AltScore

Making Credit Access Fair for everyone

Team Name : TECH CRUSADERS

Team Members : Nithilan S
Muskan Kumari V
Charumadhi M

Department : Information Technology

Problem Statement – "*Why Traditional Credit Scoring Fails Millions*"

Imagine you're 25 years old. You've been responsibly paying your rent, utilities, and phone bills for years. You've never missed a payment. But when you walk into a bank for a small loan, they say no.

Why? Because you have no "credit history".

This affects 1.7 billion people globally.

The traditional system looks for:

- Previous loans (which you don't have)
- Credit cards (which you couldn't get without credit history)
- Collateral (which young people rarely have)

Our Solution: Build a credit scoring system that uses *different* data:

- Your transaction history (how you spend money)
- Your utility bill payments (do you pay on time?)
- Social indicators (where you live, your community)

Our Solution + The Data – “*How We're Doing It (and What Data We Have)*”

Instead of asking "Do you have credit history?“

We ask:

- Do you pay your bills on time?
- Are you consistent with payments?
- What's your life context - age, education, where you live?

The Dataset That Makes This Possible:

Main data:

- **307,511 real loan applications** with outcomes (who repaid, who defaulted)
- We know exactly what happened - perfect for machine learning

What We Have	What It Shows	Records
POS/Cash Payments	Regular payment behavior (like utility bills)	**10.0 million**
Installment History	When do people actually pay? Early? Late?	13.6 million
Bureau Records	Credit relationships with other banks	1.7 million
Previous Applications	Past loan attempt patterns	1.7 million
Credit Card Behavior	How they use credit when they have it	3.8 million

Total: 58.4 million data points to learn from!

What we've done - EDA

Understanding the Data

- Loaded all 307,511 applications + 58 million supplementary records
- Profiled data quality - found missing values, understood distributions
- Mapped relationships between 7 different data tables
- Identified our target: 8.07% default rate (24,825 defaults)

Finding Patterns

- Created 10 professional visualizations
- Analysed patterns across:
  - Demographics (age, education, gender)
  - Financial behavior (income, credit, leverage)
  - Payment behavior (timing, regularity, amounts)
  - Geographic factors (region, housing, employment)
- Ran statistical tests to validate findings

Generating Insights

- Identified top predictive features
- Validated that alternative data actually works
- Found key patterns that will drive our model
- Documented everything (12 detailed reports)

Inference 1 – Alternative Data Works

The Experiment: We split all 307,511 applications into groups based on what alternative data was available:

Data Available	Count	Default Rate	Risk Level
NO alternative data	2,470 (0.8%)	**7.25%**	🔴 Baseline
Only bureau history	13,984 (4.6%)	**5.72%**	🟢 **21% LOWER!**
Only previous apps	41,550 (13.5%)	10.16%	🔴 Higher
Both sources	249,507 (81.1%)	7.84%	🟡 Moderate

