INDIAN DATA CLUB databricks CODE BASICS

DAY 11

THE PIVOT POINT

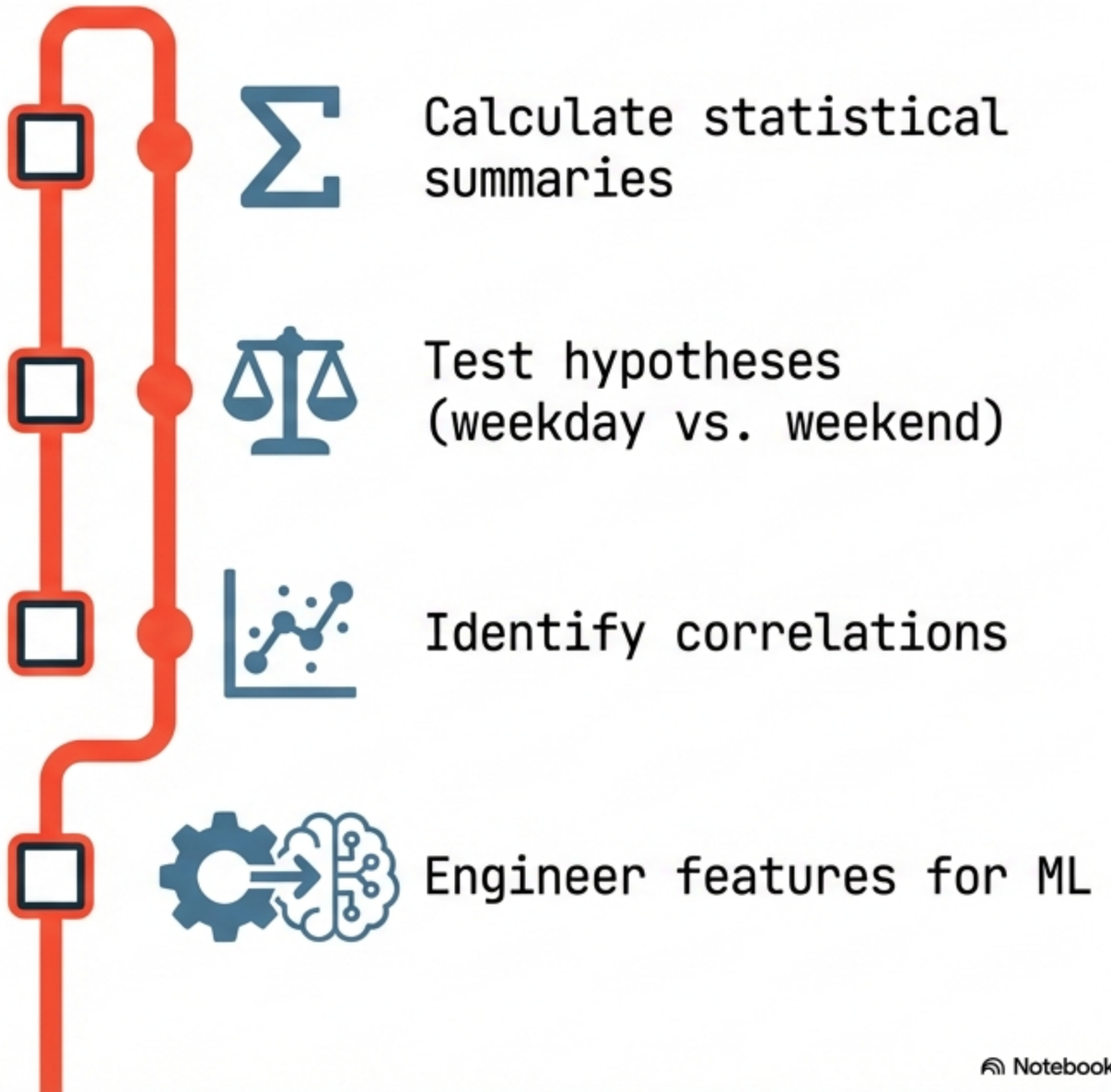
Statistical Anvalysis & ML Prep

14 DAYS AI CHALLENGE // PHASE: BRIDGE TO INTELLIGENCE

Bridging Engineering and Intelligence

We have spent the last 10 days ingesting and cleaning data. Today, we stop managing the data and start questioning it. Day 11 is about rigor—sharpening our tools to ensure the models we build tomorrow are based on truth, not noise.

