Meta-Labeling: Calibration and Position Sizing

Michael Meyer, Illya Barziy, and Jacques Francois Joubert

Michael Meyer
is a quantitative researcher at Hudson and Thames Quantitative Research in London, UK.
michael@hudsonthames.org

Illya Barziy
is a quantitative researcher and developer at Abu Dhabi Investment Authority (ADIA) in Abu Dhabi, United Arab Emirates.
illya.barziy@adia.ae

Jacques Francois Joubert
is the chief executive officer of Hudson and Thames Quantitative Research in London, UK.
jacques@hudsonthames.org

KEY FINDINGS

- Model calibration is a necessary step in the meta-labeling pipeline that significantly improves fixed position sizing methods.
- Practitioners are provided with clear insights and guidance concerning which algorithms to apply under which conditions when sizing positions.
- Position sizing outcomes based on predicted probability methods are evaluated and compared in a meta-labeling setting.

ABSTRACT

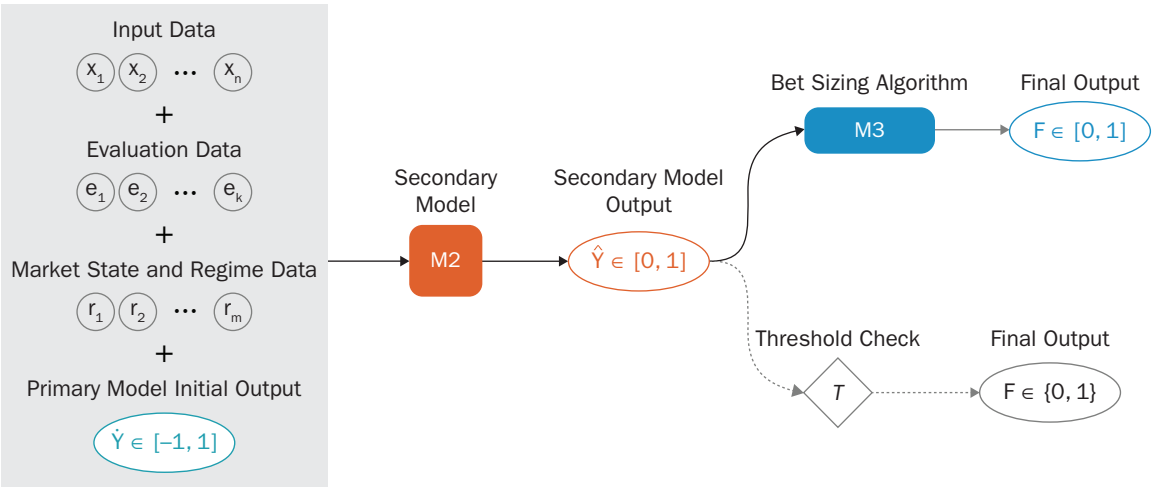
Meta-labeling is a recently developed tool for determining the position size of a trade. It involves applying a secondary model to produce an output that can be interpreted as the estimated probability of a profitable trade, which can then be used to size positions. Before sizing the position, probability calibration can be applied to bring the model's estimates closer to true posterior probabilities. This article investigates the use of these estimated probabilities, both uncalibrated and calibrated, in six position sizing algorithms. The algorithms used in this article include established methods used in practice and variations thereon, as well as a novel method called sigmoid optimal position sizing. The position sizing methods are evaluated and compared using strategy metrics such as the Sharpe ratio and maximum drawdown. The results indicate that the performance of fixed position sizing methods is significantly improved by calibration, whereas methods that estimate their functions from the training data do not gain any significant advantage from probability calibration.

As finance moves into its new era of scientific investing, and causal mechanisms become more widely available (López de Prado 2022), techniques such as meta-labeling provide investors with a systematic approach to dynamically size positions. The ability to accurately determine the probability of a successful trade can greatly improve trading strategies by providing insight into the potential risks and rewards of a position. Meta-labeling is a technique that seeks to achieve this by using a secondary machine learning (ML) model trained on top of a primary model that produces forecasts of a trade's direction.

The secondary model's target variables are meta-labels, which are binary labels $\{0, 1\}$ that indicate whether the trade forecasted by the primary model is profitable. The secondary model's output $[0, 1]$ is thus the model's confidence that it can correctly classify whether a trade will be profitable, information that can then be used to size positions (Exhibit 1). Position sizing methods that use these model outputs are monotonically increasing functions of the predicted probabilities. In other words, the higher the model's confidence in its prediction, the larger the position size will be.

EXHIBIT 1

Meta-Labeling Architecture



NOTES: Exhibit 1 was provided as an illustration of the general framework. This article specifically addresses the third modeling component (M3) as it relates to position (bet) sizes.

SOURCE: Recreated from Joubert (2022).

Meta-labeling was first introduced by López de Prado (2018) in his book, *Advances in Financial Machine Learning.* Since then, a growing body of literature has addressed the topic. Joubert (2022) provided a comprehensive guide to meta-labeling and its foundations, while Meyer, Joubert, and Alfeus (2022) explored the potential for developing different meta-labeling architectures. Man and Chan (2021) investigated how utilizing feature importance techniques could allow better meta-labeling models to be constructed that perform out-of-sample and most recently, Thumm, Barucca, and Joubert (2023) systematically investigated different ensemble methods for meta-labeling and presented a framework to facilitate the selection of ensemble learning models.

Determining the appropriate position size for a trade is a crucial risk management decision that should consider the investor's risk tolerance and reward expectations (Markowitz 1952). This study examined the use of various position sizing algorithms in the context of meta-labeling and compared their performance in a controlled experiment to evaluate their risk management characteristics.

Meyer, Joubert, and Alfeus (2022) posited that the output of the secondary model, or the model confidence, could be transformed into well-calibrated probabilities. This would then enable sophisticated techniques such as the risk-constrained Kelly (Busseti, Ryu, and Boyd 2016) to be applied to size the positions. This study, therefore, also investigated whether probability calibration can be used to transform model confidence into well-calibrated probabilities that improve the performance of position sizing methods.

The effectiveness of probability calibration has previously been demonstrated in a range of real-world prediction tasks, including predicting mortality rates in cancer patients (Fan et al. 2021), studying species distribution in ecology (Dormann 2020), and predicting water main breaks (Kumar et al. 2018).

The structure of this article is as follows. First, the relevant literature on position sizing and model calibration is reviewed. Next, the methodology of the experiments is described. The results are then presented and discussed in terms of model calibration and position sizing. Finally, the conclusions are summarized and suggestions for future research are made.

LITERATURE REVIEW

Position Sizing

The size of a position represents a trade's risk–return profile; larger position sizes come with greater risk exposure, whereas smaller position sizes result in comparatively lower returns. An investor usually aims to maximize expected returns and minimize variance or limit downside risk metrics such as the maximum drawdown and time under water.

The position sizing problem, which involves allocating the optimal fraction of capital to an investment to maximize returns based on an investor's risk tolerance, has been widely studied. Some commonly used methods include fixed position sizing (Scholz 2012), martingale-focused strategies that consider previous outcomes (Neal and Russel 2009), and various versions of the Kelly criterion (1956), such as continuous-time Kelly (Thorp 2011) and risk-constrained Kelly (Busseti, Ryu, and Boyd 2016), the latter of which considers the trade payoffs and distribution.

Joubert (2022) suggested using an empirical cumulative distribution function (ECDF) estimated from the secondary model's training dataset in a meta-labeling context. López de Prado (2018) proposed a technique that utilizes a z-statistic from predicted probabilities, as well as methods for sizing multiple concurrent trades and reducing the risk of overtrading.

Controlling risk is typically a higher priority than maximizing return. Most methods aim to minimize risk and, as a result, tend to produce higher Sharpe ratios on average. Scholz (2012) found that using smaller relative position sizes resulted in higher Sharpe ratios due to the reduction in exposure to timing risks and large drawdowns. Strub (2016) proposed limiting downside metrics like conditional value at risk and conditional drawdown at risk through the use of extreme value theory and observed improved strategy metrics. Harvey et al. (2018) studied the use of volatility targeting to size positions and found that it improved the Sharpe ratio for risk assets such as equity and credit.

