From Fog to Foresight: Predicting Mutual Fund Flows to Drive Proactive Engagement

A machine learning approach to identify, rank, and understand high-value investor behavior

RBC

The Fund Manager's Dilemma: Navigating a Sea of Uncertainty

Fund managers must optimize performance based on predicted cash flow. However, engaging with investors is often inefficient. The core challenges are:

- **Knowing Who:** Identifying which investors are most likely to subscribe or redeem in the near future.
- **Knowing When:** Anticipating the timing of these flows to manage liquidity and strategy effectively.

> Mutual fund flows can be influenced by a complex interplay of as much as 40 different factors, from macro-economics to individual investor psychology.

The High Cost of Inefficient Engagement

The traditional approach of broad-based investor contact is not just inefficient, it's counterproductive.

High Financial Cost

Resources are wasted engaging with investors who have no intention of transacting. The approach is described as "rather blindly" contacting clients.

Negative Client Impact

Frequent engagement with clients that are not considering any subscription/redemption in the near future may cause a negative impact on their relation with the fund.

Missed Opportunities

The most critical investors—those with large, irregular transaction potential—are often overlooked in favor of more predictable, lower-impact clients.

Our Approach: A Data-to-Insight Pipeline

Historical Data

2 years of transaction data from over 800,000 investors

Feature Engineering

Translating raw data into behavioral patterns and investor profiles.

Predictive Modeling

Training XGBoost models to predict next-month redemptions.

Actionable Insights

Generating ranked investor lists and explaining the “why” behind predictions

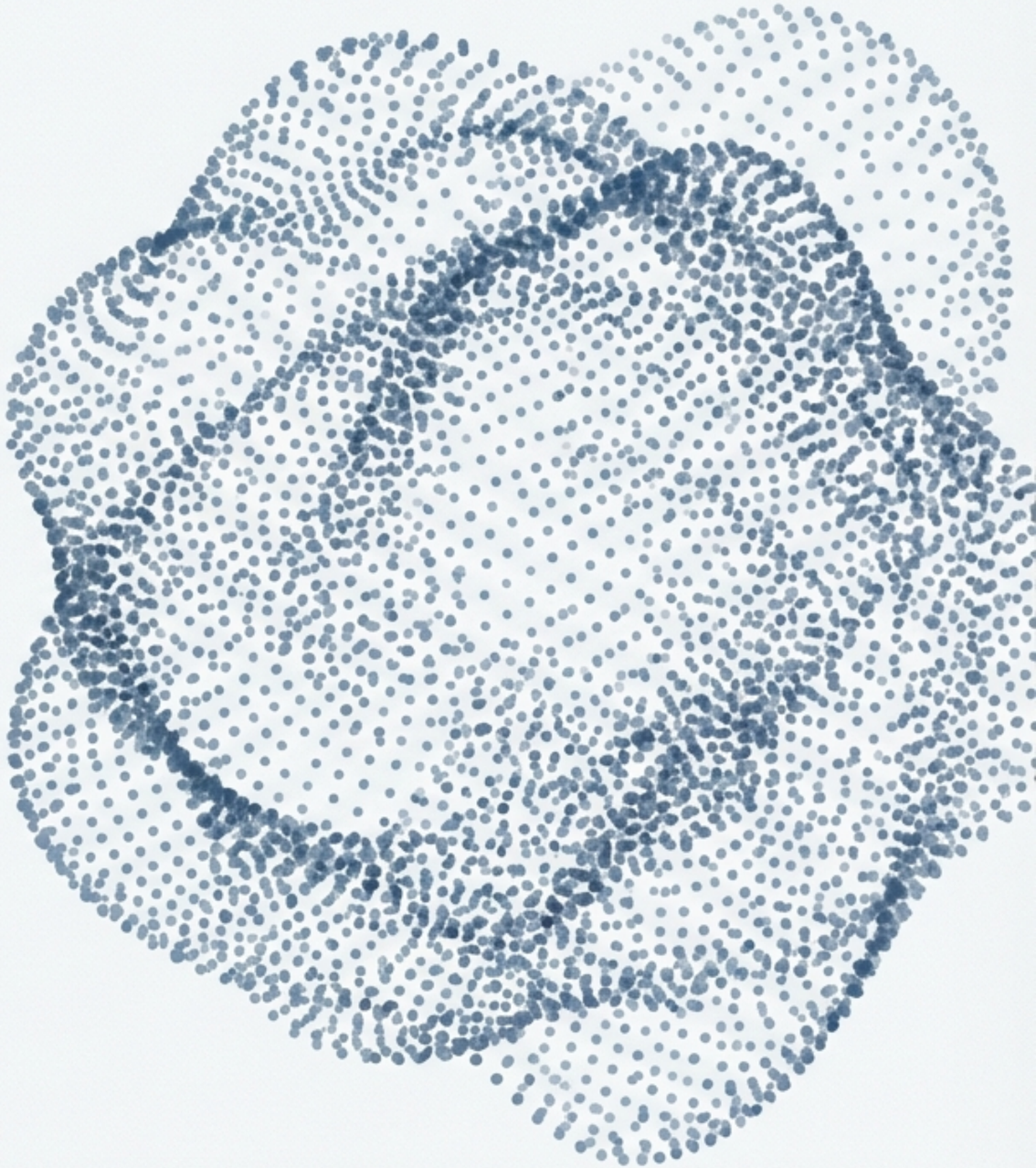
The Foundation: A Rich History of Investor Transactions