THE AGENDA



Objective 1: Establishing the Baseline

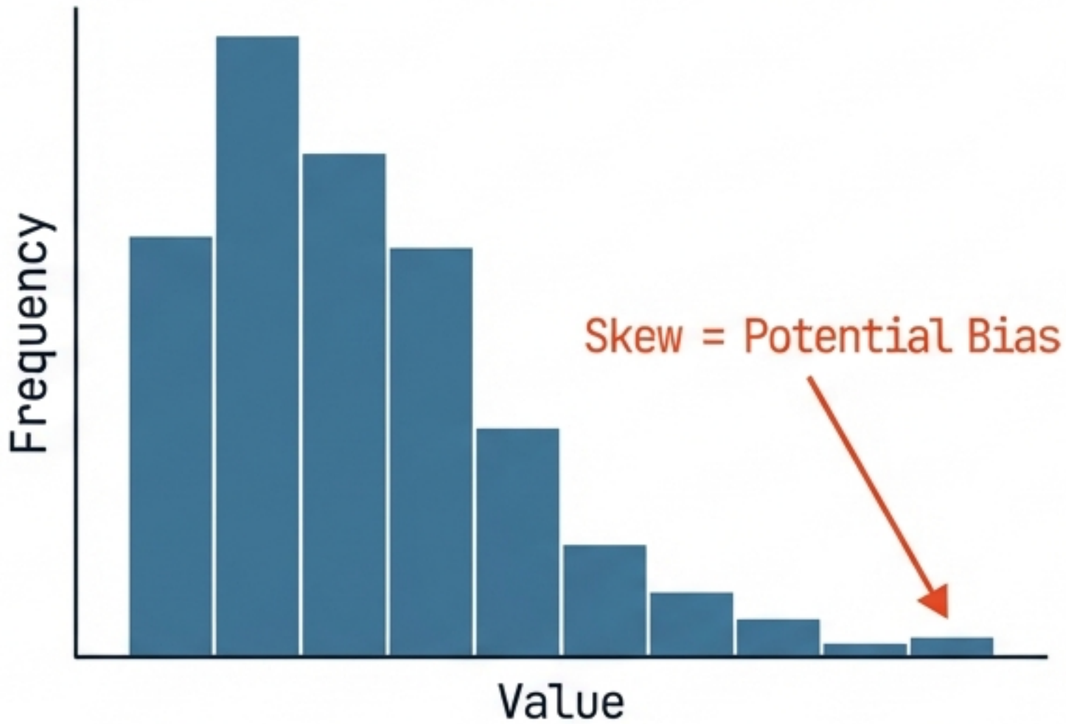
Calculate Statistical Summaries

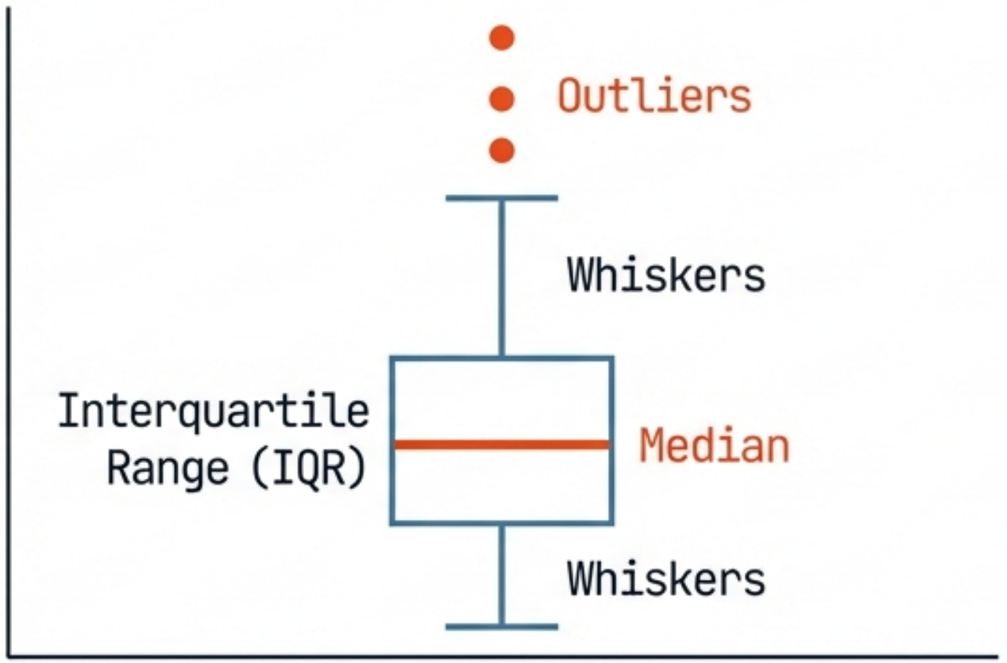
	count	mean	stddev	min	25%	50%	75%	max
transaction_value	1,000,000	120.50	45.20	10.00	85.00	110.00	150.00	5,000.00
user_age	1,000,000	35.5	10.1	18	28	35	42	75
session_duration	1,000,000	245.0	120.5	30	150	210	300	900

