Labeling

Labeling in Finance

- Virtually all ML papers in finance label observations using the fixed-time horizon method.
- Consider a set of features $\{X_i\}_{i=1,\dots,I}$, drawn from some bars with index $t = 1, \dots, T$, where $I \leq T$. An observation X_i is assigned a label $y_i \in \{-1,0,1\}$,

$$y_i = \begin{cases} -1 & \text{if } r_{t_{i,0},t_{i,0}+h} < -\tau \\ 0 & \text{if } \left|r_{t_{i,0},t_{i,0}+h}\right| \leq \tau \\ 1 & \text{if } r_{t_{i,0},t_{i,0}+h} > \tau \end{cases}$$

where τ is a pre-defined constant threshold, $t_{i,0}$ is the index of the bar immediately after X_i takes place, $t_{i,0} + h$ is the index of *h* bars after $t_{i,0}$, and $r_{t_{i,0},t_{i,0}+h}$ is the price return over a bar horizon *h.*

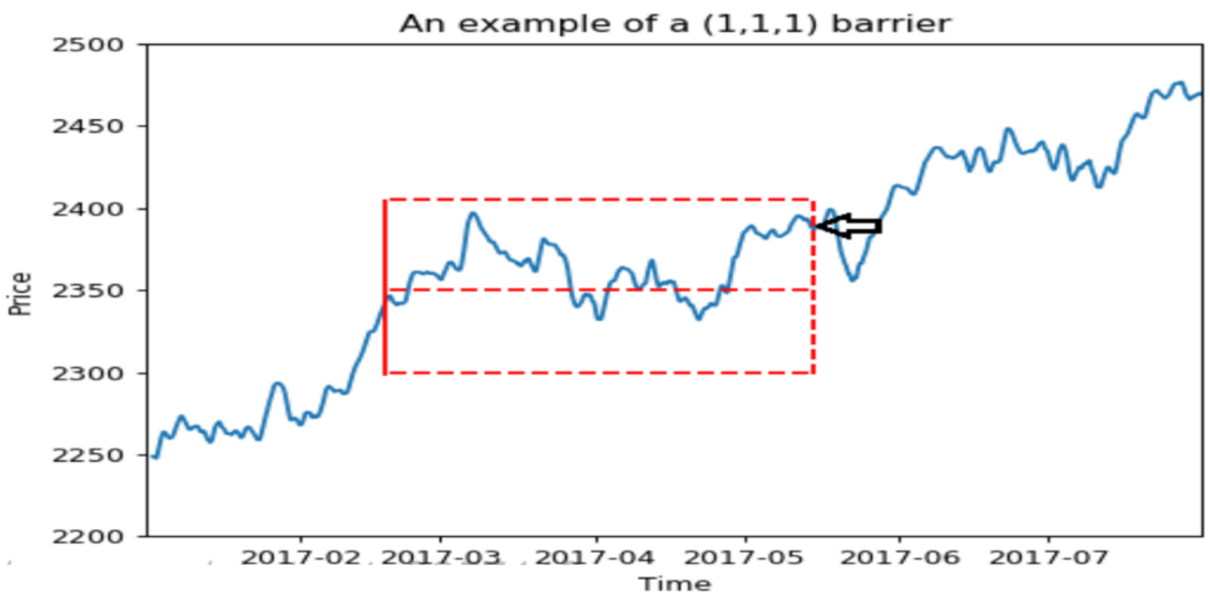
- Because the literature almost always works with time bars, *h* implies a fixed-time horizon.

Caveats of the Fixed Horizon Method

- There are several reasons to avoid such labeling approach:
  - Time bars do not exhibit good statistical properties.
  - The same threshold τ is applied regardless of the observed volatility.
    - Suppose that $\tau = 1E-2$, where sometimes we label an observation as $y_i = 1$ subject to a realized bar volatility of $\sigma_{t_{i,0}} = 1E-4$ (e.g., during the night session), and sometimes $\sigma_{t_{i,0}} = 1E-2$ (e.g., around the open). The large majority of labels will be 0, even if return $r_{t_{i,0},t_{i,0}+h}$ was predictable and statistically significant.
- A couple of better alternatives would be:
  - Label per a varying threshold $\sigma_{t_{i,0}}$, estimated using a rolling exponentially-weighted standard deviation of returns.
  - Use volume or dollar bars, as their volatilities are much closer to constant (homoscedasticity).
- But even these two improvements miss a key flaw of the fixed-time horizon method: The *path* followed by prices. We will address this with the Triple Barrier Method.

