Stocks Portfolio Construction Based on Meta-Labeling Machine Learning

PIRES FEITOZA, AFONSO[1]
[1]Instituto Tecnológico de Aeronáutica, Praça Marechal Eduardo Gomes, 50 - Vila das Acacias, São José dos Campos - SP, 12228-900 (e-mail: afonsofeitoza@ita.br)

Corresponding author: PIRES FEITOZA, Afonso (e-mail: afonsofeitoza@ita.br).

This white paper is part of the requirements for the completion of the graduate course "PO-245. Aprendizado de Máquina em Finanças Quantitativa".

ABSTRACT In quantitative financial strategies, defining the position side and size separately is essential to avoid inefficiencies and over complexities [1]. The meta-labeling approach address this need. However, due to the nature of the financial industry, which keep its edge drivers from open publications, its implementation details and effectiveness is not well-known in the academia. The goal of this article is to further the understanding of this technique and foster open discussion. To achieve this, the framework presented by Joubert [3] is extended to include real stocks' return times series, as well as portfolio construction. As a result, the observed meta-labeling premium is promising but limited, requiring further investigation.

INDEX TERMS Machine learning, meta-labeling, portfolio construction, quantitative finance

I. INTRODUCTION

Machine Learning (ML) constitute the current wave of quantitative innovation in finance. Yet, few investment firms succeed in delivering alpha to their investors through ML. According to López de Prado [1], one of the reasons why machine learning-based quantitative funds often fail is related to simultaneously learning the position side (buy, sell, or neutral) and the position size. This aspect not only makes the model complex and inefficient but also involves decisions of different natures. Choosing a position is related to price-value judgment, while determining the position size is more related to risk management [1].

The approach called Meta-Labeling, proposed by López de Prado himself, address this problem [2]. In this technique, the position signal is defined by a primary strategy, which can come from a dedicated machine learning algorithm, econometric equations, technical analysis, fundamental analysis, or even discretionary decisions. From the primary signal and additional information, a secondary model learns to predict the probability of this signal being a true positive, which can be used to define the bet sizing [3].

Although Meta-Labeling has gained traction among market practitioners [4], very few works have managed to traverse the financial industry's secrecy and make their way into academic publications. Thus, there are still many unanswered questions about implementation details and even its efficiency. Recently, a series of articles by the Hudson and Thames (H&T) quantitative research firm shed some light on the subject with the proposal of a practical the meta-labeling framework [3].

This white paper is structured as follows: Firstly, I review H&T's articles on the subject, covering the fundamentals of meta-labeling. Secondly, I outline the methodology used to reproduce and expand upon their approaches. Subsequently, following the proposed methodology, I present and discuss the results obtained. Finally, I draw conclusions and suggest potential directions for future works.

II. LITERATURE REVIEW

The researchers from H&T detailed, implemented, and extended López de Prado's ideas regarding Meta-Labeling [3]–[6]. An overview of the followed architecture is shown in Figure 1.

In these studies, the primary model (M1 in Figure 1) corresponds to an autocorrelation model that receives n directional movement indicative attributes to learn to predict the position signal (buy, sell, or neutral). The initial result of the primary model, $\tilde{Y}$, is a continuous range [+1,-1], on which a threshold τ is applied to determine the final discrete signal, $\dot{Y}$, {-1,0,+1}. This result of the primary model, its performance metrics, regime attributes, and market state attributes are then used as input to the secondary model (M2 in Figure 1), whose output $\hat{Y}$ represents the probability, [0,+1], of the primary model's signal being a true positive. In turn, this result from

