Programa de Pós-graduação em Engenharia Eletrônica e Computação (PG-EEC)

PO-245. Aprendizado de Máquina em Finanças Quantitativa. 2023.2

FICO - ITA

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.

The aim of the present white paper is to shed more light on the subject and encourage open discussion. For this purpose, the meta-labeling framework established by the Hudson and Thames quantitative research firm [3] is applied to real stocks' return time series and used to construct a portfolio consisting of those stocks with the highest true positive rates.

II. LITERATURE REVIEW

In a series of recent articles, researchers from Hudson&Thames 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 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.

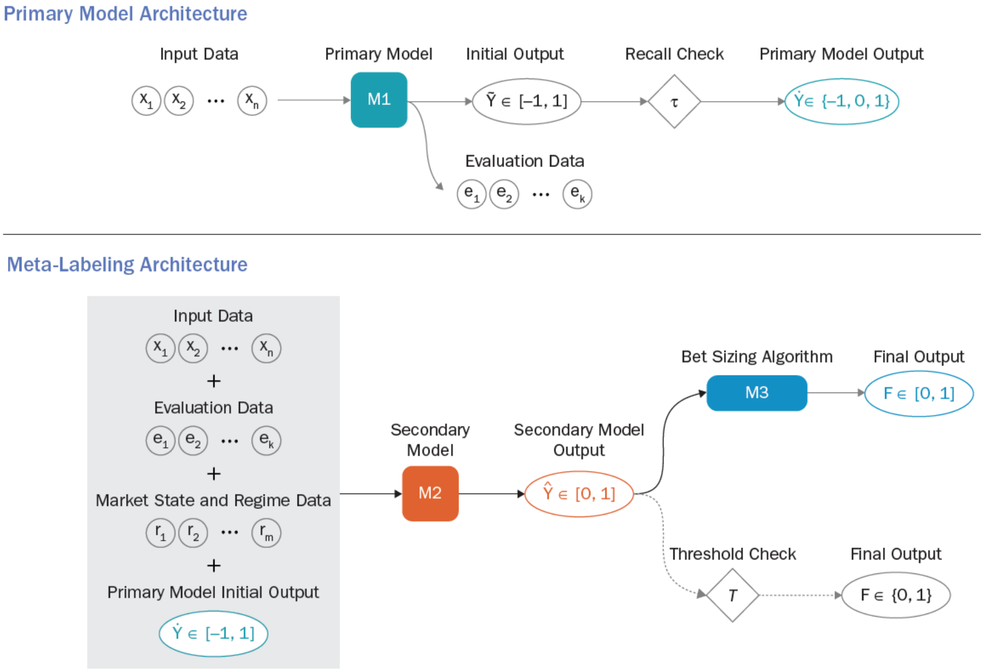

FIGURE 1. Meta-Labeling Approach Architecture [3].

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~~d~~ of synthetic return time series doesn't expose the technique to all possible stylized behaviors of real financial times series. On top of that, as the regime change is incorporated into the synthetic data construction, it is precisely known beforehand, which is rarely the case in the real world.

In the present study, some of those gaps are addressed by evaluating the framework presented by Joubert [3] with real stocks' return times series (and, in the next Bi-Month period, regime change indicators based on historical volatility).

III. METHODOLOGY

In the following subsections, it is presented the details of the meta-labeling implementation ~~spanning~~ data preparation preparation, secondary model training, position signal prediction, sizing definition and portfolio construction. The flowchart of Figure 2 provides an overview of the meta-labeling implementation sequence. It is noteworthy that the current implementation is performed in Python code and leverages on its available libraries.

A. 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 (1).

$$pmodel_t = \begin{cases} 1 & \text{, if } rets_t > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{1}$$

- 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 (2).

$$target_t = \begin{cases} 1 & \text{, if } rets_{t+1} > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

- Following the meta-labeling framework proposed by Thames&Hudson [3], the features of the secondary model are the last three returns, respectively, $rets$, $rets2$ and $rets3$ (see Figure 3). The features in both test and training sets are standardized using the means and standard deviations of the training features.
- 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 corresponds to the second table depicted in Figure 3).