The Triple Barrier Method

- It is simply unrealistic to build a strategy that profits from positions that would have been stopped-out by the fund, exchange (margin call) or investor.
- The Triple Barrier Method labels an observation according to the first barrier touched out of three barriers.
  - Two horizontal barriers are defined by profit-taking and stop-loss limits, which are a dynamic function of estimated volatility (whether realized or implied).
  - A third, vertical barrier, is defined in terms of number of bars elapsed since the position was taken (an expiration limit).
- The barrier that is touched first by the *price path* determines the label:
  - Upper horizontal barrier: Label 1.
  - Lower horizontal barrier: Label -1.
  - Vertical barrier: Label 0.

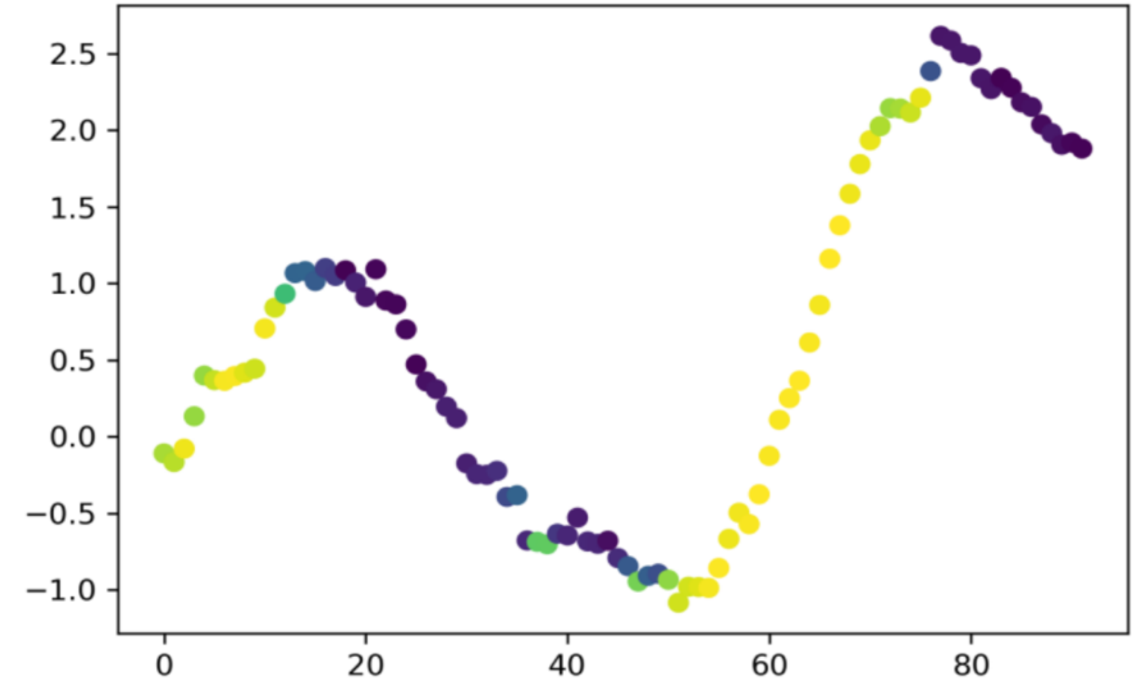


Trend Scanning Method

Consider a series of observations $\{x_t\}_{t=1,\ldots,T}$, where x_t may represent the price of a security we aim to predict. We wish to assign a label $y_t \in \{-1,0,1\}$ to every observation in x_t, based on whether x_t is part of a downtrend, no-trend, or an uptrend. One possibility is to compute the t-value ($\hat{t}_{\beta_1}$) associated with the estimated regressor coefficient ($\hat{\beta}_1$) in a linear time-trend model,

$$x_{t+l} = \beta_0 + \beta_1 l + \varepsilon_{t+l}$$
$$\hat{t}_{\beta_1} = \hat{\beta}_1 \big/ \hat{\sigma}_{\beta_1}$$

where $\hat{\sigma}_{\beta_1}$ is the standard deviation of $\hat{\beta}_1$, and $l = 0, \ldots, L-1$, and L sets the look-forward period, with $L \le t$. Different values of L lead to different t-values. To solve this indetermination, we can try a set of values for L, and pick the value that maximizes $|\hat{t}_{\beta_1}|$. In this way, we assign to x_t the most significant trend observed in the past, out of multiple possible look-forward periods.



