Meta-Labeling: Theory and Framework

Jacques Francois Joubert

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

KEY FINDINGS

- Practitioners are provided with a clear framework to support the application of meta-labeling to investment strategies.
- Meta-labeling is deconstructed into three parts, with a controlled experiment demonstrating how each component improves the Sharpe ratio and reduces drawdowns.
- This article consolidates the knowledge of various articles into a single work, laying the foundations for future research.

ABSTRACT

Meta-labeling is a machine learning (ML) layer that sits on top of a base primary strategy to help size positions, filter out false-positive signals, and improve metrics such as the Sharpe ratio and maximum drawdown. This article consolidates the knowledge of several publications into a single work, providing practitioners with a clear framework to support the application of meta-labeling to investment strategies. The relationships between binary classification metrics and strategy performance are explained, alongside answers to many frequently asked questions regarding the technique. The author also deconstructs meta-labeling into three components, using a controlled experiment to show how each component helps to improve strategy metrics and what types of features should be considered in the model specification phase.

A difficult problem to solve, for quantitative investment teams, is the forecasting of financial time series to develop trading and investment strategies. An attractive proposition is the use of machine learning (ML) as an overlay to a base primary strategy to help size positions, filter out false-positive signals, and improve strategy metrics such as the Sharpe ratio and maximum drawdown.

This process is known as meta-labeling and was first introduced by Marcos López de Prado in the textbook *Advances in Financial Machine Learning* (2018). It is a concept that has garnered the attention of practitioners; however, due to the paucity of peer-reviewed publications, there remain questions regarding how it can be implemented. This article aims to consolidate the concepts into a single work and provide new insights as well as a clear framework for the application of meta-labeling.

The article is structured as follows. The following section provides an overview of the theoretical framework of meta-labeling, including its architecture and applications. Next, the methodology of three controlled experiments is described, which are designed to break meta-labeling down into three components: information advantage, modeling for false positives, and position sizing. The subsequent section provides and discusses the results, highlighting new insights. Finally, the conclusion summarizes the findings and provides ideas for future research.

THEORETICAL FRAMEWORK

In his discussion of the 10 most common reasons ML funds fail, López de Prado (2018b) noted that pitfall number six is learning side and size simultaneously. He posited that the side decision $\{-1, 0, 1\}$ is a fundamental one that is focused on determining a fair value price for a security under a particular market state, whereas the size of the position is a risk management decision. Combining both into a single forecast is inefficient; rather, López de Prado (2018b) argued, the side and size decisions should be modeled separately because many ML algorithms exhibit high precision and low recall. The low recall means that these strategies trade less often, thus missing opportunities as they are too conservative, leading to longer underwater periods. This can be avoided by using a model with a higher F1-score.

Meta-Labeling

Meta-labeling is the process of fitting a secondary model to determine whether a primary exogenous model is correct and to size positions accordingly. The target variables in this model are meta-labels, which are defined as a binary label $\{0, 1\}$ that indicates whether the primary model's forecast was profitable or not. Thus, it makes a meta prediction. The output of this model is the probability of a positive outcome and is used to size positions in which the greater the probability, the larger the position size. It represents a trade-off, in which recall is traded for precision, leading to a better model efficiency, represented by a higher F1-score (the harmonic mean between precision and recall).

According to López de Prado (2018a), there are five reasons to incorporate meta-labeling into an investment process. First, meta-labeling can be added, as an ML layer, to any primary model. This includes ML algorithms, econometric equations, technical trading rules, fundamental analyses, factor-based strategies, and even forecasts produced by humans based on intuition. Second, it prevents the risk of overfitting because the primary model determines the expected profit and loss, and the secondary model controls the precision and the number of trades taken (López de Prado 2019). Third, by separating the side from the size prediction, sophisticated strategy structures can be developed. For example, the features driving a market rally may be different from those of a panic sell-off. Fourth, sizing positions correctly is critical to building a successful strategy. Achieving high accuracy on small bets but low accuracy on large ones will lead to ruin. Therefore, it makes sense to develop a strategy that focuses on this aspect alone. As a fifth reason, we add that meta-labeling helps to address the problem of nonstationarity in financial time series. The secondary model determines under which market conditions an algorithm is unlikely to perform well and drastically limits the positions sizes. This allows the strategy to switch off when conditions are unfavorable.

Exhibit 1 presents the architecture of a primary model. It focuses on forecasting the side of the trade, and it takes as its input features that are indicative of a directional move. It outputs the predicted side $\{-1, 0, 1\}$, where -1 is a short position, 0 to close any open positions, and 1 a long position. The model could also be designed for a strategy that is short only $\{-1, 0\}$ or long only $\{0, 1\}$.

If using an ML algorithm, then the output will be between $[-1, 1]$. An additional step can be added where a threshold value (τ) is used to adjust for precision and recall (recall check). Values greater than τ are labeled as 1, lower than $-\tau$ as -1, else 0.

