

Data Storytelling: Problematic Internet Use in Children

GROUP 1

Vũ Bùi Đình Tùng

Bùi Thị Lan Anh

Thiều Diệu Thuý

Đoàn Tùng Lâm

Trương Đức Anh

BIG IDEA

“In this project, the main character is not the ML model, but the data preparation.”

We will see how step-by-step data preparation – from raw, imbalanced, noisy data – changes the quality of models and insights for decision makers.

Before talking about models, let's talk about data

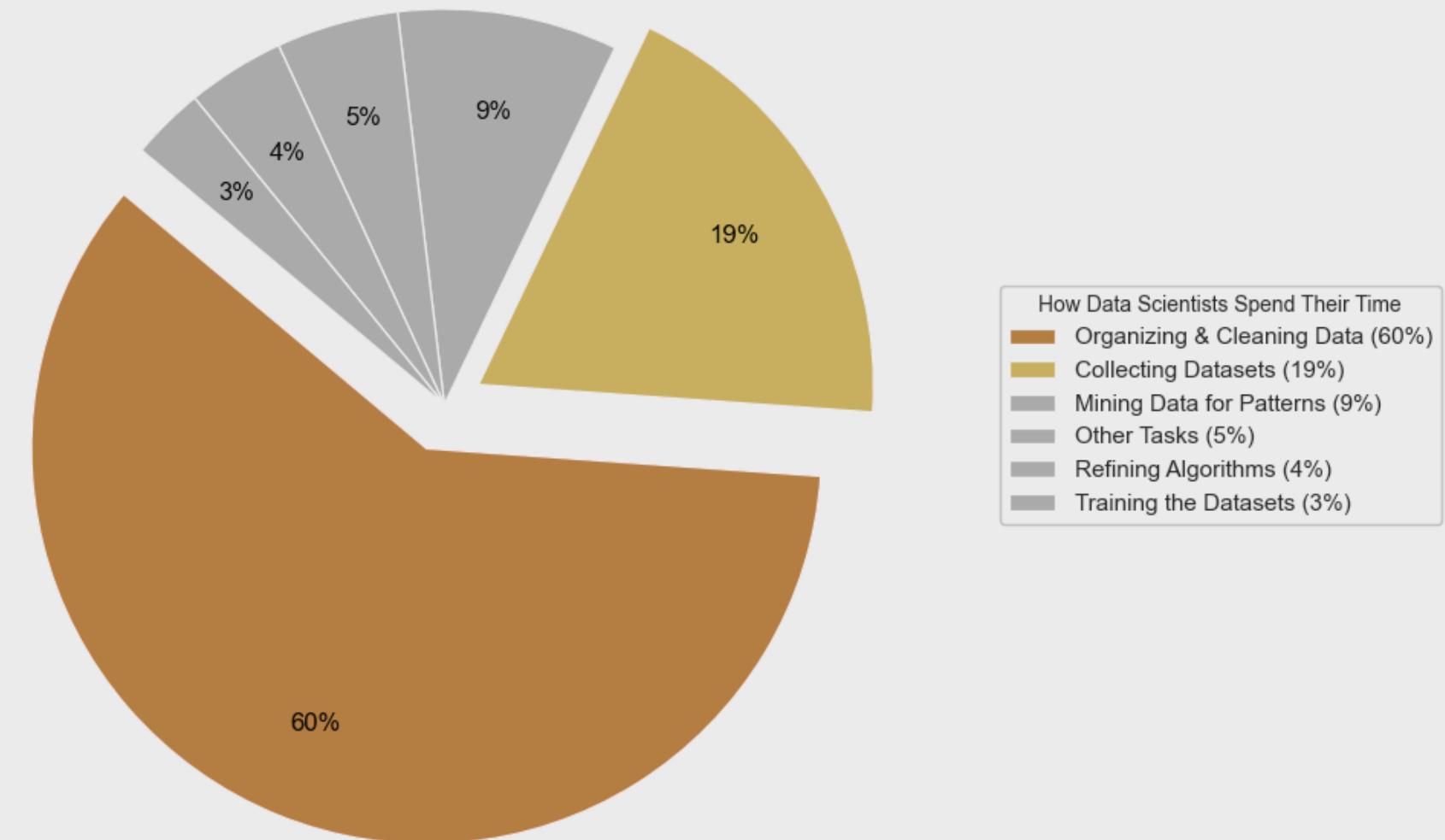
60–80% of a data scientist's time is spent on data collection & standardization

Source: multiple industry surveys on data scientist workflows

Garbage in, Garbage out

Good data preparation not only increases accuracy, but also clarifies the 'voice of the data'.

Data Preparation Dominates the Data Science Process



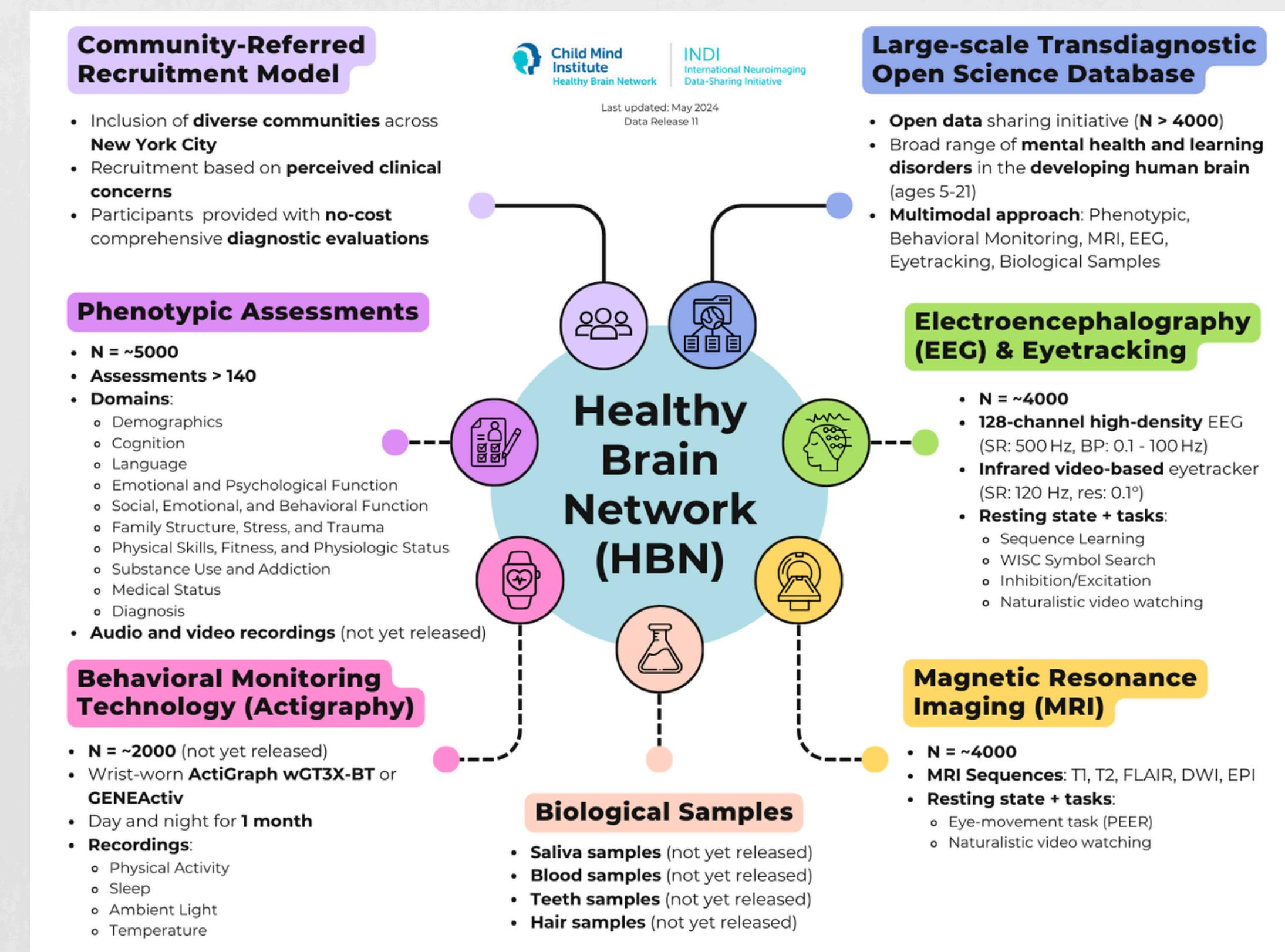
We choose dataset captures the daily lives of children and adolescents — across behaviour, routines, and mental health



**Child Mind
Institute**
Data Provider



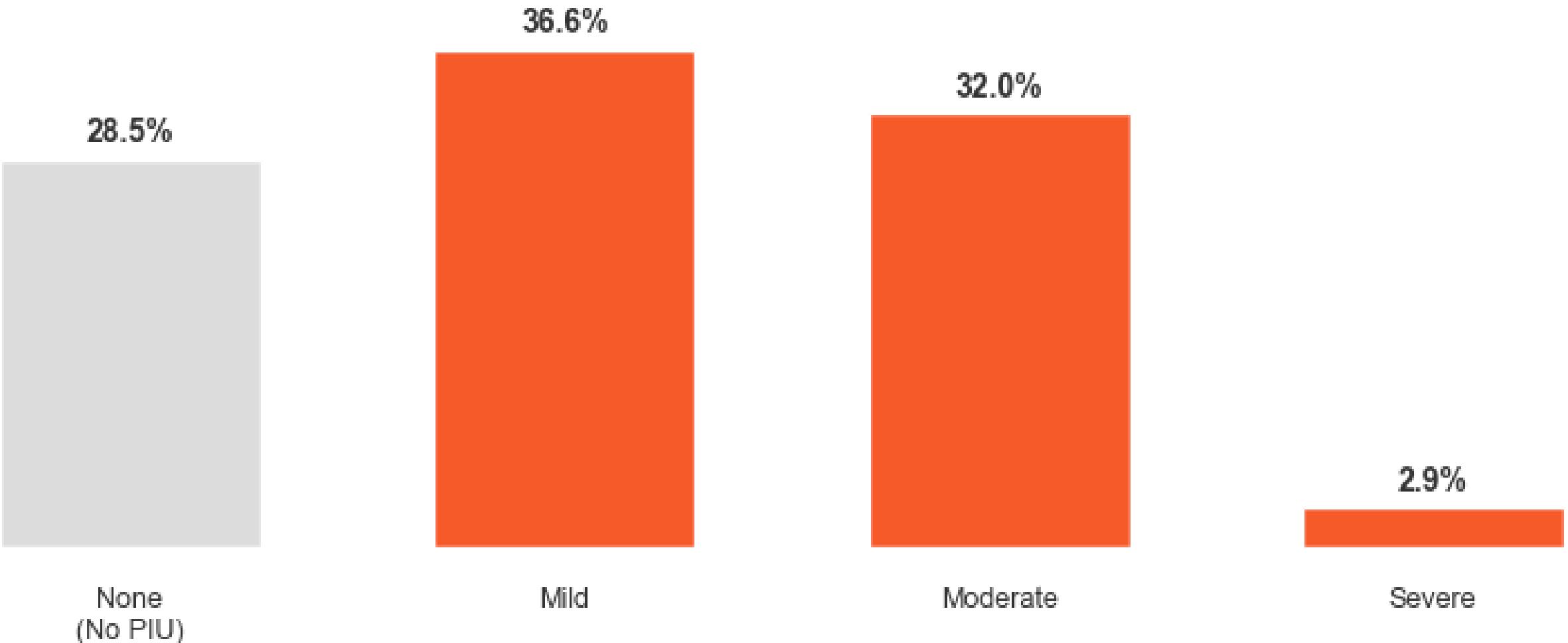
The competition



How serious is Problematic Internet Use today?

A 2024 study published in the International Journal of Indian Psychology examined a sample of 610 adolescents aged 12 to 19 and revealed an unexpected prevalence of problematic internet use (PIU).

Over 70% of surveyed students show signs of Problematic Internet Use
Mild cases are most common (36.6%), while severe cases are rare (2.9%)



Suvarma (2024), *Problematic Internet Use Amongst Adolescents: Internet Using Behaviour and Gender Differences*

What components combining to our dataset?

Dataset Overview: Two Data Sources

TABULAR DATA

82

columns

3,960

training observations

20

test observations

ACTIGRAPHY DATA

13

time-series columns

213,423

timesteps per participant
(5-second intervals)

3,960

individual time-series files

Our story is about how messy human data becomes meaningful through preparation.

WHO



Data Scientist



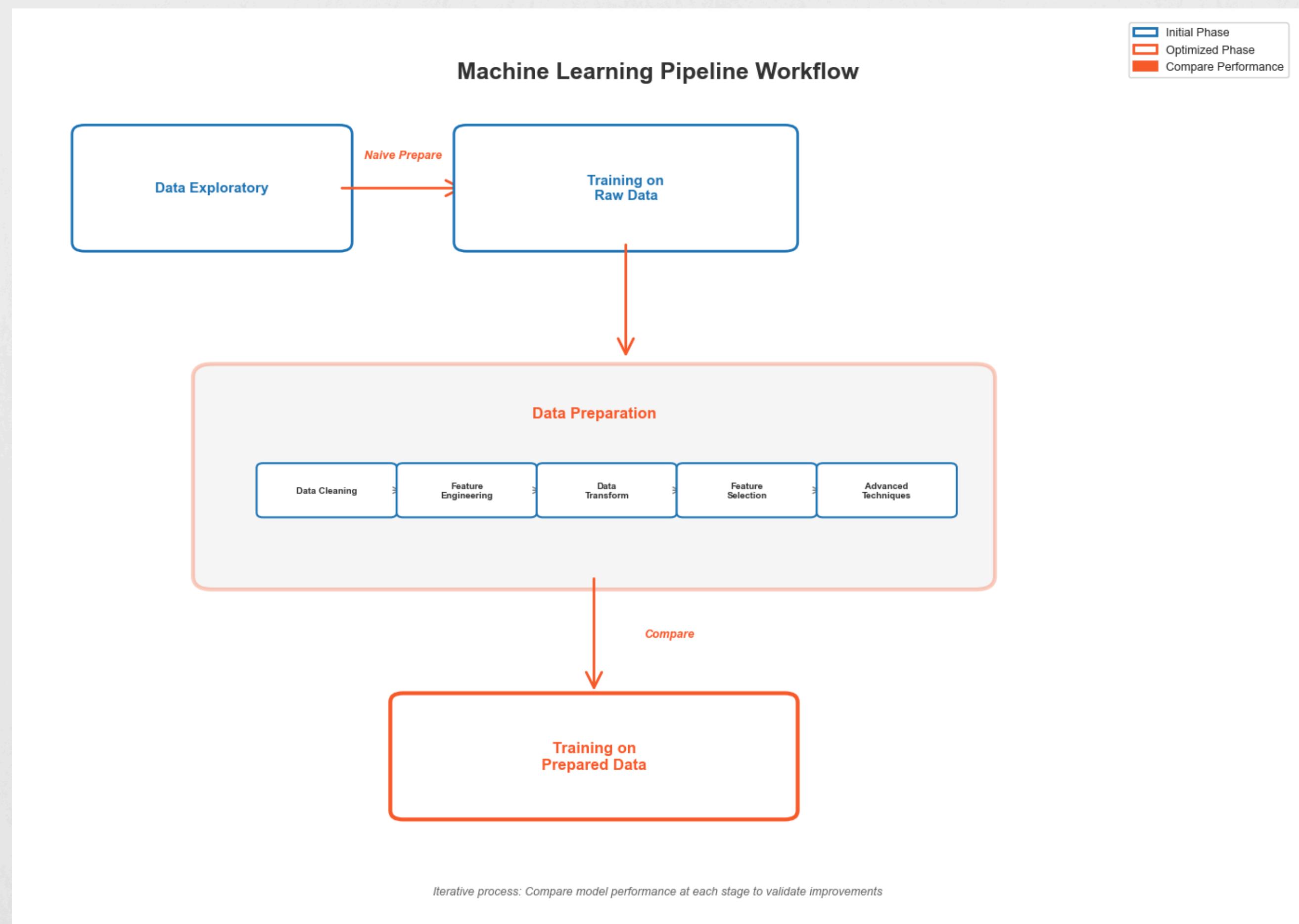
Parents

WHAT

Emphasize the difference before and after data preparation:

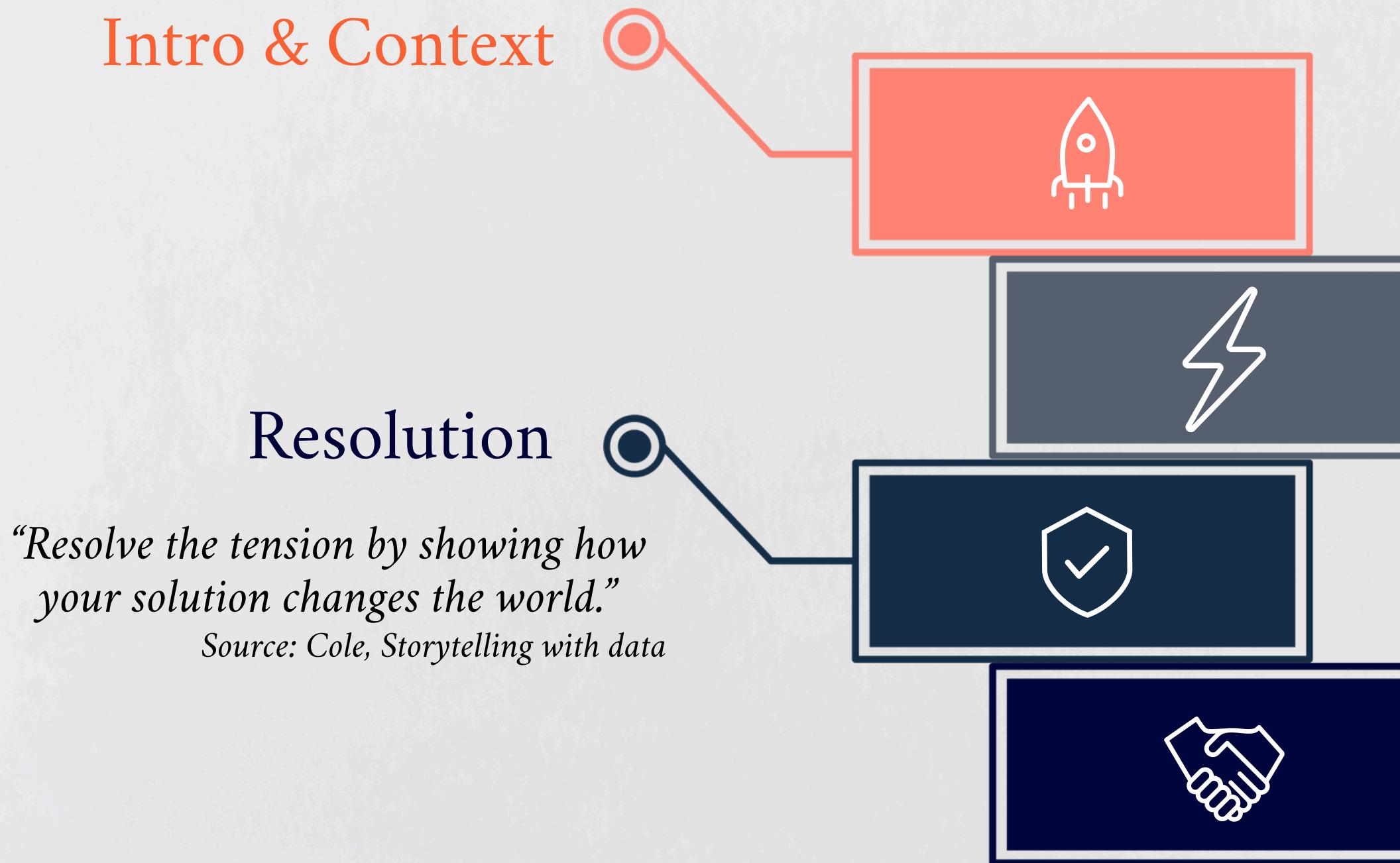
- *The same prediction task: identifying problematic internet use*
- *A model trained on raw data vs. a model trained on prepared - data*
- *How class distribution, error rates, and recall at higher severity levels change dramatically*

Our story is step-by-step HOW



Storytelling strategy: Our story combines two interwoven storylines

Main Story: The technical storyline



Additional Story: The insight storyline

Tension

"Introduce conflict to keep people engaged."

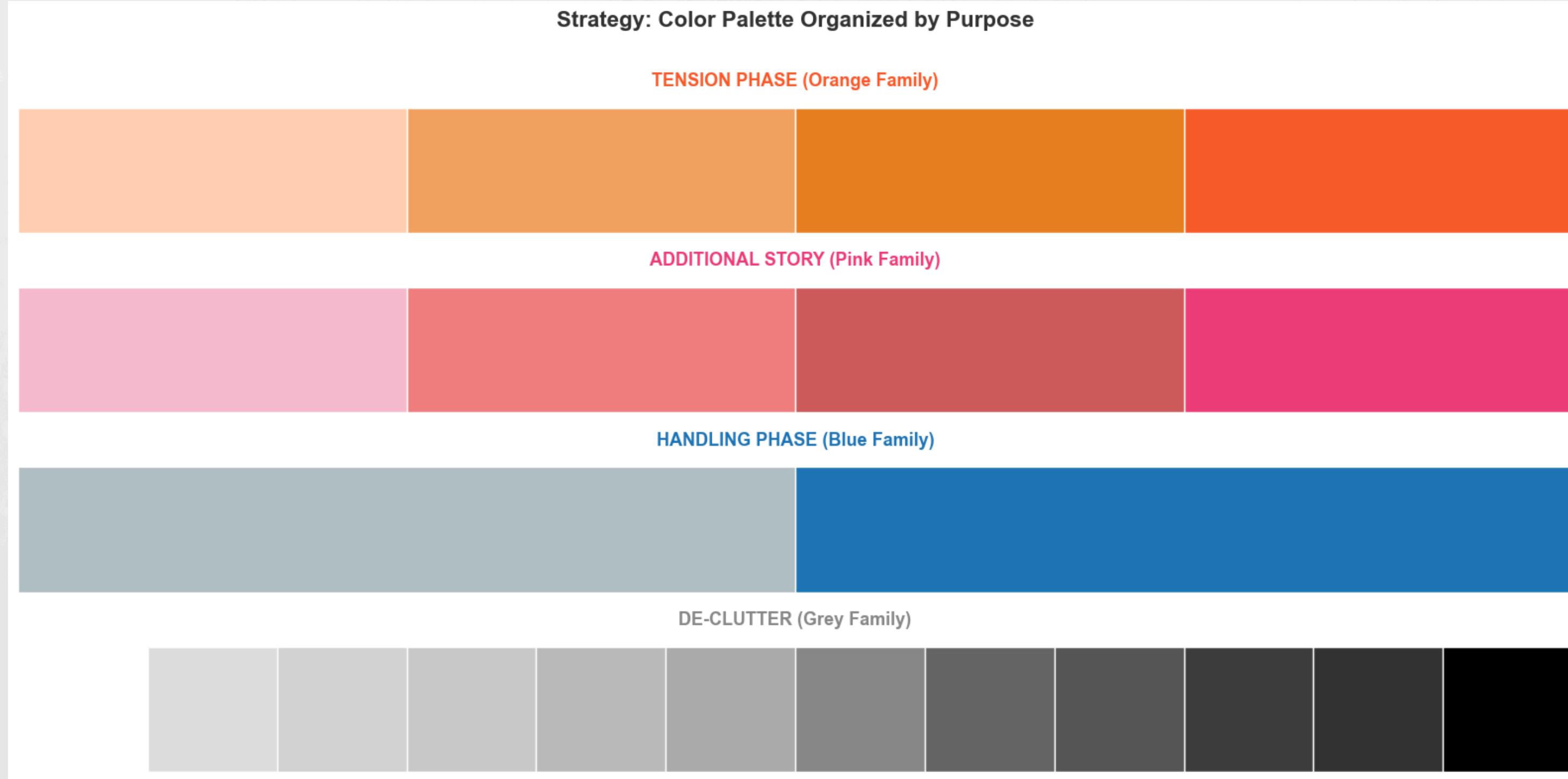
Source: Cole, Storytelling with data

Compare & End

"Humans understand contrast better than absolute numbers."

Source: Cole, Storytelling with data

Storytelling strategy: Color using



Storytelling strategy: “*Charts are not the story - they are the evidence.*”

Visual titles must contain an actionable message, answer: “So what does this show?”



Visuals should be clean and minimal, using only one highlight color to direct attention.



Storytelling strategy: Progressive Reveal

Avoid overwhelming your audience with too much information at once.

Reveal the story gradually, following a clear and logical sequence.

In our narrative, we move from
data → problem → insight → model,
and only then to solution →
transformation → outcome.