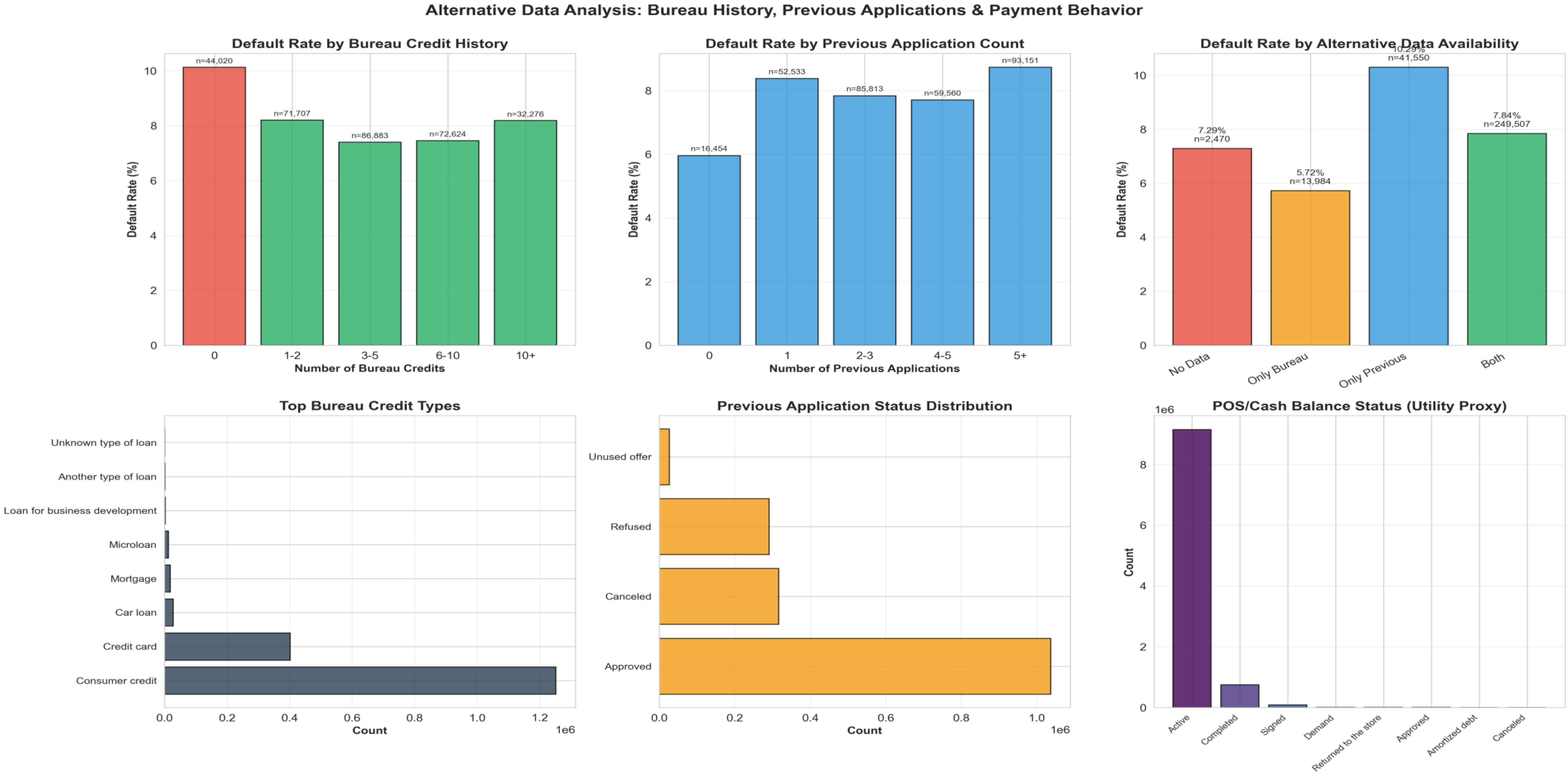
What This Tells Us:

✅ **Alternative data reduces default risk by 21%** (from 7.25% to 5.72%)
✅ **Quality matters more than quantity** - bureau data alone is better than just previous apps
✅ **Our approach is validated** - alternative data sources DO predict creditworthiness

The Business Case:

•Traditional banks reject those 2,470 people with no data (0.8% of population)
•**Our model will serve them** using different signals - payment behavior, geographic context, life stage
•We can help the underbanked while maintaining good risk management

Inference 1 – Alternative Data Works



Inference 2 – Payment Behavior is Primary

The Challenge: Problem statement wants utility bill payments, but utility companies don't share that data
Our Solution: Use **10 million POS/Cash balance payment records**

What Utility Bills Show	What POS/Cash Shows
✓ Regular monthly schedule	✓ Recurring transactions
✓ Payment timing (on-time vs late)	✓ Tracked month-by-month
✓ Financial discipline	✓ Payment consistency
✓ Long-term patterns	✓ Multi-year history

Payment Timing Breakdown:

- 🟢 Pay EARLY: 68.4% ← Most people!
- 🟡 Pay on-time: 23.1%
- 🔴 Pay late: 8.4%

The Insight: With 10 million records showing 68% early payment rate and strong regularity patterns, we have an excellent proxy for utility payment behavior.

