision tree models did not exhibit clear patterns in cumulative returns (see Figure 14). Significant positive returns are only observed for JPM and PFE in the volatile periods. Among the two extensions, the cost-sensitive approach performed better than the conformal prediction strategy. However, the plots overall suggest that the tree-based trading strategies were not able to capture directional signals robustly enough to consistently generate positive cumulative returns, which highlights the limited economic value. The return statistics are reported in Table 17 of Appendix.

6 Discussion

6.1 Confidence interval extension

The results in Chapter 5 indicate that incorporating confidence intervals into the baseline framework of Christoffersen and Diebold (2006) can provide improvements for predicting return directions on a daily frequency (see Table 6). For JPM and PFE, the confidence interval models improved predictive performance scores and detected patterns in market trends. For IBM as a more volatile asset with lees pronounced market trends, the model correctly withheld predictions. When confident predictions were issued, cumulative returns indicated robust positive yields in an out-of-sample framework.

From a methodological standpoint, a major advantage of this model approach lies in its simplicity: It only relies on two key inputs, the historical means and the one-step-ahead volatility forecasts, which makes the model tractable and interpretable. The purpose of the confidence intervals is to provide a statistical basis for evaluating whether a directional return prediction significantly deviates from random guessing (i.e. a probability of 0.5). The main strength of this approach lies in its time-varying nature, which allows the model to dynamically adjust to calm or volatile market conditions. Changes in the predicted volatility are directly reflected in the width of the confidence interval: periods of high forecasted volatility lead to wider intervals, whereas calm periods yield narrower ones. This dynamic adaptation provides additional information beyond point predictions and can help to assess the reliability of model outputs. However, a limitation of the current specification lies in the estimation of the variance and covariance structures of the estimated parameters. Currently, the variance of the volatility is approximated by the residual variance from the realized variance AR(1) model. This approach assumes homoskedastic residuals, which may underestimate variances during volatile market conditions. The presence of conditional heteroskedasticity in the residuals suggests that the model does not fully capture the time-varying volatility dynamics. In future work, the incorporation of HAR- or GARCH-type models using high-frequency data might be interesting to further capture characteristics in financial volatility modeling such as volatility clustering and leverage effects.

By construction, the model is not designed to capture abrupt large return spikes. It rather identifies stable, moderate positive returns which are typically associated with low volatility regimes. As such, the approach may be particularly well-suited for risk-averse investors who prioritize robustness and downside protection over aggressive return-seeking. In addition, the model's output might be a valuable input parameter within more advanced, non-linear predictive frameworks aimed at capturing broader market trends. Overall, the framework offers a promising and interpretable extension of the baseline

model, with potentials for further refinement and application. A possible further extension of this framework could involve its application to the skewness and kurtosis model proposed by Christoffersen et al. (2007), which incorporates higher-order moments. This extension would require the computation of variances and covariance for skewness and kurtosis and may yield improvements in predictive accuracy by capturing distributional asymmetries and fat tails that are often present in financial return series.

6.2 Logistic regression approach

The logistic regression classification approach is mainly based on statistical features of asset returns, such as past return means, volatility measures and higher order moments, as well as other calendaric, technical and macroeconomic indicators. An advantage of this methodology lies in its universality, enabling application across a wide range of assets without requiring asset-specific feature engineering. However, this generality inevitable introduces the effect of omitted variable bias, as relevant stock-specific features are excluded from the model. A strength lies in the interpretability of the Logit parameters, particularly compared to non-parametric models. Despite the simplicity of the feature set, both extensions achieved directional out-of-sample accuracy levels of $> 57\%$ for JPM and PFE in the calm market regime, reported in Table 9. However, during high-volatility periods, no clear patterns of improved precision scores were reported, indicating the models' limited robustness across different volatility regimes. The out-of-sample test results suggest that the Logit models are less prone to overfitting and more suitable for parameter calibration methods. Although the Logit models can generate economic gains, such as for JPM and PFE in the calm period (see Figure 13), they did not consistently produce statistically significant and robust positive cumulative returns in the out-of-sample evaluation. This lack of temporal consistency limits their applicability for use in fully automated, real-time trading systems.