EXHIBIT 1

Primary Model Architecture

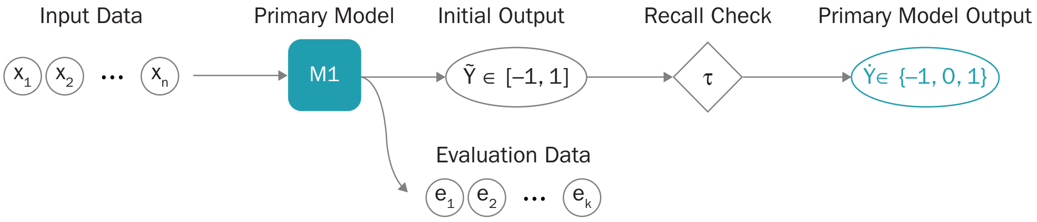
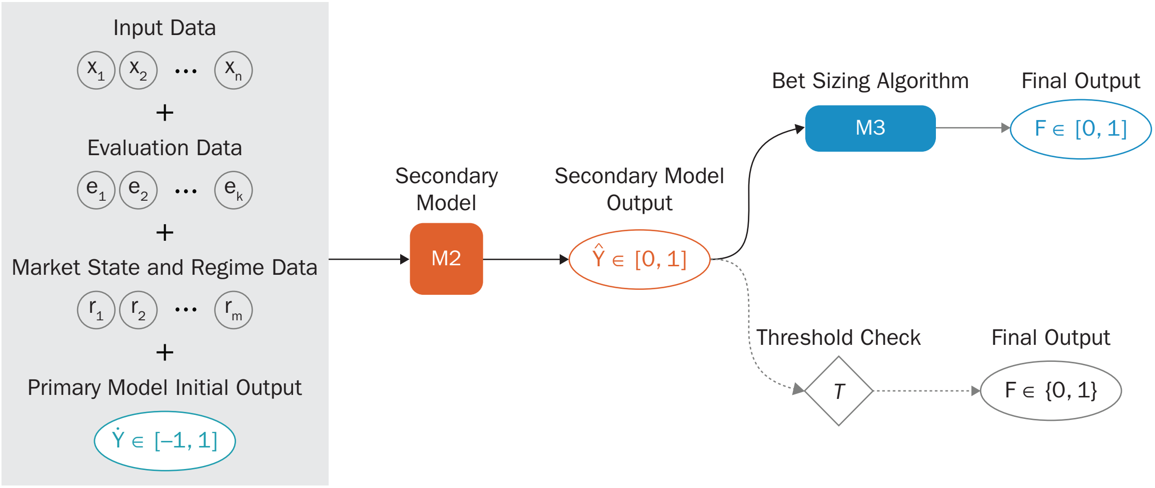


EXHIBIT 2

Meta-Labeling Architecture



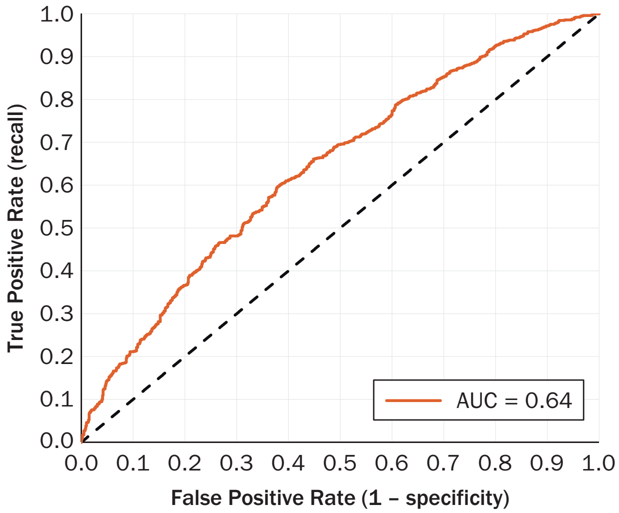
Fitting a primary model that has a high recall is an important step in the process so that the low precision can be corrected by the secondary model, resulting in a higher F1-score (López de Prado 2018b). The primary model also provides model evaluation statistics, which are used in the secondary model as a feature indicative of false positives.

Exhibit 2 presents the architecture for meta-labeling. It is a binary classification that aims to filter out false positives. The target variable are the meta-labels, and its input features can be broken down into three main categories, which are discussed in more detail in the "Methods and Results" section of this article. First, the secondary model could potentially exploit the information used in the primary model, increasing the information advantage; thus, the original features used in the primary model should be passed on to the secondary model. Second, evaluation statistics based on the primary model should be included to help indicate when recent performance is poor. Third, market state statistics and regime-related features should be included, and finally, if the primary model is an ML algorithm, then the model's forecasted probability ($\dot{y}$) should be added. The latter three points are part of the modeling for false positives component.

The output of the secondary model ($\hat{y}$) is the probability of a true positive [0, 1], which is then used to size positions. In Exhibit 2, M3 represents a position sizing algorithm that transforms the probability into a position size that aims to improve the strategy's statistics. Alternatively, an all or nothing approach could be taken, in which a threshold value (T) is added, and any probabilities greater than T result in a full investment, else none.

EXHIBIT 3

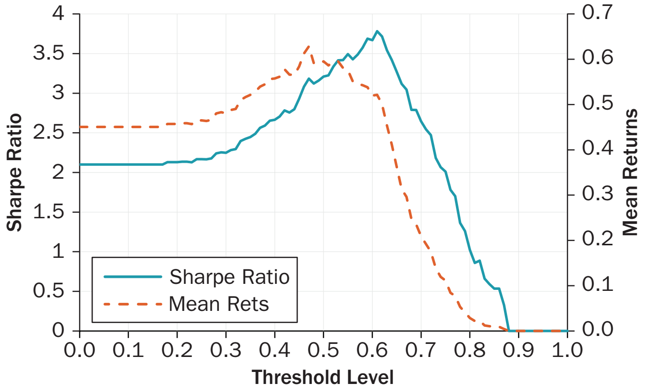
Receiver Operating Characteristic (ROC)



SOURCE: Recreated from Exhibit 4, p. 5, López de Prado (2019).