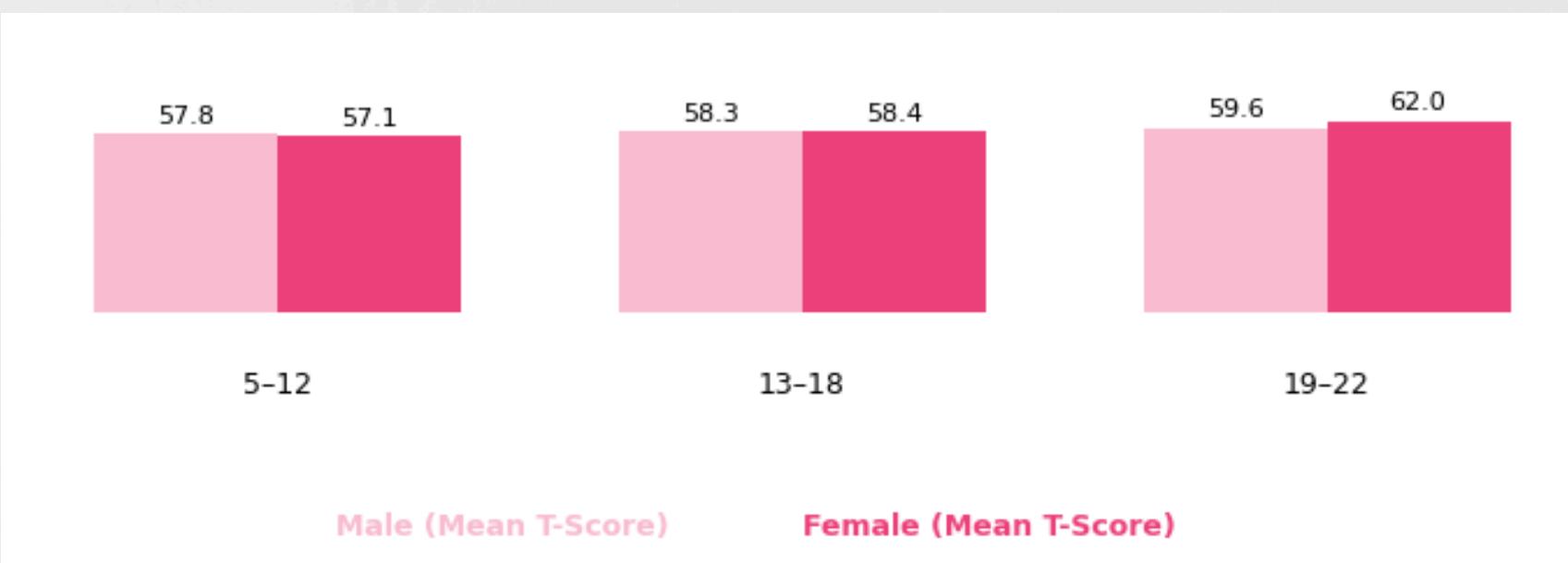
This progression reflects the
natural arc of an effective data story.



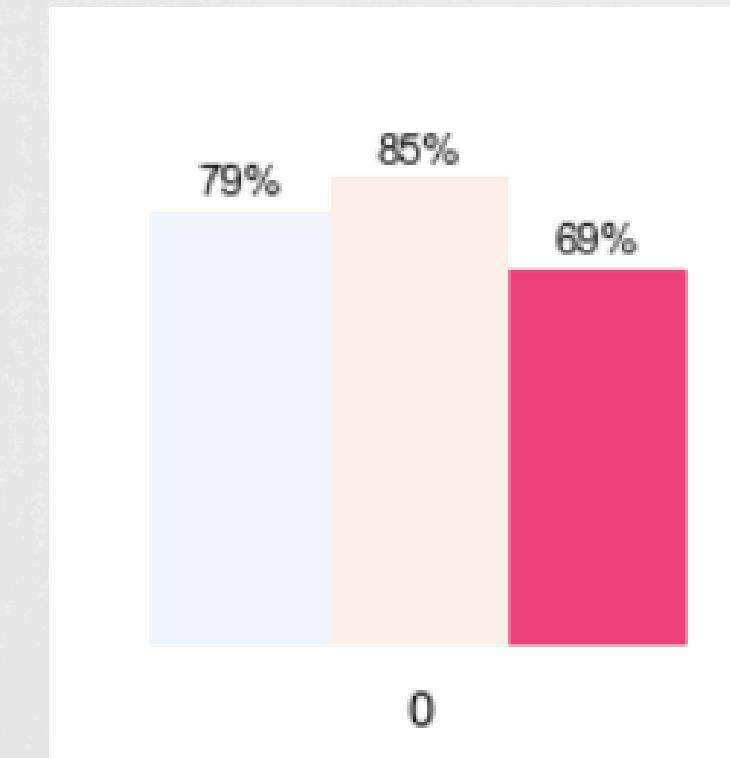
Storytelling strategy: Eliminating clutter

We intentionally remove anything that does not directly support the message

We eliminate unnecessary gridlines, borders, shadows, and decorative elements.



We also avoid heavy color usage, relying instead on a clean white background for both slides and charts, which reduces visual noise and keeps attention on the data itself.



- Only one highlight color is used to direct focus, while all other elements remain neutral.
- Labels are placed directly on the visual whenever possible to avoid forcing the audience to decode legends.
- We simplify axes, reduce ink, tighten spacing, and ensure that each chart communicates a single, unambiguous point.

Storytelling strategy: Annotation

Write the insight directly on the chart.



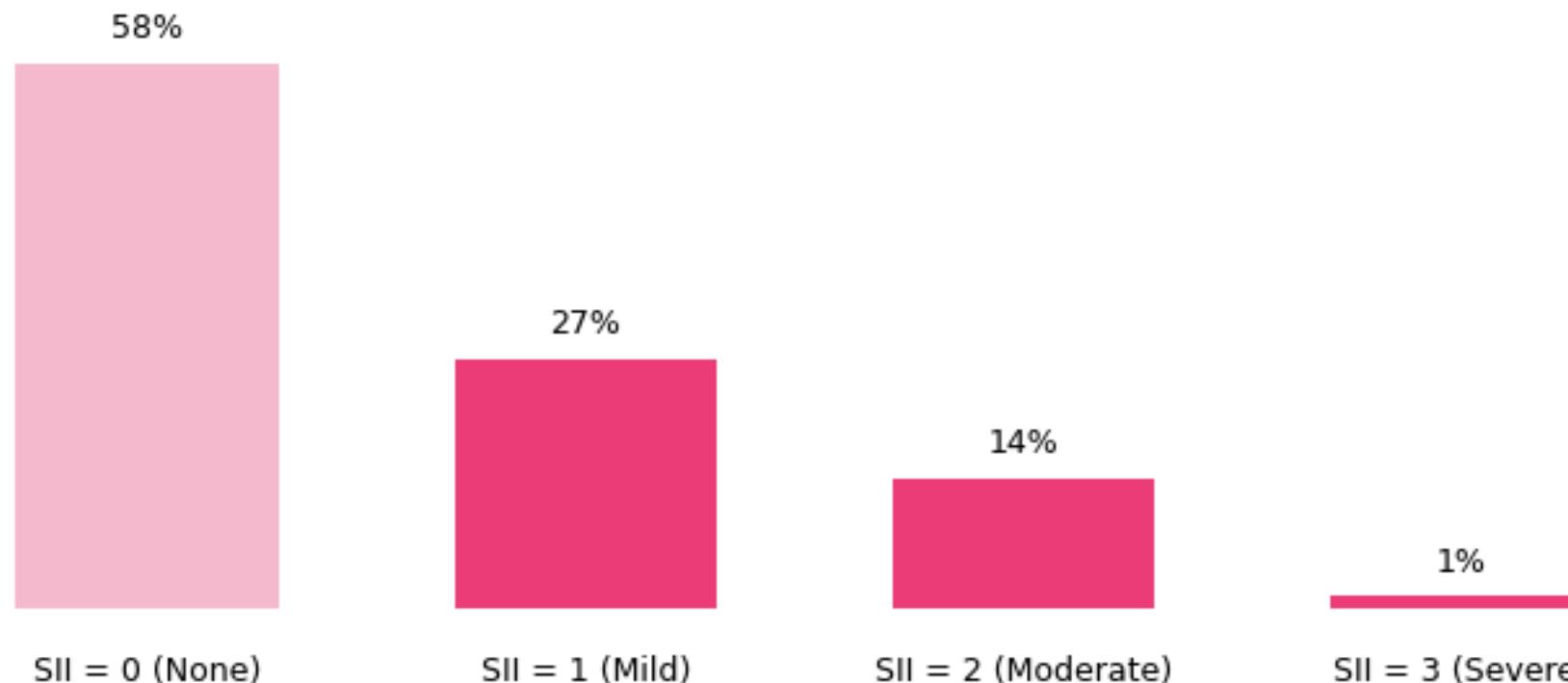
Annotations highlight the exact insight we want the audience to see.
By labeling data directly on the chart and placing explanations where the eye lands, we remove ambiguity and reduce cognitive load.

Additional storyline

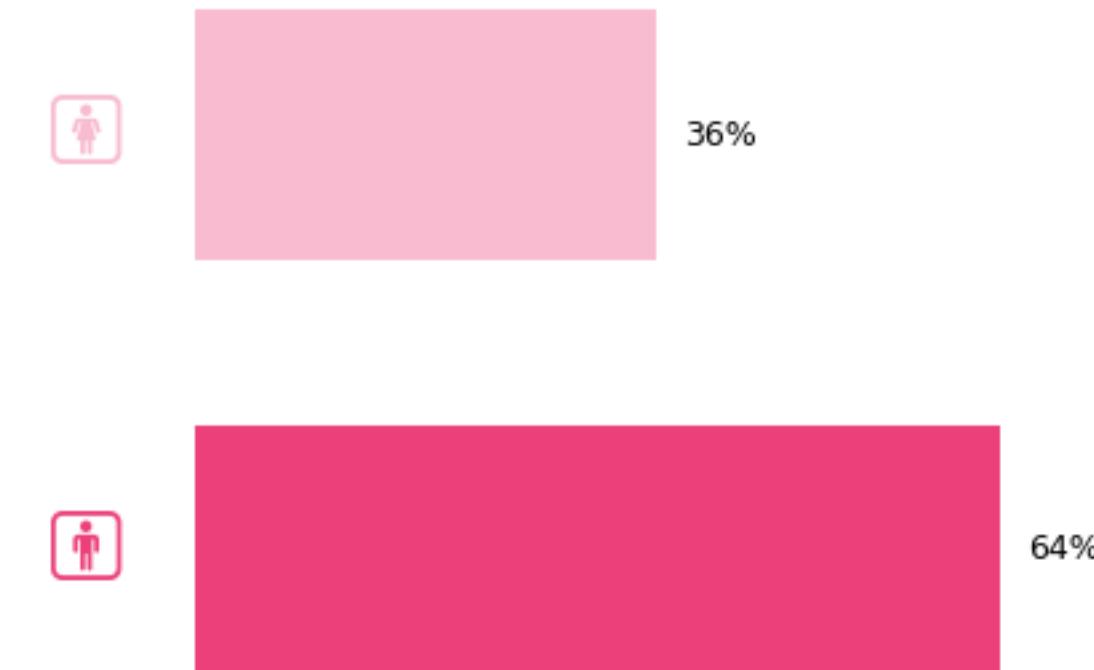
The storyline explain what the data reveals, therefore, creating the complete narrative of our work.

Problematic Internet Use patterns differ sharply by severity level, age group, and gender

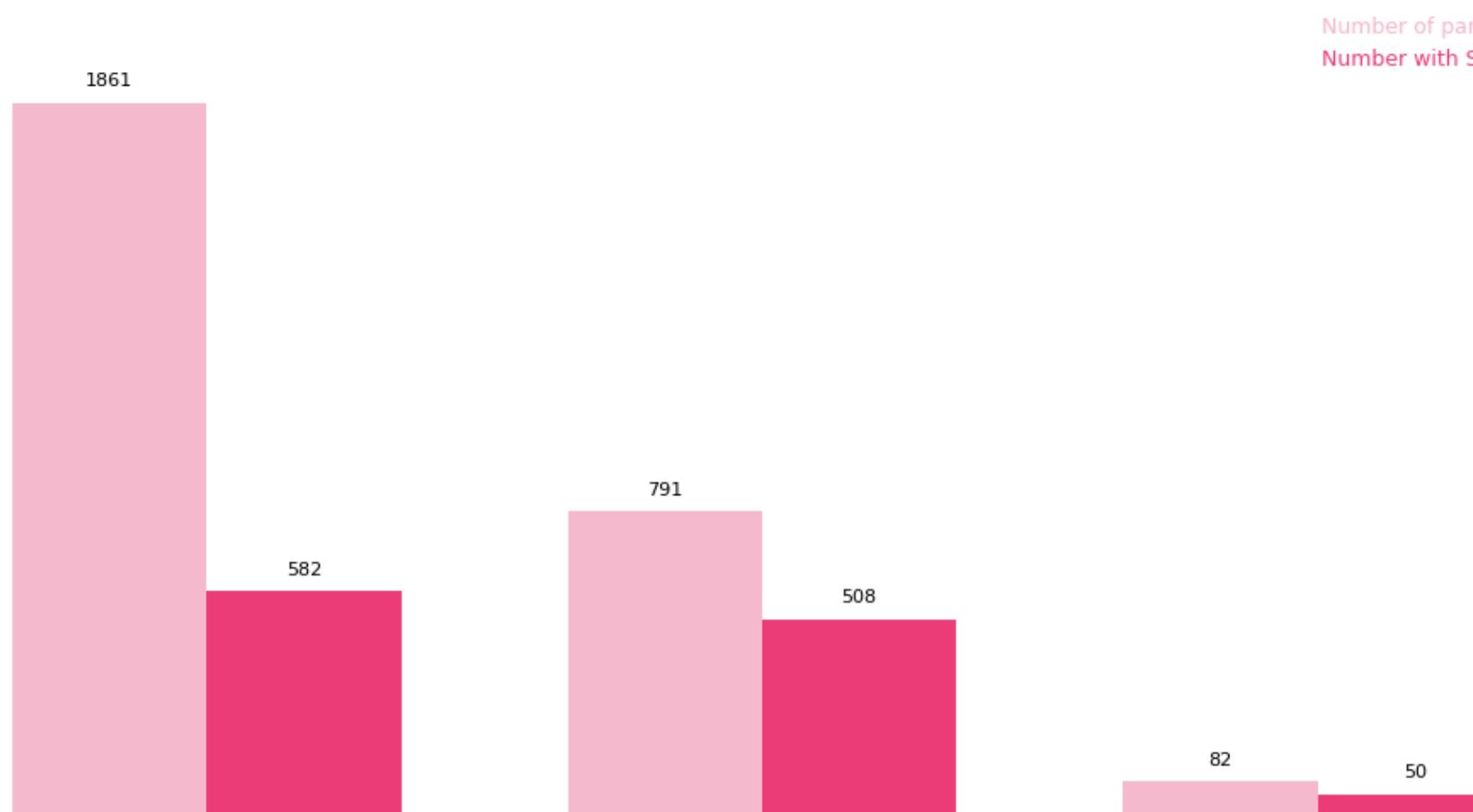
Severity distribution: higher SII levels are progressively rarer



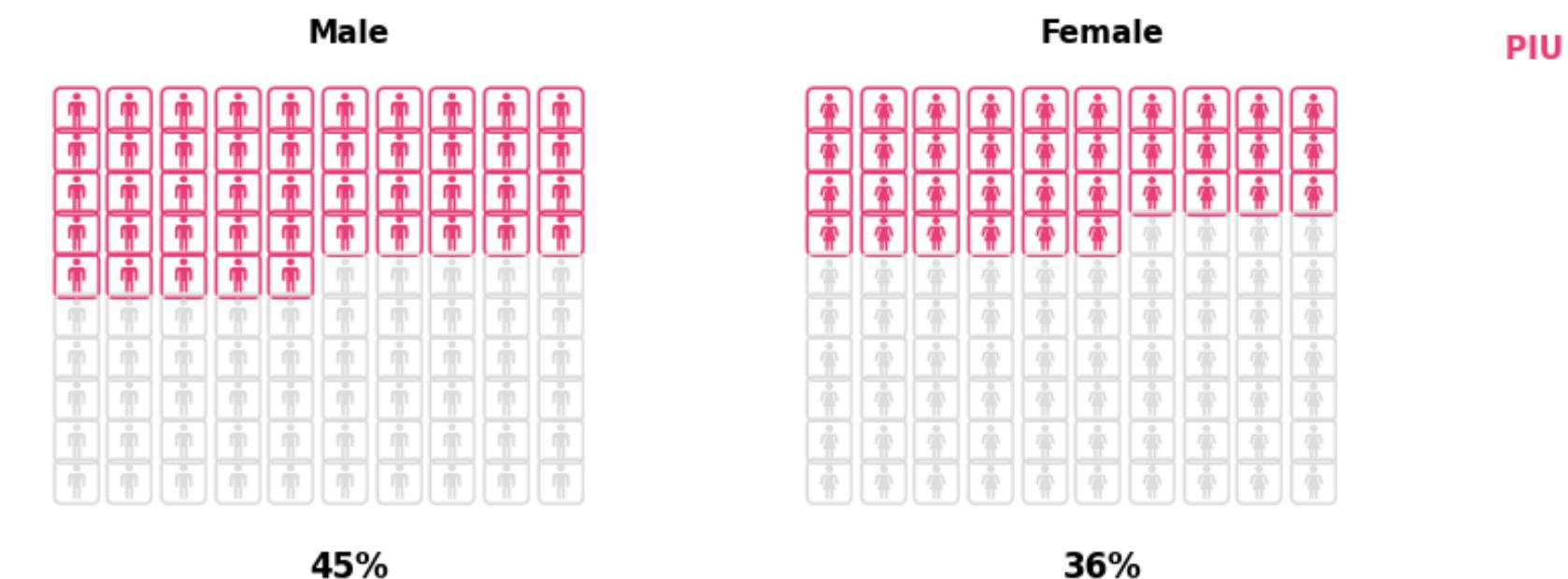
Male make up the larger share of the sample



Problematic Use Becomes More Common in Older Age Groups

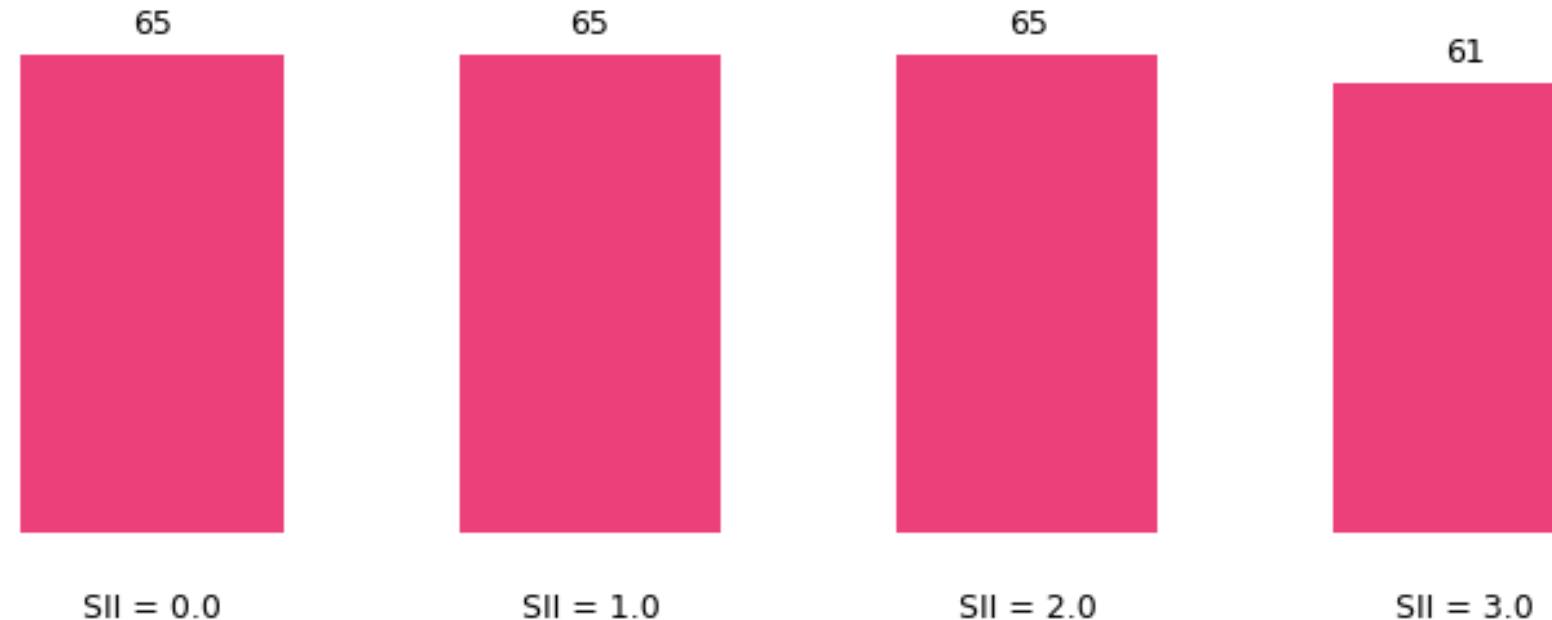


**Share of children with problematic internet use ($SII \geq 1$)
Visualised separately for boys and girls**

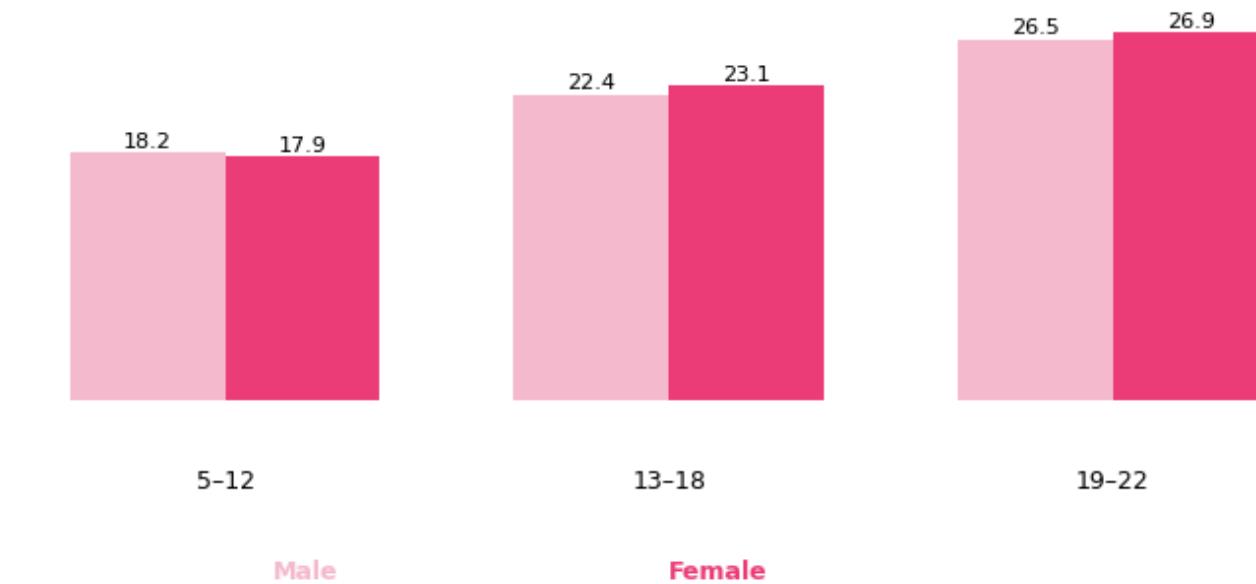


Physical and Behavioral Indicators Shift Systematically with PIU Severity — Reinforcing the Need for Reliable Measurements

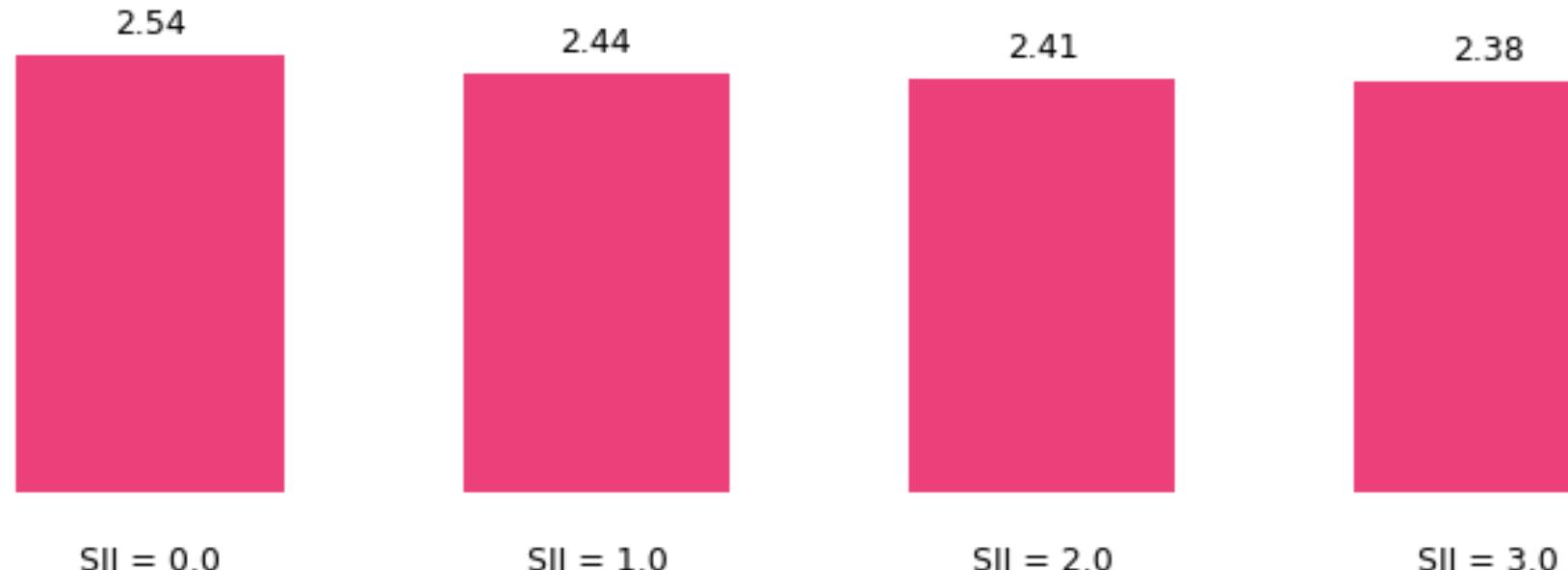
Children with higher SII levels show lower global functioning (CGAS)