Data Scope:
2 years (2016-2017) of historical transaction records.

Investor Base:
> 800,000 active investors (individuals and corporations).

Fund Coverage:
~ 400 mutual funds across different asset classes.

Core Task
To accurately predict the next month's redemption or subscription for any given investor based on their historical patterns.



Translating Transaction History into Behavioral Signals

We engineered a set of 59 features for each investor to capture a holistic view of their profile and investment patterns.

Investor Profile

High-level information such as age, province, investor type, and account type.

Recent Behavior (12-Month Sliding Window)

Statistical measures of recent activity, including average, min/max, and median net flow, plus subscription/redemption counts.

Long-Term Patterns

The same statistical measures calculated since the inception of the investor's holdings to capture their baseline behavior.

Seasonality

Monthly features (M1-M12) to account for time-of-year effects.

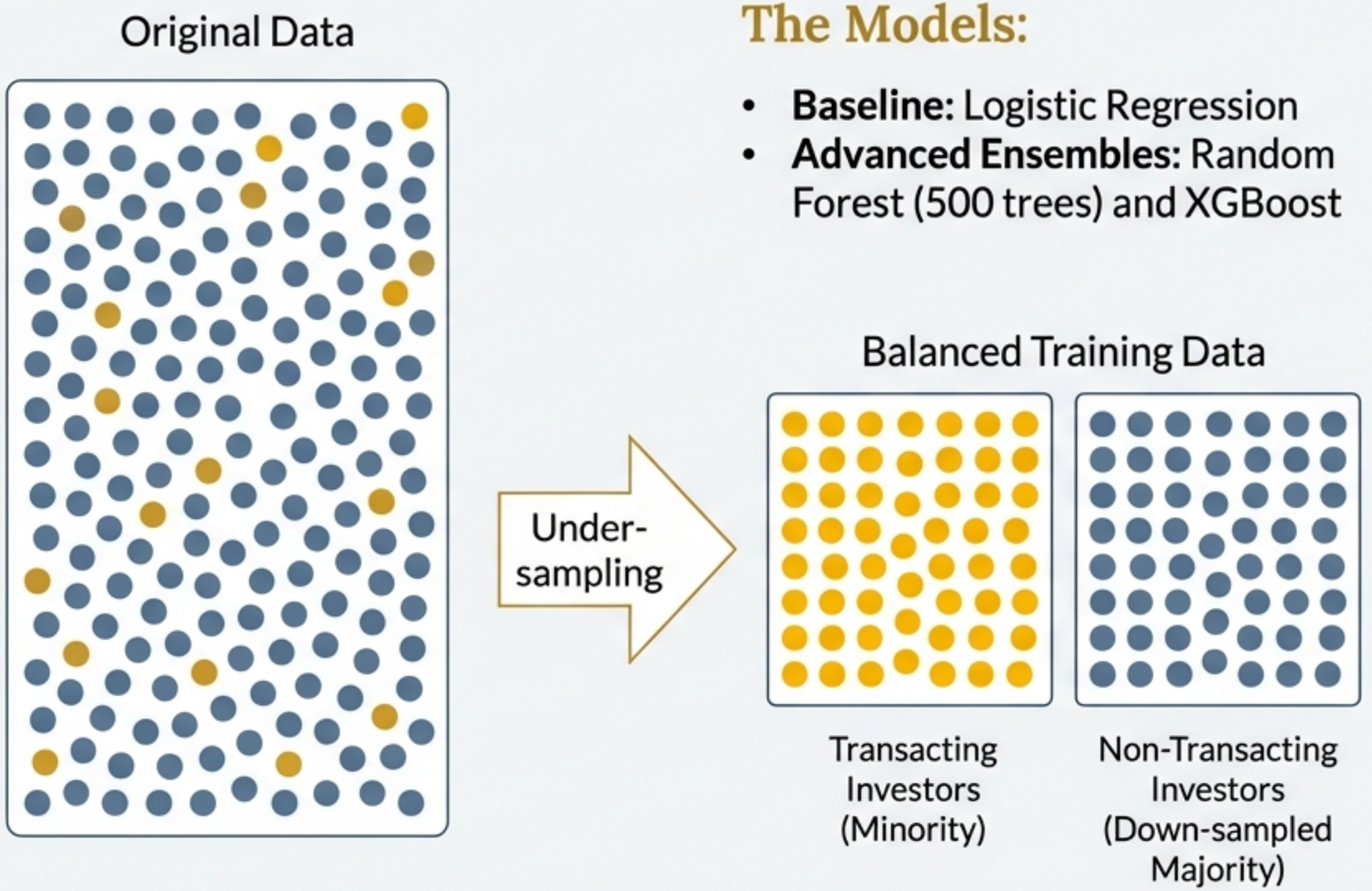
Selecting the Right Tools to Find the Signal