The Insight: Averages hide the truth.

We use Spark's `summary()` to spot zero-variance columns and extreme outliers before they break the model.

Visualizing the Distribution





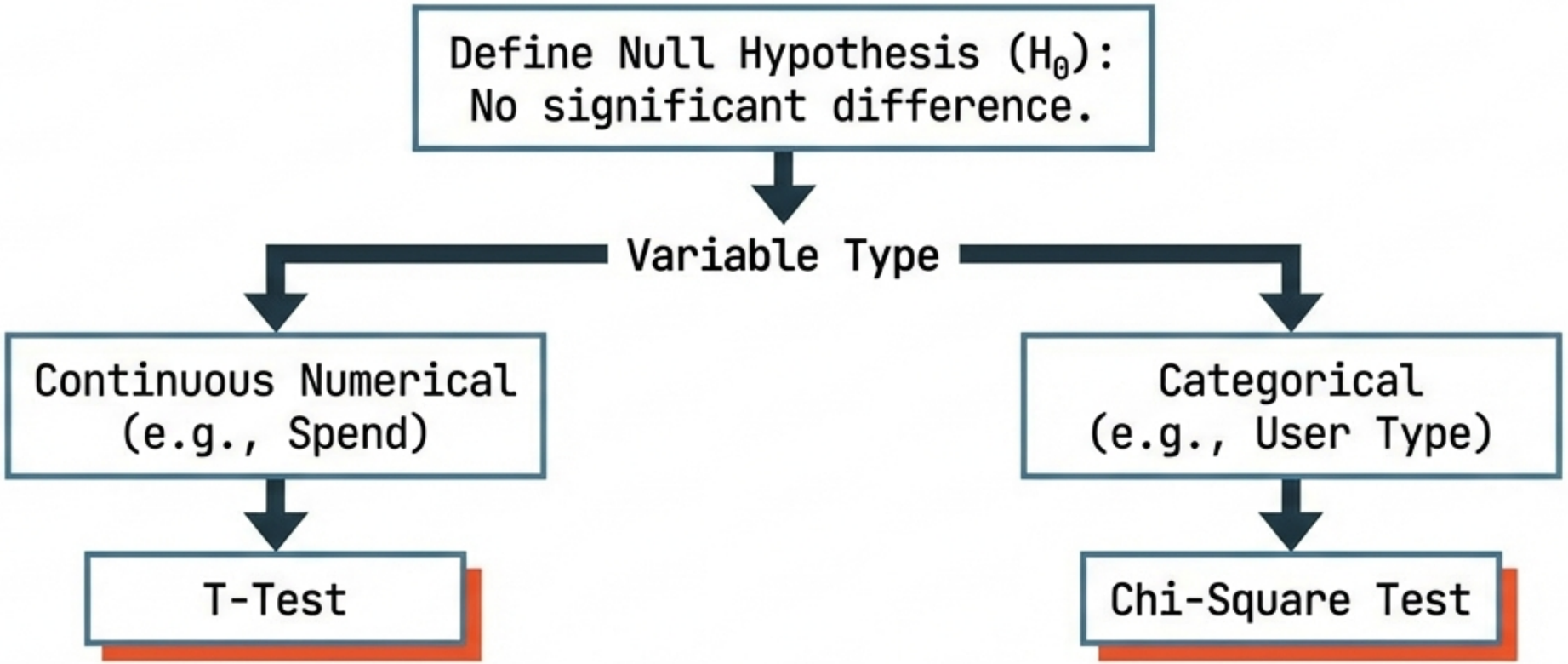
Numbers in a table can be abstract. Visualizing **central tendency** and **dispersion** brings the data's **"shape"** into focus. Skewed distributions often require **log-transformation**.