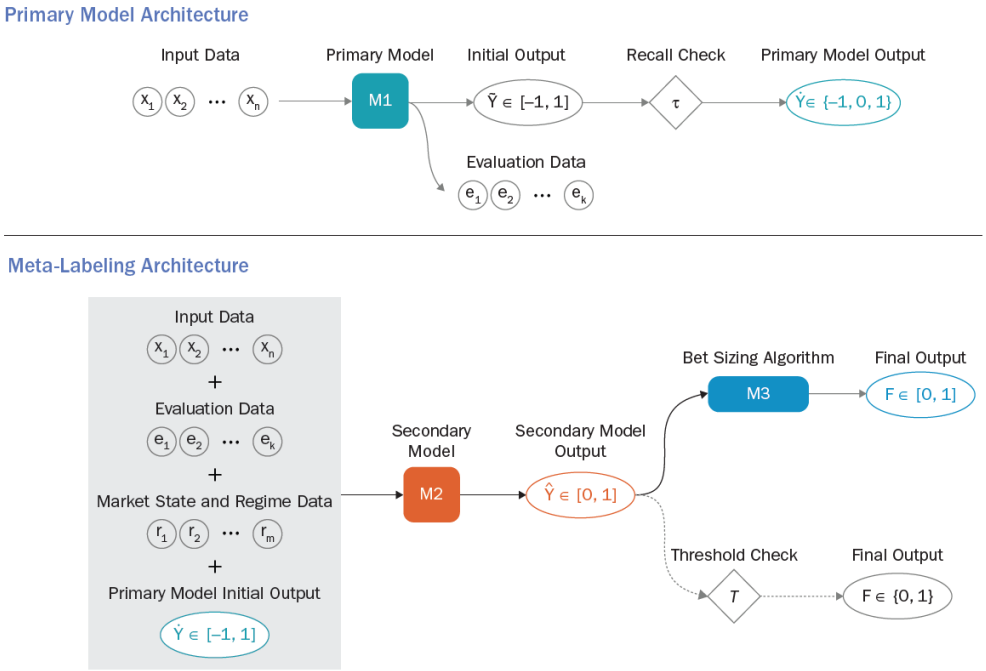

FIGURE 1. Meta-Labeling Approach Architecture [3].

the secondary model is used to determine the position size. In one method, an algorithm (M3 in Figure 1) defines larger positions when the primary model's accuracy probability is higher and smaller positions when this probability is lower. Alternatively, an all-or-nothing method, {0,1}, is used to apply all available resources in the asset if a threshold τ is reached by the output of the secondary model. Otherwise, the position is zeroed.

To evaluate the performance of the basic Meta-Labeling architecture proposed, Joubert and collaborators [3] conducted controlled experiments using synthetic data. For this purpose, linear time series were generated by a third-order autoregressive process (with returns from the current day and the two previous days). Combinations of these series were made to generate changes in the asset's regime (up and down trends; high and low volatility) and thus evaluate the models' ability to identify such changes. As the secondary model, the authors used logistic regression due to the linearity of the input data and the simplicity in adjusting its parameters. For the evaluated experiments, the authors demonstrated an improvement between the primary model and Meta-Labeling, both in terms of model metrics (precision, recall, total accuracy, F1 score, and AUC) and strategy metrics (Sharpe ratio, maximum drawdown, average return, and maximum volatility) [3].

Concerning the secondary model, Thumm, Barruca, and Joubert analyzed different ensemble methods [5]. In this study, the authors compared the base secondary model (logistic regression) with the following ensembles: (i) LightGBM (Microsoft's light gradient boosted machine); (ii) homogeneous dynamic ensembles (with Random Forest); (iii) heterogeneous dynamic ensembles (with Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, and Multilayer Perceptron). They concluded that employing ensembles provided a considerable gain compared to using a single secondary model.

Finally, in the analyzed series of Hudson&Thames articles on Meta-Labeling, Meyer, Barziy, and Joubert evaluated different formulations of the sizing algorithm (M3 in Figure 1) [6]. To transform the true positive probability provided by the secondary model into the position size, the authors considered different methods, such as All-or-Nothing and Normal CDF. In all methods, the minimum threshold considered was 50% (position is zeroed when the probability does not reach this threshold). According to the authors, some methods are more aligned with the profile of investors seeking maximum return, while others are for those seeking minimum volatility and drawdown.