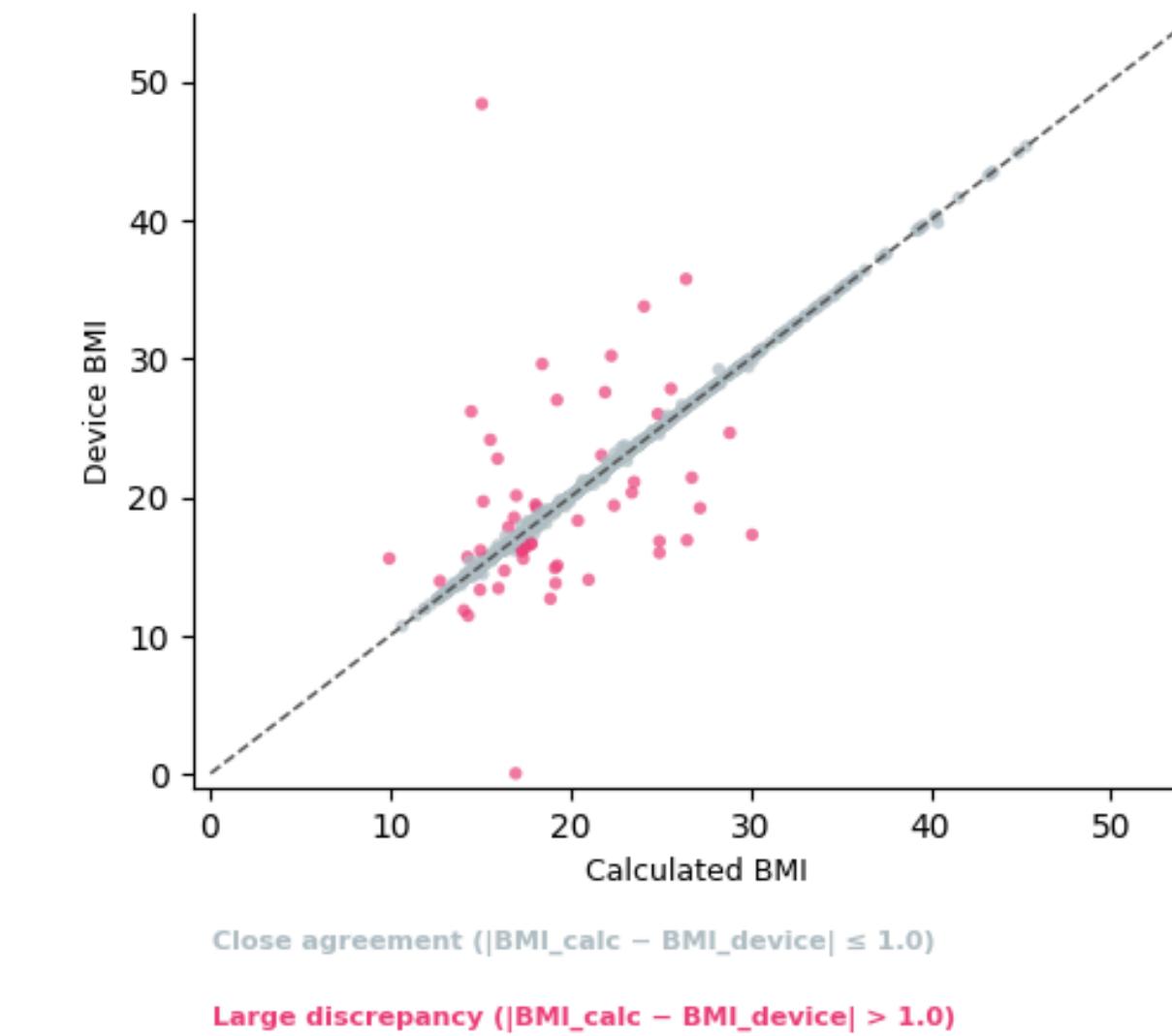
BMI increases consistently with Age, showing minimal difference between Sexes



Median physical activity (PAQ) tends to be lower for higher SII levels

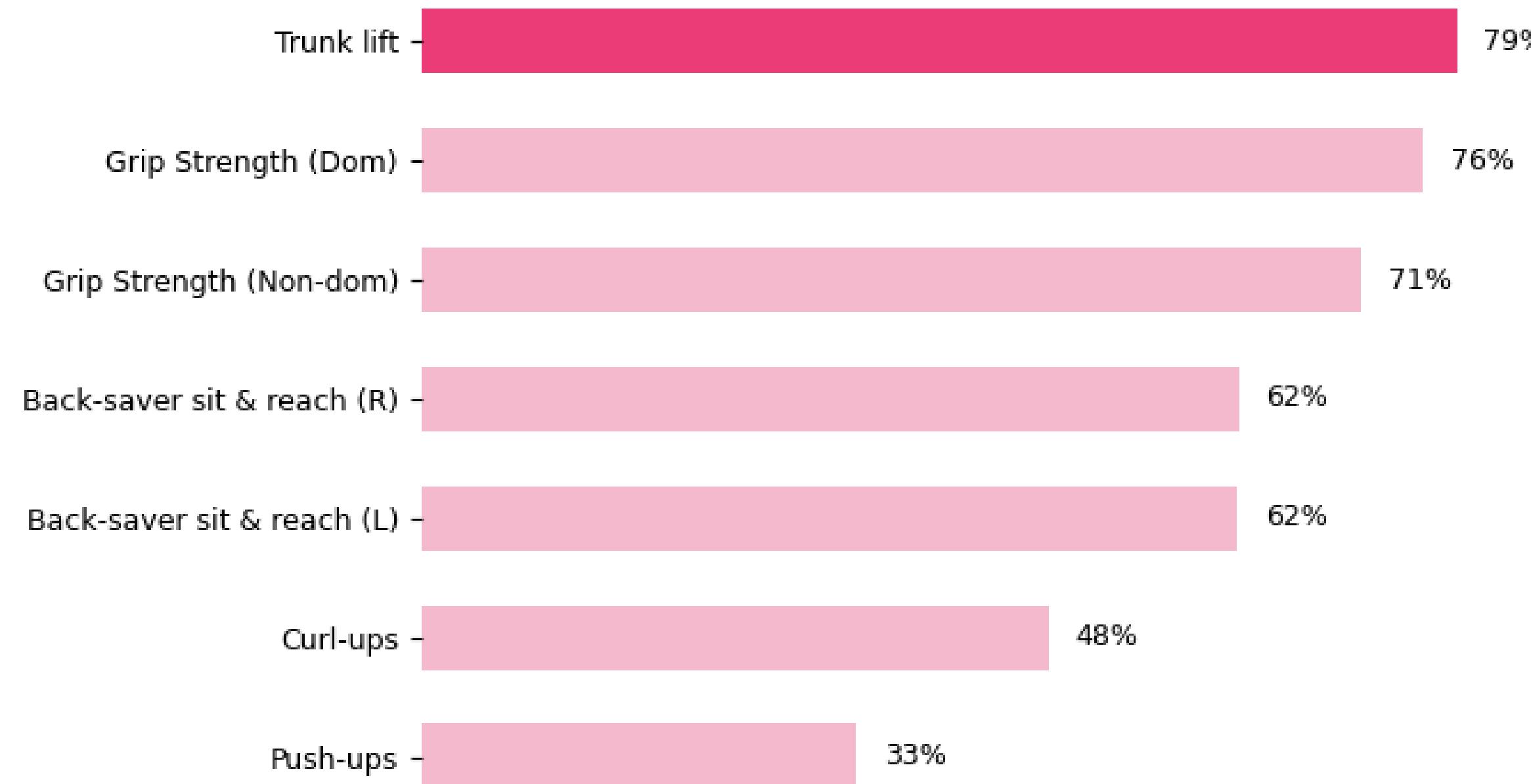


Use Device-Measured BMI due to Significant Discrepancies in Observations



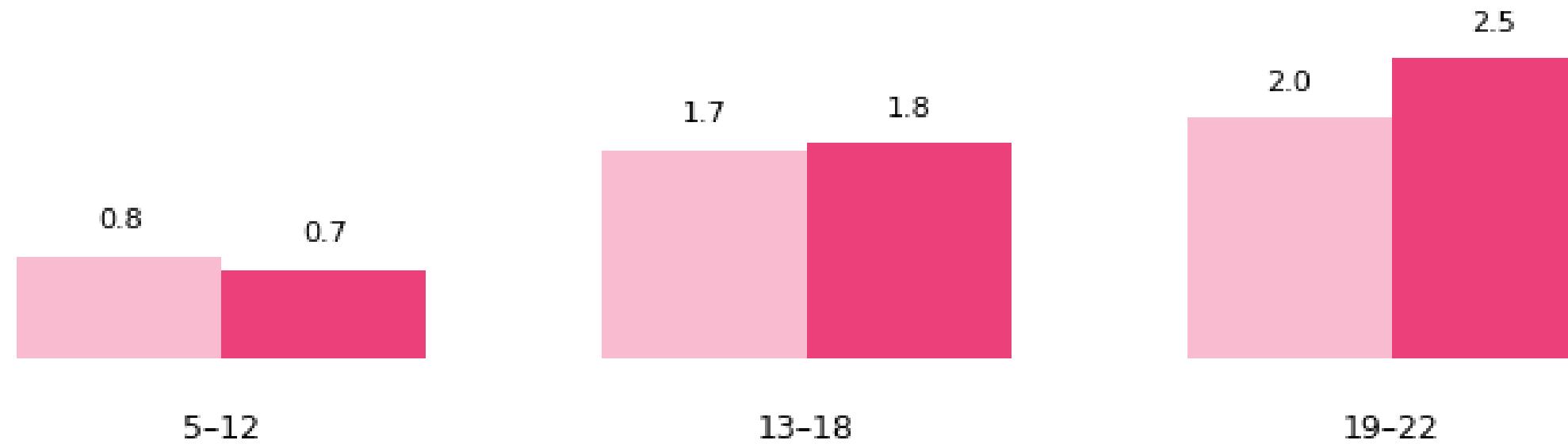
Most Participants Meet Basic Flexibility Standards, But Strength Tests Remain Challenging

Overall Ranking of Fitness Tests
Share of participants meeting Normal or Healthy Fitness Zone (HFZ) standard



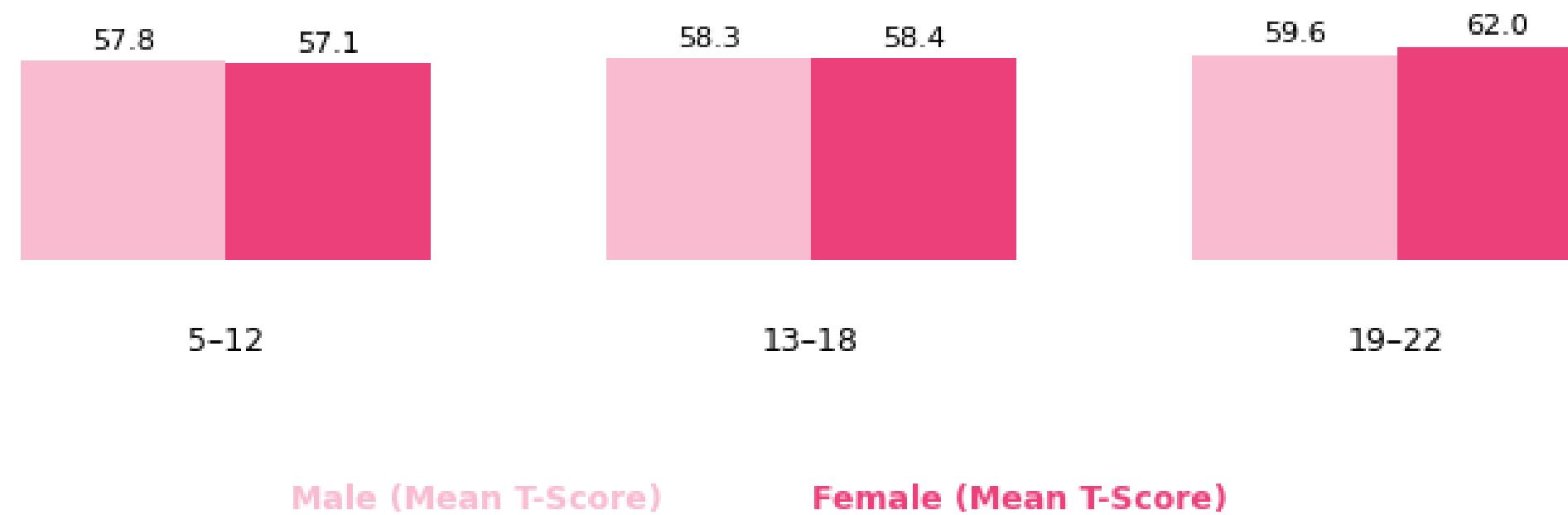
Older Adolescents Show Higher Internet Use and Greater Sleep Disturbance, Most Noticeably in Females

**Internet Use increases with Age;
Females average more hours than Males in older groups**



Male (Mean Hours/Day) Female (Mean Hours/Day)

Sleep Disturbance Risk Increases with Age, Especially in Females



Tension phase

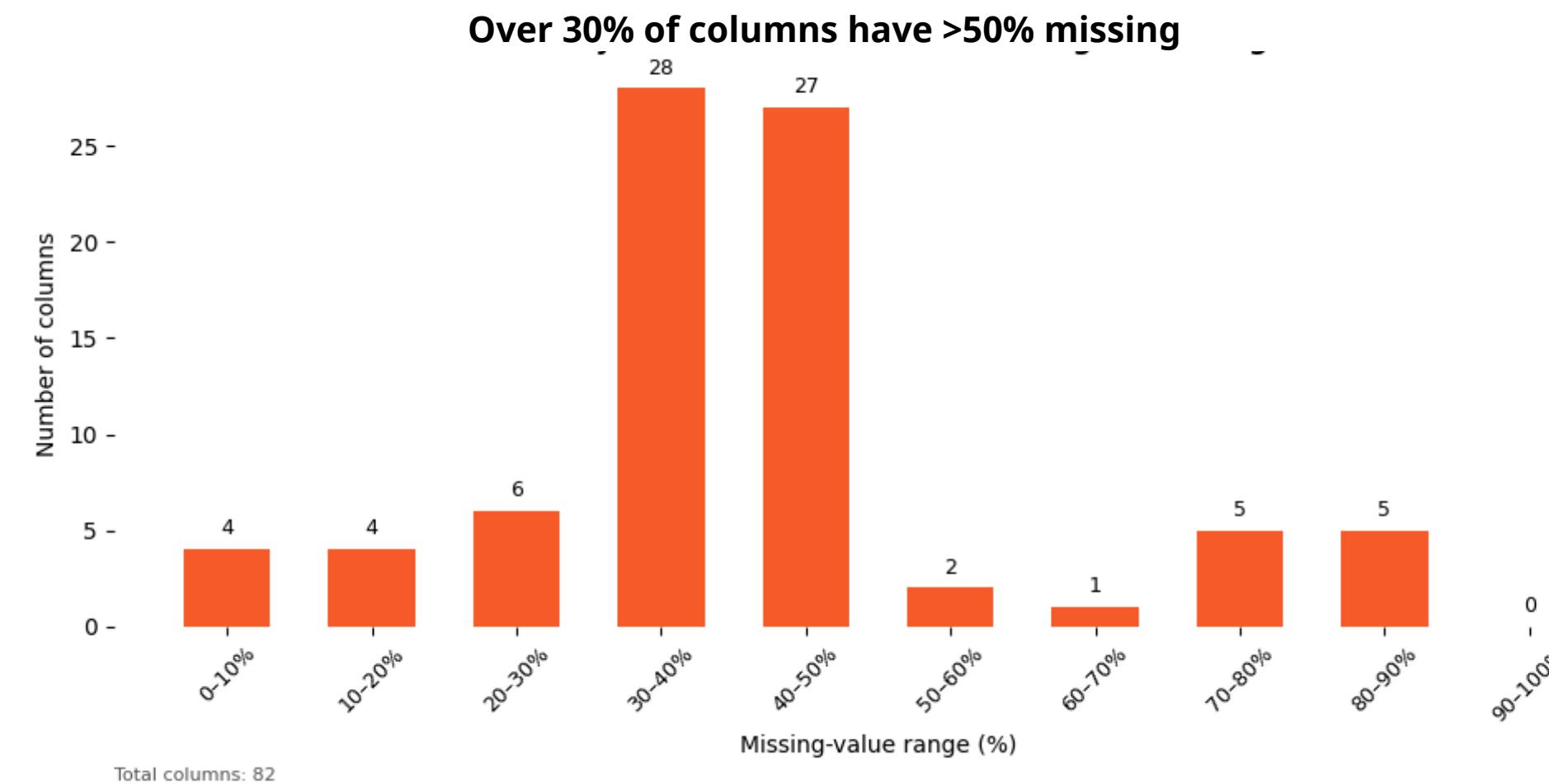
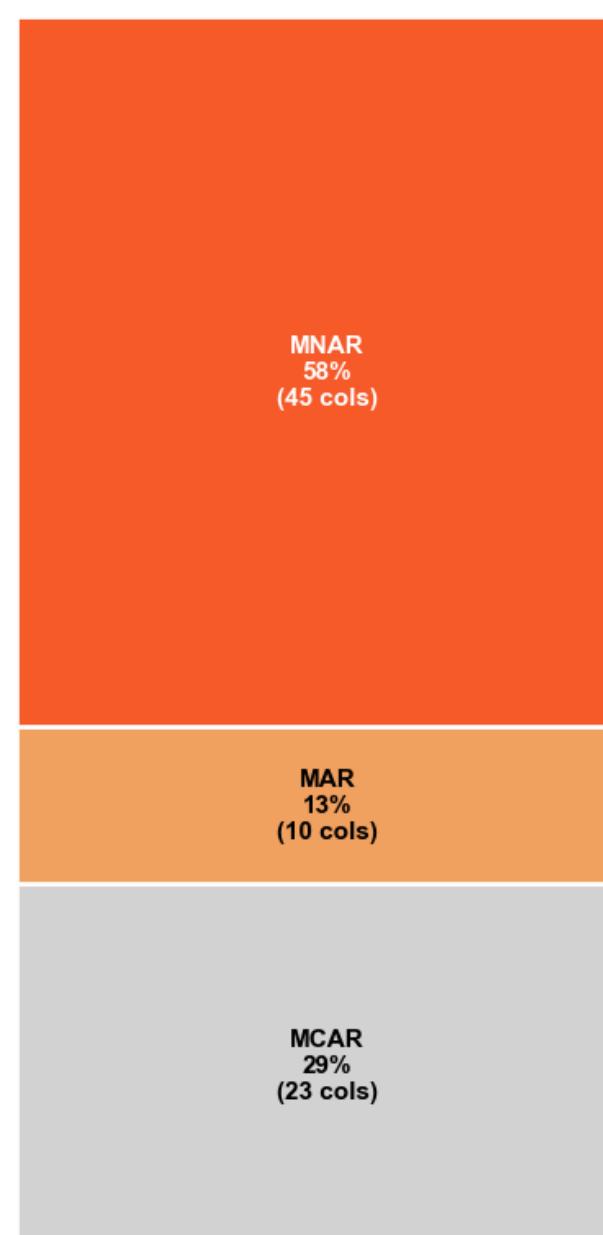
Tension drives the need for transformation.

Missingness is severe, widespread, and mostly Non-Random, creating a major modeling risk

78 /82

columns have missing values

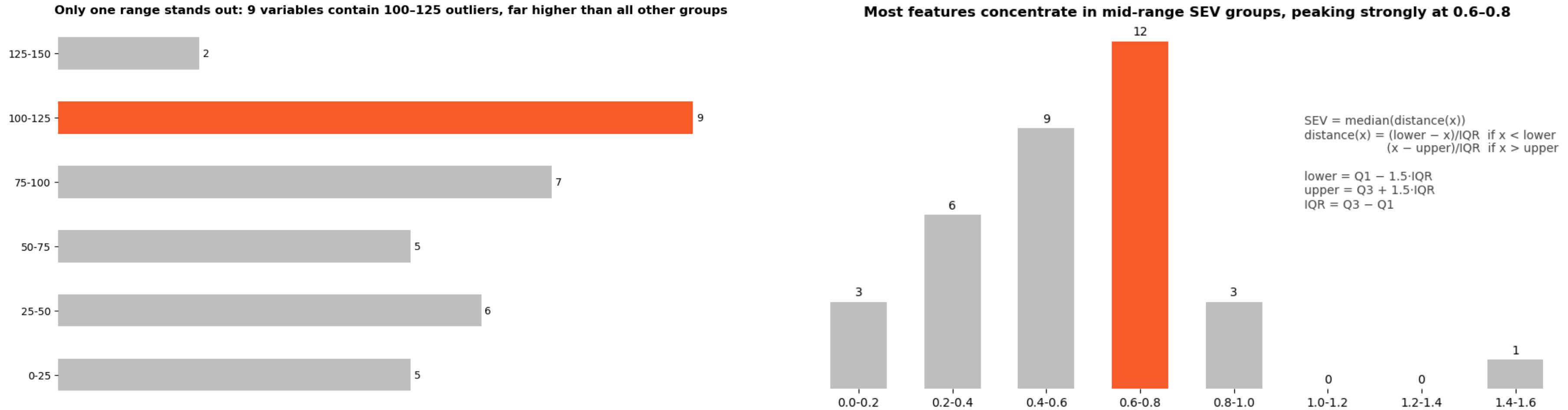
Most missing data is MNAR (58%), requiring careful handling
MAR accounts for 13%, while MCAR is 29%



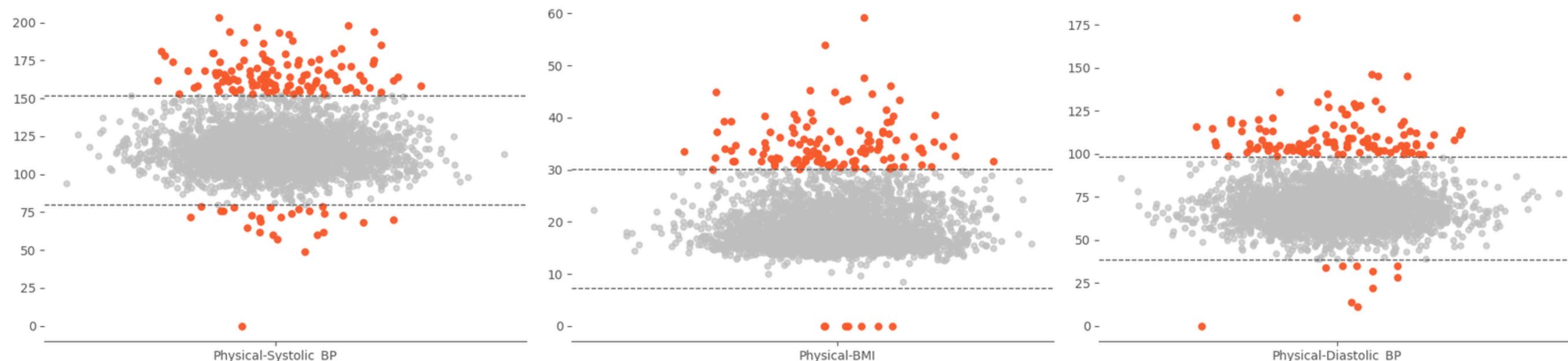
6 columns with more than 75% missing values