Objective 2: Challenging Assumptions

Test Hypotheses (Weekday vs. Weekend)

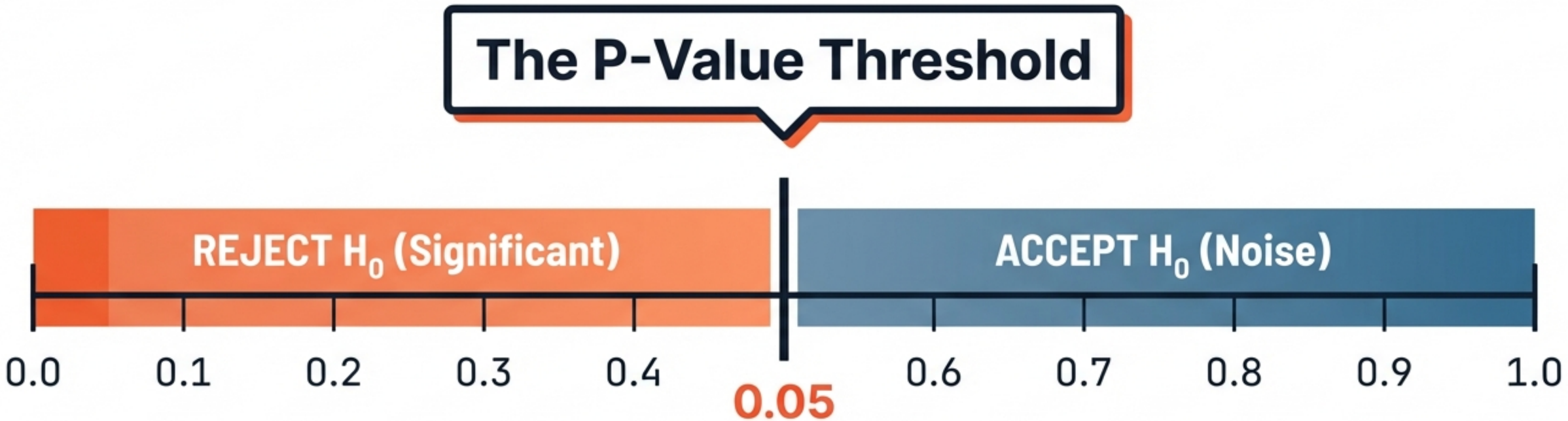
The Scenario: Does user behavior fundamentally shift on weekends? We don’t guess; we test. Is the diference in means statistically signifcant or just random noise?