EXHIBIT 4

Effect of Changing the Threshold on Strategy Metrics



SOURCE: Recreated from Exhibit 4, p. 5, López de Prado (2019).

Exhibit 3 presents the ROC curve of a secondary model and illustrates how, as T is adjusted, the true positive and false positive rates are affected. As T approaches 1, the true positive rate, also known as recall, approaches 0. Conversely, as T approaches 0, the recall increases at the expense of an increase in false positives.

Exhibit 4 illustrates the importance of adjusting T to maximize the Sharpe ratio. In this case, the default value for T is 0.5, which produces a Sharpe ratio of 3.21. When T = 0.62, the Sharpe ratio increases by 0.5 to 3.71, a significant impact. This is the result of filtering out the losing trades/false positives.

Note that the secondary model cannot produce any new signals; rather, it filters out poor ones. Therefore, it is vital to build the best possible primary model. The technique also differs fundamentally from the use of an ensemble model with new features or the use of stacking in the primary model. This is because the target variable of the secondary model are the meta-labels and not a side prediction, and it offers the opportunity to benefit from the trade-off between precision and recall to size positions.

Understanding a Strategy's Risks

Before applying meta-labeling to a strategy, the first step is to understand the strategy's risks (López de Prado 2018, Chapter 15) and to determine what the impact on the Sharpe ratio will be of increasing the precision at the cost of reducing the number of trades. The Sharpe ratio is a function of precision rather than accuracy, as true positives are rewarded, false positives are punished, and negatives are not rewarded as a position was never taken. Each strategy will have its own risk profile, and not all strategies will benefit from this trade-off; however, strategies with a high number of trades will generally be more sensitive to a change in precision.

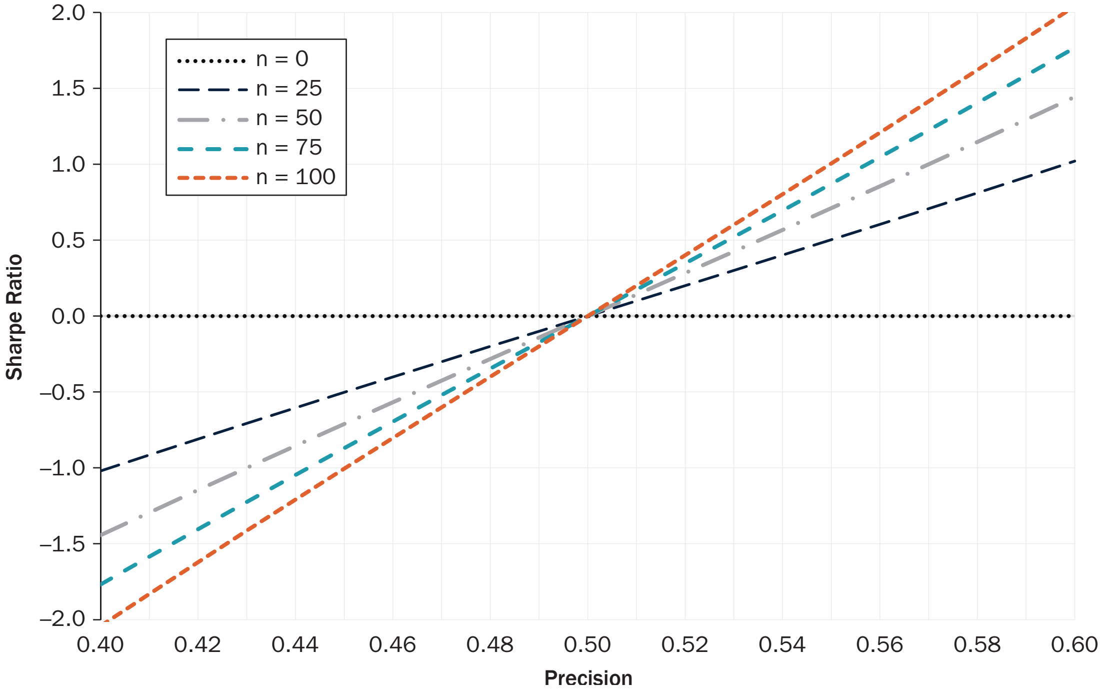
Exhibit 5 illustrates this relationship for different numbers of trades, given symmetric payouts. Note that as the number of trades increases, the slope steepens, indicating a higher sensitivity to a change in precision. The relationship between the Sharpe ratio, precision, payouts, and the number of trades is recreated in the equations that follow (López de Prado 2018a).

The equation for the Sharpe ratio under asymmetric payouts is

$$\theta(p, n, \pi_-, \pi_+) = \frac{n\mathbb{E}[X_i]}{\sqrt{n\mathbb{V}[X_i]}} = \frac{(\pi_+ - \pi_-)p + \pi_-}{(\pi_+ - \pi_-)\sqrt{p(1-p)}}\sqrt{n} \tag{1}$$

EXHIBIT 5

The Relationship between Precision and Sharpe Ratio for Different Numbers of Trades (symmetric payout)



NOTES: 1) *n* represents the number of trades in a year. 2) For *n* = 0, the Sharpe ratio is zero. This is because no trades were taken, and thus, no performance was recorded. 3) This exhibit is recreated from Figure 15.1 in López de Prado (2018a).