Outliers are concentrated and severe — Distorting key physical measures



Top 3 Features with Most Outliers: Highlighting Data Anomalies



ILLOGICAL DATA

BIA-BIA_BMC



BIA-BIA BMR



BIA-BIA FAT



BIA-BIA FFM



PHYSICAL-HEIGHT

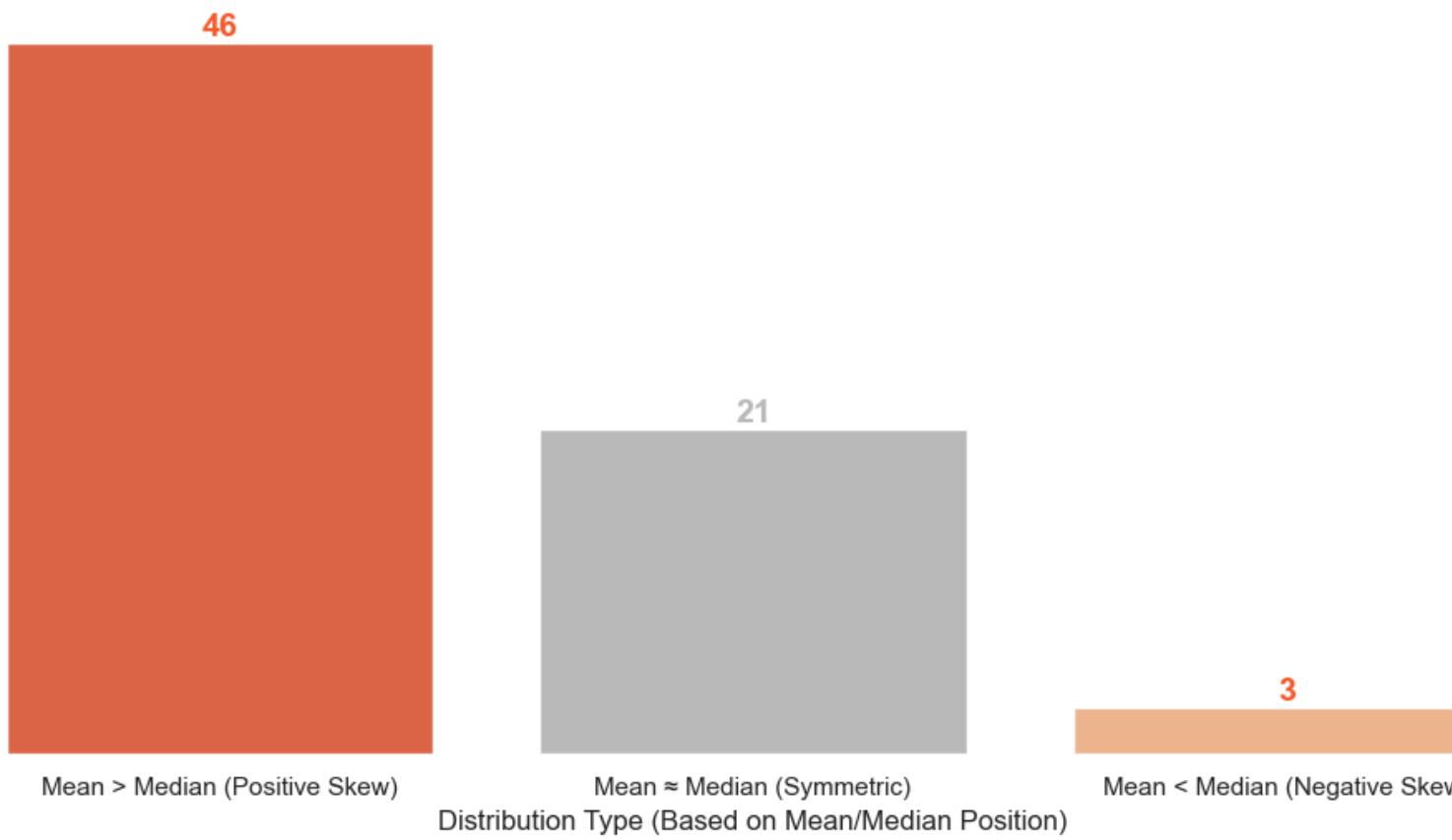


PHYSICAL-HEARTRATE

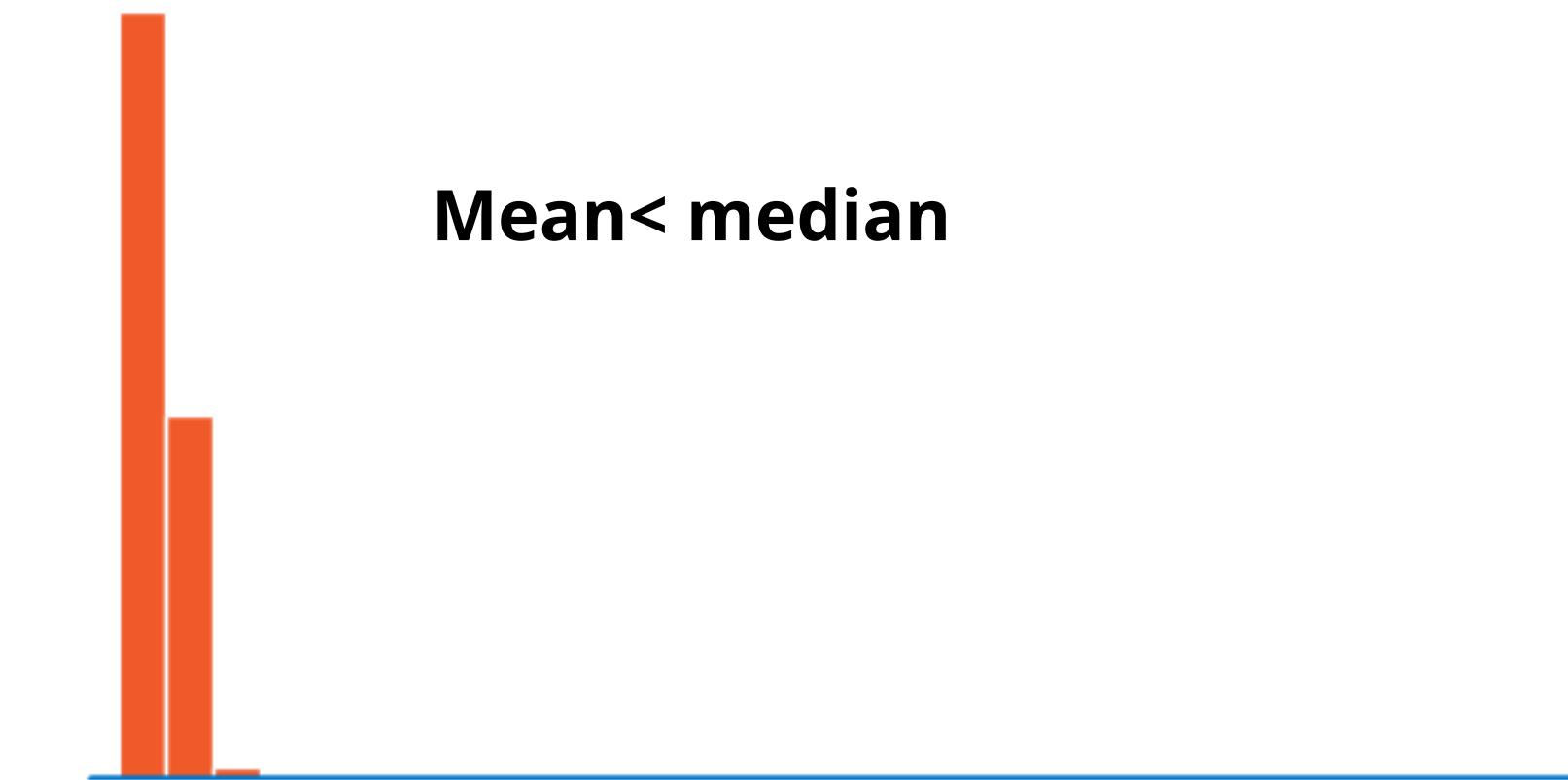


Most features are strongly positively skewed - requiring transformation before modeling

Feature Distribution Based on Mean vs. Median Relationship

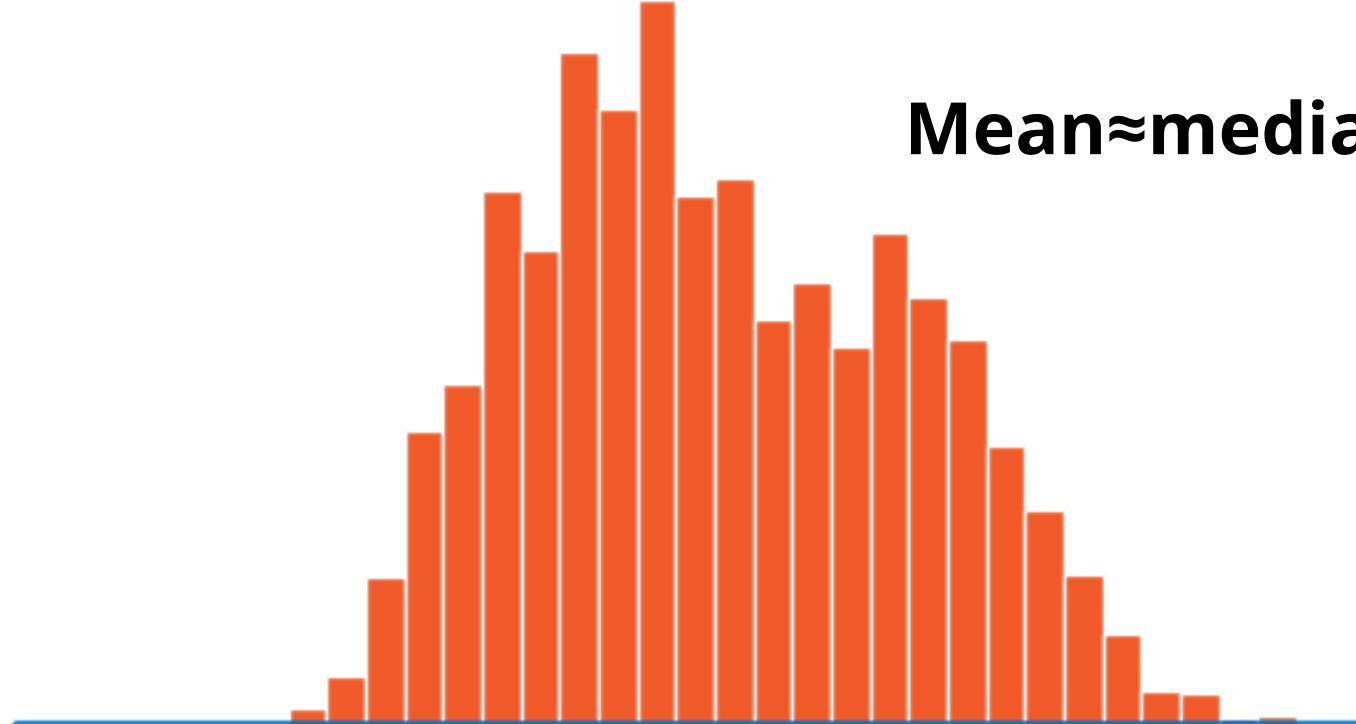


Distribution of BIA-BIA_FFMI



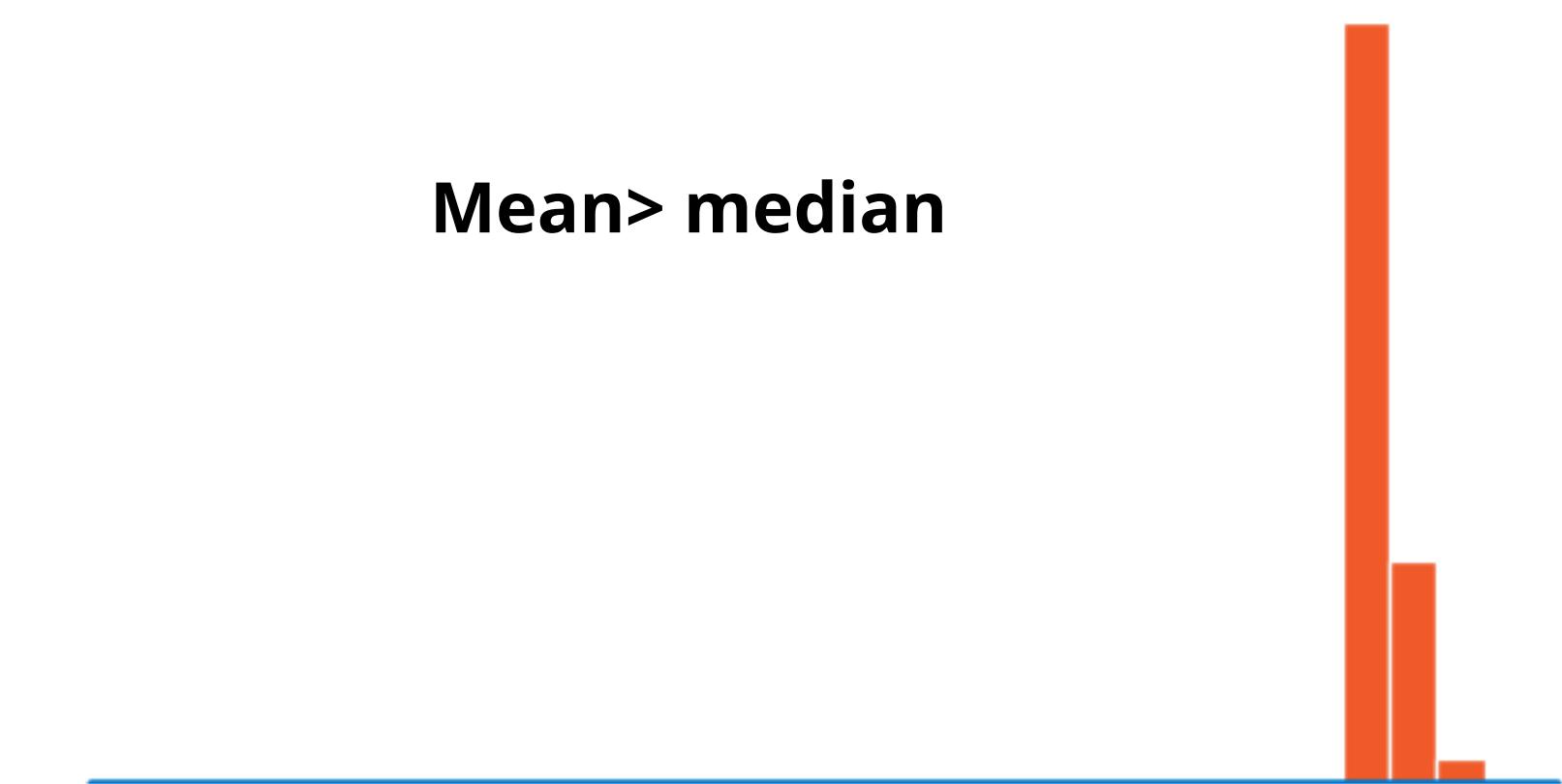
Distribution of Physical-Height

Mean≈median



Distribution of BIA-BIA_FMI

Mean> median



Most participants have no actigraphy data — Leaving a critical signal missing

This missingness is non-recoverable: no sequences exist for 1,740 participants.

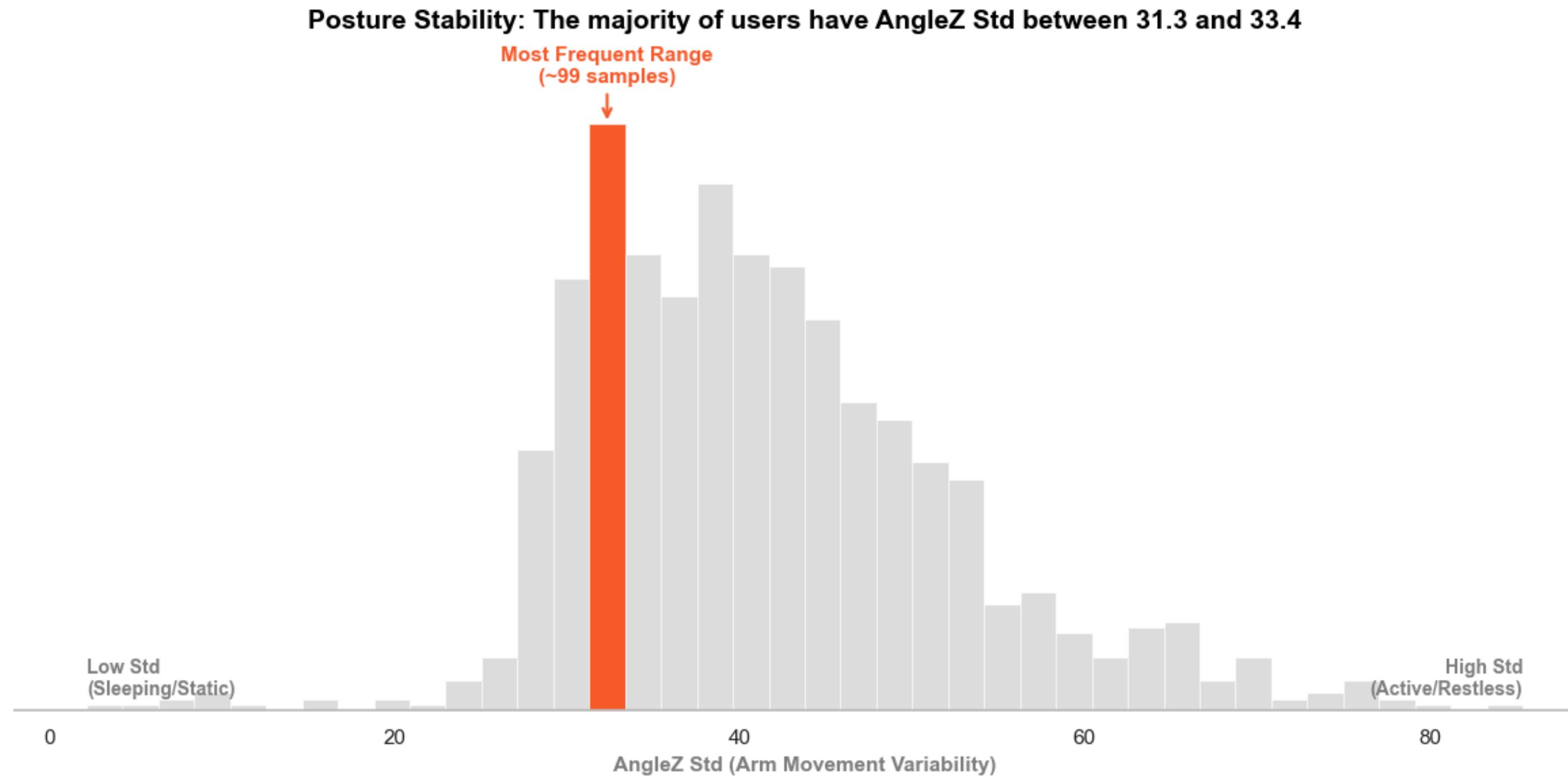
63.6%

1,740 out of 2,736 IDs

of participants have no actigraphy data

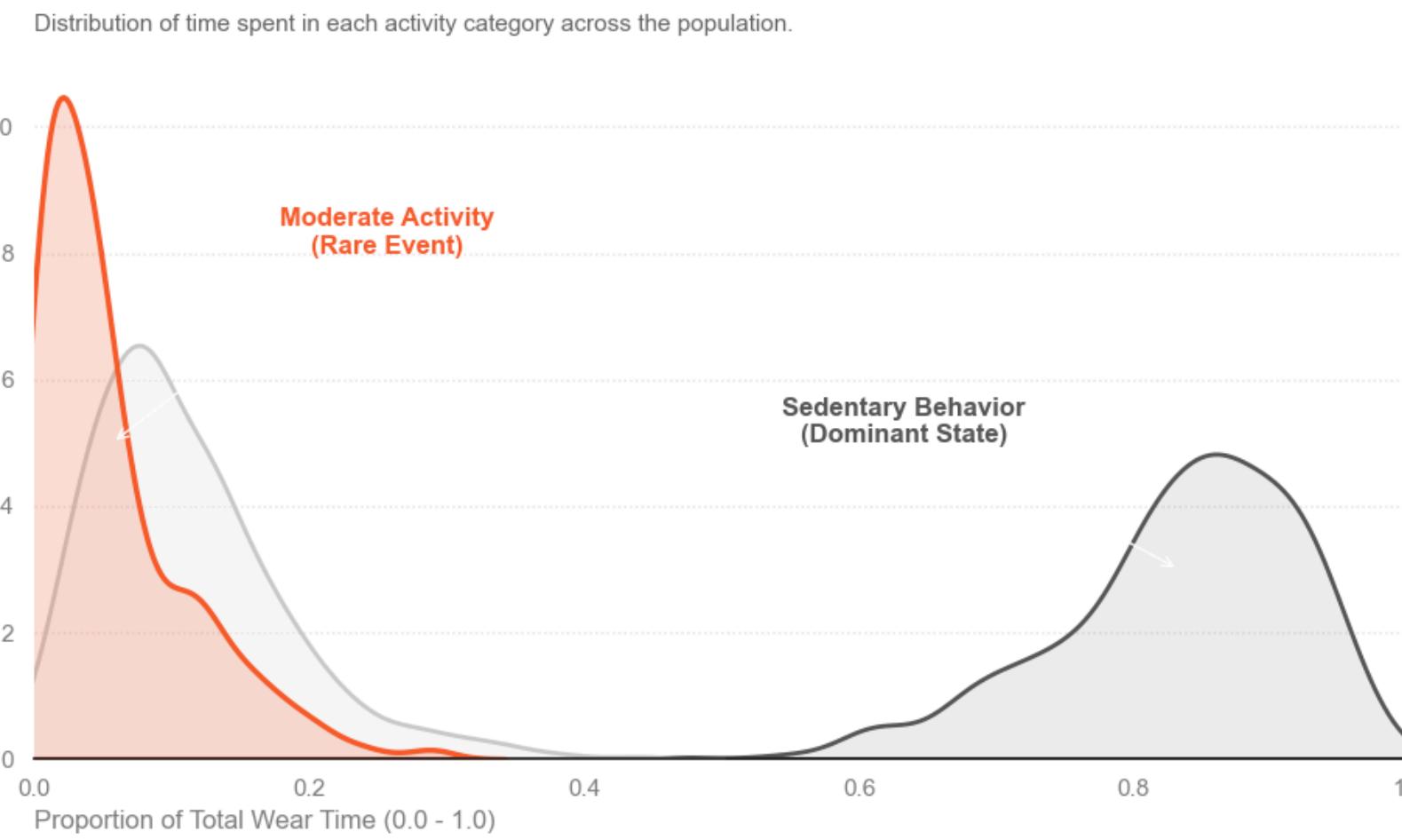
are missing all actigraphy measurements

Most participants show moderate movement variability — Reflecting stable daily posture patterns

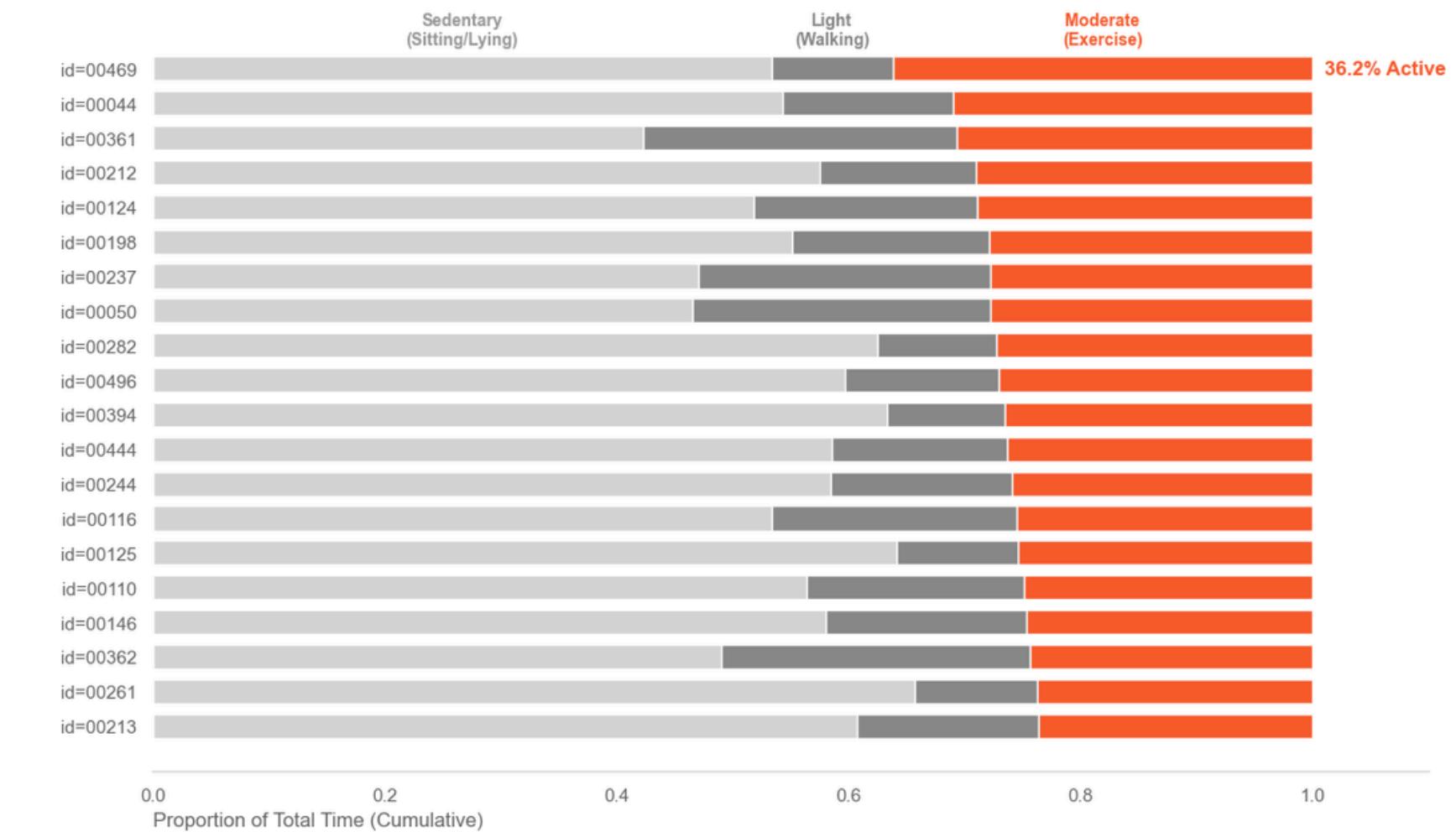


Sedentary Behavior Dominates: Moderate Activity Is Rare in Raw Actigraphy Data.

Participants spent about 85% of wear-time in sedentary states



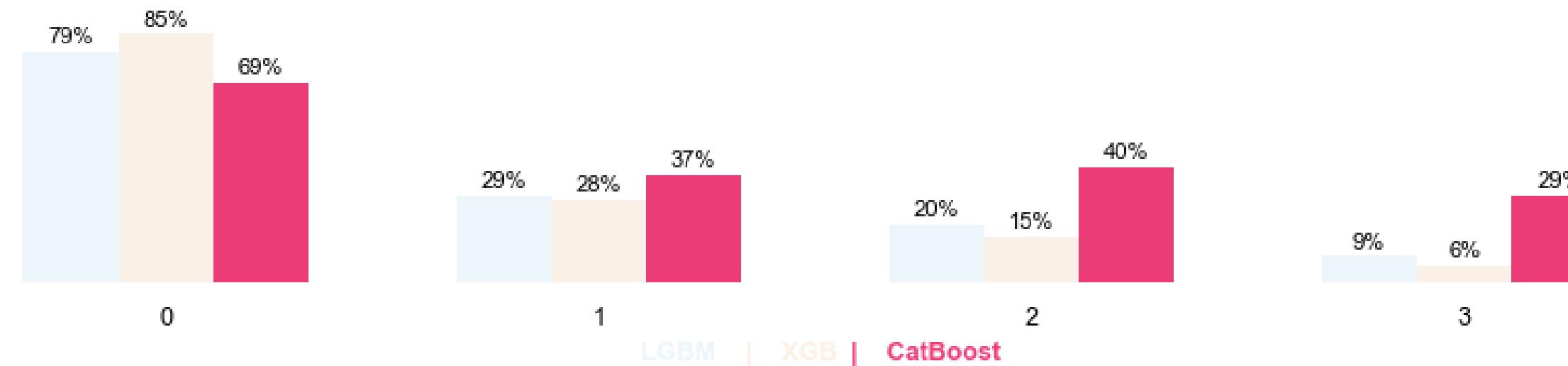
Even the most active participants are predominantly sedentary
Activity breakdown for Top 20 participants. Labels align with the top bar's segments.