Given the limited robustness of the results, an application of the cost-sensitive or conformal prediction Logit models in trading models would require the formulation of adaptive strategies. For example, one could implement portfolio-based allocation strategies, selecting only assets with the highest prediction confidence.

7 Conclusion

This work provided an overview of different model specifications of binary response models for directional prediction of daily returns. Two specifications were implemented: An extension of the direction-of-change model of Christoffersen and Diebold (2006) using delta-method based confidence intervals and a logistic regression, extended by cost-sensitive and conformal prediction frame-

works. The results indicate that linear model approaches are capable of identifying a subset of days with elevated directional predictability in calm periods, although predictive accuracy remains modest and sensitive to market regimes.

In the first specification, incorporating uncertainty via confidence intervals enabled formal testing for significance of the predicted probabilities. The resulting high-confidence predictions were typically concentrated in regimes with clear market trends and lower volatility and were able to identify directional trends during stable periods. Cumulative return curves based on the predictions showed a robust differentiation between periods with no significant movements and those with consistent upward trends, although the overall return levels remained modest.

The second specification investigated the applicability of supervised classification methods to directional return predictions. The models are subject to omitted variable bias due to the exclusion of asset- and market-specific controls. Although precision exceeded the 70% level under calm market regimes, classification performance was not consistent across periods. Moreover, the approach exhibited a high sensitivity to hyperparameter tuning, with signs of overfitting reflected in unstable out-of-sample accuracy and inconsistent cumulative returns.

Both specifications benefit from statistical interpretability and are straightforward to implement. Predictive performance was more stable in low-volatility regimes but deteriorated in volatile periods, indicating sensitivity to regime shifts. Neither model accounts for volatility clustering or leverage effects yet, which remain important aspects for future extensions within this modeling framework. The assumption of i.i.d errors limits the adaptability of the models to market regimes, as financial return series typically exhibit serial dependence and conditional heteroskedasticity. To evaluate the robustness and external validity of the results, it is essential to test the models on a broader and more diverse cross-section of assets.

Despite these limitations, both specifications provide evidence that directional predictability in the direction of returns exists. In particular, the confidence interval model produced promising and relatively robust forecasts. The results of the confidence interval model extend the findings of Christoffersen and Diebold (2006) by showing that directional predictability of direction-of-change models can be present on a daily frequency. The overall findings support further investigation into probabilistic classification frameworks for short-horizon directional forecasting of returns.

References

Ahoniemi, K., & Lanne, M. (2013). Overnight stock returns and realized volatility. *International Journal of Forecasting*, *29*(4), 592–604.

Arian, H., Mobarekeh, D. N., & Seco, L. (2024). Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment. *Knowledge-Based Systems*.

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring. *2014 13th International conference on machine learning and applications*, 263–269.

Bajgrowicz, P., & Scaillet, O. (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, *106*(3), 473–491.

Campisi, G., Muzzioli, S., & De Baets, B. (2024). A comparison of machine learning methods for predicting the direction of the us stock market on the basis of volatility indices. *International Journal of Forecasting*, *40*(3), 869–880.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd). Duxbury.

Christoffersen, P. F., & Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, *52*(8), 1273–1287.

Christoffersen, P. F., Diebold, F. X., Mariano, R. S., Tay, A. S., & Tse, Y. K. (2007). Direction-of-change forecasts based on conditional variance, skewness and kurtosis dynamics: International evidence. *Journal of Financial Forecasting*, *1*(2).

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196.

Corsi, F., Mittnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, *27*(1-3), 46–78.

Elkan, C. (2001). The foundations of cost-sensitive learning. *International joint conference on artificial intelligence*, *17*(1), 973–978.

Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383–417.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Leung, M. T., Daouk, H., & Chen, A.-S. (2000). Forecasting stock indices: A comparison of classification and level estimation models. *International Journal of forecasting*, *16*(2), 173–190.

Linton, O., & Whang, Y.-J. (2007). The quantilogram: With an application to evaluating directional predictability. *Journal of Econometrics*, *141*(1), 250–282.

Ma, C., Liu, Z., Cao, Z., Song, W., Zhang, J., & Zeng, W. (2020). Cost-sensitive deep forest for price prediction. *Pattern Recognition*, *107*.

Nyberg, H. (2011). Forecasting the direction of the us stock market with dynamic binary probit models. *International Journal of Forecasting*, *27*(2), 561–578.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*(3).

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

Xu, J., & Long, J. S. (2005). Using the delta method to construct confidence intervals for predicted probabilities, rates, and discrete changes. *The Stata Journal*, *5*(4), 537–559.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial innovation*, *5*(1), 1–20.

Appendices

Table 13: Descriptive statistics of selected features across all periods

Feature	JPM					IBM					PFE				
	Mean	SD	Q10	Q50	Q90	Mean	SD	Q10	Q50	Q90	Mean	SD	Q10	Q50	Q90
mean_20day	0.00	0.00	-0.00	0.00	0.01	0.00	0.00	-0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00
sd_20day	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02
skew_20day	0.09	0.63	-0.65	0.07	0.88	-0.07	0.87	-1.16	-0.00	0.80	0.03	0.66	-0.71	0.04	0.80
kurt_20day	-0.08	1.16	-1.14	-0.40	1.39	0.35	1.96	-1.13	-0.38	3.08	-0.08	1.31	-1.13	-0.43	1.28
mean_5day	0.00	0.01	-0.01	0.00	0.01	0.00	0.01	-0.01	0.00	0.01	0.00	0.01	-0.01	0.00	0.01
sd_5day	0.02	0.02	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02
rv_for	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rs_for	0.05	0.13	-0.09	0.04	0.20	0.09	0.16	-0.09	0.08	0.29	-0.01	0.15	-0.20	0.01	0.15
rk_for	2.47	0.92	1.50	2.26	3.63	2.53	0.98	1.30	2.52	3.74	2.64	1.14	1.45	2.45	3.83
rv	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rs	0.053	0.822	-0.821	0.035	0.955	0.082	0.863	-0.825	0.054	1.017	-0.005	0.892	-0.954	0.018	0.917
rk	2.386	3.266	-0.046	1.358	6.006	2.486	3.573	-0.084	1.387	6.151	2.650	3.879	-0.026	1.436	6.574
daily_mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
daily_range	0.021	0.018	0.008	0.016	0.037	0.015	0.010	0.007	0.012	0.026	0.016	0.010	0.008	0.014	0.028
overnight_gap	0.003	1.868	-1.645	-0.034	1.725	-0.041	1.154	-1.289	-0.049	1.214	0.030	1.246	-1.338	0.014	1.444
daily_range_var	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	3.022	1.398	1.000	3.000	5.000
z_score	0.07	1.17	-1.39	0.02	1.59	0.05	1.17	-1.39	0.03	1.51	0.02	1.06	-1.29	-0.02	1.40
day_of_week	3.02	1.40	1.00	3.00	5.00	3.02	1.40	1.00	3.00	5.00	3.02	1.40	1.00	3.00	5.00
is_month_end	0.05	0.21	0.00	0.00	0.00	0.05	0.21	0.00	0.00	0.00	0.05	0.21	0.00	0.00	0.00
first_trading_day_of_month	0.05	0.21	0.00	0.00	0.00	0.05	0.21	0.00	0.00	0.00	0.05	0.21	0.00	0.00	0.00
turn_of_month	0.21	0.41	0.00	0.00	1.00	0.21	0.41	0.00	0.00	1.00	0.21	0.41	0.00	0.00	1.00
RSI	52.48	11.75	37.07	52.48	67.64	51.15	12.33	35.14	51.30	67.14	50.25	11.92	34.39	50.35	66.01
MACD	0.14	2.13	-2.24	0.35	2.27	0.02	1.61	-1.71	0.15	1.81	-0.07	1.52	-1.91	-0.03	1.78
5PROC	0.33	6.94	-6.99	0.67	7.37	0.06	5.13	-5.48	0.32	5.75	-0.08	4.87	-5.88	0.16	5.43
econ_unc	111.987	80.426	40.124	91.190	203.140	111.987	80.426	40.124	91.190	203.140	111.987	80.426	40.124	91.190	203.140
vix	19.649	8.741	11.990	17.310	30.040	19.649	8.741	11.990	17.310	30.040	19.649	8.741	11.990	17.310	30.040
ten_year_yield	3.038	1.159	1.610	2.910	4.608	3.038	1.159	1.610	2.910	4.608	3.038	1.159	1.610	2.910	4.608