We can solve for $0 \leq p \leq 1$, to obtain

$$p = \frac{-2n\pi_- + (\pi_+ - \pi_-)\theta^2 + \theta\sqrt{(\pi_+ - \pi_-)^2\theta^2 - 4n\pi_-(\pi_+ - \pi_-) - 4n\pi_-^2}}{2(\pi_+ - \pi_-)(\theta^2 + n)} \tag{2}$$

where:

- θ = Sharpe ratio
- p = precision
- n = number of trades in a year
- π_- = expected payout of a negative outcome
- π_+ = expected payout of a positive outcome

Equation 2 provides a value for the precision required to generate a target Sharpe ratio for a primary model characterized by parameters $\{\pi_-, \pi_+, n\}$. For example, to achieve a Sharpe ratio of 2, given $n = 252$, $\pi_- = -0.01$, and $\pi_+ = 0.05$, a precision of 0.214 is required.

The number of trades plays an important role in the success of meta-labeling for two reasons. First, the secondary model needs enough observations to be trained on and to find a good fit. Second, a high number of trades increases the probability that a small change in precision will have a larger impact on the Sharpe ratio. López de Prado (2019) noted that, when $\pi_+ \gg \pi_-$, the strategy may also be a good candidate for meta-labeling.

The next section will describe the methodology of three controlled experiments that are designed to break meta-labeling down into the three components of information advantage, modeling for false positives, and position sizing.

METHODOLOGY

Three separate experiments are conducted to highlight how the performance metrics improve in each step of the meta-labeling process.

1. Improving the information advantage.
2. Modeling for false positives.
3. Sizing positions according to model confidence.

Throughout the article, the methodology is kept as simple as possible to maintain the focus of the research on the theory of meta-labeling and to assist practitioners' understanding of the process. This section details the individual components that the experiments consist of.

Data

For each experiment, a linear time series is generated using an autoregressive process of order 3, AR(3):

$$r_{t+1} = \phi_0 + \phi_1 r_t + \phi_2 r_{t-1} + \phi_3 r_{t-2} + a_t \quad (3)$$

where the time series r_{t+1} is the forecasted daily return, with r_t, r_{t-1}, r_{t-2} being the current and lagged 1 and 2 daily returns, respectively. ϕ_0, ϕ_1, ϕ_2, ϕ_3 are the model coefficients, and a_t is a white noise series with a mean of zero and a variance of 0.000211 (the variance was taken from a random sample of IBM's daily returns).

There are four benefits to using a well-understood process such as an AR(3) to run a controlled experiment, rather than using real-life financial time series. First, by using a simple process, a researcher can more easily understand the working of the technique itself. Second, by understanding how the data are generated, simple, informative features can be used rather than introducing complex feature engineering and proprietary datasets. Third, some of the problems associated with forecasting financial time series can be avoided, such as nonlinear and interactive relationships (Gu, Kelly, and Xiu 2020), heteroskedasticity (Engle 1982), and nonstationarity and heavy tails (Cont 2001). Fourth, the risk of backtest overfitting is reduced because the model is not repeatedly trained and scored on the same, single, path-dependent series (Bailey and López de Prado 2014).

In each experiment, 10,000 sequential daily returns are generated using either the single or dual autoregressive process. The primary model is then applied, and the dataset is filtered to include only observations in which the primary model has forecast a positive change in direction. This is because the chosen trading strategy is long only. The data are then split into a train/test set of 60/40 without shuffling, that is, the most recent 40% of the data are held in the test set.

The algorithms to generate the data are as follows:

Single autoregressive process: generates a time series from a single AR(3) process:

$$r_{t+1} = 0.25 r_t - 0.20 r_{t-1} + 0.35 r_{t-2} + a_t \quad (4)$$

Dual autoregressive process: generates a time series that includes data generated from two separate AR(3) processes to simulate regime shifts as well as their persisting nature (Ang and Timmermann 2012):

$$r_{t+1} = 0.25 r_t - 0.20 r_{t-1} + 0.35 r_{t-2} + a_t \quad (5)$$

$$r_{t+1} = -0.0001 - 0.25r_t + 0.20r_{t-1} - 0.35r_{t-2} + a_t \tag{6}$$

At initialization, a uniformly randomly generated number between 0 and 1 is created, and if it is above a 0.80 threshold, then the next 30 observations are generated using Equation 6. This process repeats every 30 observations until the desired sample size is reached. The goal of sampling from Equation 6 is to simulate a regime in which the primary model performs poorly and has difficulty achieving a good fit. The use of a negative drift component also ensures that a long-only model has a lower probability of success. The initial values for r_t, r_{t-1}, r_{t-2} are set to 0.032, 0.020, and –0.042, respectively.

Primary Model

The primary model is based on the German power autocorrelation strategy proposed by Narro and Caamano (2020). Adding the constraint of long only, an end-of-day buy signal is generated using a simple rule: If the difference in price from yesterday's close to today's is positive, then go long at today's close, else exit the position.

$$\dot{y}_t = \begin{cases} 1, \; if \;\; \Delta r_t > 0 \\ \dot{y}_{t-1}, \;\; if \;\; \Delta r_t = 0 \\ 0, \;\; if \;\; \Delta r_t < 0 \end{cases} \tag{7}$$

where $\dot{y}_t$ is the side of position, $\dot{y}_t \in \{0, 1\}$, and Δr_t is the difference in returns between $r_t - r_{t-1}$.