In all those studies, the primary model remained the same, i.e., a simple autocorrelation model [3]–[6]. Thus, it is not clear how more sophisticated primary models would affect the meta-labeling premium. Besides, while evaluating ensemble methods in the secondary model [5], it was not possible to identify the isolated effect of each algorithms used, such as their bias-variance trade-off, on the final meta-labeling performance. Additionally, although justified for experiment controlling purposes, the use of synthetic return time series doesn't expose the technique to all possible stylized behaviors of real financial times series.

III. METHODOLOGY

In order to thoroughly understand H&T's methodology and ensure its accurate implementation, the first step involves replicating the results from their reference paper [3], which utilize synthetic data. Once these results are successfully reproduced, confidence is built for extending the methodology to real stock data and the construction of stock portfolios.

In the subsequent subsections, it is presented the details of the meta-labeling implementation spanning from raw data curation, data preparation, secondary model training, position signal prediction, sizing definition, portfolio construction and up to backtesting. The flowchart of Figure 2 provides an overview of the meta-labeling implementation sequence. The current implementation is performed in Python code and leverages on its available libraries.

A. RAW DATA CURATION

Synthetic Data

In H&T's series of articles [3]–[6], controlled experiments are run through the use of synthetic daily stock returns. Namely, a linear time series is generated employing an auto-regressive process of order 3, AR(3):

$$rets_{t+1} = \phi_0 + \phi_1 rets_t + \phi_2 rets_{t-1} + \phi_3 rets_{t-2} \quad (1)$$

where the time series r_{t+1} represents the forecasted daily return. Here, r_t, r_{t-1}, and r_{t-2} refer to the current day, one day lagged and two day lagged returns, respectively. The models coefficients are ϕ_0, ϕ_1, ϕ_2, ϕ_3, and a_t represents a white noise series. This noise series has a mean of zero and a variance of 0.000211 (derived from a random sample of IBM's daily returns).

In H&T experiments, 10.000 sequential dailly returns are generated using two variations of equation (1) to simulate regime changes. Those two variations have the same coefficients, but with opposite signs. Every 30 data points, if a random variable between 0.0-0.1 is greater than 0.8, the time series is generated by the second variation of auto-regressive process. A regime indicator tracks which variation of equation (1) was used to create a data point.

Real Stocks Data

For this project, S&P500 stock prices are retrieved from FICO's database. Pre-pandemic data (2014-2029) was used to avoid severe time series breaks, which are beyond the scope of the present study. Using FICO's proprietary modules these stock price time series are converted into stocks log-returns time series.

B. DATA PREPARATION

For each individual stock, a training dataframe ($iModelData$ in Figure 2) is build-up as follows:

- Based on the time series of the stock's logarithmic returns ($rets$ in Figure 3), the signal of the primary model ($pmodel$ in Figure 3) is "buy" if the current training return is positive; otherwise, the position is null (2).

$$pmodel_t = \begin{cases} 1 & \text{, if } rets_t > 0 \\ 0 & \text{, otherwise} \end{cases} \quad (2)$$

