

Weighted Statistical Methods for Novel Feature Selection:

Improving Binary Classification in RWE Studies Through Prevalence-Aware Metrics

Presenters:

Ashwin Anand, MPH

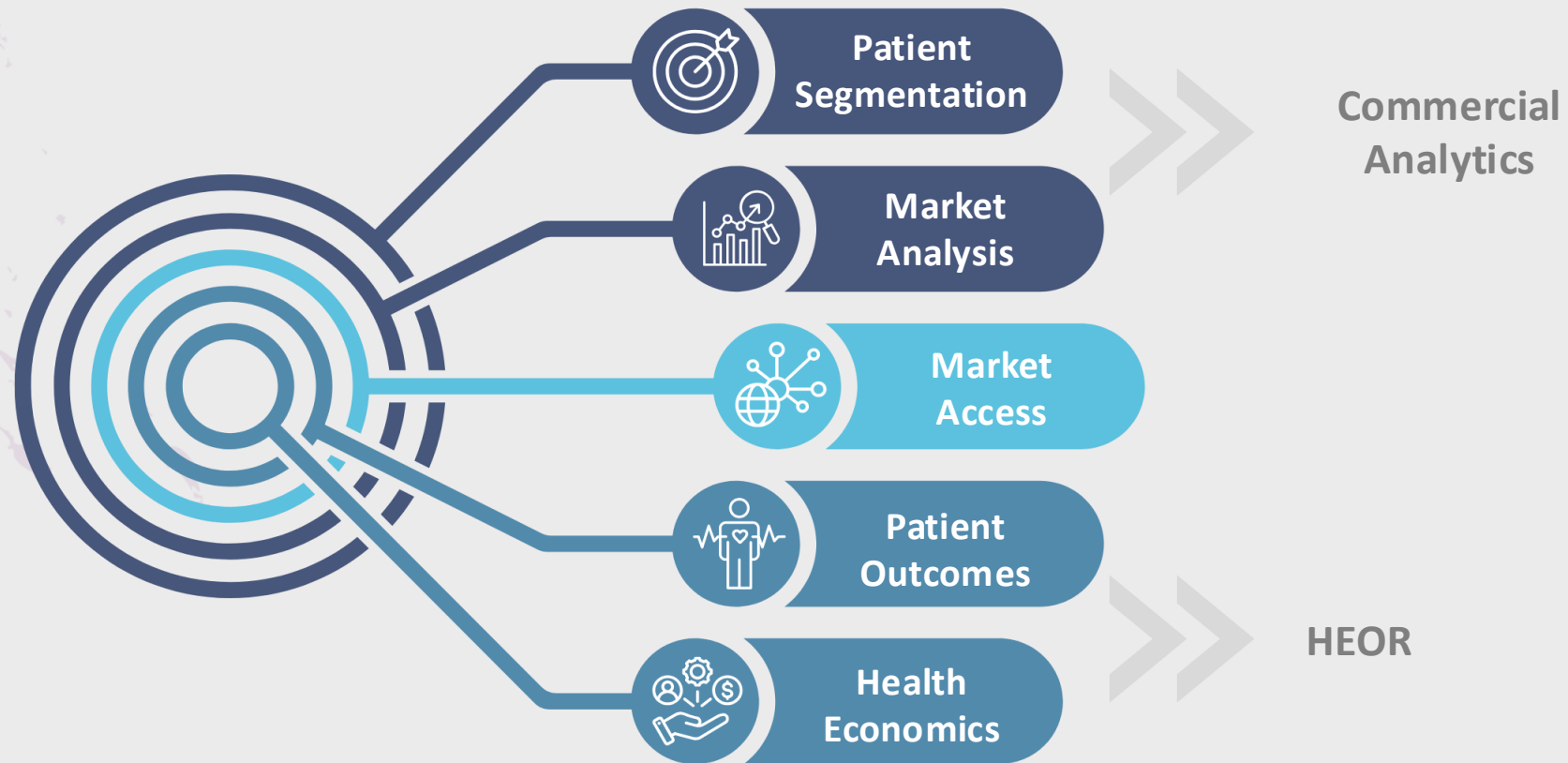
Mike Sicilia

Agenda / Roadmap

1. The Feature Selection Problem in Claims Data
2. Limitations of Traditional Methods
3. Uni-Directional Metrics
4. Bi-Directional Metrics
5. Case Study: Suspected MASH Identification