Secondary Model

The goal of the secondary model is to determine whether the side prediction from the primary model is correct and whether a position should be taken. A logistic regression in a binary classification setting is used as it provides good explainability and few parameters to tune. Because the data generation process is linear, a linear classification model such as a logistic regression is suitable.

For its features, the model makes use of the last three returns r_t, r_{t-1}, r_{t-2}. In cases in which the dual autoregressive process has been used, an additional regime detection indicator is added. The regime indicator tracks whether a data point was created using Equation 5 or 6. If created by Equation 6, it is labeled as 1 and 0 otherwise. This is then lagged by five observations to simulate the lag effect of a statistic that has a window size in its calculation. The features in both the train and test sets are standardized by removing the mean and scaling to a unit variance, based on the training dataset for each feature.

The target variable is the meta-label, whether the primary model's forecast is correct, that is, did the prediction lead to a profit or a loss.

$$y_t = \begin{cases} 1, \; if \;\; \dot{y}_t = sign(\Delta r_{t+1}) = 1 \\ 0, \; otherwise \end{cases} \tag{8}$$

where y_t is the meta-label, $y_t \in \{0,1\}$, and $\dot{y}_t$ is the side forecast from the primary model as shown in Exhibit 1.

None of the experiments include regularization, hyperparameter tuning, or cross-validation. This is to avoid overcomplicating the experiment, allowing the reader to gain a more thorough understanding of the process.

Position Sizing

Two position-sizing algorithms are used in the experiments. The first is an all-or-nothing algorithm in which, if a positive outcome is forecasted, then 100% of the portfolio is invested, else exit/hold a position size of 0%.

$$f_t = \begin{cases} 1, & if\ \hat{y}_t > 0.5 \\ 0, & otherwise \end{cases} \tag{9}$$

where f_t is the position size at time t, $f_t \in \{0,1\}$, and $\hat{y}_t$ is the secondary model's forecast of a positive outcome, $\hat{y}_t \in [0,1]$.

The second position sizing algorithm tries to maximize expected returns by placing larger positions on bets that have a higher likelihood of a positive outcome, conversely placing smaller positions on outcomes with a lot of uncertainty. To do this, an empirical cumulative distribution is fitted to the output of the secondary model, based on the training set. This is similar to the sizing algorithm described in the work of López de Prado (2018a):

$$f_t = ECDF(\hat{y}_t) \tag{10}$$

where f_t is the position size at time t and $\hat{y}_t$ is the secondary model's forecast at time t, $f_t \in [0,1]$.

Strategy Creation

A new trading strategy that filters out false positives is then created by only taking a long position when the primary model says to go long and the secondary model confirms the prediction with a positive outcome. The size of the position is determined by applying either Equation 9 or 10.

Metrics

The experiment tracks the classical binary classification metrics: precision, recall, F1 weighted score, accuracy, and area under the curve (AUC). The strategy metrics of expected returns, standard deviation, maximum drawdown, and Sharpe ratio are used. All reported performance metrics are based on the test set.

Number of Trials

One thousand trials are run for each experiment to ensure that the results are not distorted by a single path. We evaluate the distributions of each metric and report on the performance in the results section.

EXHIBIT 6

Summary Statistics of the Changes in Strategy Metrics for the 1,000 Trials in Experiment 1: Informational Advantage

	ΔSharpe^{ip}	ΔMean^{ip}	ΔStd Dev^{ip}
Mean	1.640	0.194	–0.035
Median	1.631	0.194	–0.035
Standard Deviation	0.237	0.042	0.004
Min	**0.770**	**0.026**	**–0.044**
Max	2.384	0.330	–0.026

NOTES: 1) The difference in Sharpe ratio, mean returns, and standard deviation between the secondary model and primary model are all scaled to an annualized level. 2) ip denotes a metric that is the difference between a secondary model trained only on price data (i.e., the information advantage) and the primary model.

Experiments

This section describes the components used in each experiment, making it clear how the results can be replicated.

Experiment 1: improving informational advantage. The goal of this experiment is to show empirically that meta-labeling adds value when a primary model fails to exploit the information in the data to its full potential. This answers the question: Should the features used in the primary model be included when training the secondary model?

- Data generating process: single autoregressive process
- Total sample size: 10,000
- Features: r_t, r_{t-1}, r_{t-2}
- Target variable: meta-labels
- Position sizing: all-or-nothing

Experiment 2: modeling for false positives. Financial time-series data are plagued by nonstationarity and various changing persistent regimes. The goal of this experiment is to highlight the fact that features informative of false positives provide value to the secondary model. This experiment introduces the concept of regime switching by sampling the subsequent 30 observations from a different AR(3) process when a probability threshold is crossed.

- Data generating process: dual autoregressive process
- Total sample size: 10,000
- Features: r_t, r_{t-1}, r_{t-2}, and regime indicator
- Target variable: meta-labels
- Position sizing: all-or-nothing

Experiment 3: position sizing. Combining experiments one and two, the third and final goal is to validate that sizing positions based on the likelihood of a positive outcome will lead to improved strategy performance.