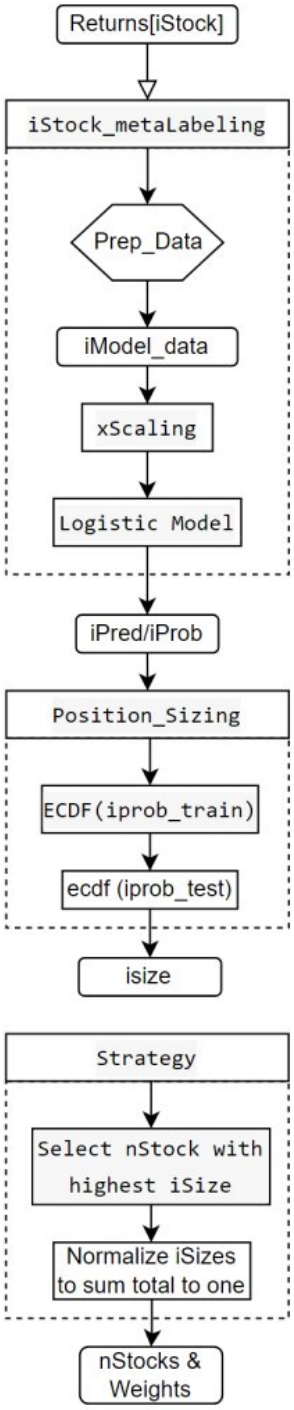

FIGURE 2. Meta-Labeling implementation flowchart

- Still based on the time series of the stock's logarithmic returns ($rets$ in Figure 3), the signal of the secondary model ($target$ in Figure 3) is "buy" if the next training return is positive; otherwise, the position is null (3).

$$target_t = \begin{cases} 1 & \text{, if } rets_{t+1} > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{3}$$

- Following the meta-labeling framework proposed by H&T [3], the basic features of the secondary model are the last three daily returns, respectively, $rets$, $rets2$ and $rets3$ (see Figure 3). As explained in section IV, such lagged returns are the only features used in what is called the informational advantage experiment. It is important to note that these are exactly the same features used to create the synthetic daily stock returns.
- The regime change indicator described in section III-A(Synthetic Data) is also used as a feature in what is fancily termed the false positive experiment. This indicator is lagged by five observations to simulate the effect of a statistic that utilizes a window size in its calculation. When dealing with real stock data, this feature is not readily available.
- Standardization of features across the entire dataset is performed using the means and standard deviations of the training features to prevent data leakage.
- Since the purpose of the meta-labeling procedure is to verify if the primary model's signal is correct, the training dataset of the secondary model comprises only datapoints which $pmodel$ is equal to one.
- The final training dataset for the secondary model for informational advantage experiment corresponds to the second table depicted in Figure 3).

Secondary model target variable

Primary model signal

Date	Ret[t] rets	target	pmodel	Ret[t-1] rets2	Ret[t-2] rets3
2011-06-03	0.006446	0.0	1	NaN	NaN
2011-06-06	-0.017041	1.0	0	0.006446	NaN
2011-06-07	0.016089	0.0	1	-0.017041	0.006446
2011-06-08	-0.005733	0.0	0	0.016089	-0.017041
2011-06-09	-0.007213	0.0	0	-0.005733	0.016089
2011-06-10	-0.025664	1.0	0	-0.007213	-0.005733
2011-06-13	0.003213	1.0	1	-0.025664	-0.007213
2011-06-14	0.011776	0.0	1	0.003213	-0.025664
2011-06-15	-0.018711	1.0	0	0.011776	0.003213
2011-06-16	0.005946	1.0	1	-0.018711	0.011776

Meta-Label

Date	x1_train rets	Y_train target	pmodel	x2_train rets2	x3_train rets3
2011-06-07	0.016089	0.0	1	-0.017041	0.006446
2011-06-13	0.003213	1.0	1	-0.025664	-0.007213
2011-06-14	0.011776	0.0	1	0.003213	-0.025664
2011-06-16	0.005946	1.0	1	-0.018711	0.011776

FIGURE 3. Data preparation steps

C. SECONDARY MODEL TRAINING AND TEST

Within the basic architecture of H&T's meta-labeling framework [3], the dataframe described in section III-B is used to train a logistic regression classifier (secondary model) from $scikit-learn$ package, whose output are the predicted label and the corresponding true positive probability. When the predicted class is 1 it means that the primary model correctly predicted next day positive return, i.e., the primary model's signal has a high confidence and can be followed. Once trained, the secondary model is applied over the test datapoints to predict next return signal and its probability.

As mentioned in section II, Joubert's team implemented the secondary model based not only using Logitic Regression (RL), but also with Light Gradient Boosted Machine (GBM), Homogeneous ensembles based on Random Forest (Des) and Heterogeneous ensembles with a myriad of machine-learning algorithms (Pool) [5]. In the present work, ensemble models were also considered in order to grasp the impact of more sophisticated secondary models.