Recently, ML models have been proposed for position sizing. For example, Lim, Zohren, and Roberts (2019) used deep neural networks to learn trend estimation and position sizing for momentum-based trading strategies. Zhang, Zohren, and Roberts (2018) applied Bayesian deep convolutional neural networks to the limit order book and demonstrated that uncertainty information from posterior predictive distributions can be used to size positions.

In this study, we evaluated and compared six methods for converting the predicted probability of a profitable trade to a position size:

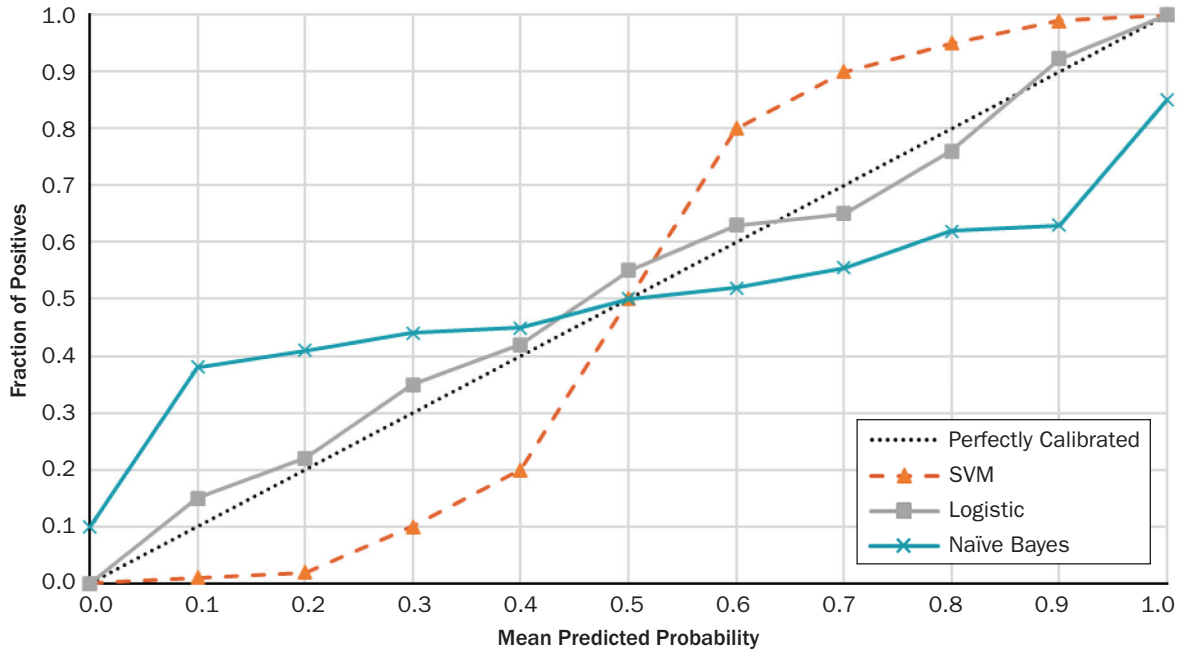
1. Model confidence
2. All-or-nothing
3. Linear scaling
4. Normal cumulative distribution function (NCDF)
5. ECDF
6. Sigmoid optimal position sizing (SOPS)

The first three methods are simple approaches that are commonly used in practice. NCDF is a variation of a method proposed by López de Prado (2018). The SOPS method is a novel technique that is discussed in detail in the position sizing section of the methodology.

It is worth noting that the Kelly criterion was not included in our analysis due to its impracticality because it generated leveraged amounts that were not feasible for any investor. An overview of the results for the Kelly criterion can be found in the appendix. Although there may be other techniques that can be applied in a

EXHIBIT 2

Calibration Plots for Logistic Regression, Support Vector Machine, and Naïve Bayes



SOURCES: Recreated from Pedregosa et al. (2011), "Probability Calibration Curves"; Niculescu-Mizil and Caruana (2005).

meta-labeling context, we specifically focused on evaluating methods that use the estimated probability output of the secondary model.

Model Calibration

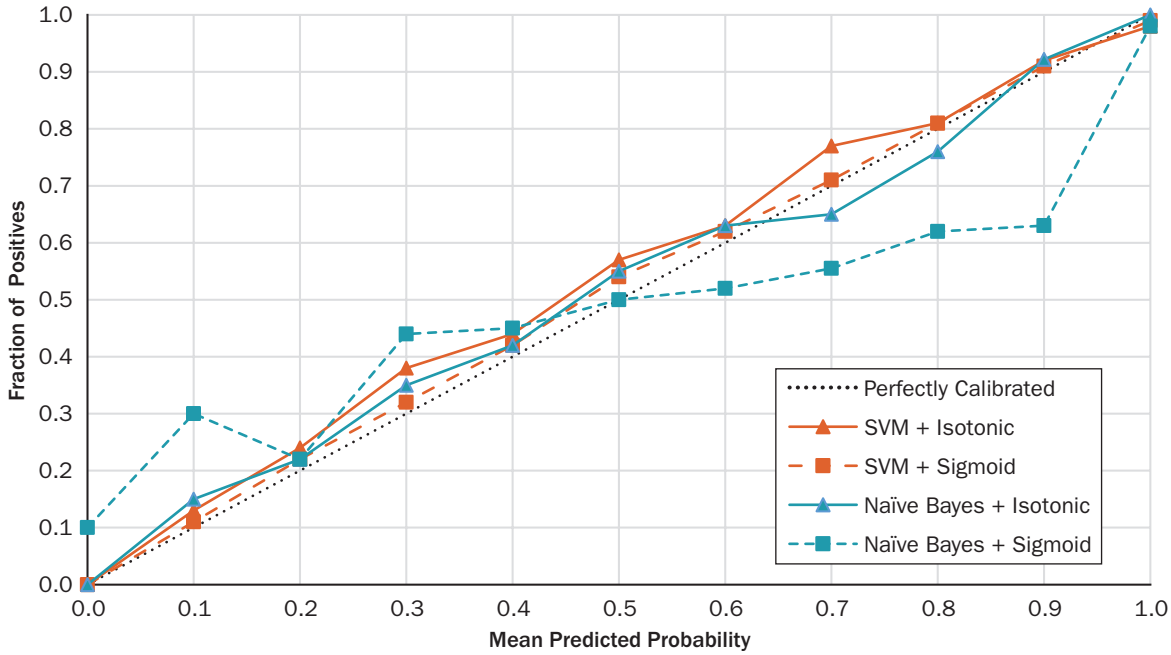
ML classification models produce outputs that signify the model's confidence that the target label is correctly classified. This confidence is often interpreted as the probability of correctly classifying the target variable. At best, these are estimated probabilities, and furthermore, not all classifiers provide probabilities. For example, logistic regression is fit based on estimating a probability, whereas a support vector machine (SVM) is not. The practice of putting the score functions from SVM into a logistic function to obtain probabilities is only done for convenience. Similarly, although the posterior probabilities of a naïve Bayes model are often biased, the model is still commonly used because this bias does not always lead to misclassification.

Niculescu-Mizil and Caruana (2005) studied the outputs of various ML models and observed that they often display characteristic distortions. They found that maximum margin methods like SVMs tend to push probabilities away from 0 and 1, resulting in a sigmoidal calibration plot, whereas other methods such as naïve Bayes tend to push predictions toward 0 and 1 (Exhibit 2). Some other ML models, such as neural networks and bagged trees, produce well-calibrated probabilities, although this also depends on the relationship between the features and the target variable. For example, if the relationship is nonlinear and a logistic regression is fitted, the estimated probabilities will not be accurate. Probability calibration can correct for these distortions and transform the model outputs into values closer to true posterior probabilities.

Calibration adjusts measured values in accordance with reference values by drawing a comparison to them. In probability calibration, the reference values are the true probabilities of observing an event. From a frequentist perspective, if the model

EXHIBIT 3

Calibration Plots for SVM and Naïve Bayes with Calibration



SOURCES: Recreated from Pedregosa et al. (2011), "Probability Calibration Curves"; Niculescu-Mizil and Caruana (2005).

output for a given event is *p*, then the model's predicted probability of *p* true positives should be observed asymptotically to be considered a true probability. Estimated probabilities from ML models that are not calibrated can be adjusted using various calibration methods to enable their interpretation as true probabilities.