- Data generating process: dual autoregressive process
- Total sample size: 10,000
- Features: r_t, r_{t-1}, r_{t-2}, and regime indicator
- Target variable: meta-labels
- Position sizing: Empirical Cumulative Distribution Function (ECDF)

RESULTS

Experiment 1: Information Advantage

Experiment 1 focuses on improving the information advantage and seeks to answer the question: Should the features used in the primary model be used in the secondary model? The primary model only made use of Δr_t when predicting the side. This is a simple strategy that fails to consider that r_{t-1} and r_{t-2} are also informative. By fitting a secondary model that includes this data, a more accurate forecast can be made. Exhibit 6 presents the summary statistics of the meta-premium, which is defined as the difference in the trading strategies metrics between the secondary and the primary model.

Exhibit 6 shows that all the trials had a positive premium across all metrics. This is highlighted by the values in the minimum row. The Sharpe ratio was increased by higher expected returns and a lower expected standard deviation.

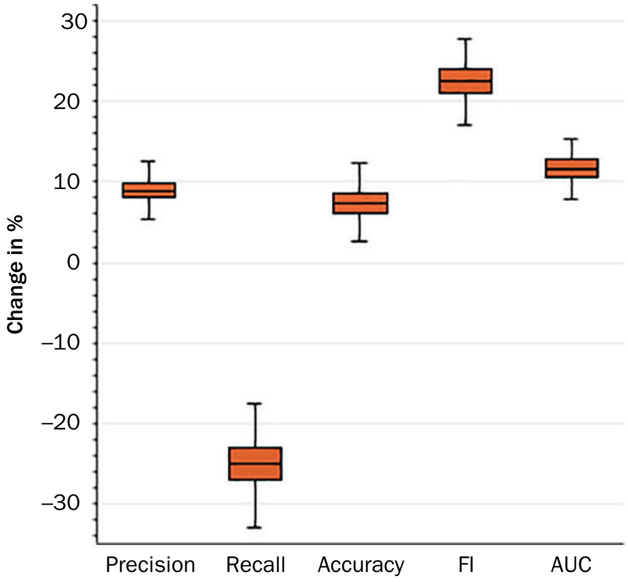
Thus, meta-labeling works when it can exploit the information in the data better than the primary model. Additionally, different models exploit information in the data in different ways (Li, Turkington, and Yazdani 2020); therefore, by introducing model diversity, additional performance may be gained.

Another example of an information advantage is when the primary model is linear and a nonlinear secondary model is used, thus benefiting from the nonlinear and interactive relationships. The information advantage component of meta-labeling may also indicate that a better primary model could be built.

Exhibit 7 illustrates the change in classification metrics between the secondary and primary models. A net positive change indicates an increase in performance and a negative change a decrease. Most notably, we can observe a considerable increase in the F1-score, indicating a more efficient classifier, capable of filtering out false positives; this is done by trading some recall for a higher precision (López de Prado 2020).

EXHIBIT 7

Boxplot of the Changes in Classification Metrics for the 1,000 Trials in Experiment 1: Informational Advantage



Experiment 2: Modeling for False Positives

Experiment 2 sets out to show how the meta-premium is increased when modeling for false positives. To do this, two regimes were introduced that made it harder for the secondary model to find a good fit. This is evidenced by the mean of ΔSharpe^{ip}, which has a lower value in Exhibit 8 than in Exhibit 6.

The useful features for identifying false positives can be split into two groups: model evaluation statistics and market state statistics. The former focus on measuring the health of the primary model's recent performance. Relevant concepts include rolling accuracy, F1, recall, precision, and AUC scores. It is also possible to measure for structural breaks in the error terms, apply a Kalman filter to the log loss, and evaluate the acceleration of various statistics. If the primary model is also an ML algorithm, then $\dot{y}$ can be added as a measure of the model's confidence.

Market state statistics, on the other hand, reflect that the primary model is likely to perform differently in various market regimes. A toy example is the trend-following strategy of crossing moving averages. These are known to suffer when volatility increases and markets stop trending, leading to many false signals (whipsaws) and a loss in profits. The secondary model could identify that an increase in volatility and a lack of autocorrelation may be a poor environment. Gu, Kelly, and Xiu (2020) found that the most successful predictors are price trends, liquidity, and volatility; we recommend starting with those.

A market state can also be represented by the four moments of distribution: mean, variance, skew, and kurtosis, as well as momentum and acceleration. We also recommend testing market regime statistics by employing, for example, structural break tests (Phillips, Shi, and Yu 2015), change point detection (Aminikhanghahi and Cook 2017), and directional change and regime tracking frameworks (Chen and Tsang 2020).

EXHIBIT 8

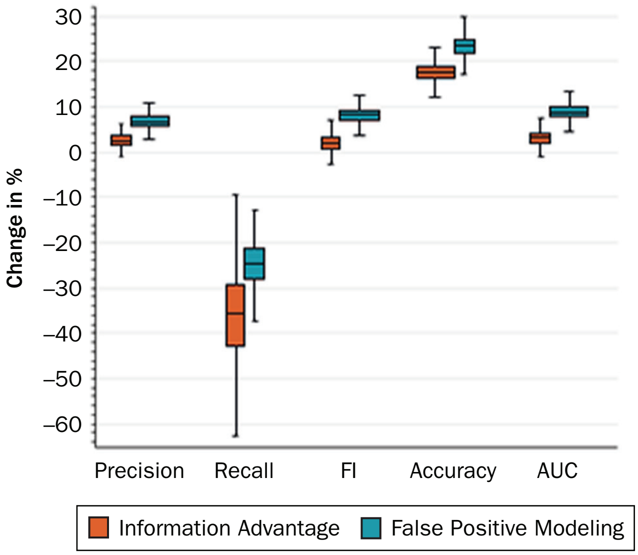
Summary Statistics of the Changes in Strategy Metrics for the 1,000 Trials in Experiment 2: False-Positive Modeling