D. SIZING DEFINITION

In the previous section, the position side was defined by the secondary model. In this step, the position size is derived from the probability related to the positive class. The idea is to assign larger position sizes for higher true positive probabilities. Although it does not reflect the optimal position sizing [3], the use of Empirical Cumulative Distribution Function (ECDF) serves this purpose in a simple manner.

Fitting the ECDF from statsmodels package to the probabilities obtained for the training dataset results in a typical cumulative distribution as presented in Figure 4. As one can see, class 1 probabilities near 0.5 are assigned a size close to $zero$, whereas probabilities around 0.7 receive a size of almost one.The size position for an individual stock ($iSize$ in Figure 2) is obtained by applying the fitted ECDF on the features of the Test dataset.

E. PORTFOLIO CONSTRUCTION

Once the position sizes for all stocks have been determined, the portfolio is constructed using the stocks with the highest individual sizes, which typically correspond to the highest positive class probabilities. Final weights are obtained by normalizing selected stocks' size so that they sum up to the unity.

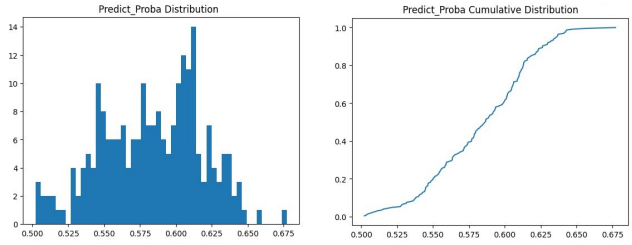

FIGURE 4. Empirical Cumulative Distribution of training returns

F. PORTFOLIO BACKTESTING

Upon implementing the Meta-Labeling strategy, a backtest is performed using FICO's $StrategySimulator$ module to get portfolio's historical weights and its next day returns. Then $quantstats$ package was used to calculate various performance metrics and create a comprehensive TearSheet.

IV. RESULTS AND DISCUSSION

In this section I present the results of reproducing the procedures outlined by H&T for synthetic data and extend then for real stocks data and portfolio construction.

A. INDIVIDUAL STOCK

Synthetic Data

To demonstrate the performance enhancement of the secondary model over the primary model, H&T first built an experiment named as "improving informational advantage"(if). In this experiment, in addition to the primary models feature (r_t), the secondary model is fed with lagged returns (r_{t-1} and r_{t-2}). In a second experiment, "modeling for false positives"(fp), the regime change indicator is used by the secondary model to verify whether features informative of false positive add value to the meta-premium. Lastly, the 'Position Sizing Experiment' explores how bet sizing based on the accuracy probability of the secondary model's prediction impacts strategy performance.

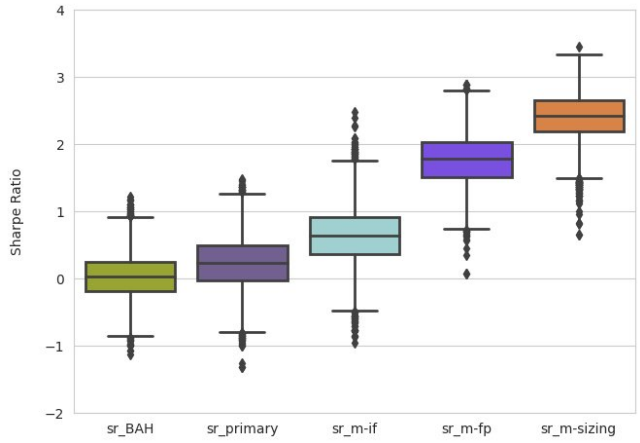

FIGURE 5. Individual Stocks - Synthetic Data - Basic Architecture

Figure 5 summarizes the results of such experiments in terms of the Sharpe Ratio metric. Notably, in these controlled experiments utilizing synthetic data, a significant improvement in the long-only strategy is evident.

Subsequently, the "improving informational advantage" experiment was conducted with more elaborated secondary models (see section III-C.Secondary Model Training and Test). Figure 6 shows the positive impact of secondary models based on ensemble methods when the generated synthetic data are employed.

