ETFs Returns Modeling:
Binned Probability Curves for 6-Month Forward Returns

Miguel Merlin

February 9, 2026

Abstract

This document presents a simple but powerful framework for understanding how cross-sectional and time-series metrics of Exchange-Traded Funds (ETFs) relate to the probability of achieving a positive 6-month forward return. Instead of directly predicting future returns as a continuous variable, we model a binary indicator I_t which records whether the forward return over the next six months is positive. For a given metric M_t (such as volatility, drawdown, momentum, or Sharpe ratio), we estimate the conditional probability

$$P(I_t = 1 \mid M_t = m)$$

in a robust and interpretable way by discretizing M_t into bins and computing empirical success rates within each bin. We complement these estimates with Wilson score confidence intervals to quantify statistical uncertainty, and with scalar summary statistics—information gain, probability range, and chi-square tests—to rank metrics by their predictive power. The resulting probability curves and statistics provide a transparent way to evaluate which metrics meaningfully shift the odds of a positive 6-month return for ETFs.

1 Introduction

A central question in quantitative portfolio analysis is whether certain characteristics (or "metrics") of an asset today carry information about its future performance. For ETFs, examples of such metrics include trailing volatility, maximum drawdown, past returns over various horizons, and risk-adjusted measures such as the Sharpe ratio.

In this work, we focus on the following question:

> *Given the value of a metric M_t at time t, how does it affect the probability that the ETF realizes a positive 6-month forward return?*

Formally, for an ETF with price process (P_t), we define the forward 6-month return as

$$R_t^{(6m)} = \frac{P_{t+126}}{P_t} - 1,$$

where 126 trading days is used as an approximation to 6 calendar months. We then define a binary outcome

$$I_t = \mathbf{1}\{R_t^{(6m)} > 0\},$$

which takes value 1 if the 6-month return is positive and 0 otherwise.

Rather than predicting $R_t^{(6m)}$ directly, we study the conditional probability

$$P(I_t = 1 \mid M_t = m),$$

where M_t is a metric computed at time t. Directly estimating this quantity as a function of the continuous argument m can be difficult and unstable, especially in finite samples. To address this, we discretize M_t into bins and work with empirical conditional probabilities within those bins.

The rest of this document describes the methodology in detail:

- Section 2 formalizes the data and notation.

- Section 3 explains the binning procedure and construction of empirical probabilities.
- Section 5 introduces Wilson score confidence intervals to quantify uncertainty.
- Section 6 defines information gain and the chi-square test as measures of predictive power.
- Section 7 describes the complete pipeline and illustrates how to rank metrics.
- Section 8 discusses potential extensions and connections to more complex models.

2 Modeling Setup

2.1 Forward 6-Month Return and Indicator Variable

Let P_t denote the adjusted closing price of a given ETF at trading day t. We define the (approximate) 6-month forward return as

$$R_t^{(6m)} = \frac{P_{t+126}}{P_t} - 1, \tag{1}$$

where 126 trading days stands in for six months.

We are not directly modeling $R_t^{(6m)}$ as a continuous variable. Instead, we encode the sign of this return as a binary indicator:

$$I_t = \mathbf{1}\{R_t^{(6m)} > 0\} = \begin{cases} 1, & \text{if } R_t^{(6m)} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This formulation turns the problem into a probabilistic classification task: we are interested in the probability that the forward return is positive.

2.2 Metrics of Interest

Let M_t be a real-valued metric computed at time t from the ETF's historical price process, returns, or other features. Examples include:

- **Rolling volatility**, e.g. 252-day annualized volatility.
- **Maximum drawdown**, e.g. 252-day trailing max drawdown.
- **Momentum**, e.g. 12-month return skipping the most recent month.
- **Sharpe ratio**, e.g. 252-day return divided by volatility.