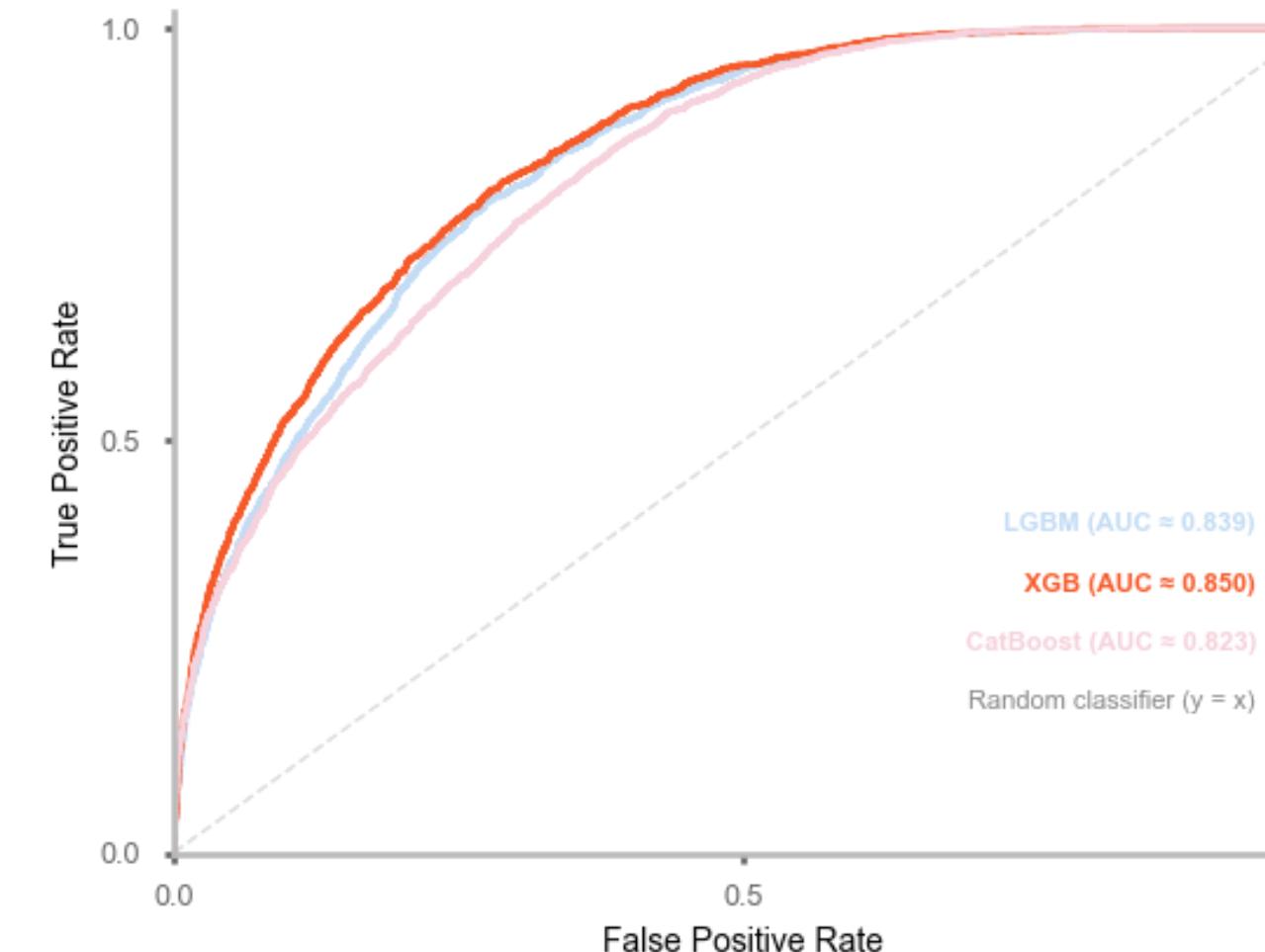
Raw actigraphy reveals imbalanced daily activity patterns — a behavioral signal that requires careful preparation

CatBoost Excels on the Most Critical Class (SII=3), While XGBoost Leads in Overall AUC.

**CatBoost achieves the best recall on severe class (SII=3),
while LGBM and XGB lag behind**

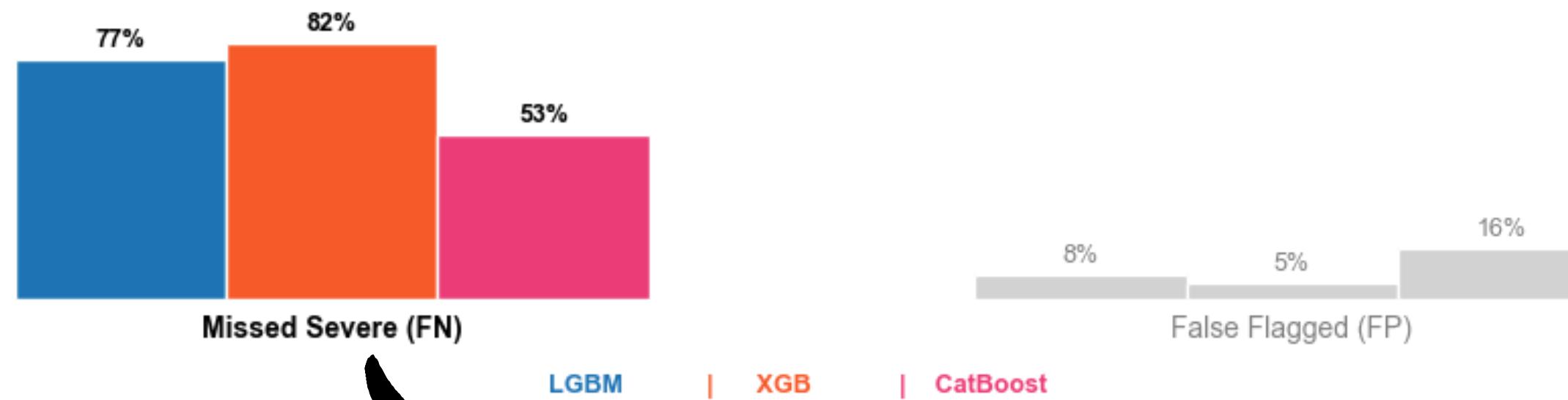


**Micro ROC curves for three models
XGBoost achieves the highest micro AUC (CatBoost & XGB behind)**

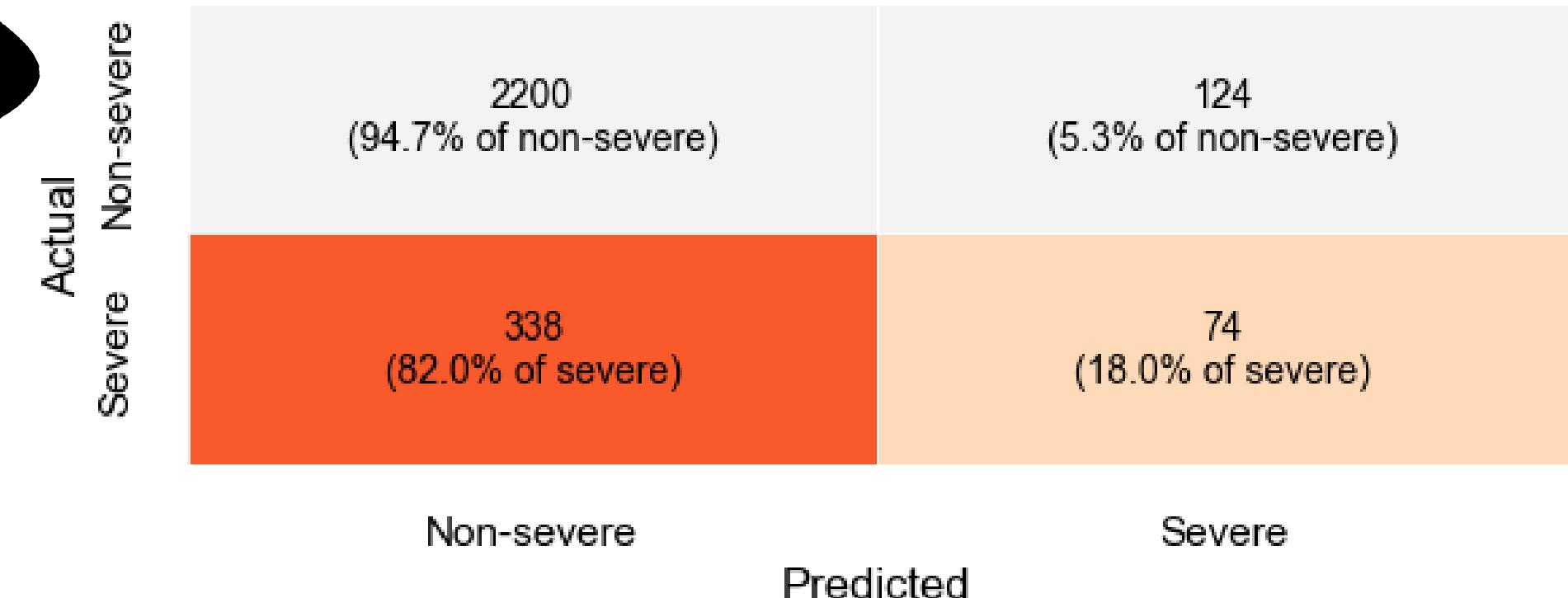


The Most Dangerous Errors Are the Ones We're Making the Most: Missing the Severe Cases

Diagnostic Risk: Models are missing too many Severe cases, while "False Alarms" are low and manageable.



XGBoost misses a meaningful share of severe cases

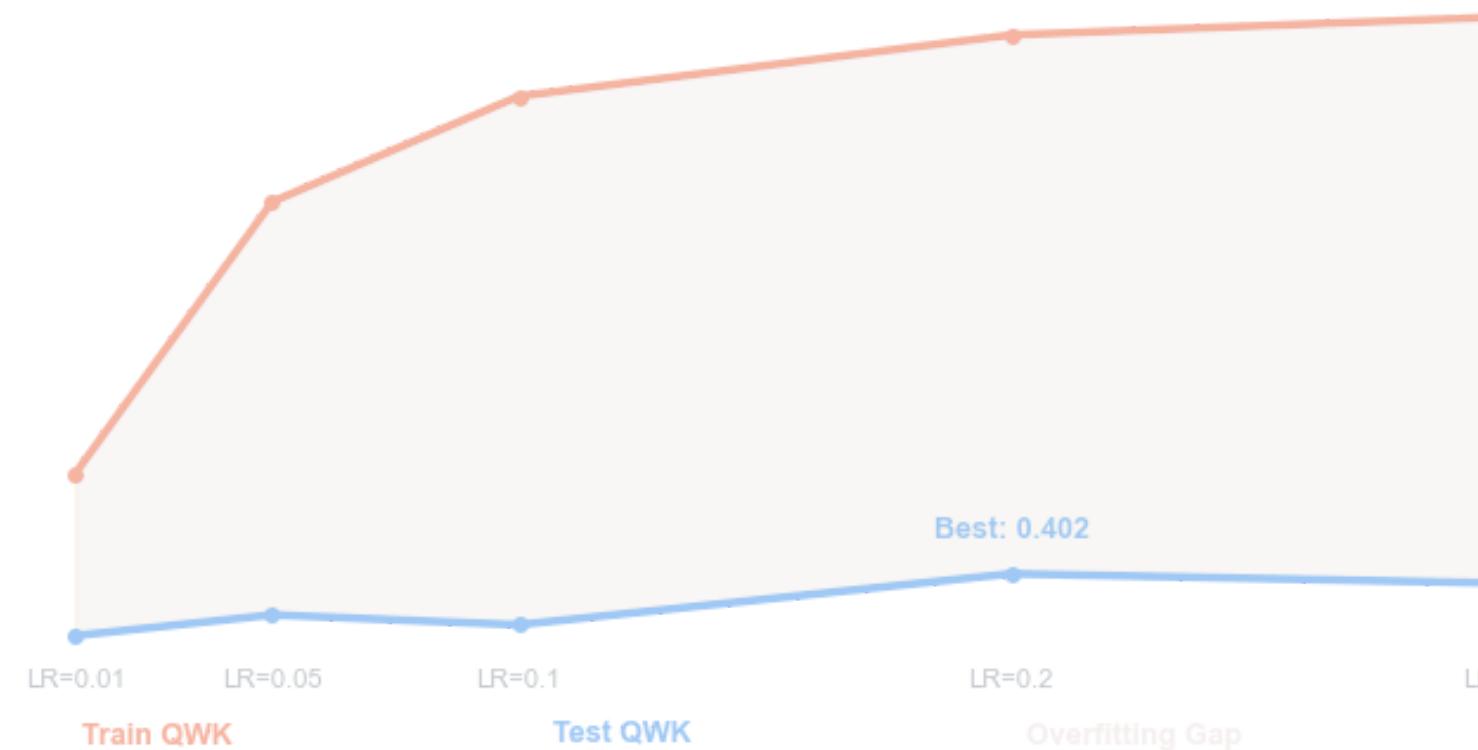


338 severe children missed → 82.0% of true severe (False Negatives)

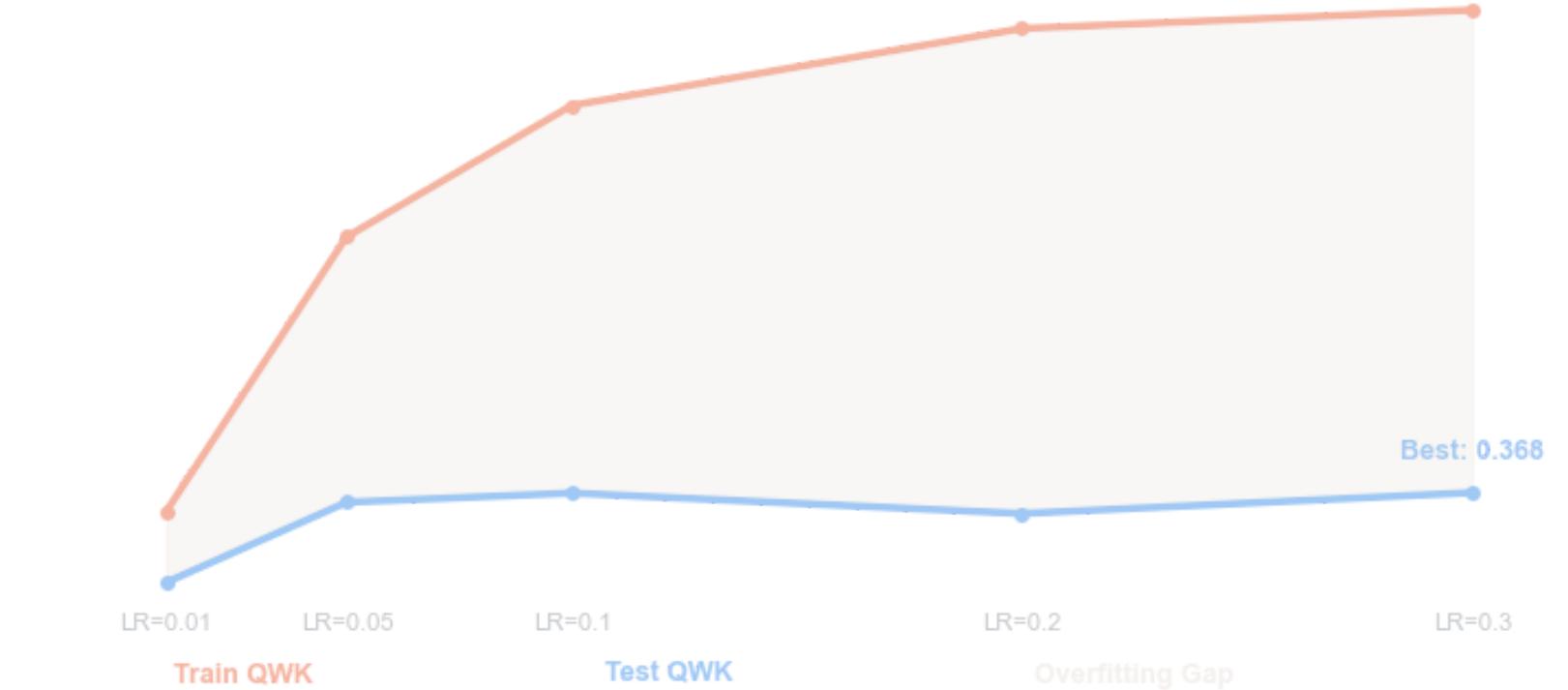
124 normal children flagged severe → 5.3% of true non-severe (False Positives)

All Models Suffer From Significant Overfitting, Especially as Learning Rate Increases.

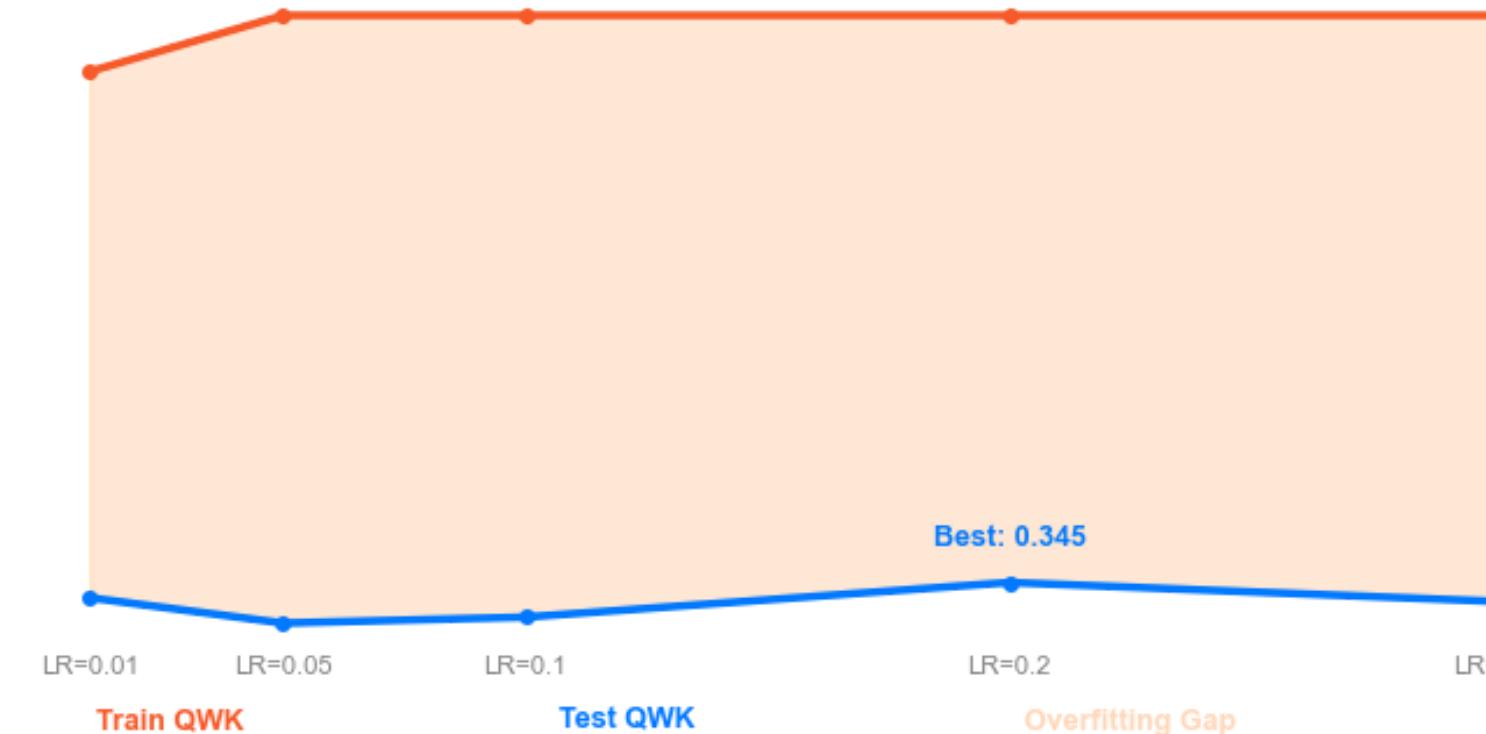
XGBoost is truly overfitting: Overfitting increases significantly when LR is too high



CatBoost: also overfitting



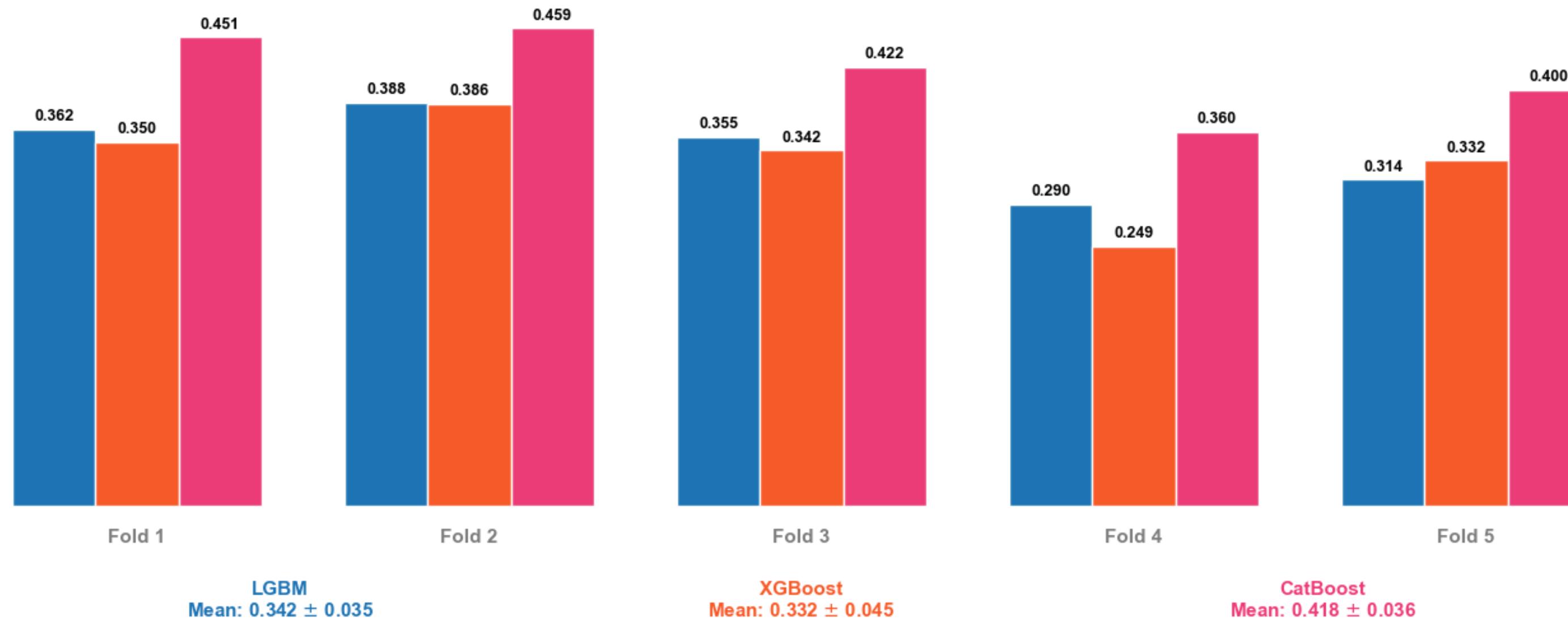
LGBM: Perfectly overfitting!!!



Stability & Performance concerning

Stability & Performance: CatBoost leads in both score and consistency

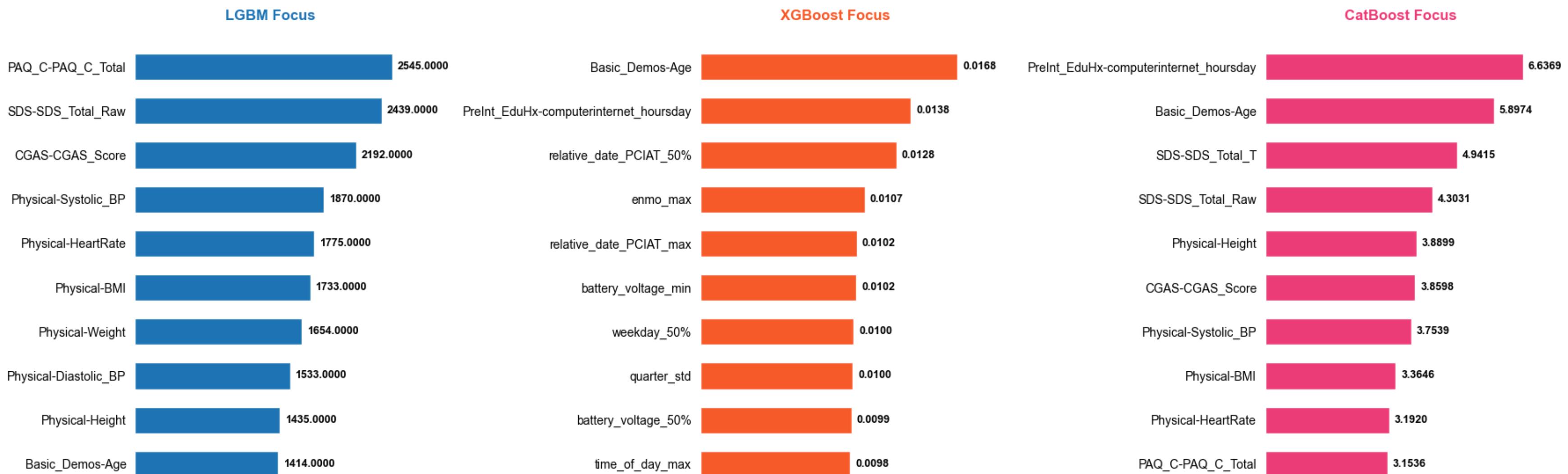
Side-by-side comparison per fold. Note how all models dip together at difficult folds (e.g., Fold 4).



Feature Importance concerning

Feature Divergence: Models prioritize significantly different features

Top 10 most important features for LGBM, XGBoost, and CatBoost do not overlap consistently.



Resolution phase

Where the problems in the data begin to resolve, and the story becomes clearer.