The Methodology: H_0 vs. H_1




```
from scipy import stats
stat, p_value = stats.ttest_ind(weekday_data, weekend_data)
```

Interpreting the Verdict



If p < 0.05: The difference is real.

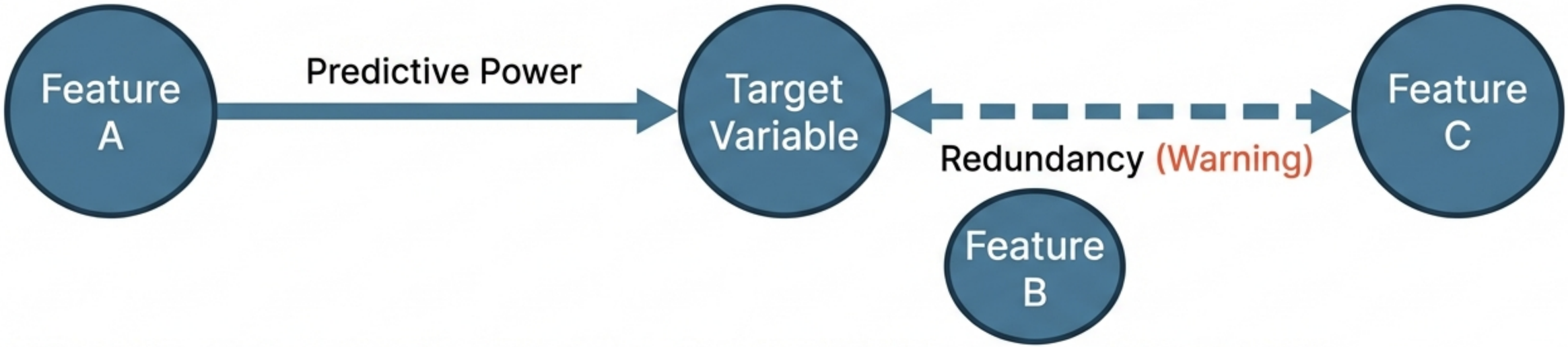
Action: Include "Is_Weekend" as a model feature.

If p > 0.05: The difference is chance.

Action: Discard to prevent overfitting.