Our goal is to understand how the value of M_t influences the probability that $I_t = 1$. More precisely, we would like to characterize the function

$$m \mapsto P(I_t = 1 \mid M_t = m). \tag{3}$$

2.3 Why Binning Is Necessary

In finite data, we observe at most one outcome I_t for each observed metric value $M_t = m$. Estimating the full conditional probability function (3) at the level of exact m is impossible without imposing a parametric model (e.g. logistic regression) or a strong smoothing assumption.

Instead, we discretize the real line into a finite set of bins and study

$$P(I_t = 1 \mid M_t \in \mathcal{B}_b),$$

for each bin $\mathcal{B}_b$, $b = 0, \dots, B-1$. This yields a stepwise approximation to (3), which is:

- **Non-parametric**: it does not assume a particular functional form.
- **Data-efficient**: each bin aggregates many observations.
- **Interpretable**: each bin corresponds to a range of metric values.

3 Binning the Metric

3.1 Bin Definitions

We partition the real line into B bins:

$$\mathbb{R} = \bigcup_{b=0}^{B-1} \mathcal{B}_b, \quad \mathcal{B}_b = [a_b, a_{b+1}), \tag{4}$$

where

$$a_0 < a_1 < \cdots < a_B.$$

Two practical choices for the bin boundaries (a_b) are:

- **Quantile bins**: the $\{a_b\}$ are chosen such that each bin contains (approximately) the same number of observations of M_t. For example, $B = 5$ (quintiles) or $B = 10$ (deciles).
- **Uniform-width bins**: the real line (or a relevant range of the metric) is divided into intervals of equal width.

Quantile bins are often preferable because they ensure reasonably balanced sample sizes in each $\mathcal{B}_b$, which is beneficial for computing stable frequency estimates and statistical tests.

3.2 Assigning Observations to Bins

For each observation at time t, we compute the metric M_t and determine the corresponding bin index:

$$\text{bin}_t = b \quad \text{if } M_t \in \mathcal{B}_b. \tag{5}$$

We then group all outcomes I_t by their bin index. For a particular bin b, define:

$$n_b = \#\{t : \text{bin}_t = b\} = \text{number of observations falling in bin } b, \tag{6}$$

$$s_b = \sum_{t:\text{bin}_t = b} I_t = \text{number of positive outcomes } (I_t = 1) \text{ in bin } b. \tag{7}$$

3.3 Empirical Conditional Probabilities

The natural estimator of the conditional probability $P(I_t = 1 \mid M_t \in \mathcal{B}_b)$ is the empirical fraction of positive outcomes in that bin:

$$\hat{p}_b = \frac{s_b}{n_b}. \tag{8}$$

Interpreting $\hat{p}_b$:

- $\hat{p}_b$ is the fraction of times the 6-month forward return was positive among all historical instances where the metric M_t fell into bin $\mathcal{B}_b$.
- If the bins are ordered from low metric values to high metric values, the sequence $(\hat{p}_0, \ldots, \hat{p}_{B-1})$ describes how the probability of a positive 6-month return changes as the metric increases.

4 Base Rate and Conditional Probabilities

4.1 Overall Base Rate

Before conditioning on the metric, it is useful to compute the overall positive rate across all observations. Let

$$\bar{p} = \frac{\sum_t I_t}{\text{total number of observations}}. \tag{9}$$

This quantity can be interpreted as the unconditional probability

$$\bar{p} \approx P(I_t = 1).$$

It serves as a baseline with which we compare the bin-specific conditional probabilities $\hat{p}_b$.

4.2 Interpretation

The comparison between $\hat{p}_b$ and $\bar{p}$ is central:

- If $\hat{p}_b \approx \bar{p}$ for all b, then the metric appears to carry little predictive information about the sign of the 6-month forward return. Regardless of the metric value, the probability of a positive outcome remains near the unconditional base rate.
- If the $\hat{p}_b$ vary substantially with b (e.g. ranging from 0.50 to 0.75), then the metric meaningfully shifts the odds of a positive return. Higher or lower values of the metric may correspond to noticeably different probability regimes.