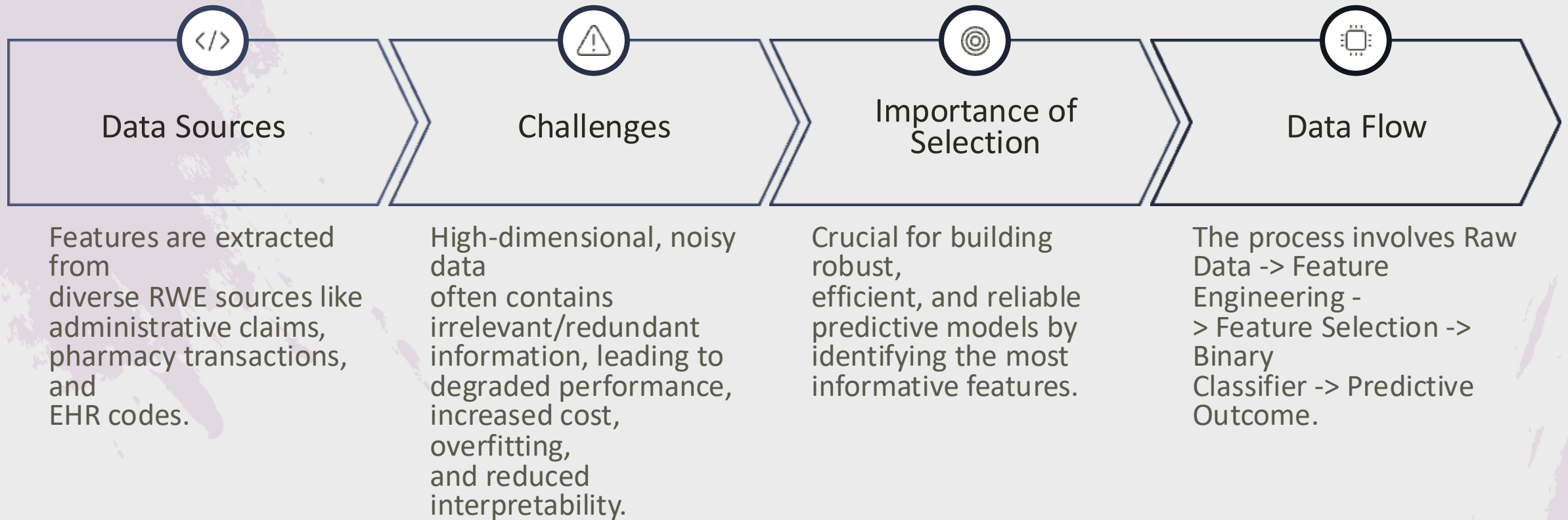
Context: Real-World Evidence (RWE) & Commercial Analytics

Feature Selection methodology can be applied to any patient cohort level analyses



The Feature Selection Problem in RWE

Correct feature selection is vital for building effective predictive models using Real-World Evidence.



The Feature Selection Problem in RWE

Needle in a Haystack

- Traditional statistical methods of cohort comparison, such as T/C Prevalence Ratios, and Mutual Information, are prone to surfacing "needle in a haystack" features that have low ubiquity in the positive class, making them potentially ungeneralizable.

Sample Size Sensitivity

- P-values are too sensitive to large sample sizes, leading to widespread 'significance' that may not be practically meaningful.