Dollar Imbalance Bars (2/2)

- Then, $\mathrm{E}_0[\theta_T] = \mathrm{E}_0[T](v^+ - v^-) = \mathrm{E}_0[T](2v^+ - \mathrm{E}_0[v_t])$
- In practice, we can estimate $\mathrm{E}_0[T]$ as an exponentially weighted moving average of *T* values from prior bars, and $(2v^+ - \mathrm{E}_0[v_t])$ as an exponentially weighted moving average of $b_t v_t$ values from prior bars.
- We define a bar as a T^*-contiguous subset of ticks such that the following condition is met

$$T^* = \arg\min_T \{|\theta_T| \geq \mathrm{E}_0[T]|2v^+ - \mathrm{E}_0[v_t]|\}$$

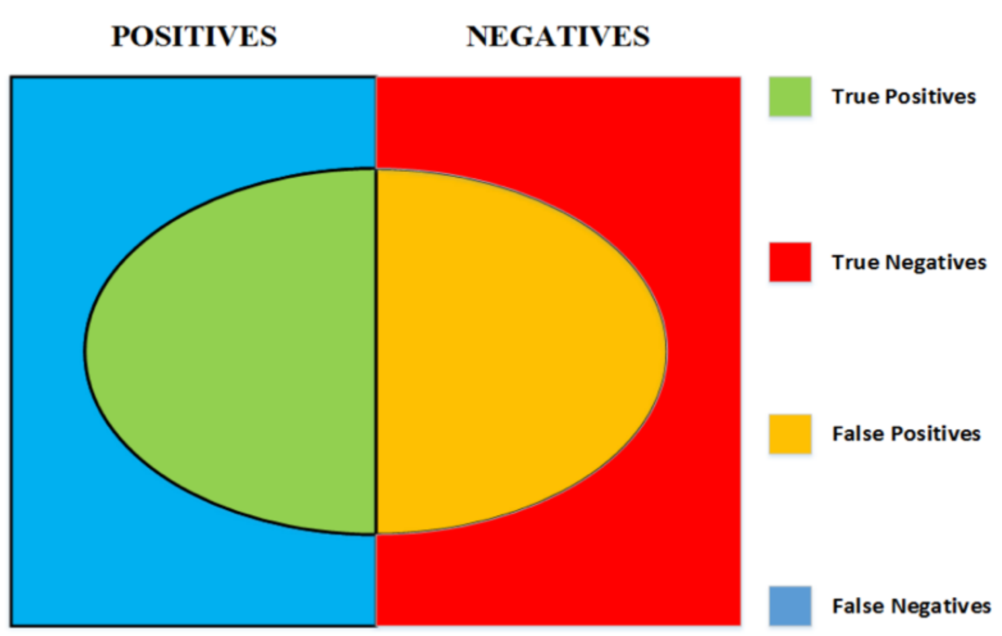
where the size of the expected imbalance is implied by $|2v^+ - \mathrm{E}_0[v_t]|$.

- When θ_T is more imbalanced than expected, a low *T* will satisfy these conditions.

Meta-Labeling

Turning a Weak Predictor to a Strong Predictor

- Suppose that you have a model for making a buy-or-sell decision:
  - You just need to learn the ***size*** of that bet, which includes the possibility of no bet at all (zero size).
  - This is a situation that practitioners face regularly. We often know whether we want to buy or sell a product, and the only remaining question is how much money we should risk in such bet.
  - Meta-labeling: Label the outcomes of the primary model as 1 (gain) or 0 (loss). See Sections 3.6-3.8 of AFML.
  - The goal is not to predict the market. Instead, the goal is to predict the success of the primary model.



- Meta-labeling builds a secondary ML model that learns how to use a primary exogenous model.
- The secondary model does not learn the *side*. It learns only the *size*.
- Meta-labeling is particularly useful when outcomes are asymmetric. In those cases, giving up some recall in exchange for improving the precision can yield a significant improvement in Sharpe ratio.