B. SECONDARY MODEL TRAINING AND TEST

The dataframe described in section III-A 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.

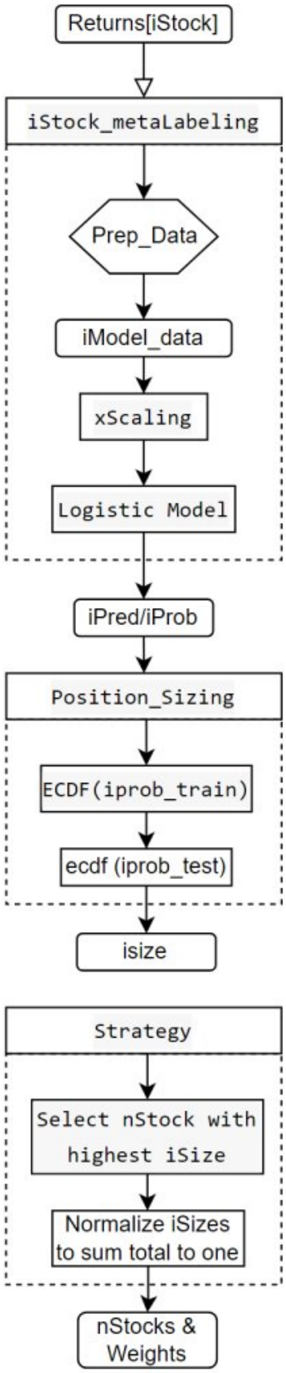

FIGURE 2. Meta-Labeling implementation flowchart

C. 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.

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

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.

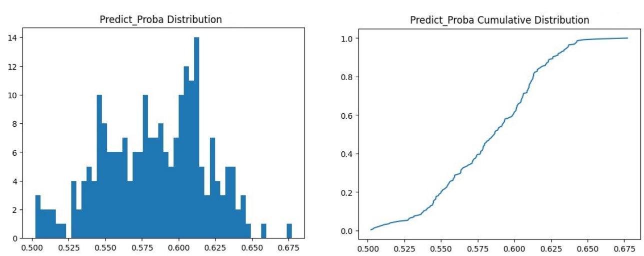

FIGURE 4. Empirical Cumulative Distribution of training returns

D. 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 scaling selected stocks' size so that they sum up to the unity.

IV. RESULTS

While no backtest has been conducted in the current phase of this study, Figure 5 displays the confusion matrix and the Roc Curve obtained for the training dataset. As can be seen, predictions are better than random guessing but to a limited extent.

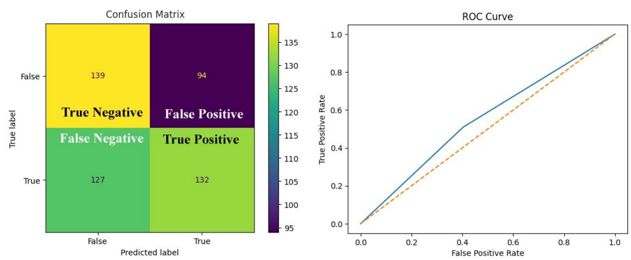

FIGURE 5. Preliminary Results

V. CONCLUSION

This article presented the application of the meta-labeling framework proposed in [3] on real stocks' return times series. Although promising, this implementation have shown limited meta-labeling premium. The main reason might be that the project design choices of the reference paper were reasonable for the employed synthetic data but ~~to~~ simplistic for real stocks' return time series. Additionally, only information advantage and bet sizing advantage were considered in this version of the white paper. Further studies could (i) include regime change indicators based on historical volatility (to account for false positive modeling), (ii) use more sophisticated primary and secondary ML techniques and (iii) evaluate a Long&Short strategy that uses a secondary model for long positions and another secondary model for short positions.

ACKNOWLEDGMENT

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

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 Journal of Financial Data Science. v. 5, n. 3, p. 23-40, 2023.