Table 14: Logit model coefficient estimates with significance levels

feature	JPM P1	IBM P1	PFE P1	JPM P2	IBM P2	PFE P2
(intercept)	0.222	0.493	-0.208	0.833	-0.764	0.097
mean_20day	11.975	14.114	-33.165	4.780	-27.596	-37.779
sd_20day	-10.829	-2.268	4.834	-6.361	-30.736*	4.861
skew_20day	0.113	0.097	-0.049	0.154**	-0.022	-0.018
kurt_20day	0.021	0.021	-0.020	0.007	0.061**	-0.008
mean_5day	-8.571*	4.208	-11.435	-5.767	8.666	-8.744
sd_5day	12.803***	6.760	3.010	10.570**	6.509	5.591
rv_for	-4.706	-21.757	-16.501	-3.516	-5.866	-6.507
z_score	-0.080	0.023	0.004	-0.106**	0.049	-0.052
rs_for	0.178	0.280	-0.484	0.581	-0.139	-0.493
rk_for	0.016	0.100	-0.027	-0.016	0.082	0.001
rs	0.064	0.108	0.069	0.129**	0.083	0.069
rk	-0.016	0.014	-0.020	-0.024*	0.017	-0.022*
daily_mean	-2735.124	3892.603	-23876.175	-4751.623	1477.877	-27356.946
rv	46.930	78.829	87.441	37.979	-96.403	-78.190
daily_range	-1.238	9.280	3.474	-1.698	2.201	-9.975
overnight_gap	-0.340	0.591	-3.076	-0.616	0.275	-3.560
daily_range_var	19.641	-148.775	99.970	108.706	-356.192	585.992
day_of_week	-0.036	-0.017	-0.022	-0.055*	0.013	-0.006
is_month_end	0.486**	0.186	0.595***	0.207	0.240	0.414*
turn_of_month	-0.012	0.039	0.162	-0.085	-0.076	0.018
month	0.003	0.022*	0.002	0.005	0.009	0.006
RSI	0.001	-0.008	0.008	-0.002	0.001	0.003
MACD	-0.008	-0.028	0.018	-0.032	-0.042	0.018
econ_unc.x	0.001	-0.001	0.000	0.000	-0.001	0.000
vix.x	-0.003	0.012	0.008	-0.013	0.033***	0.000
ten_year_yield.x	-0.020	-0.108	-0.050	-0.084	0.063	-0.068