	ΔSharpe^{ip}	ΔSharpe^{fp}	ΔMean^{ip}	ΔMean^{fp}	ΔStd Dev^{ip}	ΔStd Dev^{fp}
Mean	0.410	1.495	0.054	0.240	–0.039	–0.038
Median	0.408	1.478	0.054	0.236	–0.037	–0.038
Standard Deviation	0.235	0.287	0.042	0.053	0.014	0.006
Min	**–1.308**	**0.481**	–0.115	0.052	–0.155	–0.060
Max	1.163	2.514	0.214	0.465	–0.002	–0.019

NOTES: 1) The difference in Sharpe ratio, mean returns, and standard deviation are all scaled to an annualized level. 2) ip denotes a metric that is the difference between a secondary model trained only on price data (i.e., the information advantage) and the primary model. 3) fp denotes a metric that is the difference between a secondary model (trained on both returns data [i.e., the information advantage] and features indicative of false positives [i.e., the false positive advantage]) and the primary model.

EXHIBIT 9

Boxplot of the Changes in Classification Metrics for the 1,000 Trials in Experiment 2: False-Positive Modeling



López de Prado and Fabozzi (2020) noted that macroeconomic variables are also useful for regime detection.

Feature importance algorithms such as the model fingerprint (Li, Turkington, and Yazdani 2020), clustered mean decrease accuracy (López de Prado 2020), and Shapley values should be used to guide research and allow uninformative features to be dropped. Man and Chan (2021) showed that meta-labeling improves the Sharpe ratio and that a careful selection of features can boost this even further.

Exhibit 8 presents the meta-premiums from both the informational advantage and false-positive modeling, denoted *ip* and *fp*, respectively. In all cases, the mean of *fp* outperforms *ip*, indicating that the addition of modeling for false positives adds value to the meta-premium.

Note how the minimum value of the ΔSharpe^{ip} is negative, highlighting that, at times, modeling only for informational advantage leads to worse returns than the primary model. This is due to the shifting regimes that make it hard for the logistic regression to find a good fit. Adding to this we can see the minimum of ΔSharpe^{fp} is a positive value, indicating that once modeling for false positives is added, even losing strategies became profitable.

Exhibit 9 illustrates the changes in the classification metrics, again demonstrating the improvement of all the metrics when compared to just the information advantage. The same behavior can be observed as in Exhibit 7, where recall is traded for an increase in precision, F1-score, accuracy, and AUC.

Experiment 3: Position Sizing

Experiment 3 enhances the all-or-nothing sizing strategy by sizing positions based on the model's forecasted predictions, that is, high probability outcomes are given a larger size. Exhibit 10 compares the strategy performance metrics for buy and hold (BAH), the primary model, meta-labeling with an informational advantage, meta-labeling with the addition of modeling for false positives, and finally, the position sizing from Equation 10, termed meta-sizing in the exhibit.

EXHIBIT 10

Boxplots of Strategy Performance Metrics for the 1,000 Trials in Experiment 3: Position Sizing