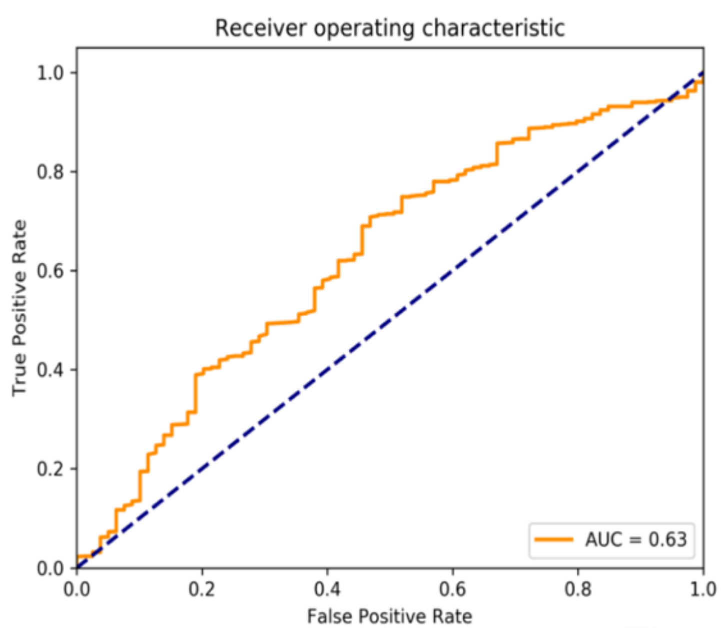
Why Meta-Labeling Works

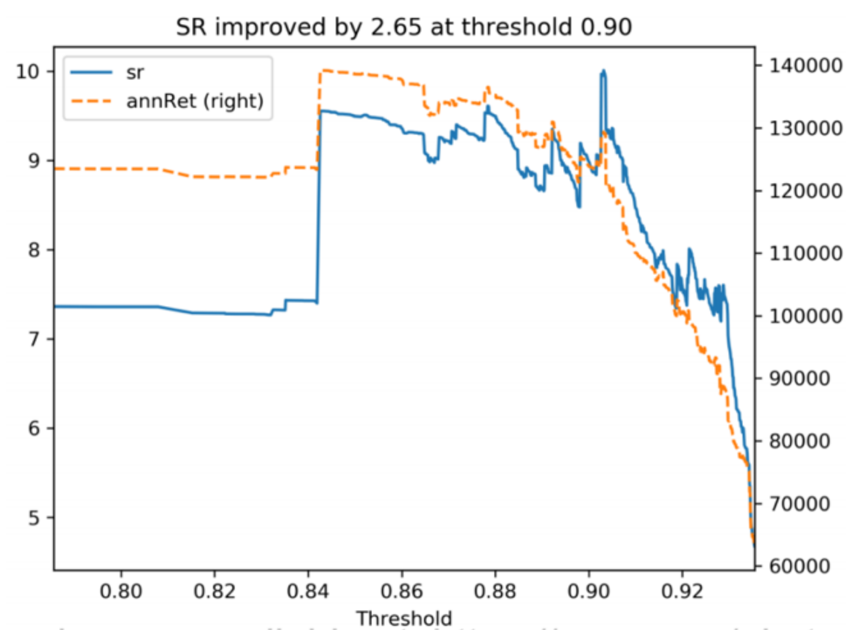
- The Sharpe ratio associated with a binary outcome can be derived as

$$\theta[p, n, \pi_-, \pi_+] = \frac{(\pi_+ - \pi_-)p + \pi_-}{(\pi_+ - \pi_-)\sqrt{p(1-p)}}\sqrt{n}$$

where $\{\pi_-, \pi_+\}$ determine the payoff from negative and positive outcomes, p is the probability of a positive outcome, and n is the number of outcomes per year (see Section 15.3 of AFML).

- When $\pi_+ \gg -\pi_-$, it may be possible to increase $\theta[.]$ by increasing p and the expense of n.





The primary model determines $\{\pi_-, \pi_+\}$, and the secondary model regulates $\{p, n\}$.

In this example, a strategy's Sharpe ratio increased by 2.65 thanks to Meta-Labeling's ability to avoid the largest losses.

How to Use Meta-Labeling

- Meta-labeling is particularly helpful when you want to achieve higher F1-scores:
  - First, we build a model that achieves high recall, even if the precision is not particularly high.
  - Second, we correct for the low precision by applying meta-labeling to the positives identified by the primary model.
- Meta-labeling is a very powerful tool in your arsenal, for three additional reasons:
  - ML algorithms are often criticized as *black boxes*. Meta-labeling allows you to build a ML system on a white box.
  - The effects of *overfitting* are limited when you apply meta-labeling, because ML will not decide the side of your bet, only the size.
  - Achieving high accuracy on small bets and low accuracy in large bets will ruin you. As important as identifying good opportunities is to *size bets* properly, so it makes sense to develop a ML algorithm solely focused on getting that critical decision (sizing) right.
- Meta-labeling should become an essential ML technique for every discretionary hedge fund
  - It allows the seamless combination of discretionary inputs (primary model) with a quantitative overlay (secondary model).