

# **Reverse-Engineering ETF Portfolio Construction**

**Inference by Logistic Regression**

**James Kidd**

# Purpose

- Demonstrate investment-analytics approach
- Explore ETF holdings construction
- Test a finance-style inference with simple ML
- Highlight limitations

# Starting Point

- **Original Idea**      Use Bloomberg Terminal to analyze CI GAM funds  
Look for inference in PM behaviour
- **Constraint**      Academic license blocks holdings export  
Could not retrieve CI GAM PM portfolios
- **Solution**      Pivoted to iShares public holdings

# Data Sources

- iShares Screener API (full holdings)
- Yahoo Finance (returns, vol, Sharpe)
- Python ETL + ML
- **Important:** No security fundamentals (time/API limits)

# Data Engineering Workflow

- Selected 30 U.S. equity ETFs
- Automated holdings downloads
- Cleaned ticker/sector/region/weight/value
- Added ETF performance metrics
- Final dataset  $\approx$  1,600 rows

# Modeling Question

- Predict if a security appears in top-25 ETF holdings
- Only ETF-level metrics + sector used
- **Goal:** detect systematic preferences

$$P(S_i \in Top_k - holdings \mid E)$$

or

$$P(\text{rank}(S_i) \leq k \mid \text{ETF data}_i)$$

# Model Setup

- **Model:** Logistic Regression (interpretable)
- **Features:** ETF return, vol, Sharpe, sector
- **Pipeline:** imputation, scaling, OHE, class balancing
- **Output:** top-25 indicator

Is security  $S_i$  inside this ETF's top-25 holdings?

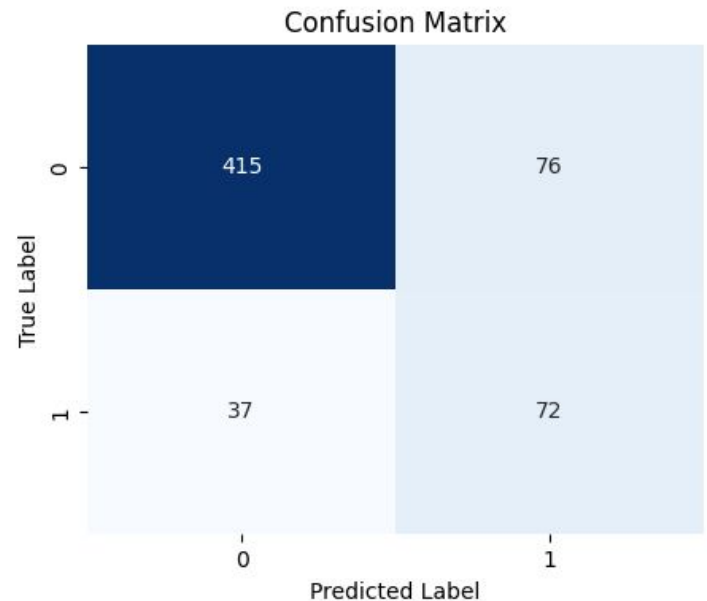


model determines

$Y_i \in \{\text{Yes}, \text{No}\}$

# Results

- Weak model performance (expected)
  - **ROC-AUS:** 0.83
  - **Log Loss:** 0.51
  - **Accuracy:** 0.81
- Interpretation
  - High-growth or low-vol securities match certain ETF styles
  - Security performance explains placement within the ETF
- **Main Roadblock:** missing fundamentals



# Missed / Inaccurate Opportunities

- Missing fundamentals (market cap, beta, valuations)
- No factor exposures (value/growth, size, quality)
- No PM-level information (mandate, constraints, benchmark)
- **Small Dataset:** limited generalization

# Why This Matters

- Structured reasoning with real investment data
- Experience with data pipelines & Bloomberg
- Ability to turn technical analysis into client narratives
- Reflects how I'd support PMs, and advisory team