Objective 3: The Hunt for Relationships

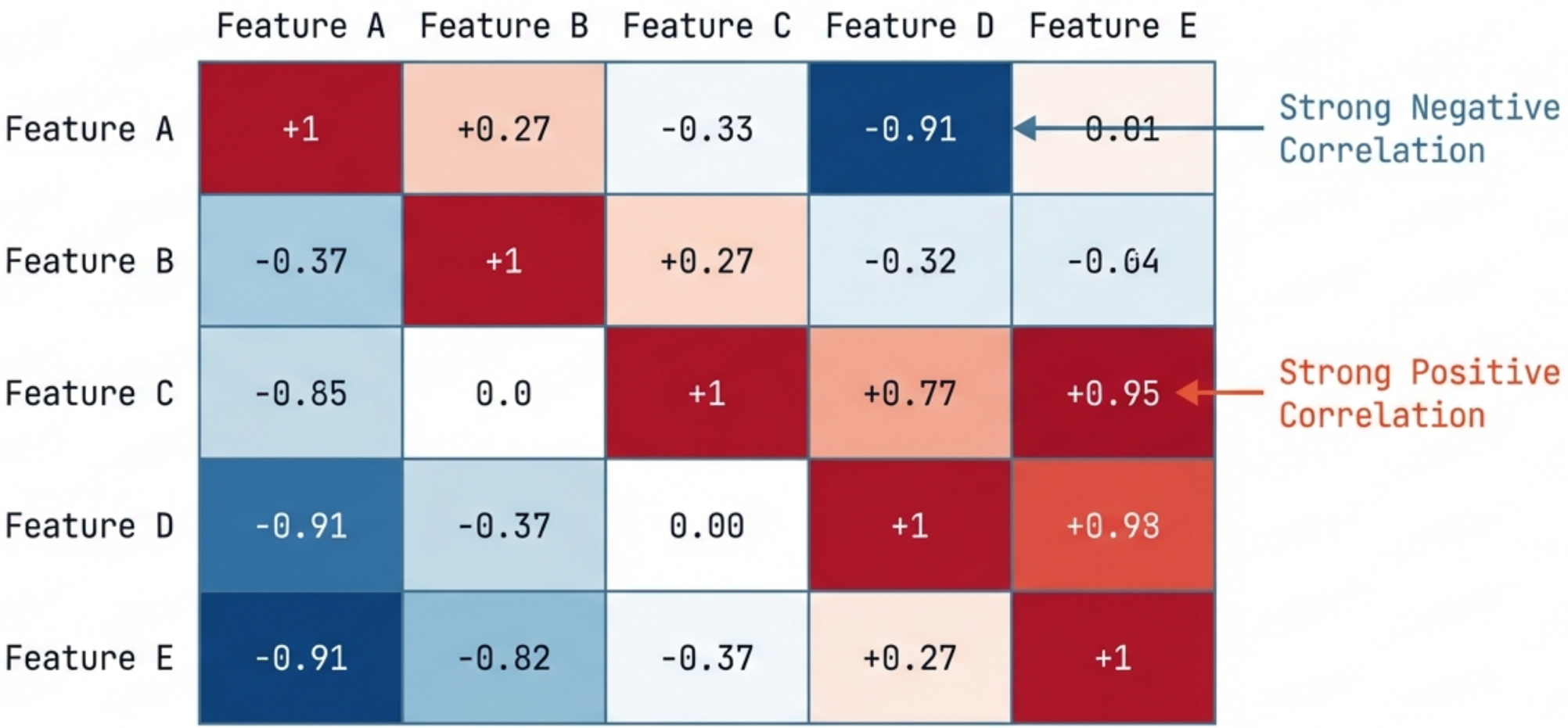
Identify Correlations



We use the Correlation Matrix to answer two questions:

1. Which features drive the target?
2. Which features are redundant clones of each other?

Mapping the Connections



CAUTION: Correlation ≠ Causation. We identify patterns, not causes.

The Trap of Multicollinearity

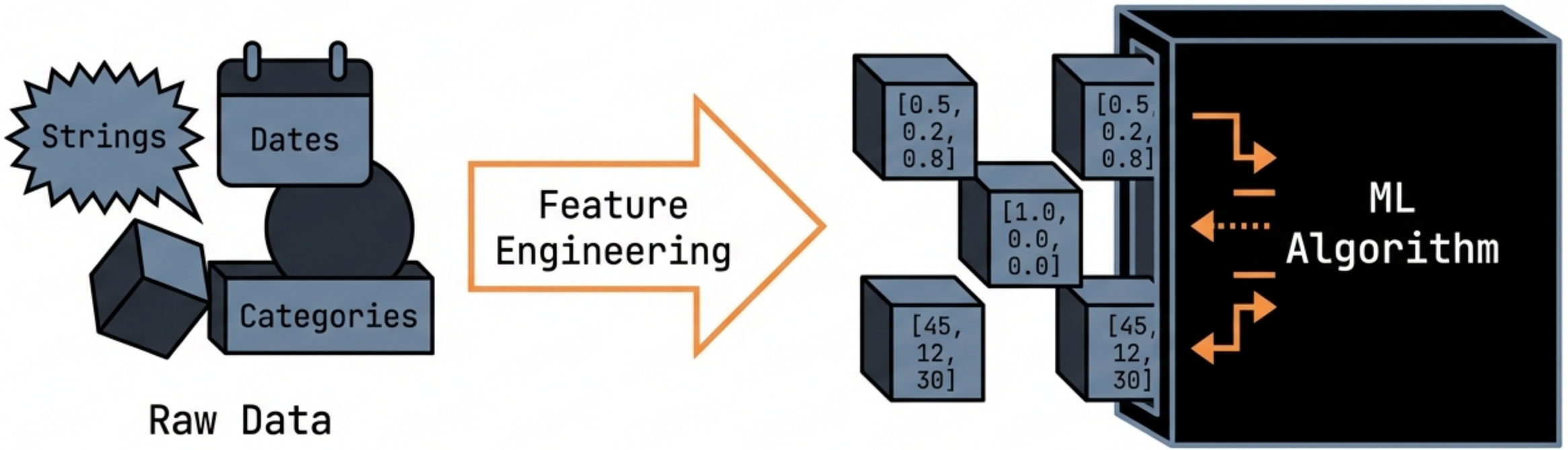
The Problem:
Highly correlated input features provide the same information.