Model calibration can be visualized using reliability diagrams or so-called calibration plots (DeGroot and Fienberg 1983). Exhibit 2 is an example of a calibration plot for logistic regression, SVM, and naïve Bayes models. The *y*-axis shows the fraction of positives correctly classified, while the *x*-axis displays the mean predicted probabilities. The dotted line shows the hypothetical case in which the predicted outputs are perfectly calibrated. In this case, the conditional probabilities estimated by the logistic regression model were more accurate because the true model was linear.

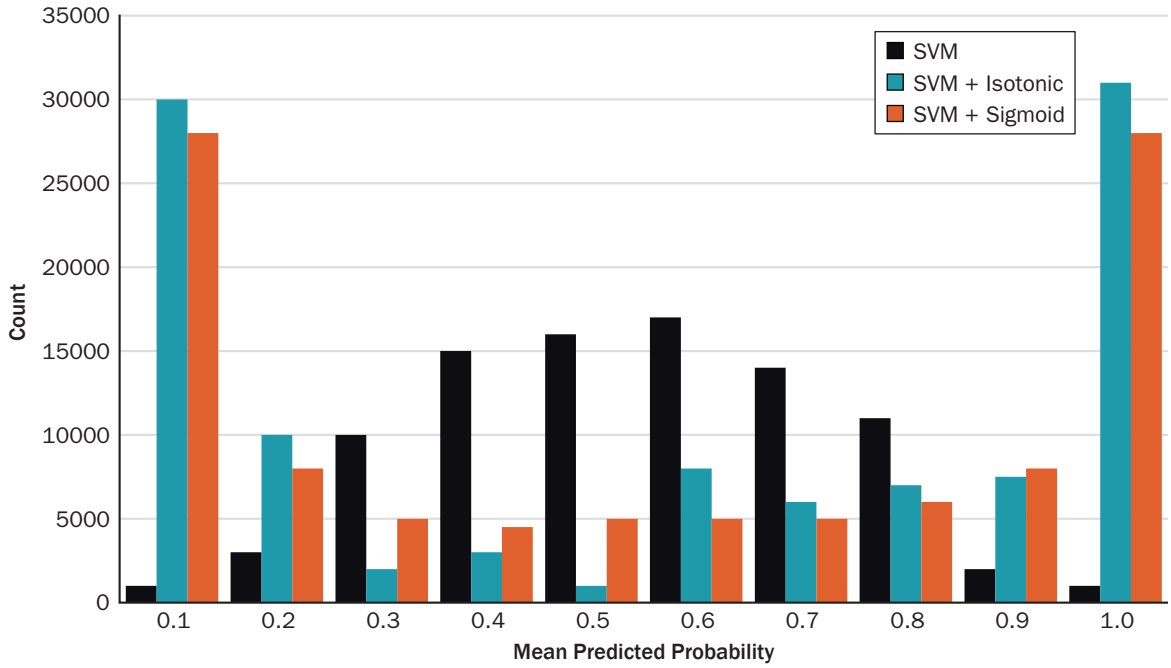
Without calibration, the SVM shows an S-shaped calibration plot. This means that the model outputs underestimate the real fraction of positives when below 0.5 and overestimate the fraction of positives when above 0.5. Take, for example, the mean predicted probability at point 0.3 of the SVM. If it were a true probability, a 0.3 fraction of positives would be observed; however, we only observe 0.1. The probability curve for naïve Bayes shows the opposite tendency, with an inverted S-shape. It should be noted that, if the assumptions of the naïve Bayes held, then it most likely would have produced accurate probabilities.

Various probability calibration methods have been proposed, both in the literature and in practice. Some of the most widely established methods are Platt scaling (Platt 1999), which applies a logistic transformation, isotonic regression (Zadrozny and Elkan 2001), which fits a piecewise constant nondecreasing (steplike) function to the data binning, and probability calibration trees (Leathart et al. 2017), a modification of logistic model trees that identifies regions of the input space in which different probability calibration models are applied to improve performance.

Exhibit 3 shows the SVM and naïve Bayes calibration plots with isotonic regression and sigmoid (Platt) scaling. The SVM with calibration produces well-calibrated

EXHIBIT 4

Histogram of SVM Model Outputs with and without Calibration



SOURCES: Recreated from Pedregosa et al. (2011), "Probability Calibration Curves"; Niculescu-Mizil and Caruana (2005).

probabilities, with sigmoid scaling slightly outperforming isotonic regression. Nevertheless, both calibration methods transform the outputs into well-calibrated probabilities. In contrast, the naïve Bayes with isotonic regression produces well-calibrated probabilities, while calibration with sigmoid scaling shows some improvement, though the distortion remains. This is similar to the results obtained by Niculescu-Mizil and Caruana (2005), which revealed that sigmoid scaling is the preferred method for S-shaped calibration plots and that isotonic regression is generally better with large datasets but comes at the cost of potential overfitting.

Exhibit 4 displays the histograms for the SVM when uncalibrated and calibrated with sigmoid scaling and isotonic regression, respectively. The uncalibrated outputs have a normal-like distribution with a mean close to 0.5 and probabilities pushed away from 0 and 1. Both calibration methods distribute the outputs more evenly.

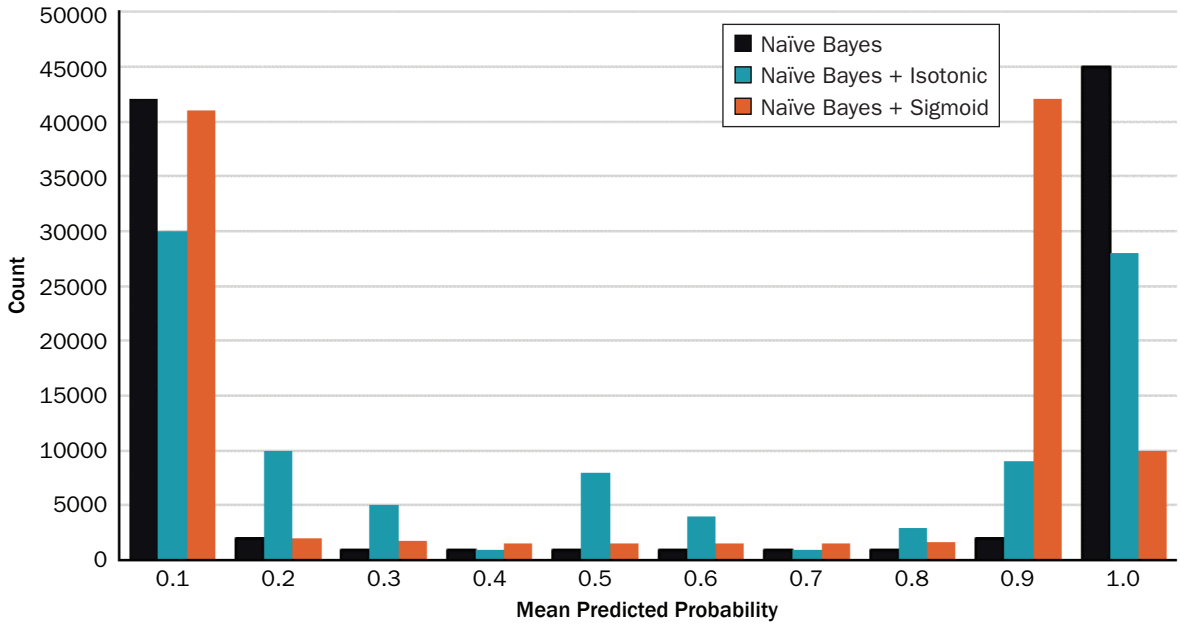
Exhibit 5 displays the histograms for the naïve Bayes when uncalibrated and calibrated with Platt scaling and isotonic regression, respectively. We see that the naïve Bayes pushes the probabilities closer to 0 and 1, whereas calibration distributes them more evenly, with the distribution obtained by isotonic regression showing a slightly better performance.