The Challenge:

A major hurdle is data imbalance. On average, only 10% of investors transact in any given month, creating a significant “needle in a haystack” problem for the model.

The Solution:

We employed an under-sampling technique to create a balanced dataset for training, ensuring the model could effectively learn the patterns of the minority class (transacting investors).

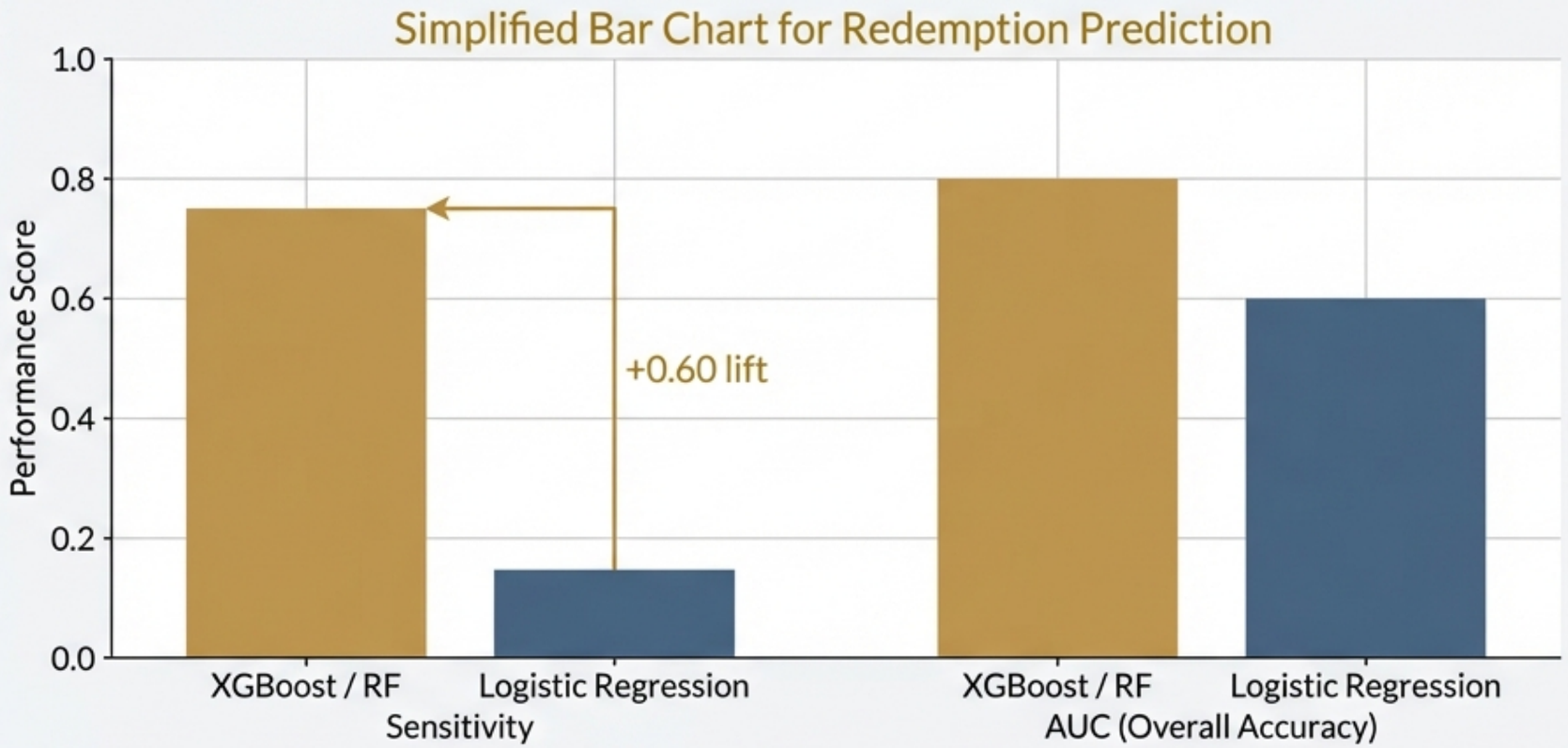


The Models:

- **Baseline:** Logistic Regression
- **Advanced Ensembles:** Random Forest (500 trees) and XGBoost

We Can Predict Future Behavior with High Confidence

Both XGBoost and Random Forest models are highly effective, significantly outperforming the baseline. The most critical improvement is in *Sensitivity*—the ability to correctly identify investors who will actually redeem.



Higher sensitivity means we miss fewer critical redemption events, allowing for more comprehensive and proactive engagement.

Our Predictions Capture the Vast Majority of At-Risk Capital