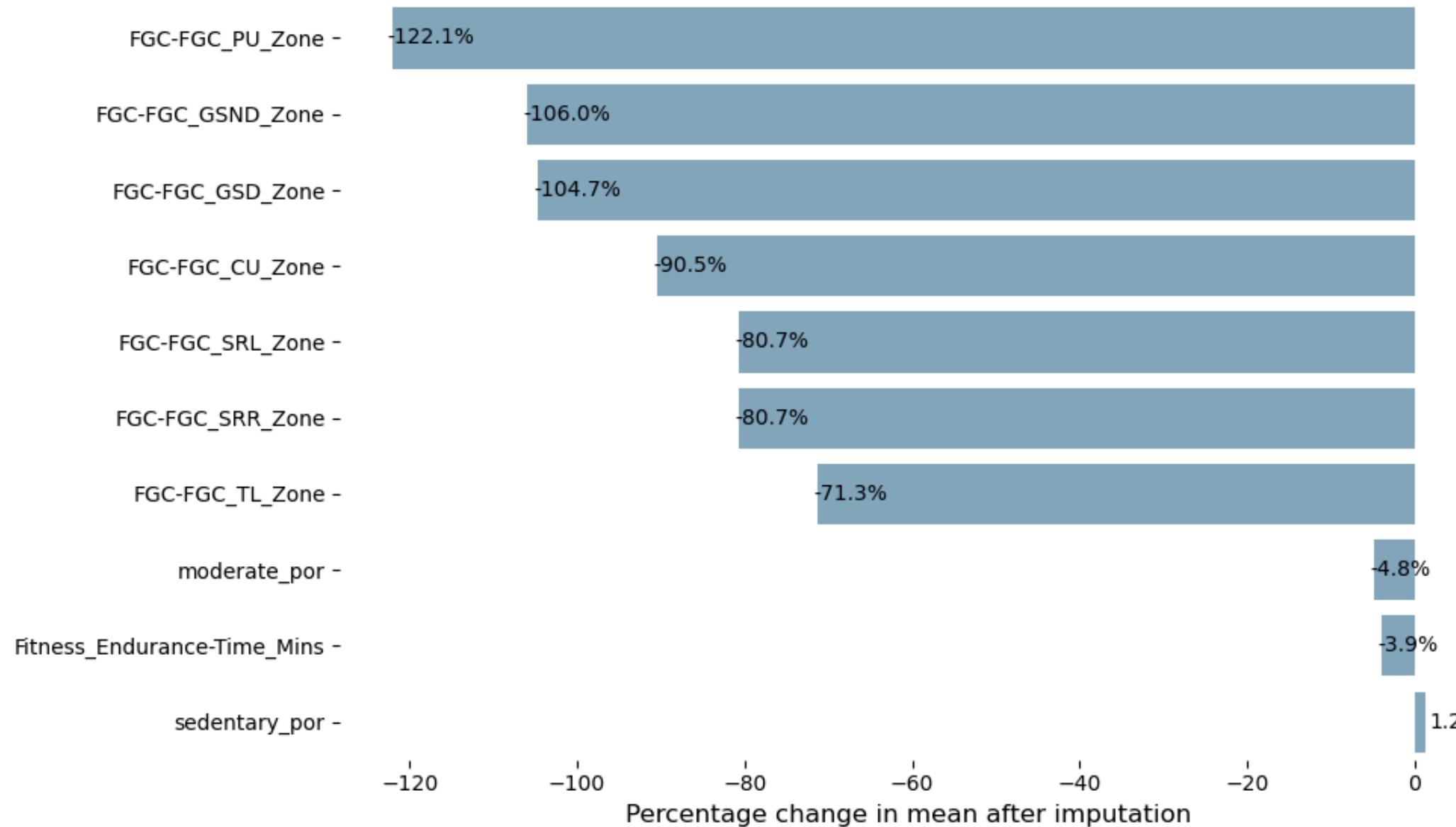
After cleaning

After data cleaning

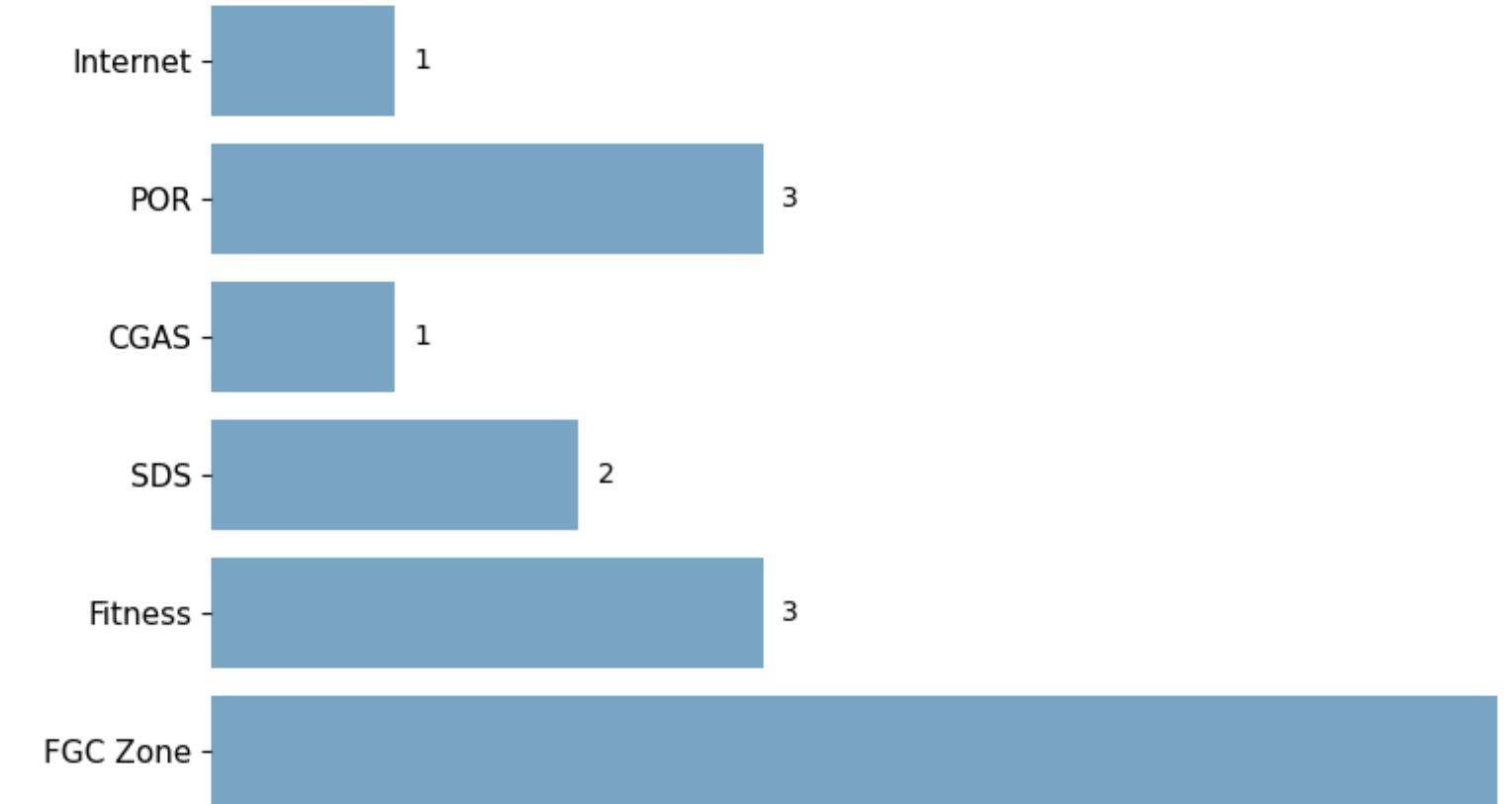
0

columns have missing values
and outliers

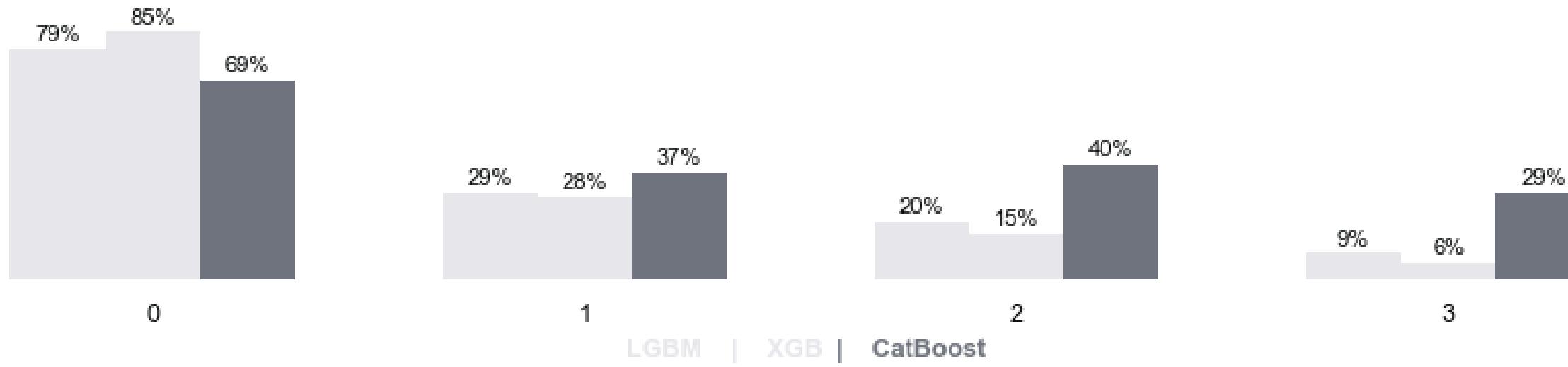
Top 10 variables most affected by imputation: Mean mostly decreases



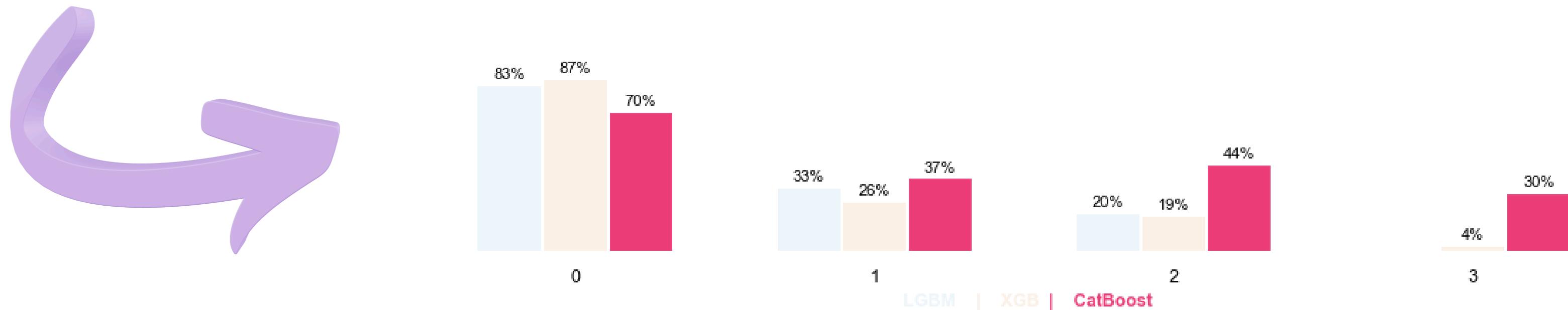
17 features use missing lags across different groups



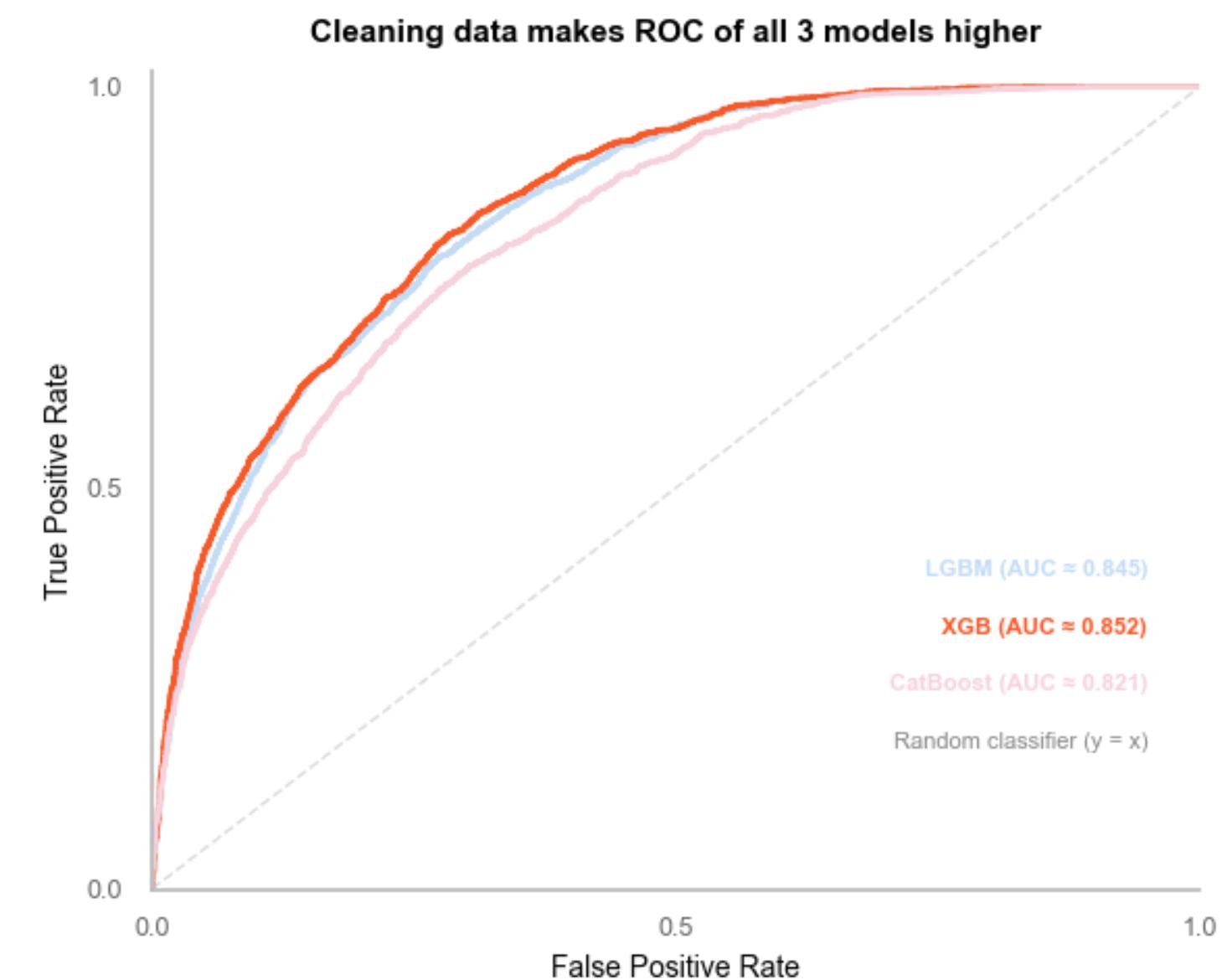
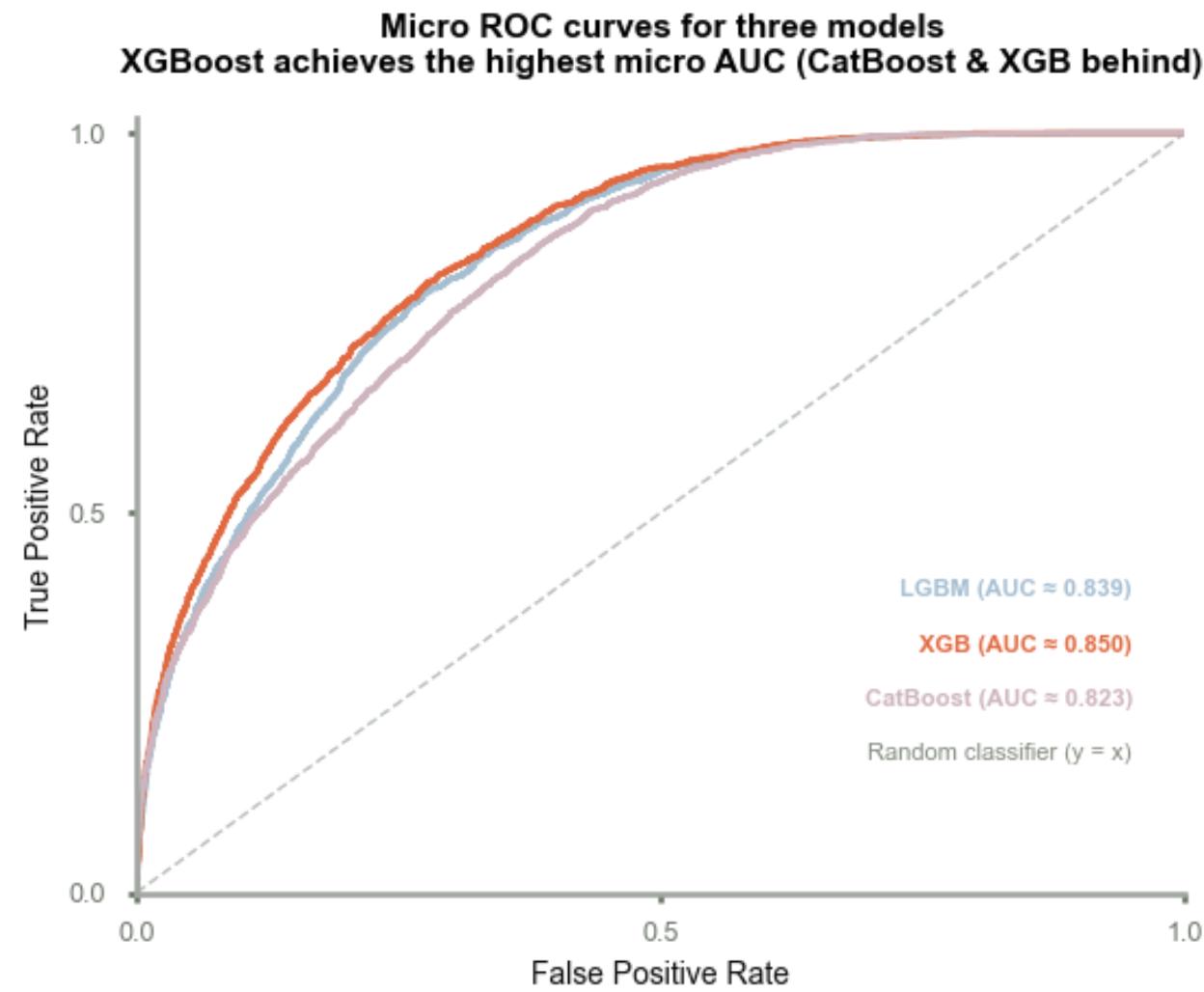
After cleaning



Recall higher after cleaning

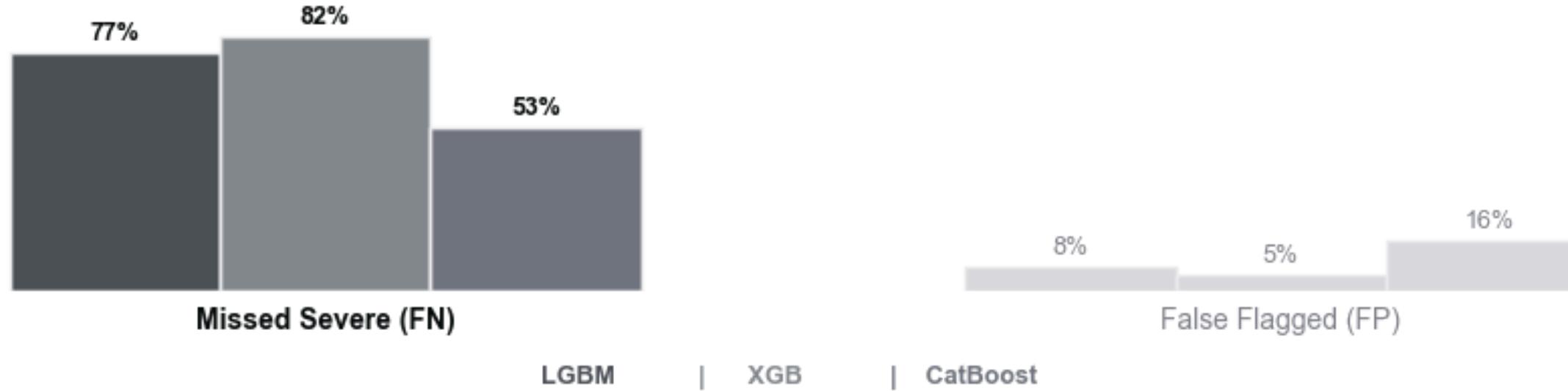


After cleaning

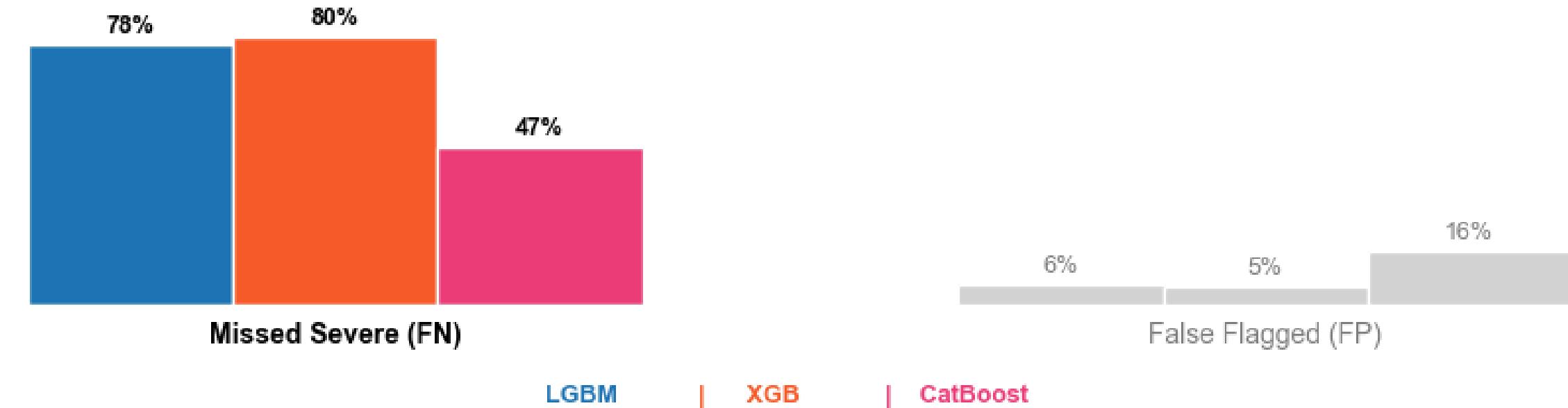


After cleaning

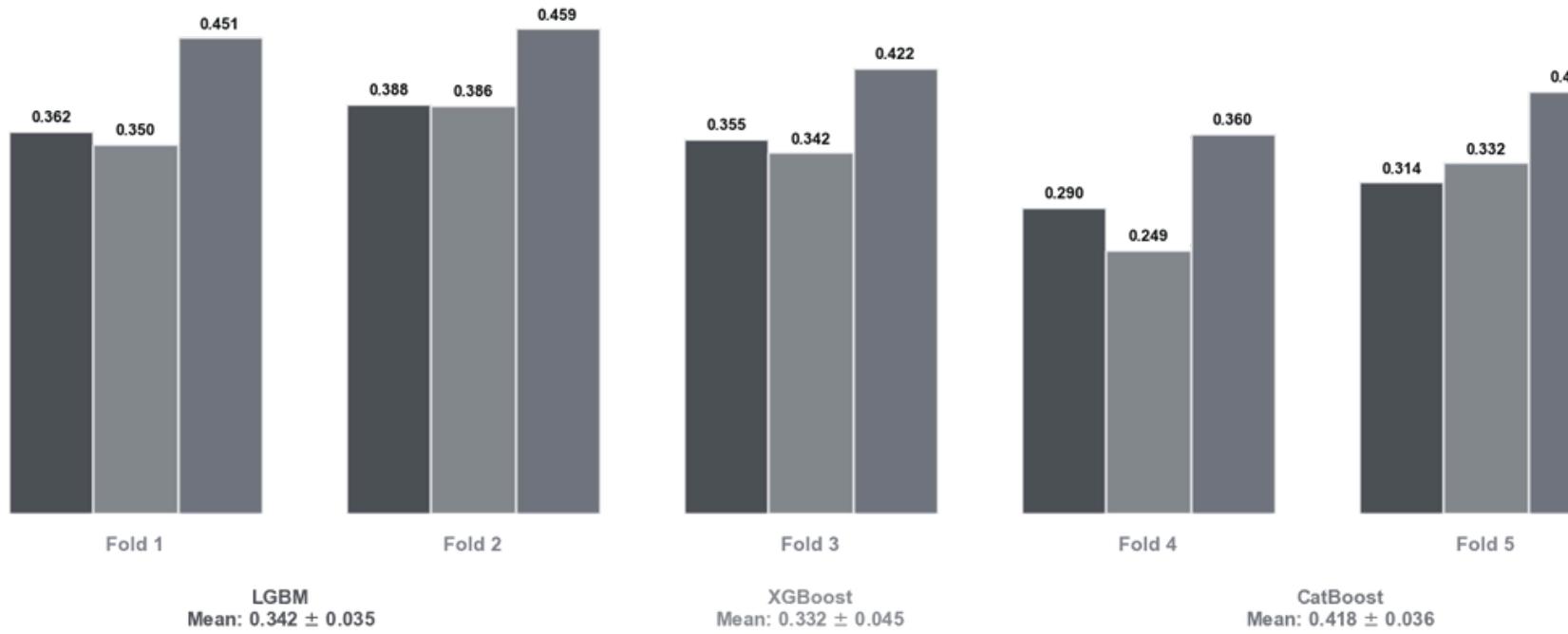
Diagnostic Risk: Models are missing too many Severe cases, while "False Alarms" are low and manageable.