METHODOLOGY

This article reports the results of two controlled experiments. First, we compared the position sizing algorithms with each other and performed model calibration to check for any significant performance gain. Second, the impact of changing the classification threshold was tested in an all-or-nothing setting, and the effects on the Sharpe ratio and maximum drawdown were measured. The adopted methodology

EXHIBIT 5

Histogram of Naïve Bayes Model Outputs with and without Calibration



`**SOURCES:** Recreated from Pedregosa et al. (2011), "Probability Calibration Curves"; Niculescu-Mizil and Caruana (2005).

closely followed the experiments conducted by Joubert (2022), and this section provides a brief overview.

Data Generation Process

A dual autoregressive process of order 3, AR(3), was used to simulate 10,000 daily price returns. The simulated AR(3) helped avoid the complex stylistic behaviors found in real-world financial time series, such as heteroskedasticity (Engle 1982), nonstationarity, and heavy tails (Cont 2001). The stochastic component of the AR(3) process was simulated by a white noise process with a standard deviation similar to that of IBM's stock returns. The entire process was repeated 1,000 times to eliminate the effect of path-dependent specific results (Bailey and López de Prado 2014), and the data were split according to a 60/40 ratio without reshuffling for training and testing.

The dual component was used to simulate regime changes. Regime 1 was slightly more favorable than Regime 2 for a long-only trading strategy. Each regime lasted 30 days, after which the next regime was randomly selected, with an 80% probability of Regime 1 and a 20% probability of Regime 2. A regime indicator was added as a feature to the model; this lagged the regime change by five days and indicated whether the data generated were from the first regime. This allowed us to simulate the impact of indicators that use a rolling window.

Primary Model

The primary model was an autoregressive strategy in which only the previous day's sign of return was used to determine the side of the position. It was a long-only strategy that closed all positions if the sign was negative.

Secondary Model

The secondary model was a logistic regression in which the target variable was meta-labels that indicated whether a trade was profitable or not, and the features fed into the model were the previous three days' returns and the lagged regime indicator. To keep the experiment as simple as possible, regularization was not applied.

By including the last three observations and the regime indicator, we were able to measure the information advantage and false positive premium, respectively (Joubert 2022).

Linear discriminant analysis, SVM, naïve Bayes, and random forests were initially also tested as secondary models; however, the results were excluded from the analysis because they provided no additional insights.

Model Calibration

Isotonic regression was used to calibrate the secondary model outputs for two reasons. First, Platt scaling is best suited to S-shaped calibration plots, whereas isotonic regression generalizes better; second, isotonic regression is better suited to datasets with many observations, as found by Niculescu-Mizil and Caruana (2005).

Cross-validation was used both to estimate the parameters of the logistic regression (secondary model) and subsequently to calibrate the model's outputs. For each cross-validation split, a copy of the base estimator (logistic regression) was fitted to the training subset and calibrated using the validation subset. The predicted probabilities were then averaged across these separately calibrated classifiers (Pedregosa et al. 2011).

Position Sizing Methods

This section describes the various position sizing methods that were evaluated on the testing data for raw uncalibrated and calibrated probabilities. The metrics used to measure the performance were the mean Sharpe ratio, return, standard deviation, and maximum drawdown for the 1,000 simulated paths.

To ensure a fair comparison of the performance of various methods, short positions and leverage were not permitted. This helped avoid certain complexities and ensure a fair assessment of the tested methods. Nevertheless, the various methods can easily be adapted to allow for shorting and leverage.

A condition was imposed on each method to only take a trade if the predicted probability of a profitable trade was greater than 50%. It should be noted that this assumption only holds when the payoffs are symmetric, which was not true in our case, but we chose 50% to simplify the process. The rationale was that, if the expected value of a trade is negative, it is not advisable to take the position. The threshold to take a position should, therefore, be chosen so that the expected value of the trade is greater than zero. Then, assuming the underlying process remains the same, each trade will be asymptotically profitable. The threshold can also be adjusted to the risk characteristics of the primary model.

A uniform notation was used across all methods, where represents the probability of profitable trade, and represents the position size determined by each method. A total of six position sizing algorithms were examined in the experiment, with the model confidence serving as the baseline method against which the other five were compared. The Kelly criterion and a novel method called the linear optimal position sizing method were also tested; however, the results revealed that these methods possessed undesirable characteristics. A discussion of these methods is included in the appendix.

There are broadly two classes of position sizing functions: fixed position sizing methods and methods that are estimated from the training data. Fixed position functions have a deterministic function that always produces the same result, whereas functions estimated from the training data assume a functional form, and the parameters are estimated from the training data. Model confidence, all-or-nothing, and NCDF are all fixed approaches, while the linear scaling, ECDF, and SOPS methods are parametric approaches.

Model confidence. A naïve method is to simply use the predicted probability as the percentage of capital to allocate to the position size when the probability is above a certain threshold. The formula is given as

$$b = \begin{cases} p, \textit{ if } p > 0.5 \\ 0, \textit{ otherwise} \end{cases} \tag{1}$$

All-or-nothing. A common approach used in practice is the all-or-nothing method. If the probability of success is greater than a certain threshold, all the available capital is allocated to the position:

$$b = \begin{cases} 1, \textit{ if } p > 0.5 \\ 0, \textit{ otherwise} \end{cases} \tag{2}$$

Linear scaling. Another straightforward approach is to scale the probabilities into a position size through min-max scaling. The only two points required for this are the minimum and maximum from the training set. The formula is given in the following, where p_{train} indicates the estimated probabilities from the training data:

$$b = \begin{cases} \dfrac{p - min(p_{train})}{max(p_{train}) - min(p_{train})}, \textit{ if } p > 0.5 \\ 0, \textit{ othewise} \end{cases} \tag{3}$$

NCDF. López de Prado (2018) described a method for bet sizing from predicted probabilities. In this study, we used the version with two possible outcomes {0, 1}, as predicted by the secondary model. The following H_0 hypothesis is tested: $p[x = 1] = \frac{1}{2}$. The test statistic is calculated as follows:

$$z = \frac{p - \frac{1}{2}}{\sqrt{p(1-p)}} = \frac{2p - 1}{2\sqrt{p(1-p)}} \sim Z \tag{4}$$

The statistic is in the range $(-\infty, +\infty)$ and Z represents a standard normal distribution. The position size is then calculated as

$$b = 2Z[z] - 1 \tag{5}$$

Here, Z[·] is the cumulative distribution function of Z. This original formula results in low mean returns; therefore, we adjusted it simply as

$$b = \begin{cases} Z[z], \textit{ if } p > 0.5 \\ 0, \textit{ otherwise} \end{cases} \tag{6}$$

ECDF. The fifth position sizing algorithm uses an empirical cumulative distribution fitted to the output probabilities of the secondary model based on the training set. The training set is filtered to only include trades with a probability above a threshold of 50%. It was proposed by Joubert (2022) and is based on scaling the position size to the distribution of positive probabilities. Probabilities in the tail end of the distribution get the largest allocations, and smaller sizes are allocated to those with higher levels of uncertainty.

$$b = \begin{cases} ECDF_{train}(p), \text{ if } p > 0.5 \\ 0, \text{ otherwise} \end{cases} \tag{7}$$

SOPS method. The sixth position sizing method we tested was a novel technique called sigmoid optimal position sizing (SOPS). The algorithm fits a sigmoidal transformation of probabilities to position sizes to maximize the Sharpe ratio in the training set. The sigmoidal function is used because of its similarity to the transformation in the ECDF method and because the existence of a general sigmoid function equation simplifies the fit process. The transformation results from solving the following optimization problem:

$$\min_{a,c} -SR$$

$$\text{where } SR = \frac{mean(m)}{std(m)}$$

$$m = f(p_{train}) * r_{train}$$

$$f(p) = \frac{1}{1 + e^{-a*p-c}}$$

$$b = \begin{cases} f(p), \text{ if } p > 0.5 \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

The result is a pair of values (a, c), which are used to build the sigmoidal transformation function. The minimization is carried out using the Nelder–Mead method (Nelder and Mead 1965).

Effect of Changing Threshold

In the second experiment, we changed the probability threshold for the all-or-nothing approach to analyze the effect of the threshold on the Sharpe ratio and maximum drawdown. For each threshold chosen, 1,000 simulations were run and averaged out. This experiment served to showcase the importance of the selected threshold, which should be taken into consideration when applying any of the above methods.