Uni-Directional Metrics: Weighted Prevalence, Mutual Information, and P-Value

Formulas:

Weighted Target to Control Prevalence Ratio (WTCP) = $(\text{Target Prevalence})^2 / \text{Control Prevalence}$

Weighted P Ratio = WTCP $\times$ P-Value (FDR-Adjusted)

Weighted MI Ratio = WTCP $\times$ MI-Value

Purpose:

- By weighting the target to control ratio by the prevalence in the target class (square function), we help to control the “needle in a haystack” problem of a traditional ratio. This also helps us to mitigate for the same issue with mutual information, and reduce sample size sensitivity in P-Values. Effectively, the more unique to the target class the feature, the higher the weight.

What Uni-Directional Metrics Find

Identifying Unique Identifiers of a Target Class

1

High Target Prevalence
The target class occurs frequently.

2

High Target-to-Control Ratio
Features are disproportionately present in the target class compared to the control class.

3

Statistical Significance or Information Gain
The relationship between the feature and the target class is statistically robust.

Bi-Directional Metrics — Definitions

- Maximum Ratio:

$\text{Max} (\text{Target Prev.} / \text{Control Prev.}, \text{Control Prev.} / \text{Target Prev.})$

- Maximum Prevalence:

$\text{Max} (\text{Target Prev.}, \text{Control Prev.})$

Weighted ABS MI-Ratio:

- Calculated as 'Max Ratio * Max Prevalence * MI'.

Weighted ABS P-Ratio:

- Calculated as 'Max Ratio * Max Prevalence * P-Value (FDR-Adj.)'.

What Bi-Directional Metrics Find



Maximize Separation

Identify features that maximally separate the two classes.



Dual Indication

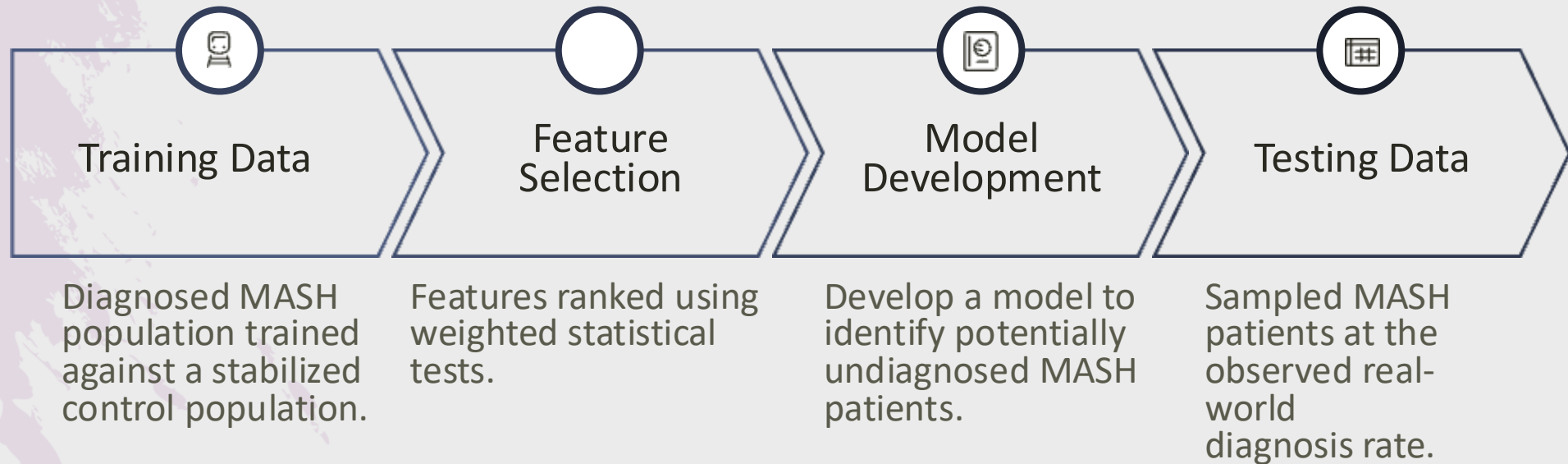
Capture features that strongly indicate the target class and strongly indicate the negative class.