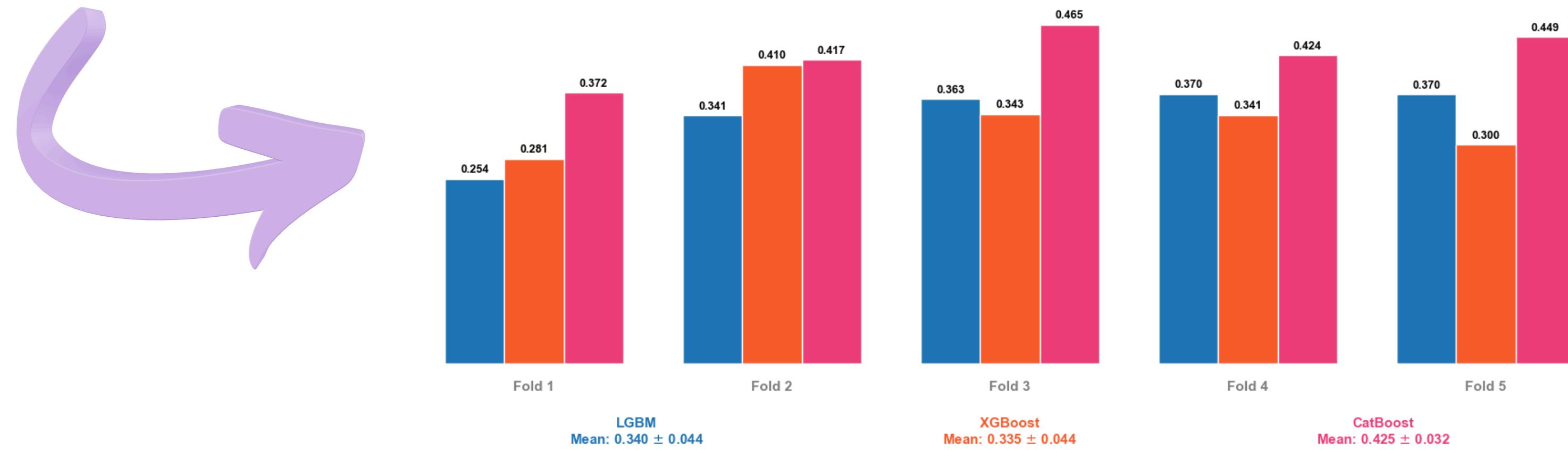
Diagnostic Risk reduces in almost models after data cleaning.



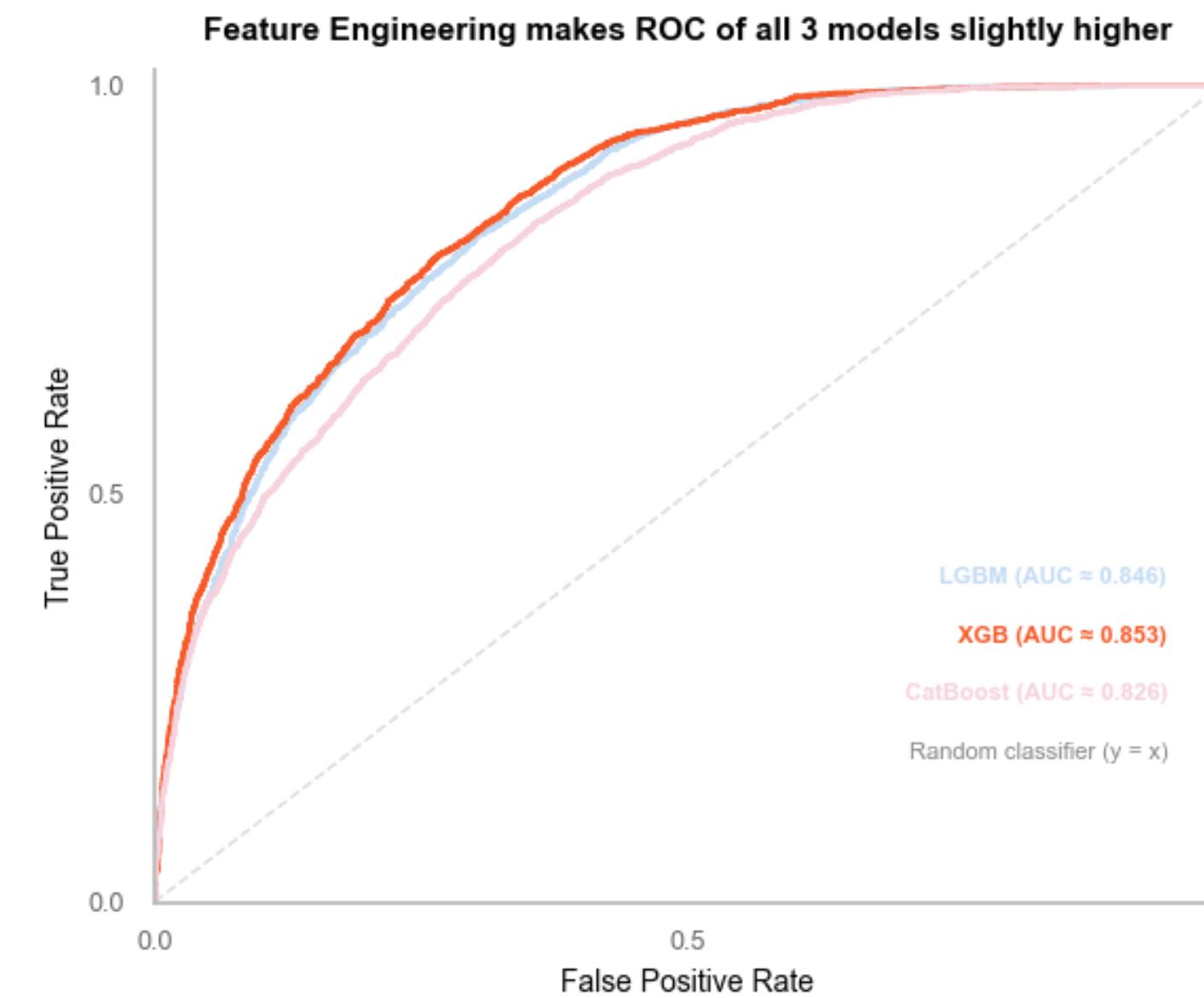
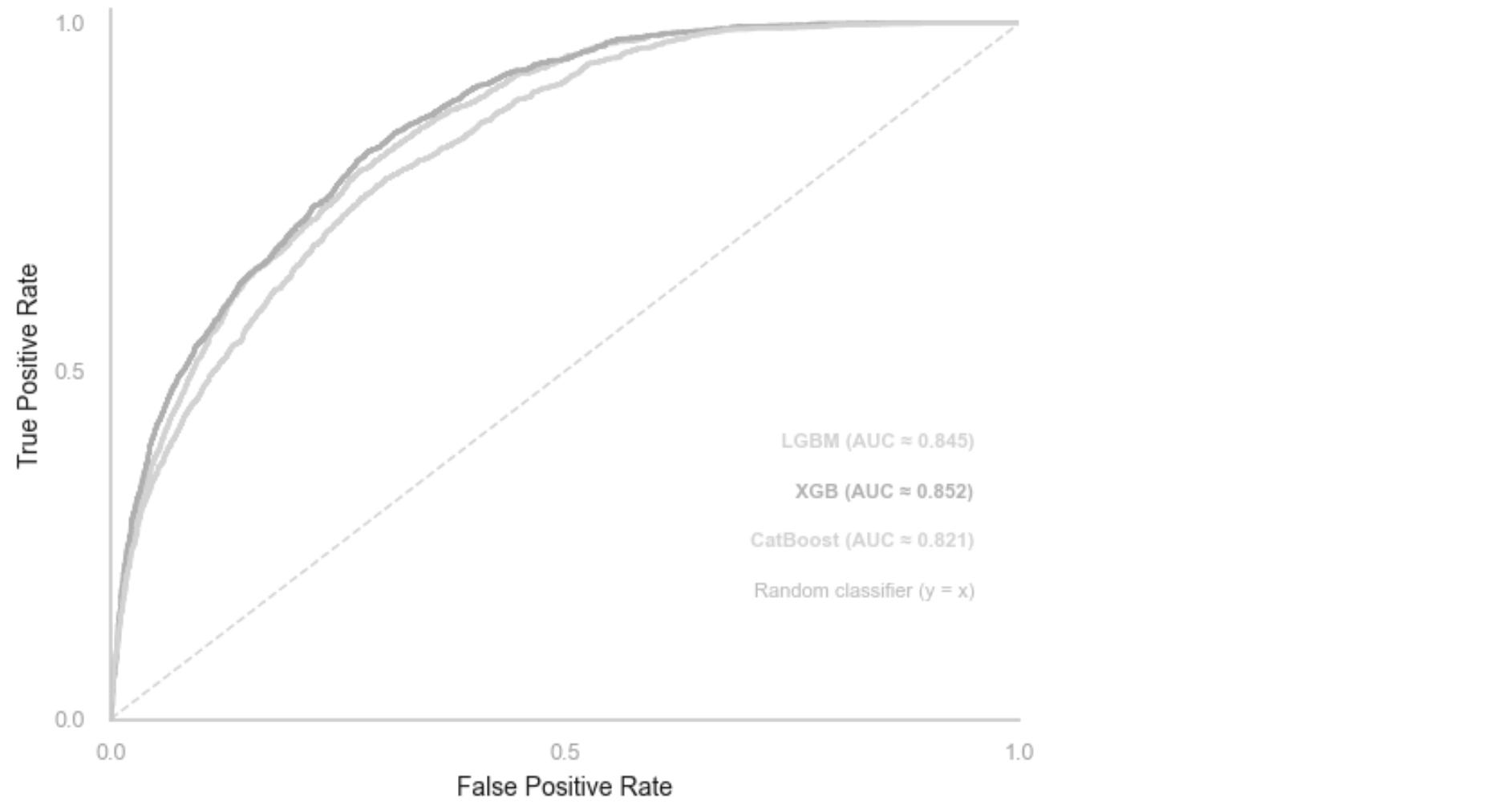
After cleaning



After data cleaning, XGBoost and CatBoost stable and better, however, LGBM faces downgrade in training



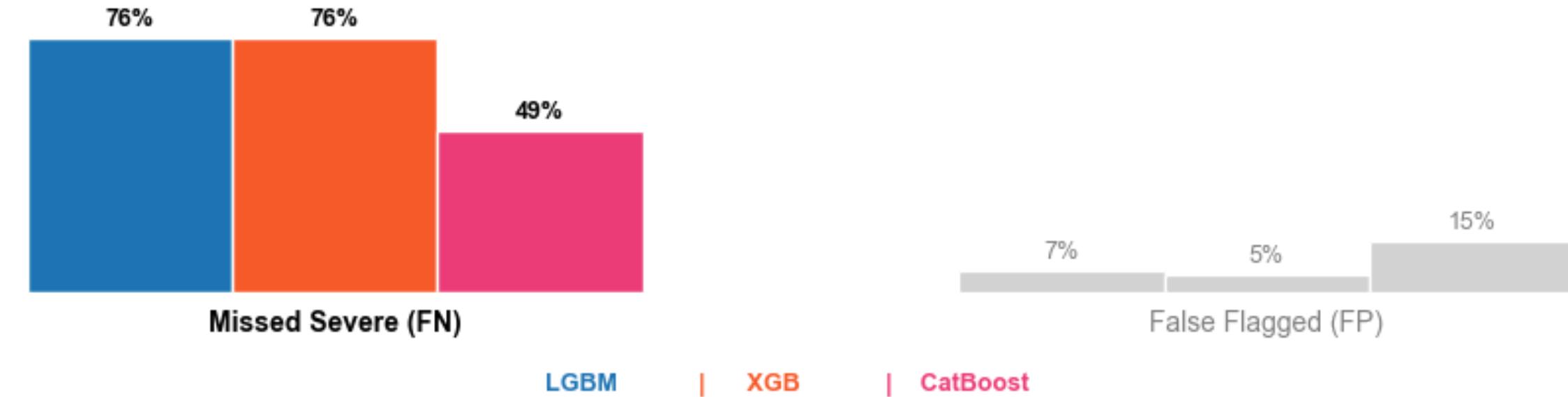
After feature engineering



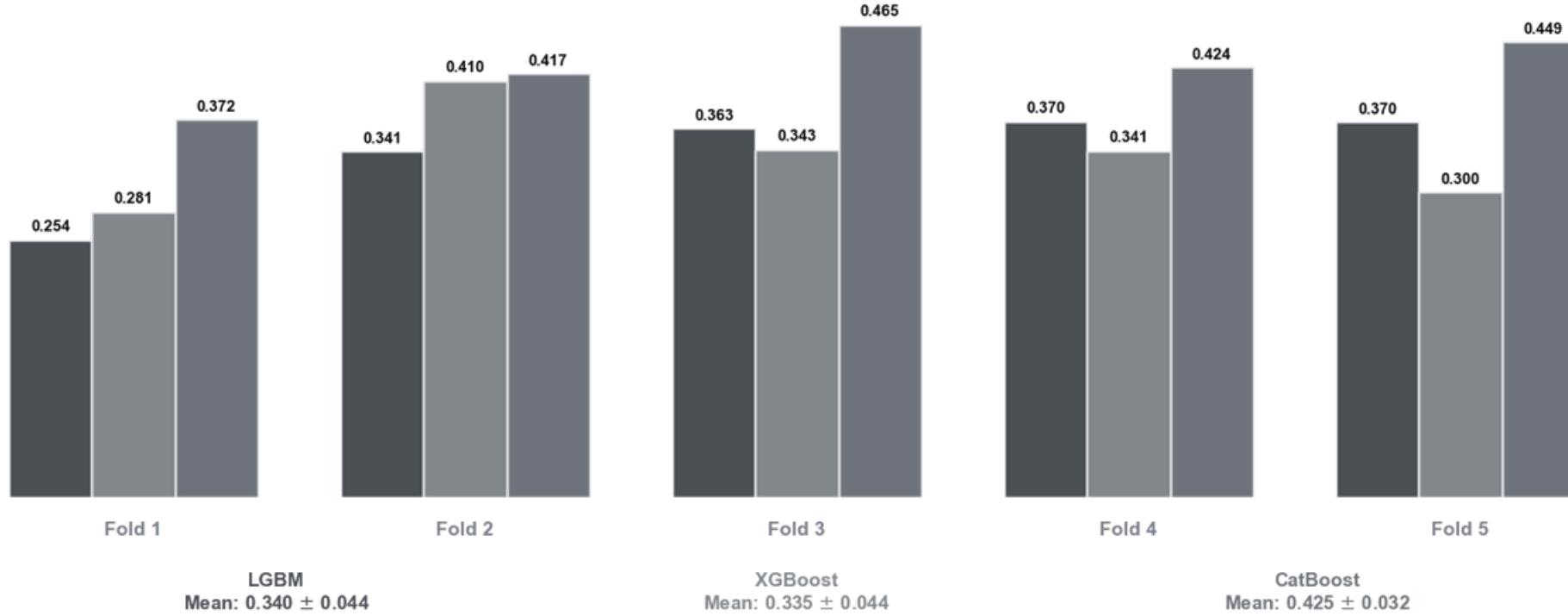
After feature engineering



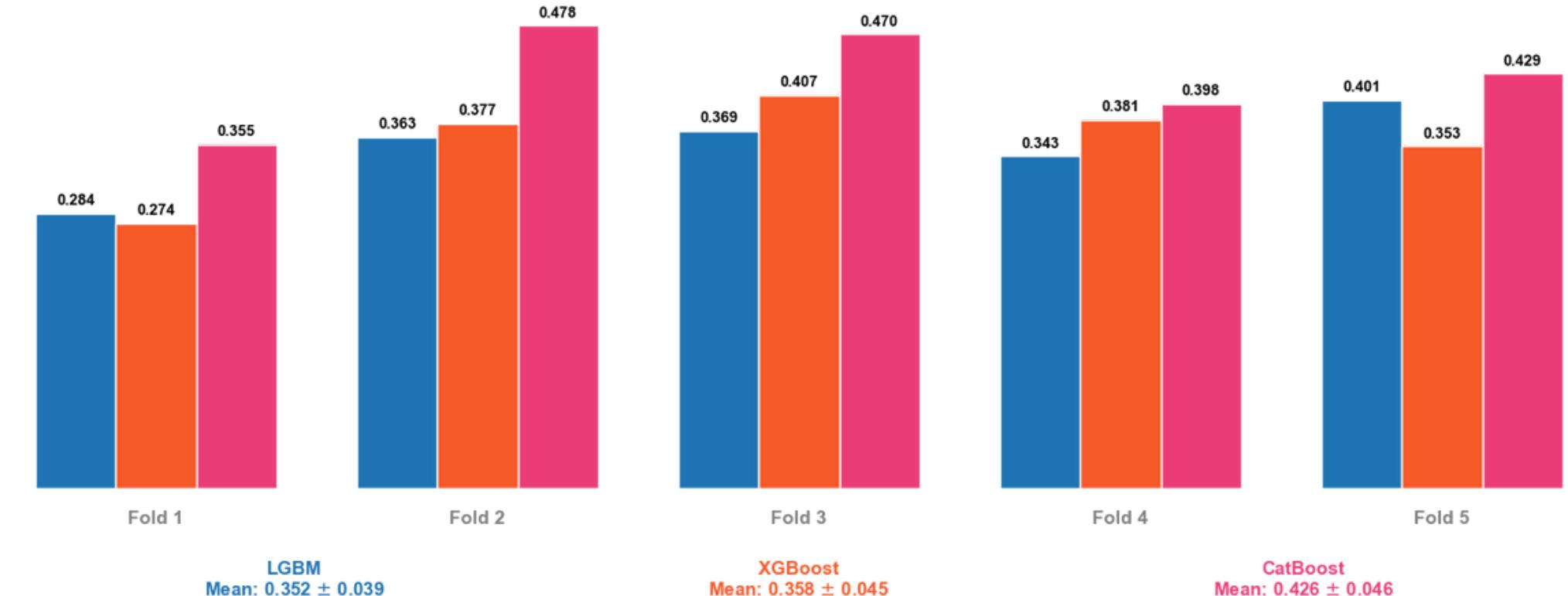
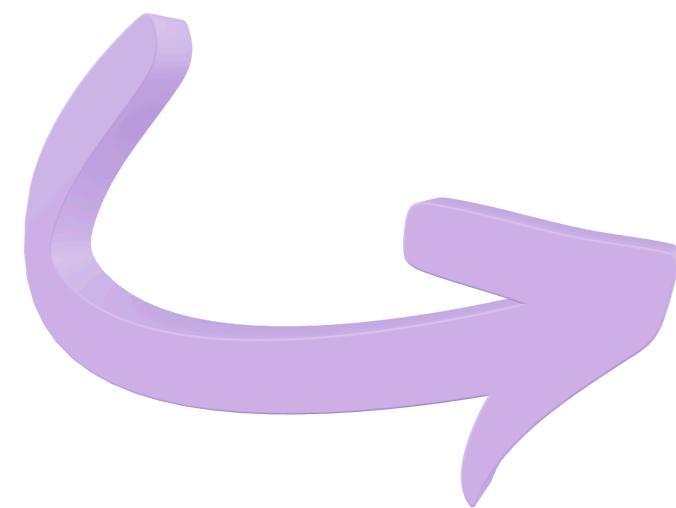
Diagnostic Risk continually reduces in LGBM and XGB models after feature engineering.