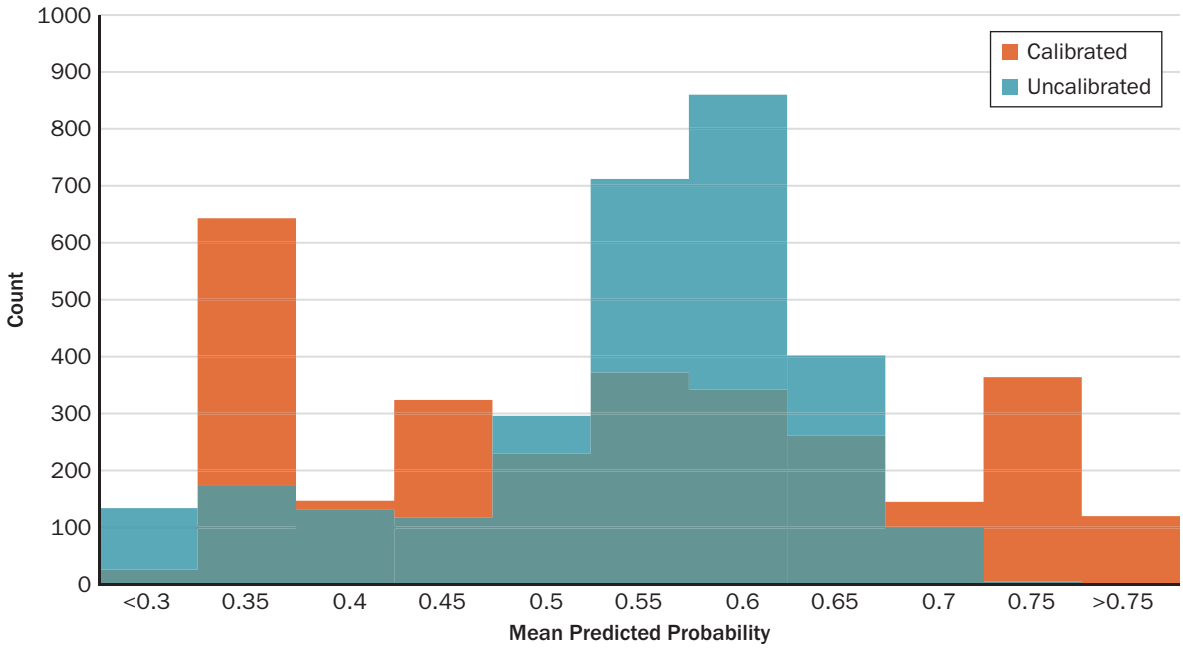
RESULTS

Secondary Model Outputs

The position sizes in this study were solely based on the predicted outputs of the secondary model for each method. The distribution of these outputs is important because it largely determines the final position sizes for the fixed position sizing

EXHIBIT 6

Predicted Probability Distribution from Logistic Regression Training Set



methods, whereas the functions estimated from the training data transform the original distribution into a new distribution that is used to size the positions. It is crucial to examine the distribution of the model outputs to understand how it impacts the position sizes.

Exhibit 6 displays the distribution of the uncalibrated and calibrated secondary model outputs from the training dataset for one simulation. A single simulation was chosen to illustrate how these outputs can be analyzed for a real-world problem. Histograms such as Exhibit 6 will likely differ depending on the type of secondary model selected and the true underlying model, as shown in Exhibit 4 and Exhibit 5 for the SVM and naïve Bayes models, respectively.

The outputs of the uncalibrated model are centered around the interval (0.45,0.65], and almost none of the outputs exceeded 0.7. On the other hand, the calibrated outputs were more evenly distributed, with over 100 values exceeding 0.75. In terms of the number of trades with an estimated probability above 0.5, the uncalibrated outputs had 2,078, whereas the calibrated outputs had 1,604. This indicates that when calibration is applied, fewer trades are taken, but they are made with larger position sizes in the case of the fixed position sizing methods.

Exhibit 7 illustrates the calibration plot for the logistic regression outputs, with the number of points per decile shown on the *x*-axis for both the uncalibrated and calibrated predicted probabilities. The uncalibrated outputs display an S-shaped curve; therefore, isotonic regression was the preferred calibration method, as suggested by Niculescu-Mizil and Caruana (2005). The plot demonstrates that calibration significantly improved the secondary model's outputs, bringing them closer to true posterior probabilities. It is clear that values between (0.4,0.5] produced a higher number of false positives, which were subsequently mapped to values between (0.3,0.35], as shown in Exhibit 6.

Exhibit 8 displays the average percentage change in strategy metrics after applying calibration to all the position sizing methods. Calibration significantly improved the

EXHIBIT 7

Calibration Plot: Isotonic Regression on Logistic Regression

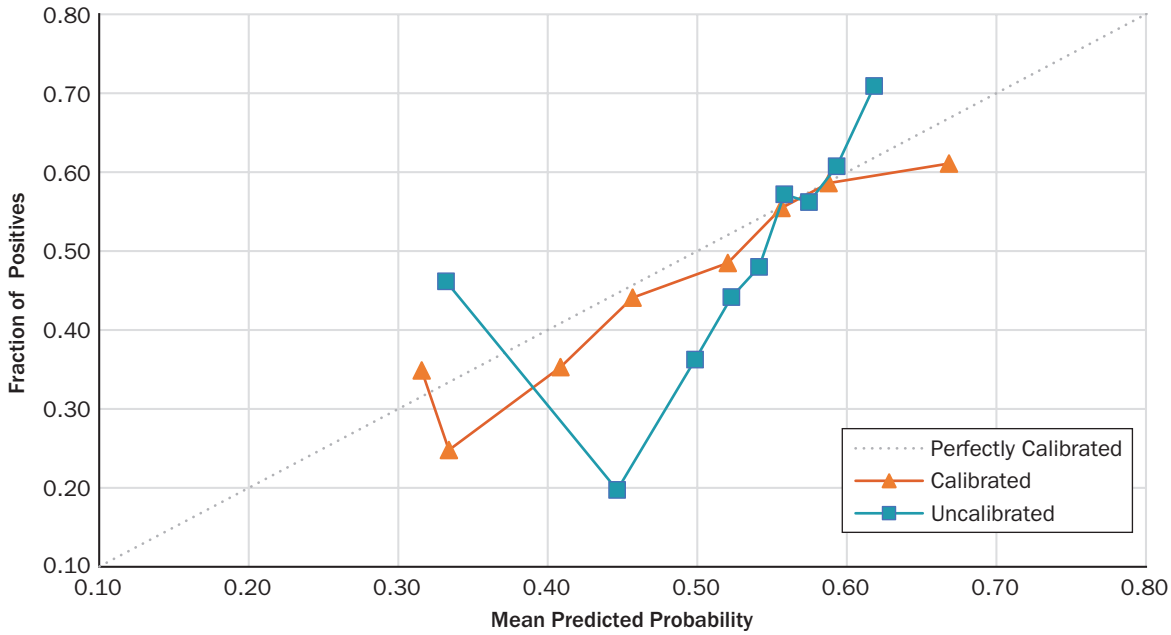


EXHIBIT 8

Average Percentage Increase/Reduction in Strategy Metrics after Calibration

	%Δ Sharpe Ratio	%Δ Mean Return	%Δ Standard Deviation	%Δ Maximum Drawdown
All-or-Nothing	25.90	13.79	**9.51**	–25.21
Model Confidence	23.79	25.78	–1.72	–18.02
Linear Scaling	1.41	9.04	–7.50	**3.89**
NCDF	24.61	25.90	–1.16	–18.58
ECDF	1.59	2.69	–1.08	–2.07
SOPS	2.18	2.73	–0.97	–0.86

NOTES: The bold face values represent values where there is a decrease in the performance of the strategy metrics (which is an increase for % change in Standard Deviation and Maximum Drawdown).