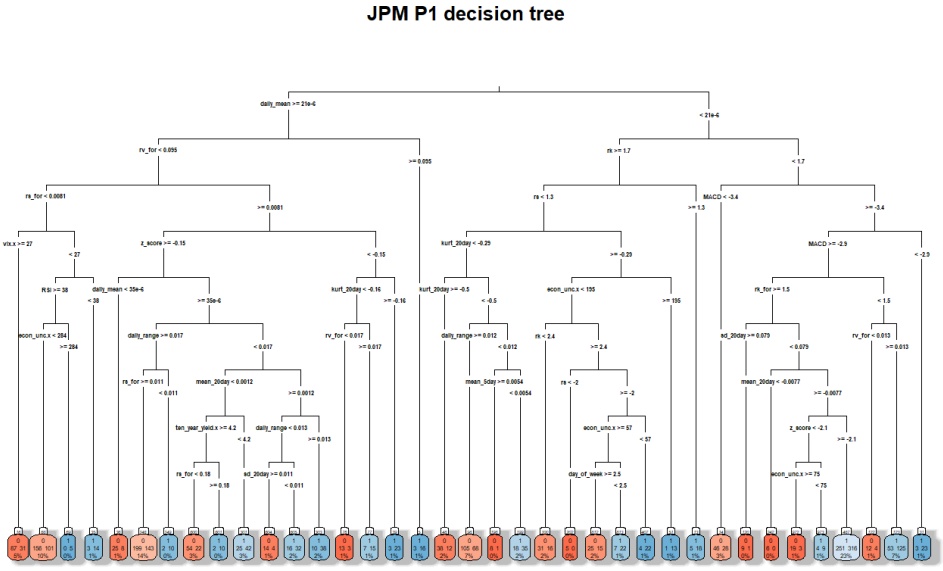
Note: Table reports coefficient estimates from Logit models fitted for each Asset. P1 and P2 denote calm and volatile periods, respectively. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 15: Logit cost-sensitive model coefficient estimates with significance levels

feature	JPM P1	IBM P1	PFE P1	JPM P2	IBM P2	PFE P2
(intercept)	-0.314	-0.336	-0.465	0.286	-0.764	-0.241
mean_20day	11.805	13.693	-32.281	4.357	-27.596	-37.148
sd_20day	-11.274*	-3.058	4.551	-6.771	-30.736*	4.928
skew_20day	0.115	0.091	-0.050	0.156**	-0.022	-0.018
kurt_20day	0.022	0.020	-0.019	0.007	0.061**	-0.009
mean_5day	-8.530*	4.194	-11.516	-5.525	8.666	-8.739
sd_5day	13.252***	7.203	3.035	11.322***	6.509	5.532
rv_for	-4.702	-21.094*	-16.936	-3.654	-5.866	-6.573
z_score	-0.083*	0.019	0.006	-0.107**	0.049	-0.050
rs_for	0.185	0.288	-0.487	0.600*	-0.139	-0.494
rk_for	0.014	0.098	-0.027	-0.018	0.082	0.000
rs	0.070	0.110**	0.072	0.132***	0.083	0.071
rk	-0.016	0.016	-0.020	-0.024*	0.017	-0.022*
daily_mean	-2732.104	-211.496	-24061.167	-4693.251	1477.877	-27577.062
rv	46.916	47.164	93.154	44.017	-96.403	-81.242
daily_range	-1.347	9.352	3.601	-1.507	2.201	-10.105
overnight_gap	-0.342	0.064	-3.098	-0.610	0.275	-3.585
daily_range_var	18.636	-128.229	95.893	96.744	-356.192	591.862
day_of_week	-0.036	-0.017	-0.023	-0.055**	0.013	-0.005
is_month_end	0.493***	0.172	0.596***	0.208	0.240	0.417**
turn_of_month	-0.013	0.042	0.161	-0.080	-0.076	0.016
month	0.002	0.021**	0.002	0.005	0.009	0.006
RSI	0.001	-0.007	0.007	-0.001	0.001	0.003
MACD	-0.007	-0.030	0.017	-0.034	-0.042	0.017
econ_unc.x	0.001	-0.001	0.000	0.000	-0.001	0.000
vix.x	-0.002	0.013	0.009	-0.013	0.033***	0.001
ten_year_yield.x	-0.020	-0.106	-0.049	-0.085*	0.063	-0.068