77%

Our model correctly identifies the investors who account for **76.95%** of the total redemption dollar value across all funds.

This level of coverage means engagement efforts can be focused on a predictable subset of investors while still addressing the most significant portion of potential outflows.

But Not All Predictions Are Created Equal

The Problem

Ranking investors solely by their prediction probability is a flawed strategy. This approach tends to surface investors with highly regular, predictable, and often small, monthly transactions.

The Missed Opportunity

It fails to capture investors with irregular patterns who often represent the largest and most impactful redemptions or subscriptions.

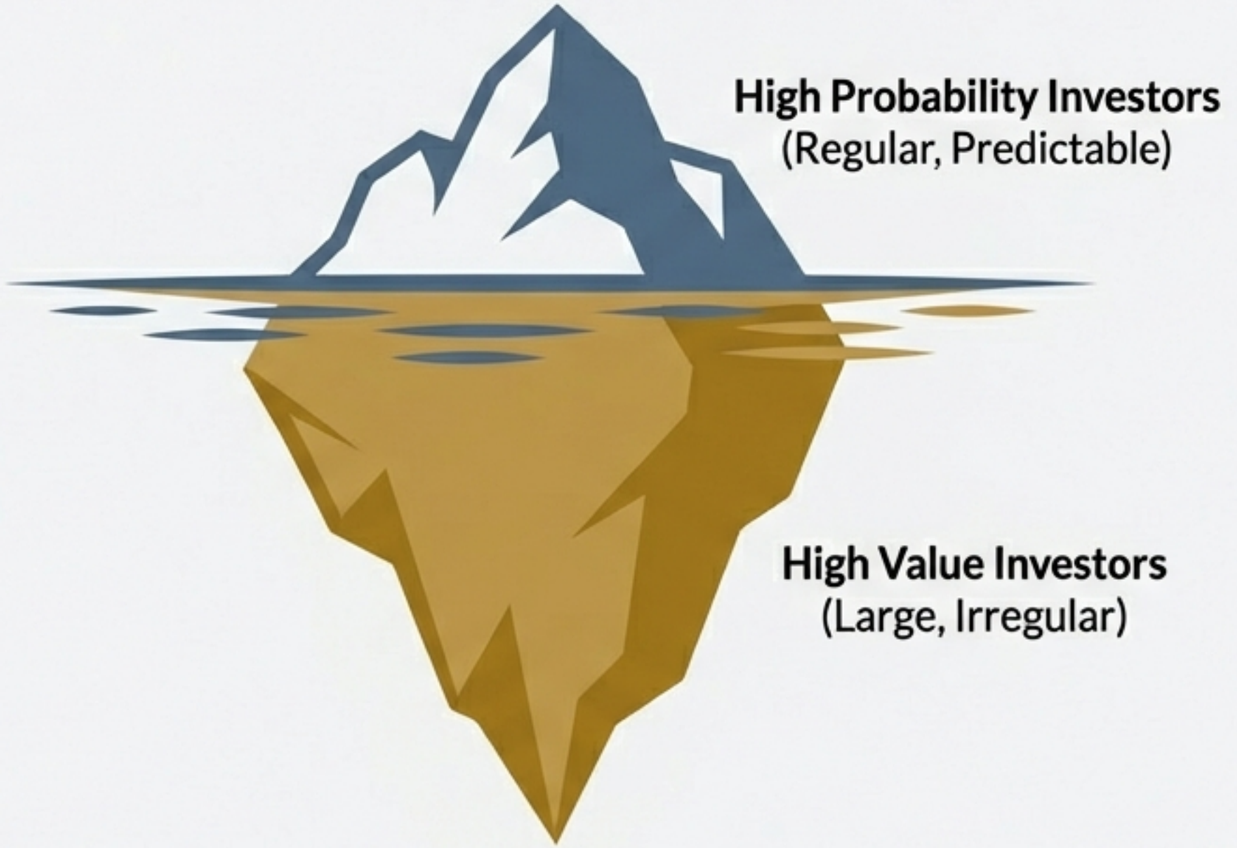
For example, our model correctly predicted large, irregular redemptions for investors like #13 and #14, who would have been missed by a simple probability ranking. We need a way to bring these high-value investors to the top of the list.