Sharpe ratio for the fixed approaches, with an average increase of 24.77%. Calibration of the all-or-nothing approach increased the annual mean return at the cost of a higher standard deviation but still led to a significant improvement in the Sharpe ratio. For both the calibrated model confidence and NCDF approaches, the mean returns increased and the standard deviation reduced. The calibrated fixed approaches also exhibited significantly reduced maximum drawdowns, with an average reduction of 20.6%.

Calibration had little effect, however, on the methods estimated from training. The Sharpe ratio only increased by an average of 1.73% for these methods. The drawdowns decreased slightly for the ECDF and SOPS methods and, notably, even increased for linear scaling. Therefore, calibrating these outputs before applying a position sizing method that is estimated from the training data appears to be an unnecessary step.

Exhibit 9 compares the Sharpe ratios for all the position sizing methods for both uncalibrated and calibrated outputs across the 1,000 simulations. The uncalibrated methods estimated from the training data displayed overall higher Sharpe ratios than the calibrated fixed methods. The ECDF approach had the highest average Sharpe ratio, and the SOPS approach had a Sharpe ratio that was only slightly smaller than that of the ECDF approach. Linear scaling produced a comparatively high Sharpe ratio for such a simple method, with the added benefit that few data points (just two) are needed to apply the method. The calibrated model confidence approach had the highest Sharpe ratio among the calibrated fixed approaches, and the all-or-nothing approach had the lowest.

Exhibit 10 displays the maximum drawdown for each of the methods across all the simulations. The all-or-nothing approach displayed a significant improvement

EXHIBIT 9

Boxplots of Sharpe Ratio for Uncalibrated and Calibrated Values

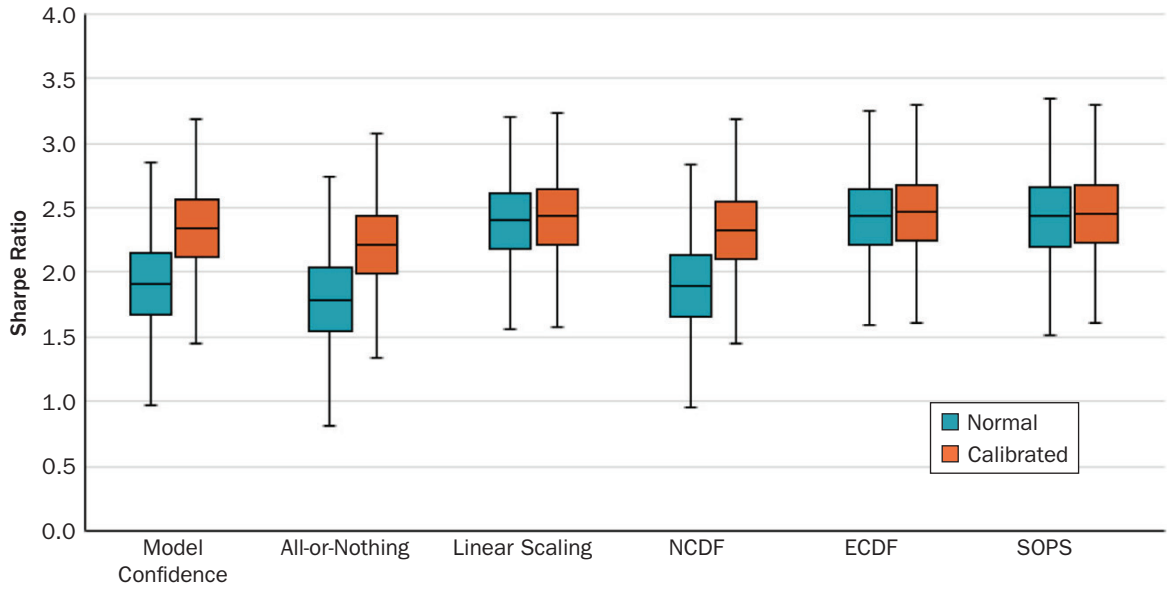
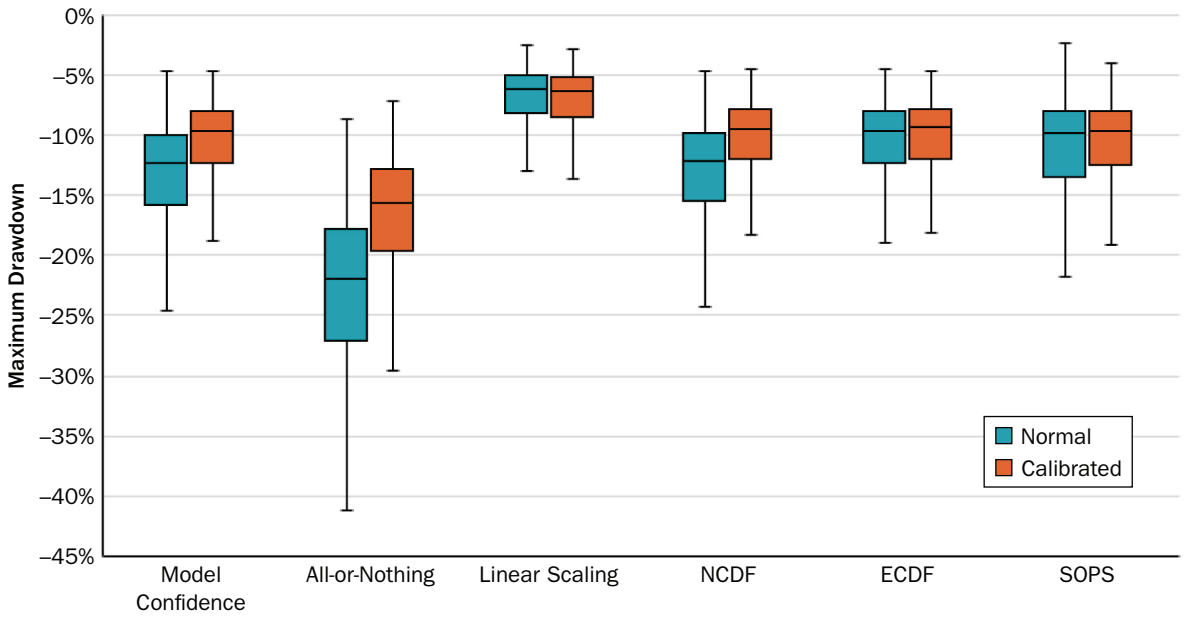


EXHIBIT 10

Boxplots of Maximum Drawdowns for Uncalibrated and Calibrated Values



once calibrated, although it still had relatively large drawdowns compared to the other methods. Both the model confidence and NCDF approaches also showed improvement in terms of drawdown when calibrated. On the other hand, the methods estimated from the training data showed little change in terms of drawdown. Linear scaling had the smallest average drawdown, while the other fixed methods showed similar average drawdowns when calibration was applied.

Exhibit 11 showcases the mean returns of all the position sizing methods. The all-or-nothing approach had the highest overall return, even before calibration was