As an example, suppose we use $B = 5$ equal-count (quintile) bins and obtain the following estimates:

Bin b	$\hat{p}_b$
0 (lowest metric values)	0.50
1	0.55
2	0.60
3	0.65
4 (highest metric values)	0.70

If the overall base rate is, say, $\bar{p} = 0.58$, then we see that low metric values (bin 0) underperform the base rate, whereas high metric values (bin 4) materially exceed it. This pattern suggests that higher values of the metric are associated with a higher probability of a positive 6-month return.

5 Quantifying Uncertainty: Wilson Score Intervals

5.1 Motivation

Each bin-specific estimate $\hat{p}_b$ is based on a finite sample of size n_b. The reliability of $\hat{p}_b$ depends strongly on n_b:

- If n_b is large, we can expect $\hat{p}_b$ to be a precise estimate of $P(I_t = 1 \mid M_t \in \mathcal{B}_b)$.
- If n_b is small, sampling variability is large, and $\hat{p}_b$ may be noisy.

To communicate this uncertainty, we compute a confidence interval for each $\hat{p}_b$. Rather than using a simple normal approximation, we employ the *Wilson score interval* for a binomial proportion, which generally has better coverage properties, especially when sample sizes are moderate and probabilities are not near 0.5.

5.2 Wilson Score Formula

Given $\hat{p}_b = s_b/n_b$ and a desired confidence level (e.g. 95%), let z denote the corresponding z-score from the standard normal distribution (e.g. $z \approx 1.96$ for 95% confidence). The Wilson score interval is defined as follows.

First compute:

$$\text{den}_b = 1 + \frac{z^2}{n_b}, \tag{10}$$

$$\text{center}_b = \frac{\hat{p}_b + \frac{z^2}{2n_b}}{\text{den}_b}, \tag{11}$$

$$\text{margin}_b = \frac{z}{\text{den}_b}\sqrt{\frac{\hat{p}_b(1-\hat{p}_b)}{n_b} + \frac{z^2}{4n_b^2}}. \tag{12}$$

The confidence interval is then

$$[\text{center}_b - \text{margin}_b,\ \text{center}_b + \text{margin}_b] = [\text{ci\_lower}_b,\ \text{ci\_upper}_b]. \tag{13}$$

5.3 Interpretation in Plots

When plotting $\hat{p}_b$ as a function of the bin index b, we also display the interval (13) as vertical error bars or shaded bands. This visualization conveys:

- **Magnitude of uncertainty**: bins with small n_b will have wider intervals, signaling that individual estimates should be interpreted more cautiously.
- **Robustness of differences**: if the intervals for two bins overlap heavily, we should be more cautious in claiming that their probabilities differ. If they are well-separated, the differences are more likely to be statistically meaningful.

6 Measuring Informativeness of Metrics

Beyond visual inspection of probability curves, we would like scalar summary statistics that quantify how informative a metric is about the outcome I_t. We discuss two such measures: information gain (based on Kullback–Leibler divergence) and the chi-square test for independence.

6.1 Information Gain via KL Divergence

6.1.1 Bin Weights

Each bin b contains n_b observations. Define the normalized bin weights

$$w_b = \frac{n_b}{\sum_j n_j}, \tag{14}$$

so that $\sum_b w_b = 1$. These weights reflect the empirical frequency with which the metric falls into each bin.

6.1.2 Definition of Information Gain

We interpret the unconditional base rate $\bar{p}$ as a "baseline" Bernoulli distribution for the outcome I_t. In bin b, the empirical distribution for I_t is a Bernoulli with parameter $\hat{p}_b$. For each bin, we can compute the Kullback–Leibler (KL) divergence between these two Bernoulli distributions.