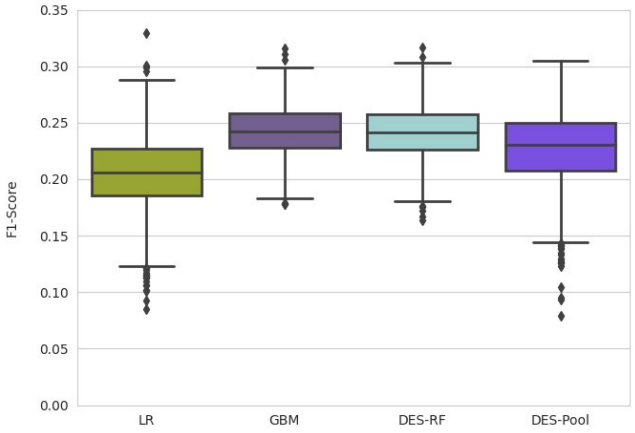

FIGURE 6. Individual Stocks - Synthetic Data - Ensembles

Both results shown in figures 5 and 6 matches exactly what is presented in references [3] and [5], instilling confidence to move forward.

Real Stocks Data

The same procedures used to reproduce H&T results were extended for the real stock data described in section III-A.

Figure 7 presents the results considering the "improving informational advantage" experiment and the simple Logistic Regression secondary model.

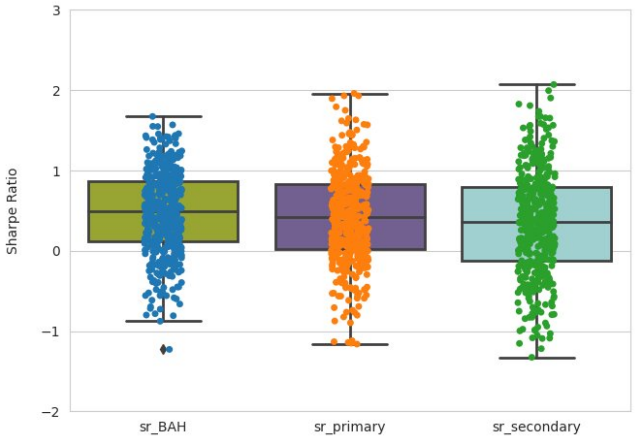

FIGURE 7. Individual Stocks - Real Stocks Data - Basic Architecture

Figure 8 depicts the results considering the "improving informational advantage" experiment and the different secondary models.

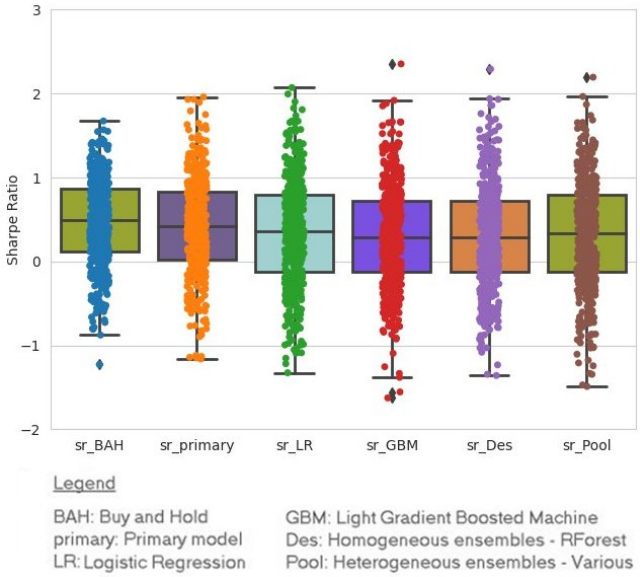

FIGURE 8. Individual Stocks - Real Stocks Data - Ensembles

In contrast to the observed improvement in the secondary model's performance over the primary one with synthetic data, there was no enhancement when real stock data were used. This lack of improvement persists even when employing the more elaborated ensemble secondary models, indicating that the absence of meta-premium cannot be attributed to the secondary model algorithms.