EXHIBIT 11

Boxplots of Mean Returns for Uncalibrated and Calibrated Values

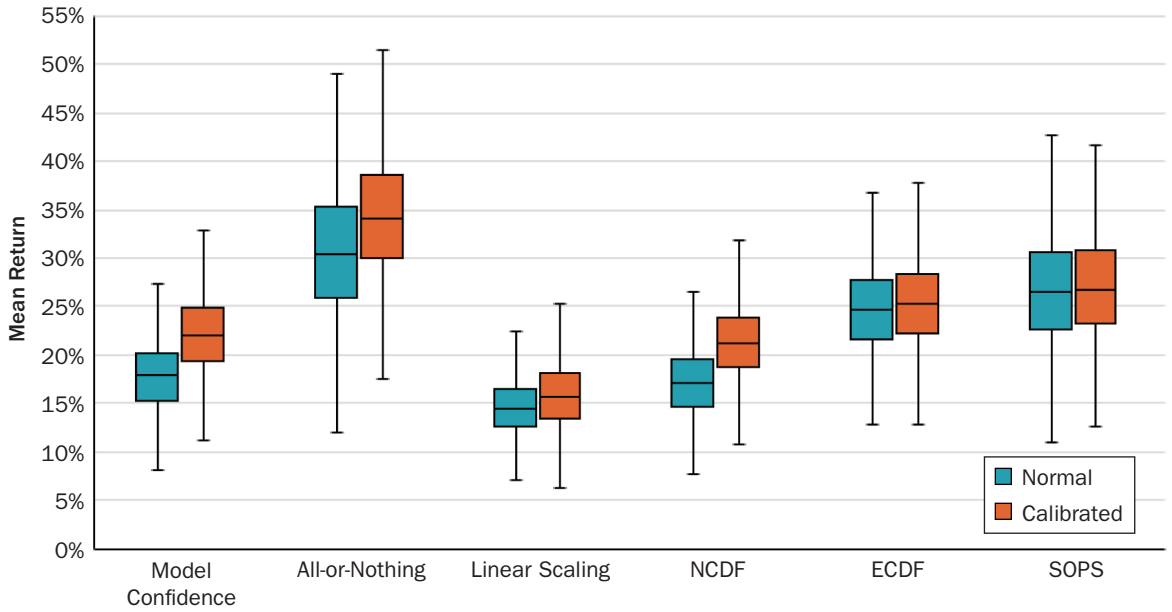
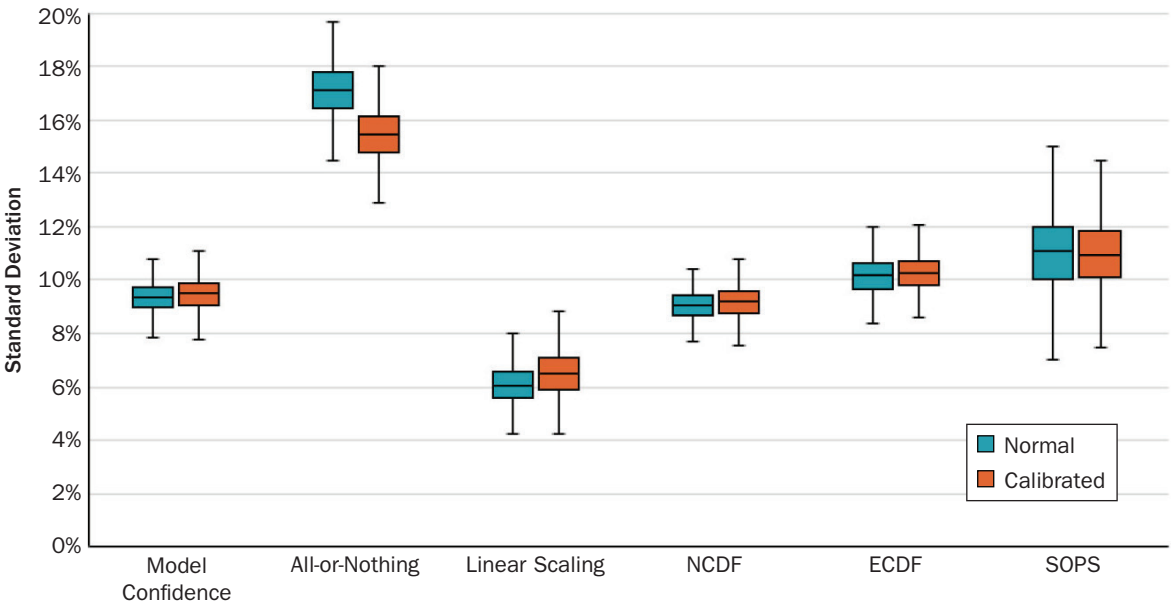


EXHIBIT 12

Boxplot of Standard Deviation for Raw and Calibrated Values



applied, whereas linear scaling had the lowest return. The SOPS method displayed the second-highest return, closely followed by the ECDF approach. Exhibit 12 illustrates the standard deviations and shows that the high return of the all-or-nothing approach came at the cost of a higher standard deviation, while linear scaling had the lowest standard deviation. Calibration had little impact on the standard deviation of the various methods, except for the all-or-nothing approach.

EXHIBIT 13

Sharpe Ratio as a Function of the Threshold for the All-or-Nothing Approach

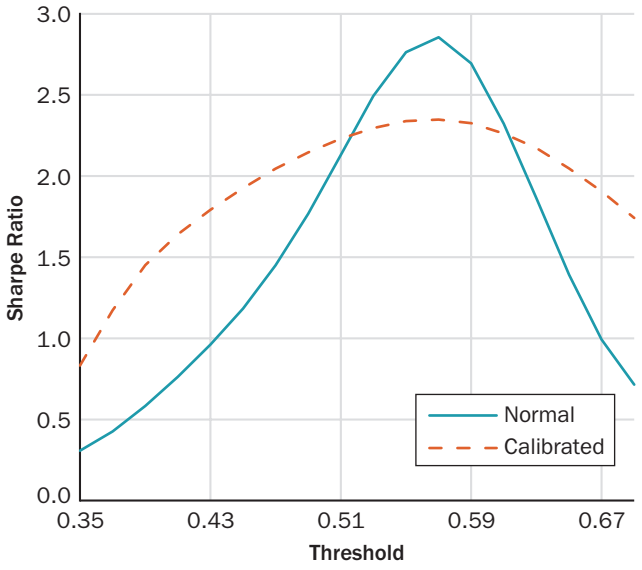
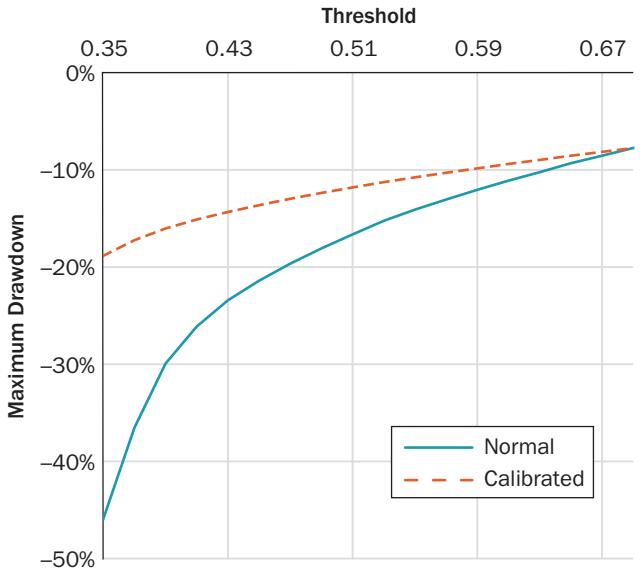


EXHIBIT 14

Maximum Drawdown for Different Thresholds for the All-or-Nothing Approach



Effect of Changing the Threshold

This section reports the effects of changing the probability threshold on the performance of one position sizing method. Specifically, the Sharpe ratio for the all-or-nothing approach was calculated using raw and uncalibrated probabilities at different thresholds. The results, shown in Exhibit 13, indicated that, for the raw probabilities, the optimal threshold for maximizing the Sharpe ratio was not 0.5, but rather 0.57. Also, although calibrated probabilities produced more stable Sharpe ratios, the raw probabilities had a higher maximum Sharpe ratio than the calibrated probabilities.

Exhibit 14 shows the maximum drawdown for both raw and calibrated probabilities. Calibrated probabilities resulted in a lower maximum drawdown, indicating that model calibration is effective at controlling downside risk. This was also evident in the stability of the maximum drawdown across different thresholds. Therefore, even if the optimal threshold is uncertain, it is generally better to use calibrated probabilities because they result in a smaller maximum drawdown when using the all-or-nothing approach.

DISCUSSION

The results of the experiment suggest that each position sizing method has its own strengths and weaknesses. If the goal is to maximize returns, the all-or-nothing approach or SOPS may be suitable options. On the other hand, if the focus is on risk management, the linear scaling or NCDF approaches may be preferable due to their low standard deviation and lower maximum drawdown. In cases in which calibration is not effective, the ECDF or SOPS methods may be preferred. It is worth noting that SOPS is designed to maximize the Sharpe ratio, which automatically takes into account the risk characteristics of the strategies, making it a convenient choice without the need for additional steps. A summary of the all-or-nothing, linear scaling, ECDF, and SOPS approaches is provided in Exhibit 15.