The *information gain* (IG) of the metric is defined as the expected KL divergence across bins:

$$\mathrm{IG} = \sum_b w_b \left[\hat{p}_b \log\left(\frac{\hat{p}_b}{\bar{p}}\right) + (1-\hat{p}_b)\log\left(\frac{1-\hat{p}_b}{1-\bar{p}}\right) \right]. \tag{15}$$

6.1.3 Interpretation

- If $\hat{p}_b = \bar{p}$ for all bins, then each bin has the same Bernoulli distribution as the unconditional base rate, and $\mathrm{IG} = 0$. In this case, the metric does not change the distribution of outcomes at all.
- Larger values of IG indicate that the bin-wise outcome distributions deviate more strongly from the base rate. Intuitively, the metric then carries more information about I_t: knowing that the metric is in a particular bin significantly changes our belief about the probability of a positive 6-month return.
- IG is measured in *nats* if natural logarithms are used (as above), or in *bits* if logarithms base 2 are used.

6.2 Probability Range

A simpler but highly interpretable measure is the range of the bin-wise probabilities:

$$\mathrm{Range} = \max_b \hat{p}_b - \min_b \hat{p}_b. \tag{16}$$

This quantity captures how much the conditional probability of a positive 6-month return varies as we move from the "worst" metric regime to the "best" one.

- A large range indicates that the metric is capable of sharply differentiating between favorable and unfavorable regimes.
- A small range suggests that, even though the metric may be somewhat informative, the variation in probability is modest.

Range is not a formal statistical test, but it is often useful for ranking and communicating results.

6.3 Chi-Square Test for Independence

6.3.1 Contingency Table

To formally test whether the metric and the outcome are statistically dependent, we build a contingency table over bins and outcomes. For each bin b:

$$\text{Positive count: } s_b, \qquad \text{Negative count: } n_b - s_b.$$

Stacking these across all bins yields a $B \times 2$ table:

$$\begin{bmatrix} s_0 & n_0 - s_0 \\ s_1 & n_1 - s_1 \\ \vdots & \vdots \\ s_{B-1} & n_{B-1} - s_{B-1} \end{bmatrix}.$$

6.3.2 Hypotheses and Test

The chi-square test evaluates:

$$H_0 : \text{Outcome } I_t \text{ is independent of bin index (and hence of the metric)} H_1 : \text{Outcome depends on bin index.}$$

Under H_0, the probability of a positive outcome is the same across all bins and equals the base rate $\bar{p}$. The chi-square statistic measures deviations between the observed counts $(s_b, n_b - s_b)$ and the counts expected under H_0.

A small p-value (e.g. $p < 0.05$) provides evidence against H_0, suggesting that the outcome distribution varies across bins. In other words, the metric is statistically associated with the probability of a positive 6-month return.

7 End-to-End Pipeline

This section summarizes the implemented pipeline in `screener/main.py`, which exposes four CLI workflows that can be run independently or in sequence.

7.1 CLI Workflows

1. **Fetch ETF price data (`--fetch-data`):**
   - Parse ticker symbols from input CSVs (default: `data/*.csv`).
   - Fetch historical price data from Yahoo Finance.
   - Save per-ticker CSVs under `data/` (or `--fetch-output-dir`).
2. **Fetch macro data (`--fetch-macro`):**
   - Fetch FRED series (default: `T10Y2Y`, `CPIAUCSL`, `GS10`, `SP500`).
   - Save macro series under `data/macro` (or `--macro-dir`).
3. **Model ETF returns (`--model-etf-returns`):**
   - Load ETF price histories from `data/etfs` (or `--etf-dir`).
   - Optionally load macro data from `data/macro` if available.

   - Create the forward-return target (default: 6 months; configurable via `--model-target-months`).
   - Compute technical and macro features.
   - Run the selected modeling mode (enumeration, logistic, or stepwise).
   - Save outputs under `results/` (or `--model-results-dir`).

4. **Rank ETFs or predictive metrics (`--rank-etfs`):**
   - Rank ETFs by descriptive performance metrics (default).
   - If `--rank-predictive-metrics` is provided, rank *metrics* by Information Gain.
   - Save consolidated rankings to `etf_rankings.csv` (or `--rankings-output`).