Note: Table reports coefficient estimates from of cost-sensitive (CS) Logit models for each asset. P1 and P2 denote calm and volatile periods, respectively. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure 15: Plot of JPM Decision Tree splits



Applied decision tree controls: minsplit = 10, cp = 0.004, maxdepth = 10

Table 16: Return statistics of Logit models

Asset	Model	Mean Return (%)	Std. Dev. (%)	Sharpe Ratio
JPM P1	Base	0.0597	1.0804	0.878
	CS	0.0150	0.3010	0.791
	CP	0.0214	0.2668	1.273
JPM P2	Base	0.0400	1.8375	0.345
	CS	-0.0162	0.9576	-0.268
	CP	-0.0172	0.9593	-0.284
IBM P1	Base	0.0388	1.2111	0.509
	CS	-0.0086	0.2682	-0.511
	CP	-0.0136	0.2369	-0.911
IBM P2	Base	-0.0071	1.0819	-0.104
	CS	-0.0020	0.7898	-0.041
	CP	-0.0134	0.8550	-0.249
PFE P1	Base	0.0109	0.7141	0.242
	CS	0.0017	0.2869	0.093
	CP	0.0308	0.3317	1.474
PFE P2	Base	0.0083	1.3668	0.097
	CS	0.0293	0.8893	0.522
	CP	0.0296	0.8558	0.550

Note: P1, P2 refer to the calm and volatile periods. Mean, std. deviation and sharpe ratio are calculated of realized returns of strategy

Table 17: Return statistics of Decision Tree models

Asset	Model	Mean Return (%)	Std. Dev. (%)	Sharpe Ratio
JPM P1	Base	0.0374	0.9826	0.603
	CS	0.0034	0.5787	0.094
	CP	-0.0133	0.3114	-0.680
JPM P2	Base	0.0557	1.4866	0.595
	CS	0.0557	1.4866	0.595
	CP	-0.0133	0.6374	-0.330
IBM P1	Base	0.0115	0.9598	0.191
	CS	0.0067	0.4623	0.230
	CP	0.0065	0.5388	0.191
IBM P2	Base	0.0179	1.4727	0.193
	CS	0.0142	0.5951	0.379
	CP	-0.0057	0.6545	-0.138
PFE P1	Base	0.0146	0.7990	0.290
	CS	-0.0004	0.2379	-0.026
	CP	0.0005	0.4618	0.017
PFE P2	Base	0.0170	1.4947	0.181
	CS	0.0719	1.1525	0.991
	CP	0.0794	1.1564	1.090

Note: P1 and P2 refer to the calm and volatile periods, respectively. Mean returns and standard deviations are daily values reported in percent. Sharpe ratios are computed accordingly.

Table 18: Return statistics of Confidence Interval models

Asset	Mean Return (%)	Std. Dev. (%)	Sharpe Ratio
JPM P1	0.0077	0.4233	0.289
JPM P2	0.0056	0.3577	0.249
IBM P1	-0.0006	0.0551	-0.160
IBM P2	0.0007	0.0225	0.501
PFE P1	0.0006	0.3083	0.029
PFE P2	0.0134	0.3697	0.576

Note: P1 and P2 refer to the calm and volatile periods, respectively. Mean returns and standard deviations are daily values reported in percent. Sharpe ratios are computed accordingly.

Eigenständigkeitserklärung

Berke, Stephan

.....

(Name, Vorname)

17.06.1999

.....

(Geburtsdatum)

5548152

.....

(Matrikelnummer)

Hiermit versichere ich, die vorliegende Arbeit ohne unerlaubte Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt zu haben. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, habe ich als solche kenntlich gemacht. Die eingereichte Masterarbeit wurde weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens. Die elektronische Version der eingereichten Masterarbeit stimmt in Inhalt und Formatierung mit den auf Papier ausgedruckten Exemplaren überein.

Freiburg im Breisgau, 01.07.2025 .....

(Signature)