CONCLUSION

This article has shown that position sizing is a crucial element of a successful strategy. By comparing six different methods for transforming the probability of a positive trade into a position size in a meta-labeling setting, we were able to identify the most effective approaches. The ECDF and SOPS methods were found to have the best overall performance. This article has provided a clear indication of which methods are suitable for different investor risk preferences and underlying conditions.

EXHIBIT 15

Position Sizing Methods: Advantages and Disadvantages

Methods	Advantages	Disadvantages
All-or-Nothing	Highest mean return Easy to use	Highest maximum drawdown Highest standard deviation Calibration needs to be applied
Linear Scaling	Low standard deviation and maximum drawdown Only two points required from training data Calibration does not need to be applied	Low mean returns
ECDF	High Sharpe ratio Calibration does not need to be applied	High standard deviation
SOPS	High Sharpe ratio Calibration does not need to be applied	High standard deviation Needs many data points to optimize

Additionally, we demonstrated that calibration can significantly improve the performance of commonly used position sizing methods such as the all-or-nothing approach.

We also provided a framework for when and how to use probability calibration for the outputs of the secondary model. These findings may differ when different primary and secondary models are used, but the analysis conducted in each case would be similar. Each method can also be adapted to capture specific characteristics of the setting in which it is applied.

Future research directions could include investigating strategy allocation using probabilities produced by various secondary models, finding a closed-form solution for the optimal probability threshold that maximizes the Sharpe ratio (or potentially other metrics) and using portfolio optimization techniques to allocate capital.

APPENDIX

Two additional position sizing algorithms were explored, namely the Kelly criterion and a novel technique called linear optimal position sizing. These were left out of our report of the main results, however, because they were found to produce outcomes that would not be suitable for practical implementation.

KELLY CRITERION

The Kelly criterion is one of the first position sizing methods to come to mind when using probabilities. The basic formula is as follows:

$$b = \frac{p}{c} - \frac{1-p}{a} \tag{A1}$$

Aside from the probability of the success of an event, this method also requires the expected percentage gain in case of a positive outcome, denoted as a, and the expected percentage loss in case of a negative outcome, denoted as c. These values are only estimated on positions taken. This formula produces values that are highly leveraged, and scaling to between $[0, 1]$ reduces it to an all-or-nothing approach. This can be demonstrated by letting $b = 1$ and rewriting the equation to solve for p:

$$p = \frac{c(1+a)}{a+c} \tag{A2}$$

Taking a sample of one training dataset where $a = 0.0131$ and $c = 0.0129$, then $p = 0.4827$, which equates to an all-or-nothing approach. Therefore, these results were excluded.

LINEAR OPTIMAL POSITION SIZING

The linear optimal position sizing approach was initially proposed to compete against methods based on linear transformations and determine whether better performance could be achieved. The algorithm was so named because it aimed to fit a linear transformation of probabilities to position sizes, maximizing the Sharpe ratio in the training set. The transformation results from the solution of the following optimization problem:

$$\min_{a,\,c} -SR$$

$$\text{where} \quad SR = \frac{mean(m)}{std(m)}$$

$$m = f(p_{train}) * r_{train}$$

$$f(p) = min(max(a * p + c, 0), 1) \tag{A3}$$

The result is a pair of values (a, c) that is used to build a linear transformation. Our experimental results, however, revealed that, although it does produce a good Sharpe ratio, it produces returns so low that no trader would use it. A more suitable adaptation of this method might be to optimize for returns instead of the Sharpe ratio.

REFERENCES

Bailey, D. H., and M. López de Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality." *The Journal of Portfolio Management* 40 (5): 94–107.

Busseti, E., E. K. Ryu, and S. Boyd. 2016. "Risk-Constrained Kelly Gambling." *The Journal of Investing* 25 (3): 118–134.

Cont, R. 2001. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues." *Quantitative Finance* 1 (2): 223.

DeGroot, M. H., and S. E. Fienberg. 1983. "The Comparison and Evaluation of Forecasters." *Journal of the Royal Statistical Society: Series D (The Statistician)* 32 (1–2): 12–22.

Dormann, C. F. 2020. "Calibration of Probability Predictions from Machine-Learning and Statistical Models." *Global Ecology and Biogeography* 29 (4): 760–765.

Engle, R. F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica: Journal of the Econometric Society* 35 (4): 987–1007.

Fan, S., Z. Zhao, H. Yu, L. Wang, C. Zheng, X. Huang, Z. Yang, M. Xing, Q. Lu, and Y. Luo. 2021. "Applying Probability Calibration to Ensemble Methods to Predict 2-Year Mortality in Patients with DLBCL." *BMC Medical Informatics and Decision Making* 21 (1): 1–12.

Harvey, C. R., E. Hoyle, R. Korgaonkar, S. Rattray, M. Sargaison, and O. Van Hemert. 2018. "The Impact of Volatility Targeting." *The Journal of Portfolio Management* 45 (1): 14–33.

Joubert, J. F. 2022. "Meta-Labeling: Theory and Framework." *The Journal of Financial Data Science* 4 (3): 31–44.

Kelly Jr., J. L. 1956. "A New Interpretation of Information Rate." *The Bell System Technical Journal* 35 (4): 917–926.

Kumar, A., S. A. A. Rizvi, B. Brooks, R. A. Vanderveld, K. H. Wilson, C. Kenney, S. Edelstein, A. Finch, A. Maxwell, J. Zuckerbraun, and R. Ghani. 2018. "Using Machine Learning to Assess the Risk of and Prevent Water Main Breaks." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 472–480. New York, United States: Association for Computing Machinery.

Leathart, T., E. Frank, G. Holmes, and B. Pfahringer. 2017. "Probability Calibration Trees." In *Proceedings of the Ninth Asian Conference on Machine Learning*. *PMLR* 77: 145–160.

Lim, B., S. Zohren, and S. Roberts. 2019. "Enhancing Time-Series Momentum Strategies Using Deep Neural Networks." *The Journal of Financial Data Science* 1 (4): 19–38.

López de Prado, M. *Advances in Financial Machine Learning*, 1st ed. New York City: Wiley, 2018.

——. 2022. "Causal Factor Investing: Can Factor Investing Become Scientific?" *SSRN* 4205613.

Man, X., and E. Chan. 2021. "The Best Way to Select Features? Comparing MDA, LIME, and SHAP." *The Journal of Financial Data Science* 3 (1): 127–139.

Markowitz, H. 1952. "Portfolio Selection." *The Journal of Finance* 7: 77–91.

Meyer, M., J. F. Joubert, and M. Alfeus. 2022. "Meta-Labeling Architecture." *The Journal of Financial Data Science* 4 (4): 10–24.

Neal, D. K., and M. D. Russell. 2009. "A Generalized Martingale Betting Strategy." *Missouri Journal of Mathematical Sciences* 21 (3): 183–197.

Nelder, J. A., and R. Mead. 1965. "A Simplex Method for Function Minimization." *The Computer Journal* 7 (4): 308–313.

Niculescu-Mizil, A., and R. Caruana. 2005. "Predicting Good Probabilities with Supervised Learning." In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632. New York City: Association for Computing Machinery.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Platt, J. "Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods." In *Advances in Large Margin Classifier*, pp. 61–74. 1999. Washington, United States: Microsoft Research.

Scholz, P. "Size Matters! How Position Sizing Determines Risk and Return of Technical Timing Strategies." Working paper no. 31, CPQF, 2012.

Strub, I. S. 2016. "Trade Sizing Techniques for Drawdown and Tail Risk Control." *SSRN* 2063848.

Thorp, E. O. "The Kelly Criterion in Blackjack Sports Betting, and the Stock Market." In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pp. 789–832. 2011. Amsterdam, The Netherland: Elsevier.

Thumm, D., P. Barucca, and J. F. Joubert. 2023. "Ensemble Meta-Labeling." *The Journal of Financial Data Science* 5 (1): 10–26.

Zadrozny, B., and C. Elkan. 2001. "Obtaining Calibrated Probability Estimates from Decision Trees and Naïve Bayesian Classifiers." In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, pp. 609–616. San Francisco: Morgan Kaufmann Publishers Inc.

Zhang, Z., S. Zohren, and S. Roberts. 2018. "BDLOB: Bayesian Deep Convolutional Neural Networks for Limit Order Books." *arXiv* 1811.10041.