7.2 Modeling Procedure (Enumeration Mode)

When `--model-type enumeration` is selected, the following steps are applied to each candidate metric M_t for the ETF universe:

1. **Compute metric time series:**
   - For each ETF and each eligible time t, compute the metric value M_t using the ETF's historical data.
2. **Compute forward return indicator:**
   - For each time t where forward data is available, compute the 6-month forward return $R_t^{(6m)}$ as in (1).
   - Compute the binary indicator I_t as in (2).
3. **Bin the metric:**
   - Choose the number of bins B (default 5; configurable via `--model-bins`) and a binning scheme (e.g. quantile bins).
   - Determine the bin boundaries $\{a_b\}$ and define $\mathcal{B}_b = [a_b, a_{b+1})$.
   - For each observation t, assign bin_t according to (5).
4. **Compute bin-wise statistics:**
   - For each bin b, compute n_b and s_b via (6)–(7).
   - Compute $\hat{p}_b = s_b/n_b$ as in (8).
   - Compute Wilson score intervals $[\mathrm{ci\_lower}_b, \mathrm{ci\_upper}_b]$ using (10)–(13).
5. **Global metrics:**
   - Compute the base rate $\bar{p}$ using (9).
   - Compute bin weights w_b via (14).
   - Compute information gain (IG) via (15).
   - Compute probability range via (16).
   - Build the contingency table and compute the chi-square statistic and corresponding p-value.

7.3 Visualization and Interpretation

If `--model-type logistic` is selected, the pipeline trains a binary classifier on the full feature set and reports model diagnostics without generating per-metric probability plots. If `--model-type stepwise` is selected, the pipeline performs forward feature selection to maximize ROC-AUC and reports the chosen features and incremental AUC improvements.

For each metric, we produce a plot with:

- x-axis: bin index b, arranged from low to high metric values.
- y-axis: estimated conditional probability $\hat{p}_b$.
- Horizontal dashed line: base rate $\bar{p}$.
- Error bars or shaded bands: Wilson confidence intervals.

From this plot, we can qualitatively assess:

- Whether the relationship between the metric and the probability of positive returns is monotonic, U-shaped, or more complex.
- Whether the range $\max_b \hat{p}_b - \min_b \hat{p}_b$ is economically meaningful.
- Whether the Wilson intervals are tight (indicating robust signal) or wide (indicating noisy estimates).

In addition, summary tables can be constructed to compare multiple metrics side by side, showing, for each metric:

- Information gain (IG).
- Probability range.
- Chi-square statistic and p-value.
- Minimum and maximum $\hat{p}_b$ across bins.

These statistics can be used to rank metrics in terms of how strongly and reliably they shift the odds of a positive 6-month return.

8 Extensions and Future Work

The binned probability framework is intentionally simple and non-parametric, making it easy to interpret and robust to model misspecification. However, it also opens the door to several natural extensions.

8.1 Expanding the ETF Universe

A larger ETF universe increases the total number of observations, which improves:

- **Statistical power:** more data per bin reduces the uncertainty of $\hat{p}_b$ and yields tighter Wilson intervals.
- **Generalizability:** results are less likely to be driven by idiosyncrasies of a small subset of ETFs.

As more ETFs are included, one can also study whether the shape of the probability curves is consistent across sectors, regions, or asset classes.

8.2 Combining Multiple Metrics

The current framework evaluates one metric at a time. In practice, we may want to combine several metrics $\text{metric}_1, \dots, \text{metric}_n$ into a joint model for

$$P(I_t = 1 \mid \text{metric}_1, \dots, \text{metric}_n).$$

There are several possible approaches:

- **Multivariate binning:** define bins in a higher-dimensional space (e.g. 2D bins over pairs of metrics). This becomes challenging as dimensionality grows due to data sparsity.
- **Parametric models:** use logistic regression or generalized additive models (GAMs) where each metric enters as a feature. This imposes structure on the conditional probability while retaining interpretability.
- **Nonlinear models:** employ tree-based methods (random forests, gradient boosting) or neural networks for richer interactions between metrics.