Holistic Feature Set

Complement uni-directional metrics for a more comprehensive understanding and selection of features.

Case Study: Identifying Undiagnosed MASH Patients



Case Study: Identifying Undiagnosed MASH Patients

End Goal: Estimate the prevalence of MASH patients, both diagnosed and undiagnosed, in the United States year over year from 2021-2024

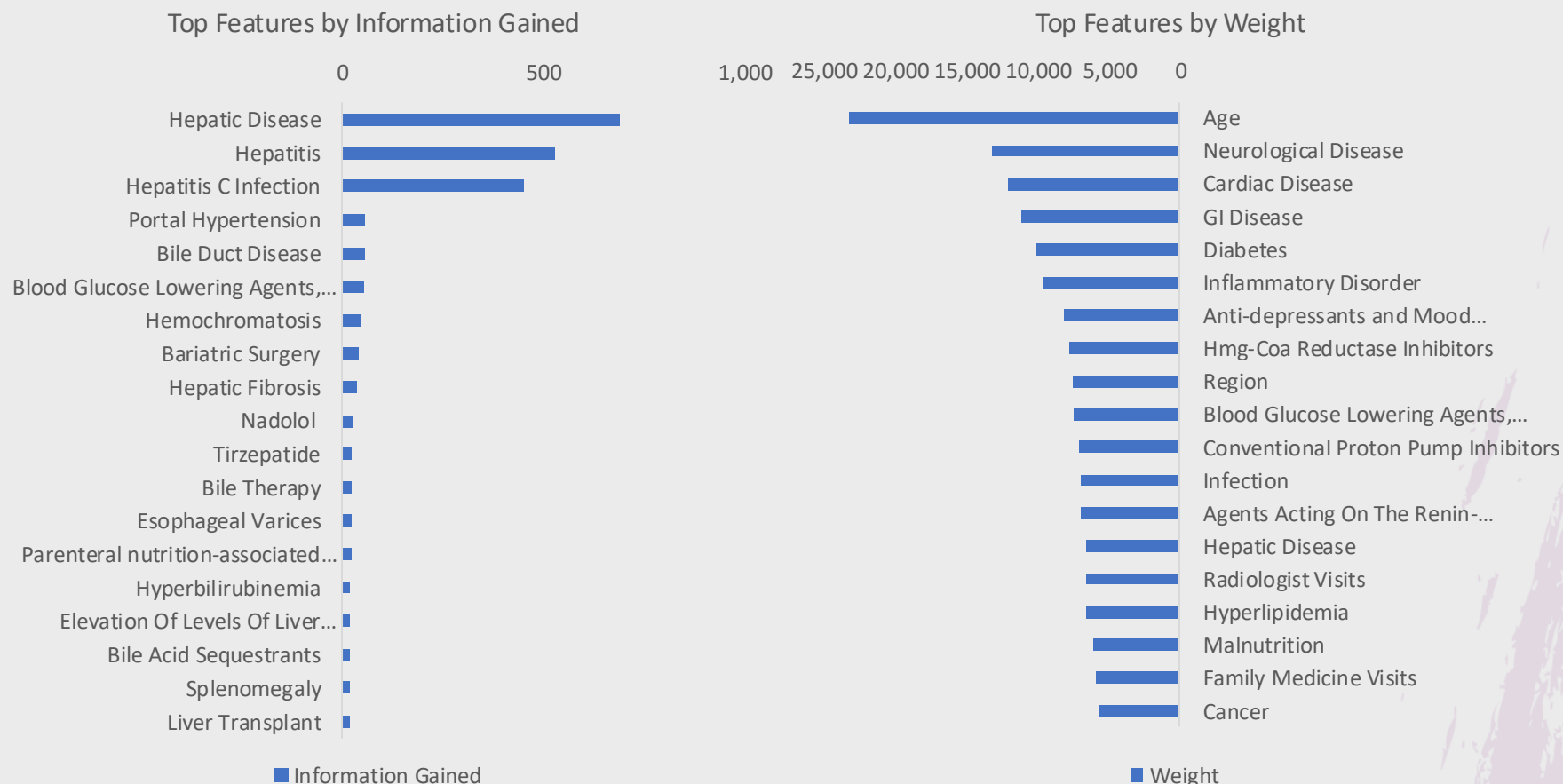
The Problems:

1. A large majority of MASH patients are under/undiagnosed in the United States, due to many factors, including:
 - Asymptomatic Nature in early stages
 - Changing Diagnostic criteria over time (NAFLD/NASH)
 - Historic lack of non-invasive testing (Liver Biopsy was the only option)
 - Historic low-emphasis on the association of liver with adverse outcomes
2. However, in more recent years, health experts have increasingly called for more intense screening. The Fib-4 blood test can help providers analyze signs of liver fibrosis, allowing more patients to be diagnosed earlier. Therefore, the patient profile of MASH patients (in claims) is shifting relatively rapidly year over year as patients are diagnosed in earlier stages.

MASH Testing Data: Performance Results

The model successfully identified 89% of MASH patients (55% precision) in the held-out test set, with key features like Hepatitis C, Age, and specific disease categories driving the predictions.

Dataset	MASH Patient Years	Control Patient Years	Notes
Training Data	486K	481K	Used to teach the model patterns within MASH patients claim data
Testing Data	186K	4.5M	MASH patients sampled to mimic real world diagnosed MASH rate



MASH Testing Data: Performance Results

The MASH-impacted population grew nearly 2.5× from 2022 to 2024 — driven by the expansion of the “suspected” population found by the model

	2022	2023	2024
Diagnosed MASH	1.3 Million	1.6 Million	1.9 Million
"Suspected" MASH	8 Million	14 Million	21 Million
Total	9.3 Million	15.6 Million	22.8 Million

Epidemiologic Context

- Strong correlation between **rising metabolic comorbidities** (diabetes, obesity) and MASH prevalence
- **GLP-1 and metabolic drug uptake** may be shifting identification patterns. More patients with metabolic disease being evaluated for MASH
- **Demographic aging** and **increased longevity** contribute to higher fibrosis progression and detection rates

MASH is emerging as a significant national health burden

- Nearly **1 in 10 U.S. adults** may now fall within the modeled MASH risk spectrum
- Model projections suggest over **22.8 million** U.S. patients may be affected in 2024, underscoring the urgency for targeted screening and intervention strategies
- Model more accurately represents the prevalence of MASH, with model growth rates flattening over time

Reducing Reliance on Manual KOL Feedback

Traditional Approach

- Heavy reliance on manual Key Opinion Leader (KOL) input.
- Identification of non-obvious features through subjective feedback or previous literature.
- Significant manual lift and time investment.
- Potentially slower model development cycles.

Automated Discovery

- Leverages weighted metrics to automate feature discovery, allowing KOLs to review higher-quality, more concise results.
- Significantly **reduces** manual effort.
- Automated discovery with weighted metrics streamlines feature identification and accelerates model development.

Summary & Key Takeaways



Limitations of Standard Metrics

MI and p-values struggle with claims-data scale.



Novel Approach

Four weighted metrics developed, incorporating prevalence ratios to address identified gaps.



Metric Types

Uni-directional metrics surface target-specific predictors, while bi-directional metrics surface class-separating predictors.



Validation

Successfully applied in a MASH identification model, achieving 89% recall.



Benefits

Weighted methods reduce KOL dependency and accelerate feature selection.

Questions?
Contact: Info@Forian.com