B. STOCKS PORTFOLIO

Figure 9 shows an extract of the TearSheet created running the backtest of the employed portfolio construction strategy, as described in sections III-E and III-F.

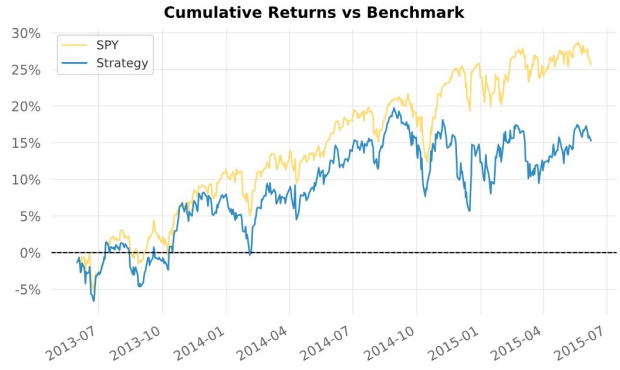

FIGURE 9. Stocks Portfolio - Real Stocks Data - Backtest

The portfolio's performance significantly lagged behind the benchmark across all metrics. This raises concerns regarding the practicality of applying H&T's approach, originally devised for synthetic data, to real stock data.

The success of the secondary model in enhancing the performance metrics in the H&T controlled experiments can be associated with the fact that the lagged returns and the regime change indicator can completely explain the synthetic data, given they are the elements used to generate it in the first place. Hence, although further investigations are need to assert the source of the observed behavior, the suspicion arises from the potential limitations of the used features (only the last three stocks' returns) in capturing the dynamics of real stock return time series.

V. CONCLUSION

The research showcased varied experiments implementing Meta-Labeling, both in controlled settings and with real stock data. While the approach demonstrated promising outcomes in controlled synthetic environments, its translation to real stock data didn't replicate the same success. Despite leveraging more sophisticated ensemble secondary models, the absence of performance improvement suggests limitations within the feature set (solely relying on the last three stocks' returns) in encapsulating real stock dynamics. The disparity between synthetic and real data outcomes necessitates further exploration to comprehend the observed behavior thoroughly.

As with any evolving strategy, there exist opportunities for enhancement and refinement. Plans are underway to further optimize the Meta-Labeling strategy throught the following next steps:

- Enhance the primary model (Random Forest and additional features) and the secondary model (performance metrics of the primary model).
- Improve the criteria used for selecting portfolio assets (e.g., by incorporating Liquidity/Volume filters).

Moreover, future works might use the insights acquired in the present study to integrate this meta-labeling framework as a performance-enhancing module atop different primary strategies.

ACKNOWLEDGMENT

ChatGPT-3.5 was used to improve text readability and flow of sentences. It was not used to generate ideas, data or results.

REFERENCES

[1] López de Prado, M. M. "The 10 Reasons Most Machine Learning Funds Fail". The Journal of Portfolio Management. v. 44, n. 6, p. 120–133, 2018a.
[2] López de Prado, M. M. Advances in Financial Machine Learning. New Jersey: John Wiley & Sons, p. 50-55, 2018b.
[3] Joubert, J. F. "Meta-Labeling: Theory and Framework". The Journal of Financial Data Science. v. 4, n.3, p. 31-44, 2022.
[4] Meyer, M., Joubert, J. F., Messias, A. "Meta-Labeling Architecture". The Journal of Financial Data Science. v. 4, n. 4, p. 10–24, 2022.
[5] Thumm, D., Barucca, P., Joubert, J. F. "Ensemble Meta-Labeling". The Journal of Financial Data Science. v. 5, n. 1, p. 10-26, 2023.
[6] Meyer, M., Barziy, I., Joubert, J. F. "Meta-Labeling: Calibration and Position Sizing". The J. Financial Data Science. v. 5, n. 3, p. 23-40, 2023.