The Risk:
In linear models, this causes instability in coefficient estimation. The math "panics" trying to attribute weight.

The Fix:
Drop one of the pair during Feature Selection.

Objective 4: The Art of Translation

Engineer Features for ML



Machine Learning speaks Math. Feature Engineering is the process of translating domain knowledge and raw types into vectors.

Temporal Features

2023-10-27 14:30:00

Hour_of_Day

Capture cyclical daily patterns.

Day_of_Week

Capture weekly routines.

Is_Weekend

Capture work vs. leisure behavior.

Decomposing time allows the model to learn context.

Encoding & Scaling

Categorical Encoding

```
["Red",
 "Blue",
 "Green"]
```

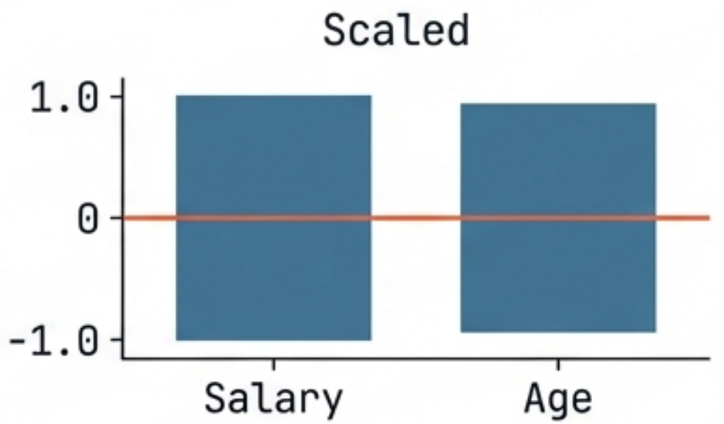
One-Hot Encoding

Color	Red	Blue	Green
Red	1	0	0
Blue	0	1	0
Green	0	0	1

Numerical Scaling

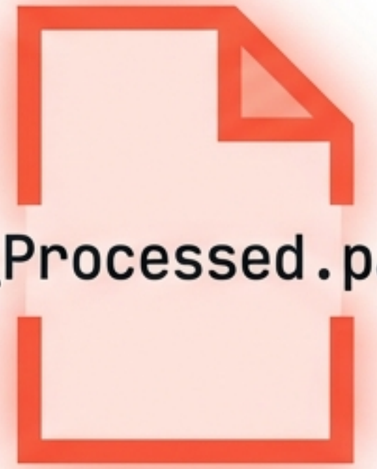
Spark MLlib: VectorAssembler

features
[1.0, 0.0, 0.0, 0.75, 0.8]
[0.0, 1.0, 0.0, 0.6, 0.65]
[0.0, 0.0, 1.0, 0.8, 0.9]



The ML-Ready Dataset

- [x] **Anomalies identified**
- [x] **Key drivers validated**
- [x] **Redundancies removed**
- [x] **Data vectorized**



`Day11_Processed.parquet`

`Status: READY FOR TRAINING`

We started with raw logs.
We now possess a refined asset.
The foundation for Day 12 is set.

CHALLENGE DAY 11: COMPLETE

We have built the bridge to intelligence.
Join us tomorrow for Model Building.

#DatabricksWithIDC

Indian Data Club

databricks

{CODE} BASICS