Optimal Ranking: A Smarter Way to Prioritize Engagement

The Approach

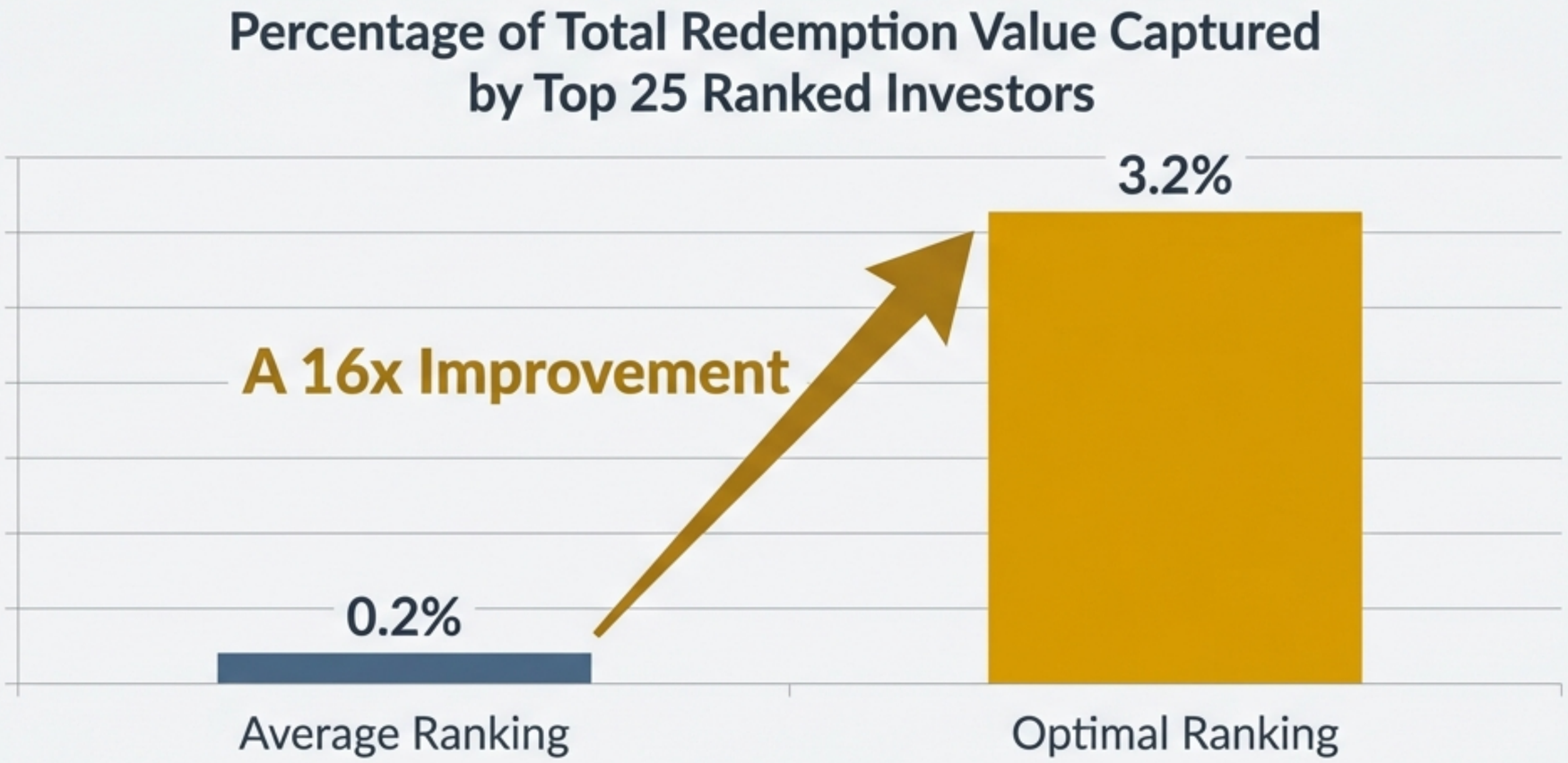
We developed a multi-factor ranking system that looks beyond simple probability. It combines four key features to create a holistic "importance" score:

- 1. Prediction Probability
- 2. Max Historical Redemption (min)
- 3. Average Net Flow (avg)
- 4. Likely Redemption Amount (min × probability)



Optimal Ranking allows us to see the whole picture, revealing the high-value investors hidden beneath the surface.

The Result: 16x More Effective Targeting



This dramatic increase in captured value is achieved with a nearly identical false positive rate (33.7% vs 32.6%), meaning the improved targeting comes with no loss in efficiency.

Moving from “What” to “Why”: Understanding the Drivers of Redemption

The Challenge

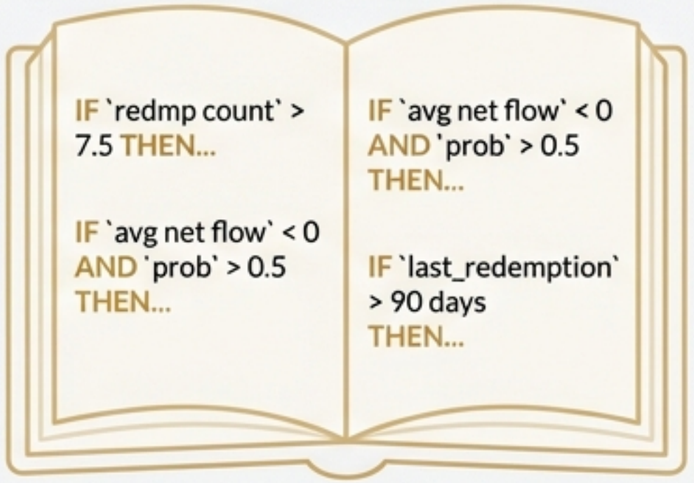
Advanced models like XGBoost are powerful but their decision-making process can be opaque. To make the insights truly actionable, we need to understand the *reasons* behind a prediction.

Black Box Model

Our Solution

We used a rule extraction method (RF+MSGL) to deconstruct the model’s logic.

Extracted Business Rules



This process generated 47 simple, high-fidelity rules that mimic the performance of the complex model (79% accuracy) but are easily understood by humans.

The Key Drivers of Investor Decisions

The number of recent redemptions (`redmp count`) is the single most influential factor in the model's decisions.

High-Risk Pattern #1

If an investor has made more than 7 redemptions in the past 12 months, they are highly likely to redeem again.

(Based on R16)

High-Risk Pattern #2

If an investor has a history of large single redemptions (‘*min* <= -1,530’) AND they redeem frequently (*redmp count* > *3.5*’), the probability of another redemption is very high.

(Based on R19)

Low-Risk Pattern

Conversely, if an investor has made fewer than 2 redemptions in the past 12 months, they are very unlikely to redeem.

(Based on R2)

A New Strategic Capability for Asset Management

This work provides an end-to-end system that transforms how we can manage fund flows and investor relationships.

PREDICT

Accurately forecast near-term subscriptions and redemptions at the individual investor level, capturing 77% of redemption value.

PRIORITIZE

Use an optimal ranking system to focus engagement on the highest-value investors, improving targeting effectiveness by 16x.

UNDERSTAND

Move beyond the ‘what’ to the ‘why’ by extracting clear, actionable rules that explain the key drivers of investor behavior.