Even when using more complex models, the binned probability plots can serve as a diagnostic tool: we can compare the model-implied probabilities to the empirical $\hat{p}_b$ curves for sanity checks.

8.3 Nonlinear and Machine Learning Models

Beyond the simple binning approach, one can consider:

- **Random forests and gradient boosting:** tree-based models naturally capture nonlinear interactions and can output calibrated class probabilities.
- **Isotonic regression:** when we believe that the relationship between a metric and the probability of success is monotonic, isotonic regression can be used to fit a monotone probability curve that respects this shape constraint.

The binned framework described here is complementary to these models. It provides a clear, model-agnostic picture of how the empirical probabilities behave, which can then guide the choice and evaluation of more sophisticated models.

9 Parametric Estimation via Logistic Regression

The binned probability framework provides a non-parametric and highly interpretable approximation of the conditional probability function

$$P(I_t = 1 \mid M_t = m).$$

However, binning introduces discretization error and can obscure fine-grained structure in the relationship between the metric and future returns. A natural parametric alternative is to model this conditional probability directly using **logistic regression**.

9.1 Single-Metric Logistic Model

For a single metric M_t, the logistic regression model assumes:

$$P(I_t = 1 \mid M_t) = \sigma(\beta_0 + \beta_1 M_t), \tag{17}$$

where:

- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic (sigmoid) function,
- β_0 is the intercept,
- β_1 is the slope coefficient controlling how strongly the metric affects the log-odds of a positive return.

This model implies a smooth, monotonic probability curve in M_t, in contrast to the stepwise curve produced by binning.

9.2 Interpretation of Coefficients

Taking log-odds, we obtain:

$$\log\left(\frac{P(I_t = 1 \mid M_t)}{1 - P(I_t = 1 \mid M_t)}\right) = \beta_0 + \beta_1 M_t.$$

Thus:

- The sign of β_1 determines whether the metric increases or decreases the probability of a positive return.
- The magnitude $|\beta_1|$ controls how sensitive the probability is to changes in M_t.
- A large positive β_1 implies that higher metric values rapidly push the probability toward 1.

9.3 Relationship to Binning

The binned estimates $\hat{p}_b$ can be viewed as a piecewise-constant, non-parametric approximation of the same underlying conditional probability that logistic regression attempts to model as a smooth function. In practice:

- The binned probabilities provide a diagnostic check on whether the logistic functional form is reasonable.
- The logistic model can interpolate between bins and extrapolate beyond observed quantiles.
- When the empirical probability curve is roughly monotonic, logistic regression often provides a compact and stable summary.

9.4 Multivariate Logistic Regression

When multiple metrics are available, we can model the joint conditional probability as:

$$P(I_t = 1 \mid \boldsymbol{M}_t) = \sigma\left(\beta_0 + \sum_{k=1}^{d} \beta_k M_{k,t}\right), \tag{18}$$

where $\boldsymbol{M}_t = (M_{1,t}, \ldots, M_{d,t})$ is the vector of metrics at time t.

This formulation enables:

- Joint conditioning on multiple signals,
- Control for correlations between metrics,
- Direct modeling of combined predictive structure.

However, once we move to multivariate models, interpretability and metric ranking become significantly more subtle—this motivates the next section.

9.5 Backtesting and Portfolio Construction

Ultimately, the goal of modeling $P(I_t = 1 \mid M_t)$ is to inform portfolio decisions. Once a set of metrics has been identified as strongly predictive (e.g. high IG, large probability range, significant chi-square), one can:

- Define trading rules that favor ETFs in bins with high $\hat{p}_b$ and avoid those in bins with low $\hat{p}_b$.
- Backtest these rules over historical data to evaluate out-of-sample performance.
- Combine probability-based signals with other considerations such as risk constraints, liquidity, and transaction costs.

The binned probability estimates thus act as a bridge between statistical analysis and actionable trading strategies.