After feature engineering

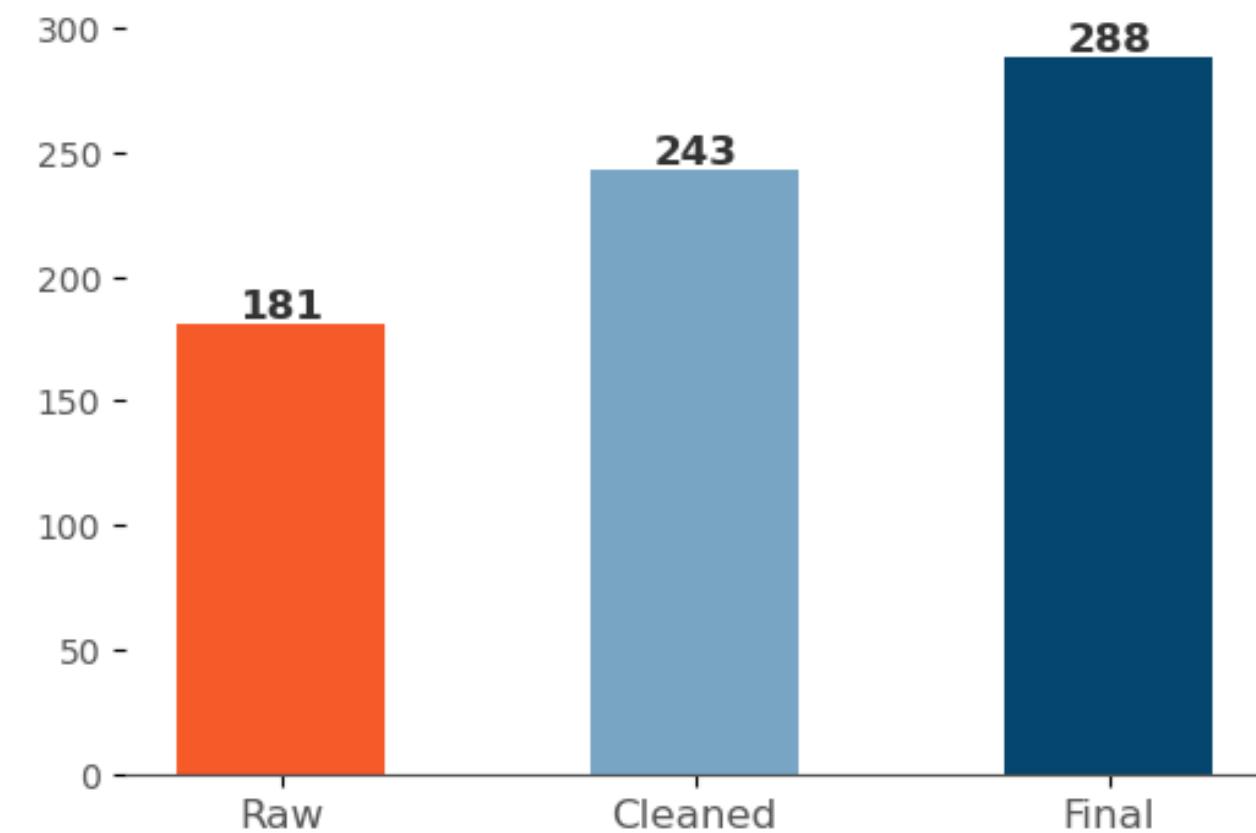


Affter feature engineering, all of 3 models more stable and better

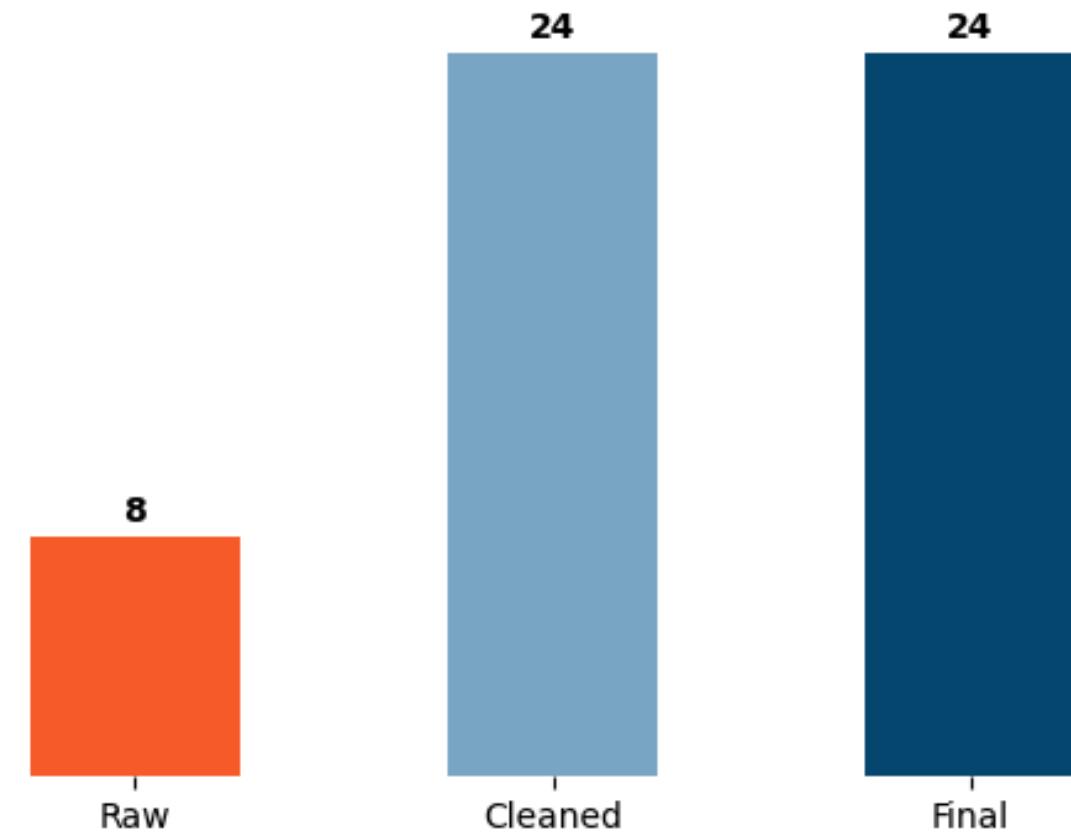


After Data Transform

Number of variables increased sequentially through data processing

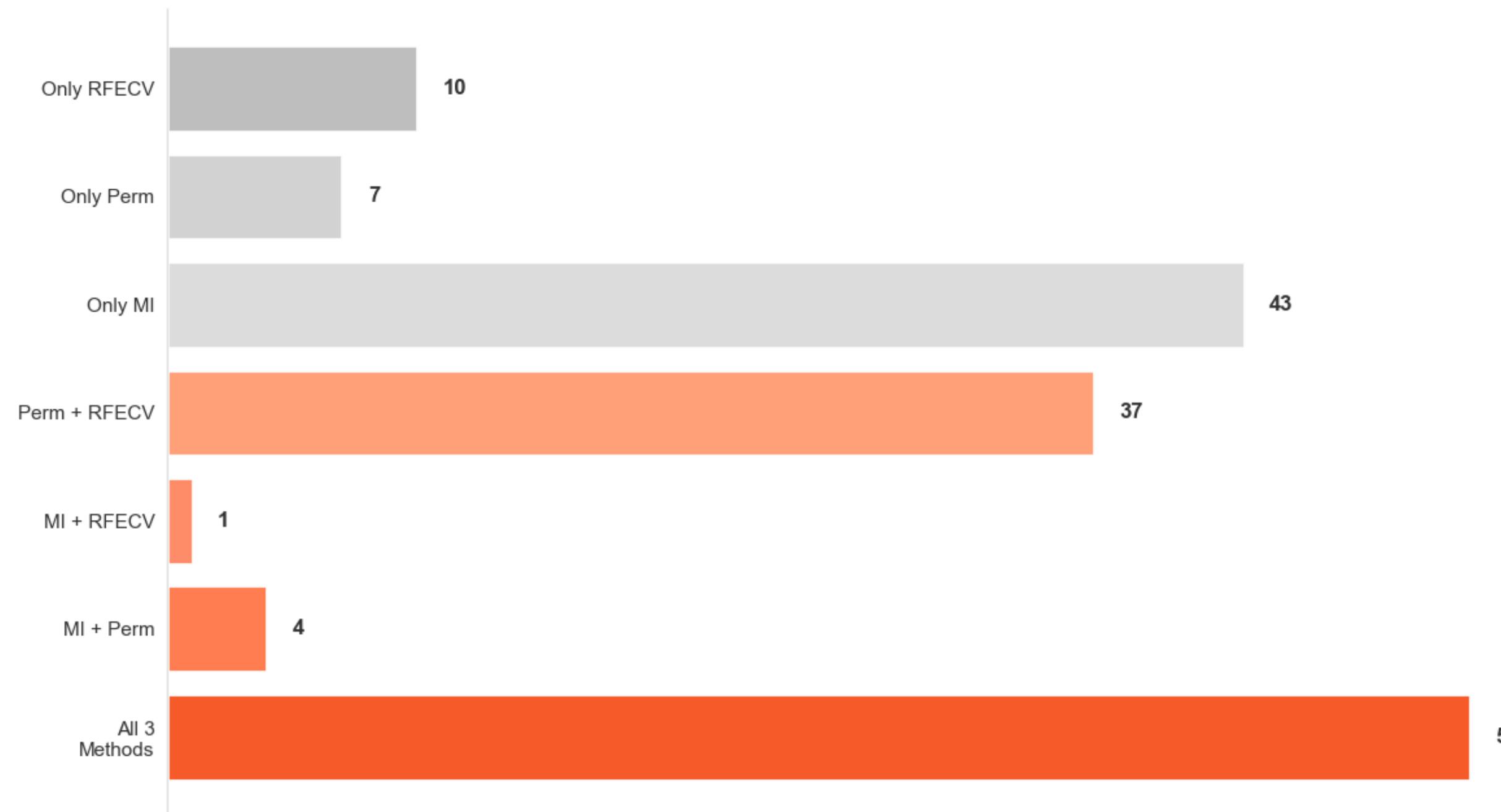


Number of variables with correlation > 0.2 increased after cleaning

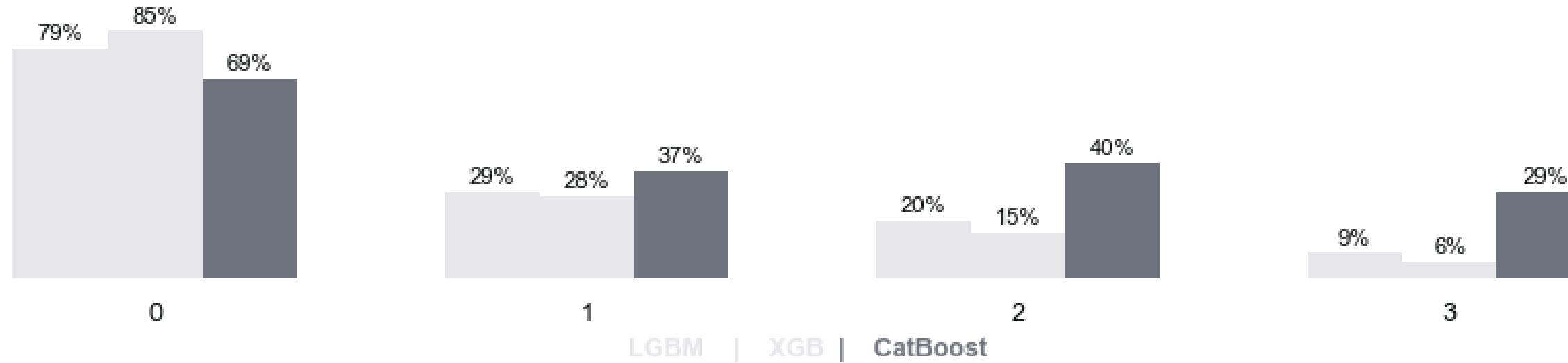


Feature Selection Results (`n_features_choose = 100`, for each method)

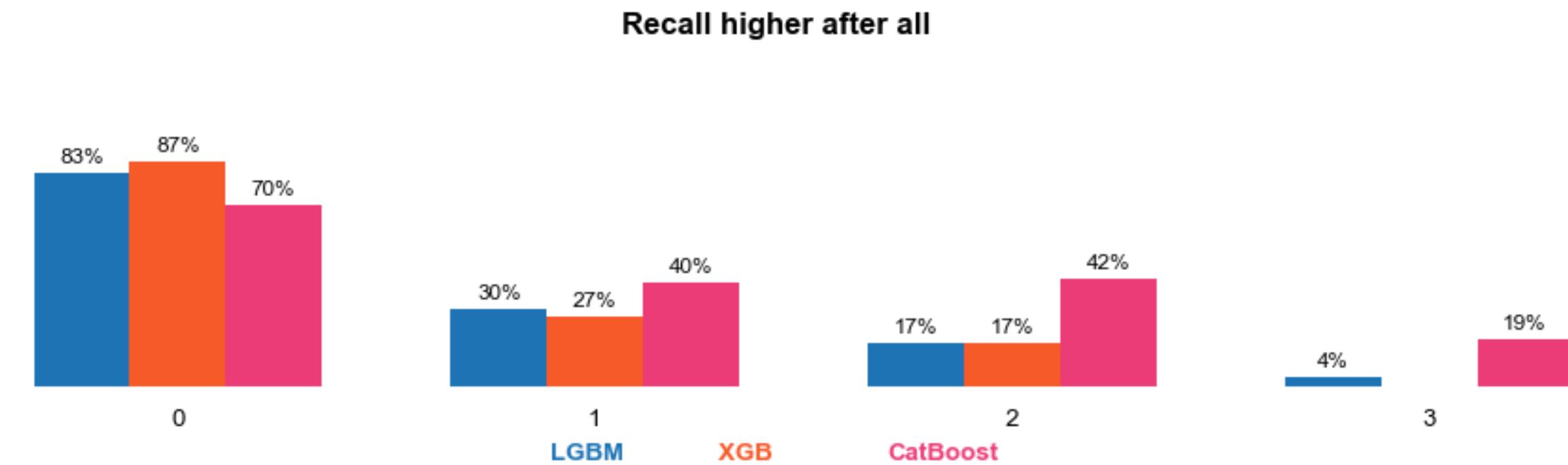
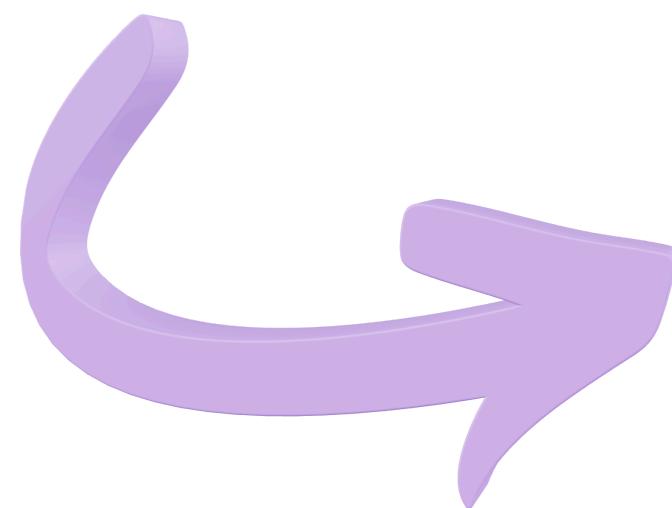
Most Features Show Method-Specific Importance Rather Than Universal Consensus



Final model after all the data preparation flow

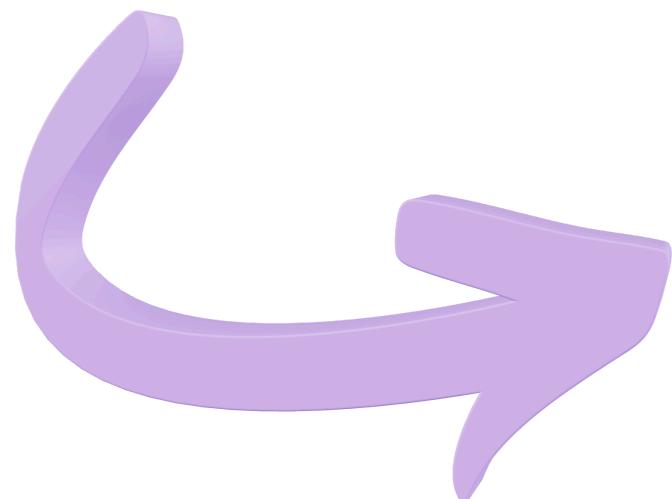
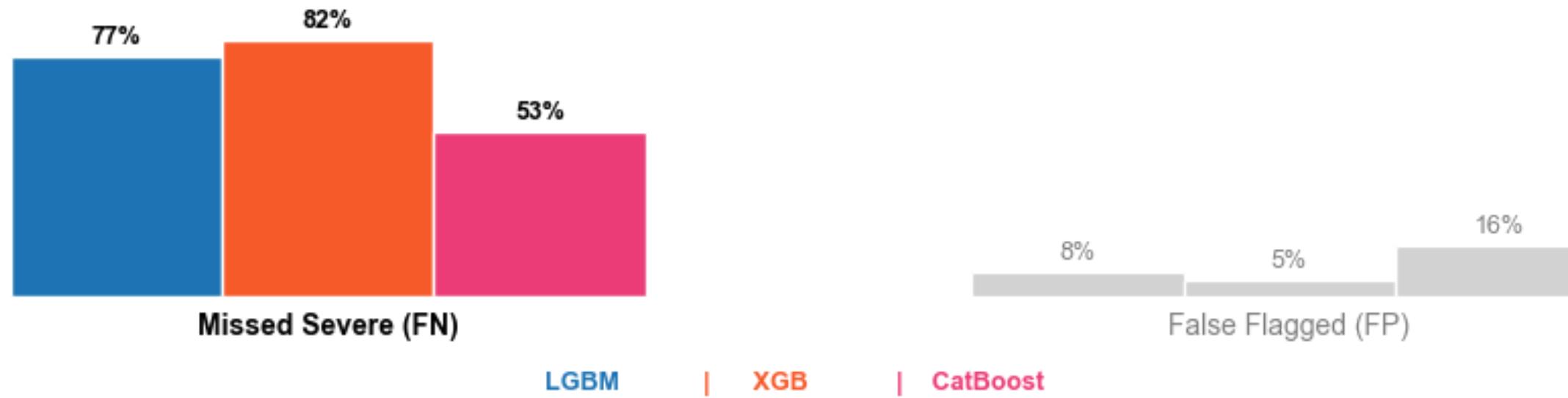


Raw training

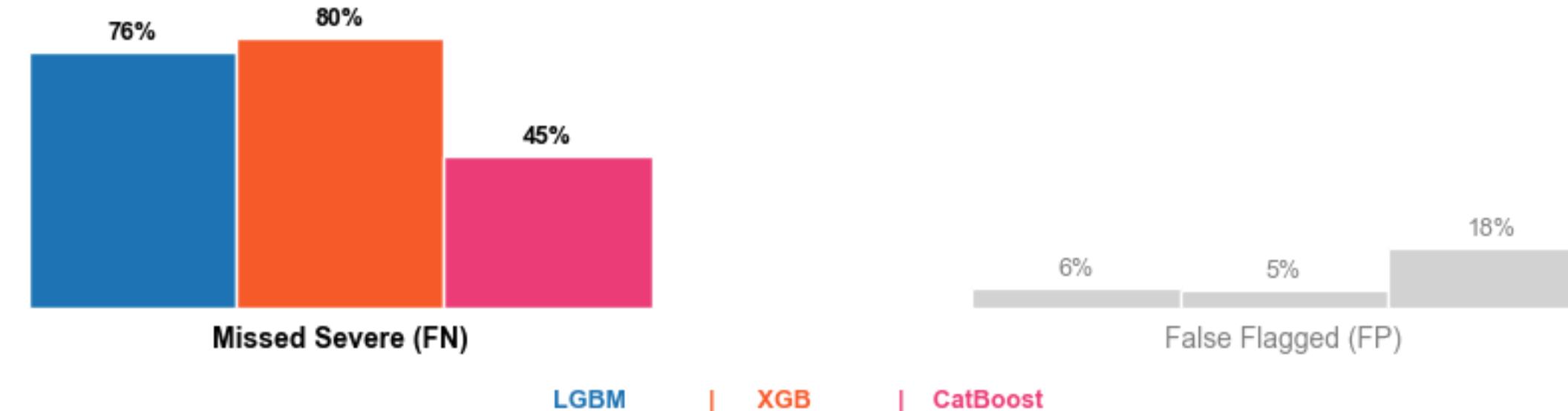


Prepared training

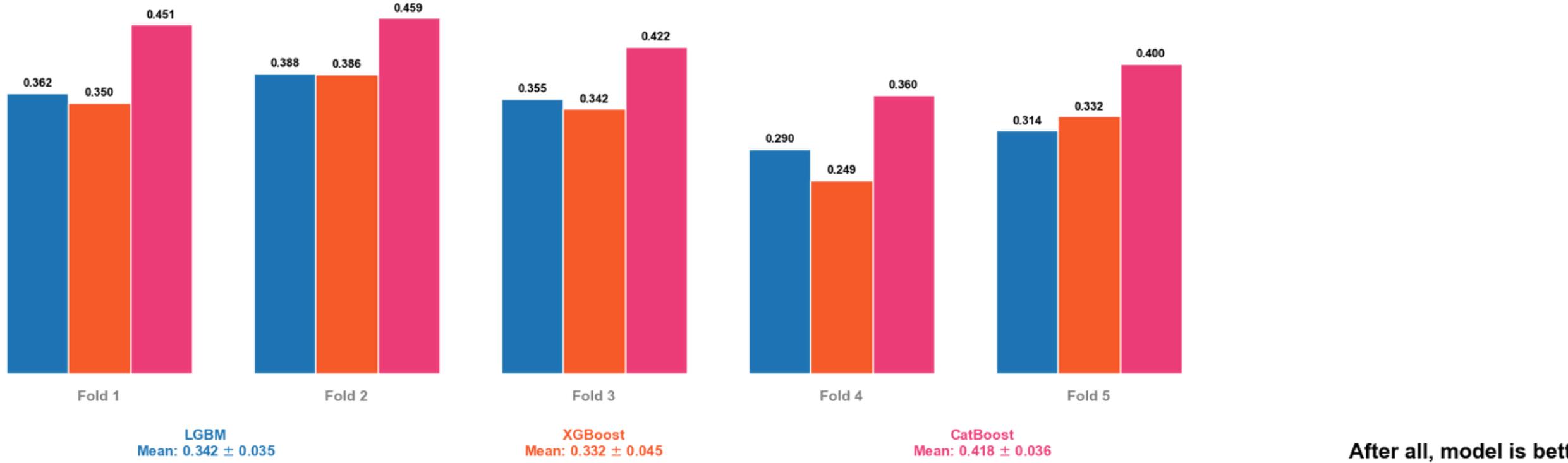
Final model after all the data preparation flow



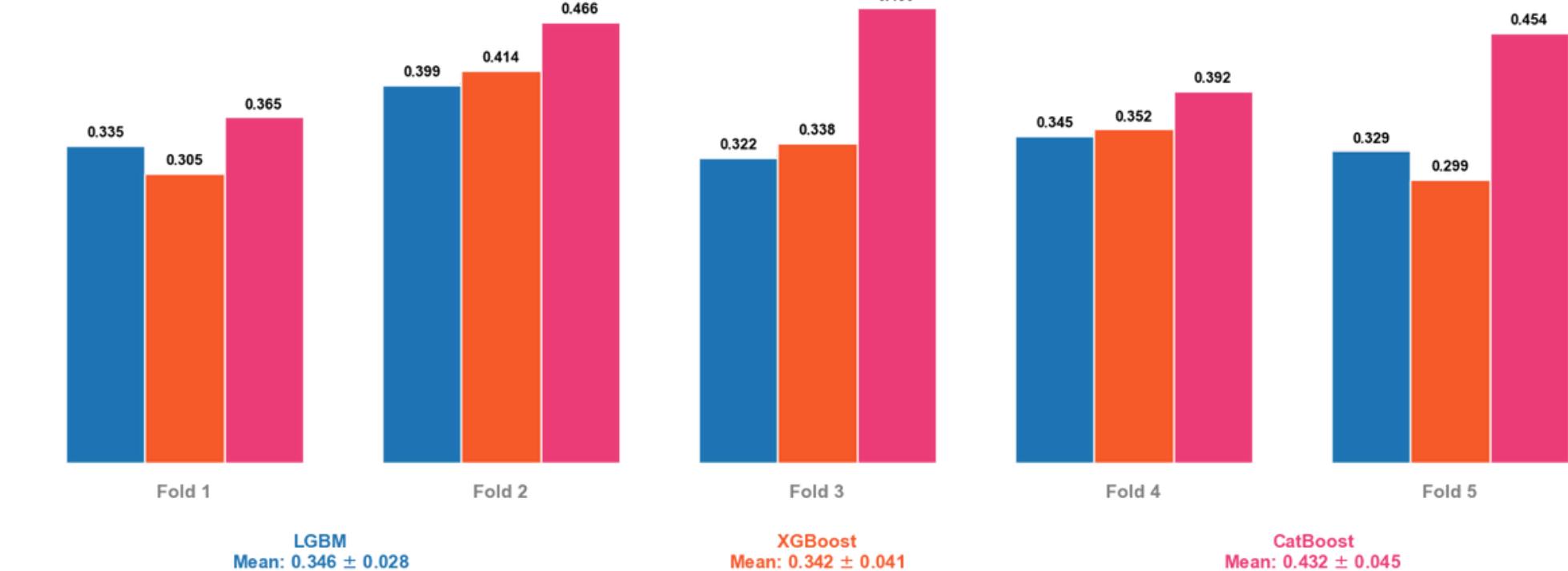
Diagnostic Risk reduces after all.



Final model after all the data preparation flow

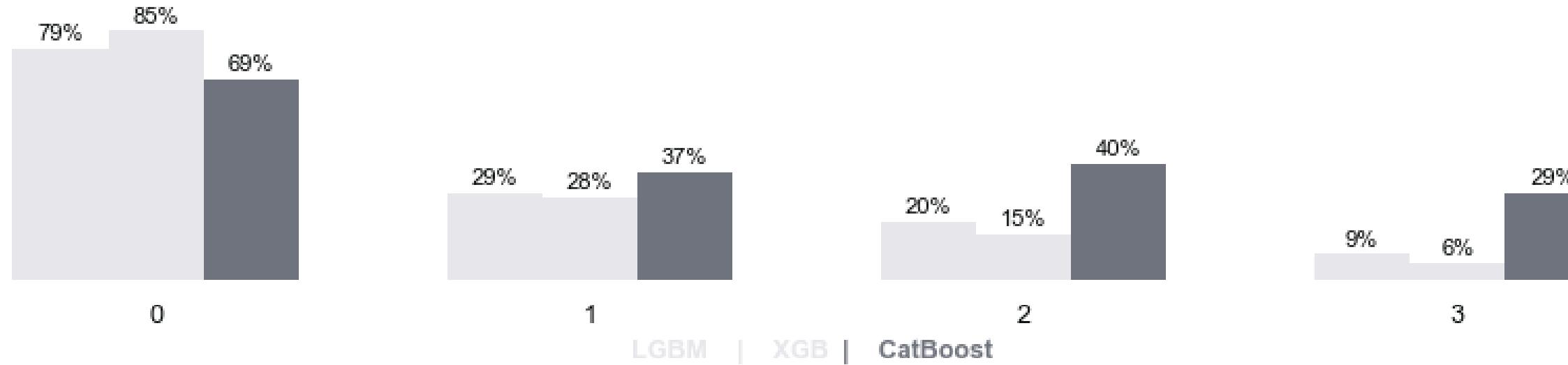


Raw training

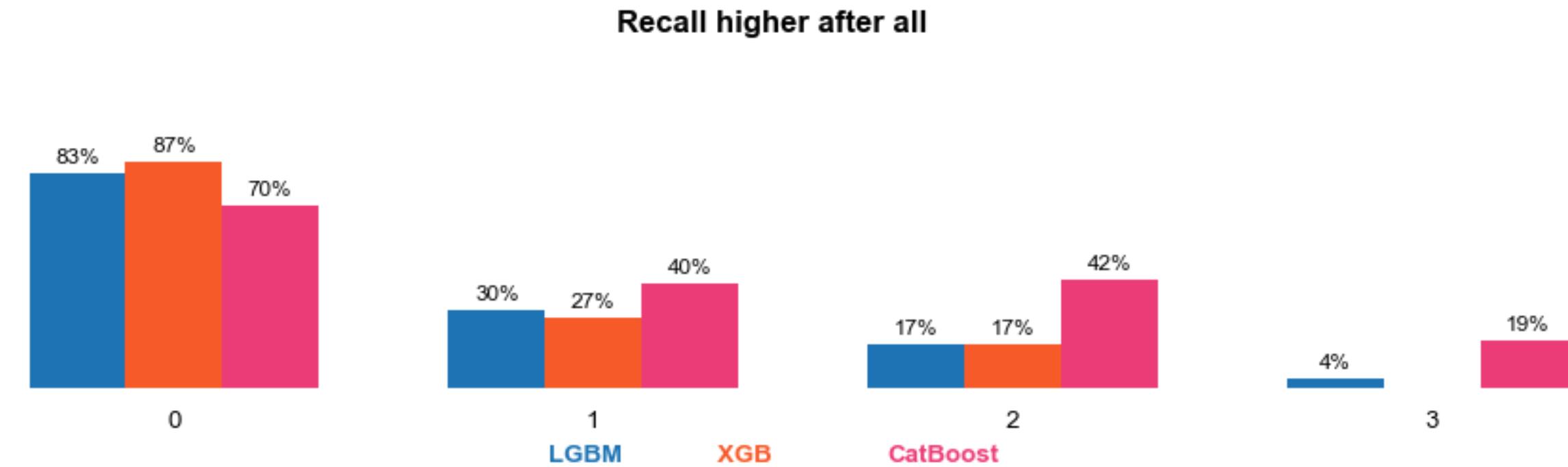
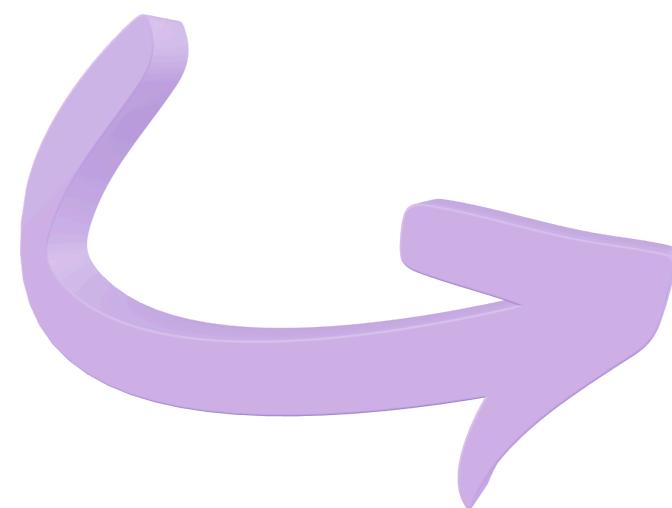


Prepared training

Final model after all the data preparation flow



Raw training

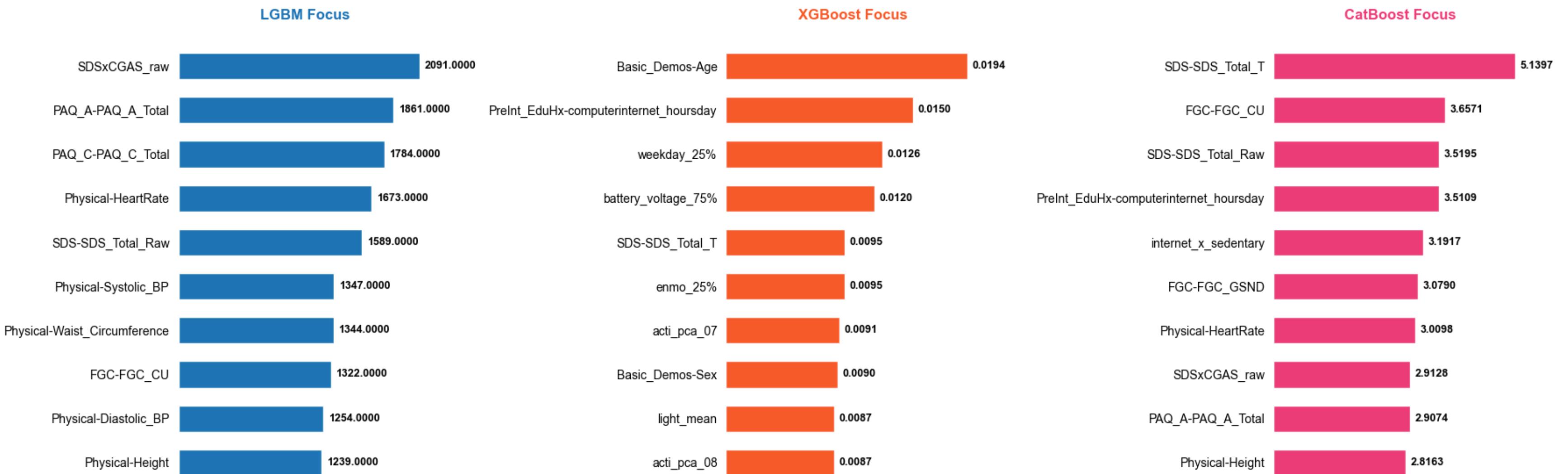


Prepared training

Final model after all the data preparation flow

Final Feature Importance after all

Top 10 most important features for LGBM, XGBoost, and CatBoost do not overlap consistently.



Thank you for
watching our story