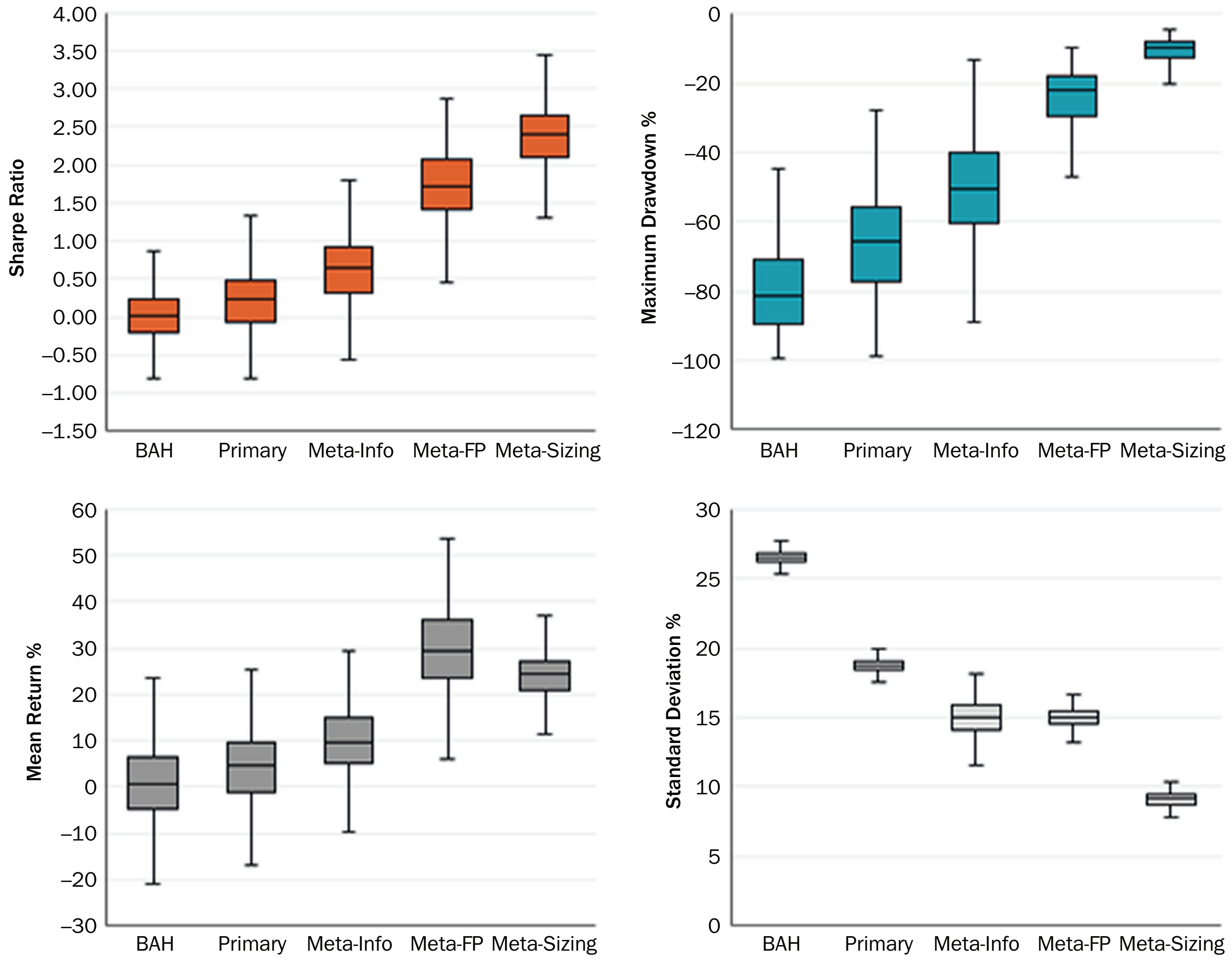


Exhibit 10 shows how, at every step of the meta-labeling process, the performance improves. The meta-sizing step sees a slight decrease in the expected mean return; however, because this is accompanied by a substantial reduction in the standard deviation, it leads to an improved Sharpe ratio.

The substantial reduction in the maximum drawdown and standard deviation shows that meta-labeling is effective not only in filtering out losing trades but also at increasing the quality of trades taken. Both a lower maximum drawdown and a higher Sharpe ratio allow strategies to assume more leverage and scale.

A promising area for further research would be to determine whether the model's probability forecasts can be transformed into a frequentist interpretation of probability so that sophisticated techniques such as the risk-constrained Kelly criterion (Busseti, Ryu, and Boyd 2016) can be applied. Equation 10 does not reflect the optimal position size.

CONCLUSION

This article has provided a clear framework for applying meta-labeling to trading and investment strategies by breaking the process down into three key areas: information advantage, modeling for false positives, and position sizing. We validated the trade-off between recall and precision, showing that meta-labeling not only improves classification metrics but also strategy metrics. Due to its ability to significantly improve the performance of various types of primary models, meta-labeling is a good case study of how ML can be applied in financial markets.

There are three areas in which further research may prove useful: first, investigation of which features are informative of false positives, across various primary models and asset classes, and to publish a curated list; second, investigation of the impact of position sizing techniques and determine an optimal algorithm; and third, development of a framework to select model architectures for meta-labeling and exploration of how ensembles can be incorporated.

ACKNOWLEDGMENTS

The author is grateful to Marcos López de Prado for reviewing this article and for his contributions to the quantitative finance community, and the author thanks Valeriia Pervushyna for her help on the visualizations.

REFERENCES

Aminikhanghahi, S., and D. J. Cook. 2017. “A Survey of Methods for Time Series Change Point Detection.” *Knowledge and Information Systems* 51 (2): 339–367.

Ang, A., and A. Timmermann. 2012. “Regime Changes and Financial Markets.” *Annual Review of Financial Economics* 4 (1): 313–337.

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.

Chen, J., and E. P. K. Tsang. *Detecting Regime Change in Computational Finance: Data Science, Machine Learning and Algorithmic Trading*. Boca Raton, Florida: CRC Press, 2020.

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

Engle, R. F. 1982. “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation.” *Econometrica: Journal of the Econometric Society* 50 (4): 987–1007.

Gu, S., B. Kelly, and D. Xiu. 2020. “Empirical Asset Pricing via Machine Learning.” *The Review of Financial Studies* 33 (5): 2223–2273.

Li, Y., D. Turkington, and A. Yazdani. 2020. “Beyond the Black Box: An Intuitive Approach to Investment Prediction with Machine Learning.” *The Journal of Financial Data Science* 2 (1): 61–75.

López de Prado, M. *Advances in Financial Machine Learning*, 1st ed. Hoboken, New Jersey: Wiley. 2018a.

——. 2018b. “The 10 Reasons Most Machine Learning Funds Fail.” *The Journal of Portfolio Management* 44 (6): 120–133.

——. 2019. "Ten Applications of Financial Machine Learning." *SSRN* 3365271.

——. *Machine Learning for Asset Managers*, 1st ed. Cambridge, UK: Cambridge University Press, 2020.

López de Prado, M., and F. J. Fabozzi. 2020. "Crowdsourced Investment Research through Tournaments." *The Journal of Financial Data Science* 2 (1): 86–93.

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.

Narro, J., and M. Caamano. *Systematic Trading in Energy Markets*. London, UK: Risk Books, 2020.

Phillips, P. C. B., S. Shi, and J. Yu. 2015. "Testing for Multiple Bubbles: Historical Episodes of Exuberance and Collapse in the S&P 500." *International Economic Review* 56 (4): 1043–1078.