10 Open Questions: How Should Metrics Be Ranked?

The binned framework provides several natural scalar quantities for ranking metrics, including:

- information gain (IG),
- the probability range $\max_b \hat{p}_b - \min_b \hat{p}_b$,
- chi-square test statistics.

These quantities work well in the discrete, bin-based setting. However, once we move to continuous probability models—such as logistic regression or more general machine learning estimators—a fundamental open question emerges:

> **Even if we have a highly accurate estimate of the full conditional probability $P(I_t = 1 \mid M_t)$, how should we formally rank metrics by "predictiveness"?**

10.1 Ambiguity in What "Best" Means

Several competing notions of what it means for a metric to be "better" arise:

- **Statistical strength:** how strongly does the metric affect the outcome in a probabilistic sense?
- **Economic significance:** how large are the induced changes in return probability?
- **Stability:** does the relationship persist across time, regimes, and ETF sub-universes?
- **Incremental value:** does the metric add information beyond what is already captured by others?

A single scalar ranking cannot generally capture all of these dimensions simultaneously.

10.2 Possible Ranking Criteria Under Continuous Models

Suppose we use a smooth model (e.g. logistic regression) to estimate the conditional probability function

$$p(m) = P(I_t = 1 \mid M_t = m).$$

Several candidate ranking criteria then become available.

10.2.1 (1) Expected KL Divergence from the Base Rate

A continuous analogue of the binned information gain is

$$\mathrm{IG}_{\mathrm{cont}} = \mathbb{E}_{M_t}\Big[p(M_t)\log\frac{p(M_t)}{\bar{p}} + \big(1 - p(M_t)\big)\log\frac{1 - p(M_t)}{1 - \bar{p}}\Big], \tag{19}$$

where $\bar{p}$ is the unconditional base rate. This measures how much, on average, conditioning on M_t changes the distribution of outcomes relative to the base rate.

10.2.2 (2) Slope Magnitude in Logistic Regression

In the single-metric logistic model

$$P(I_t = 1 \mid M_t) = \sigma(\beta_0 + \beta_1 M_t),$$

one simple ranking is by the absolute value $|\beta_1|$, which measures the change in log-odds per unit change in the metric. However:

- this ranking is scale-dependent (rescaling M_t changes $|\beta_1|$),
- it depends on how the metric is normalized,
- it does not directly measure probability separation.

10.2.3 (3) Maximum Probability Separation

A continuous analogue of the probability range is

$$\Delta p_{\max} = \sup_m p(m) - \inf_m p(m),$$

which captures how far apart the best and worst regimes are under the model.

10.2.4 (4) Out-of-Sample Predictive Power

Metrics can also be ranked by their contribution to out-of-sample predictive performance, for example via:

- improvement in log-likelihood,
- area under the ROC curve (AUC),
- reduction in Brier score,
- cross-validated information gain.

These criteria shift the focus from descriptive probability modeling to predictive performance.

10.3 Causal vs. Predictive Ranking

An even deeper unresolved issue is whether metrics should be ranked by:

- their **predictive** strength (associational),
- or their **causal** influence on future returns.

The framework developed in this document is fundamentally associational:

$$P(I_t = 1 \mid M_t = m)$$

does not identify causal effects without additional structural assumptions. Two metrics may have similar probability curves but vastly different causal interpretations.

10.4 Key Open Problem

The central unresolved research question is therefore:

> **What is a principled, model-agnostic way to rank financial metrics once we move beyond binned empirical probabilities and into continuous probabilistic estimators?**

Possible directions for future work include:

- information-theoretic rankings valid under arbitrary conditional probability estimators,
- stability-based rankings using rolling-window estimation,
- causal ranking criteria based on structural return models,
- portfolio-level utility-based rankings that connect probability curves to realized performance.

At present, no single ranking method is universally dominant, and the “correct” ranking depends fundamentally on whether the goal is explanation, prediction, or portfolio construction.