Inference 3 – Top Predictors

The Rankings (What Correlates Most with Default):

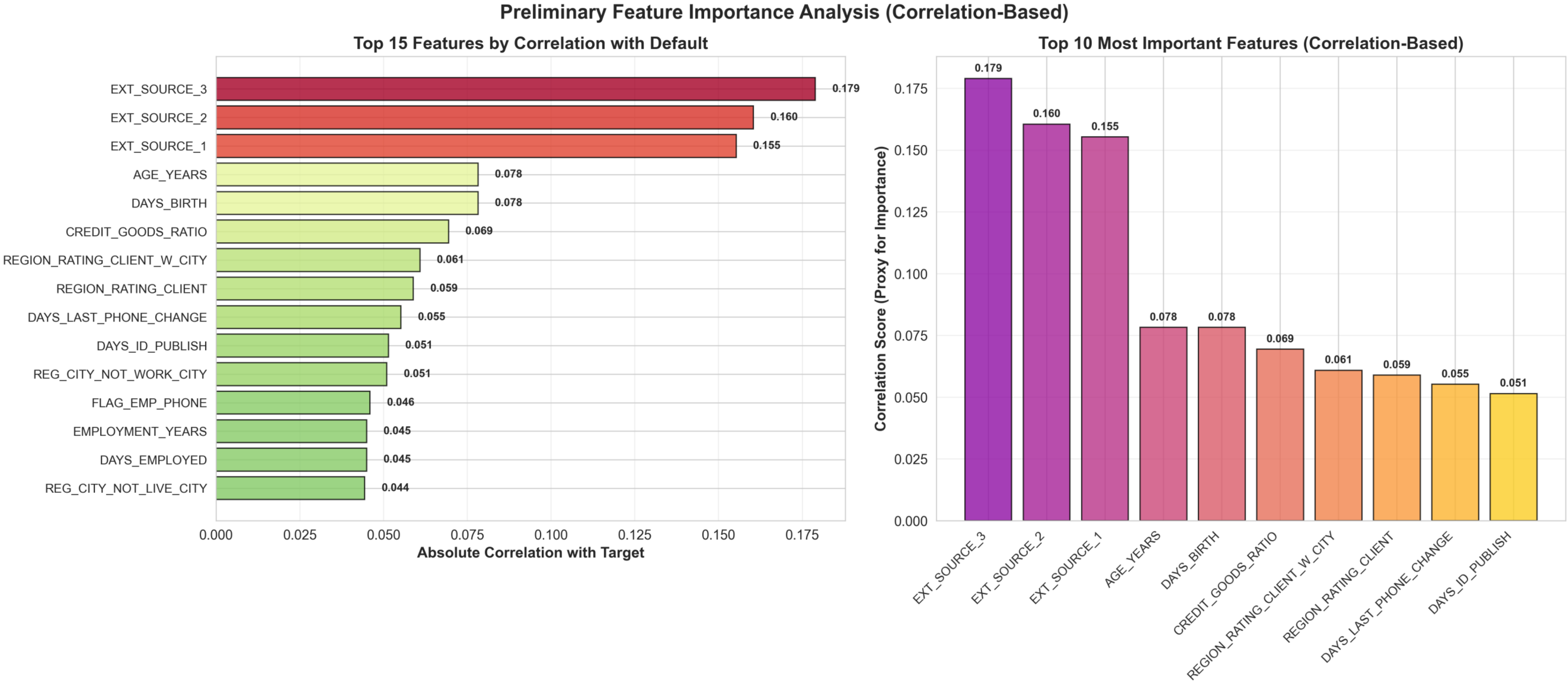
Rank	Feature	Correlation	What Is It?
🥇 1	**EXT_SOURCE_3**	**0.179**	Alternative credit score ⭐
🥈 2	**EXT_SOURCE_2**	**0.160**	Alternative credit score ⭐
🥉 3	**EXT_SOURCE_1**	**0.155**	Alternative credit score ⭐
4	AGE_YEARS	0.078	Demographics
5	CREDIT_GOODS_RATIO	0.069	Financial
6	REGION_RATING	0.061	Geographic
7	DAYS_LAST_PHONE_CHANGE	0.055	Behavioral stability

What the data shows:

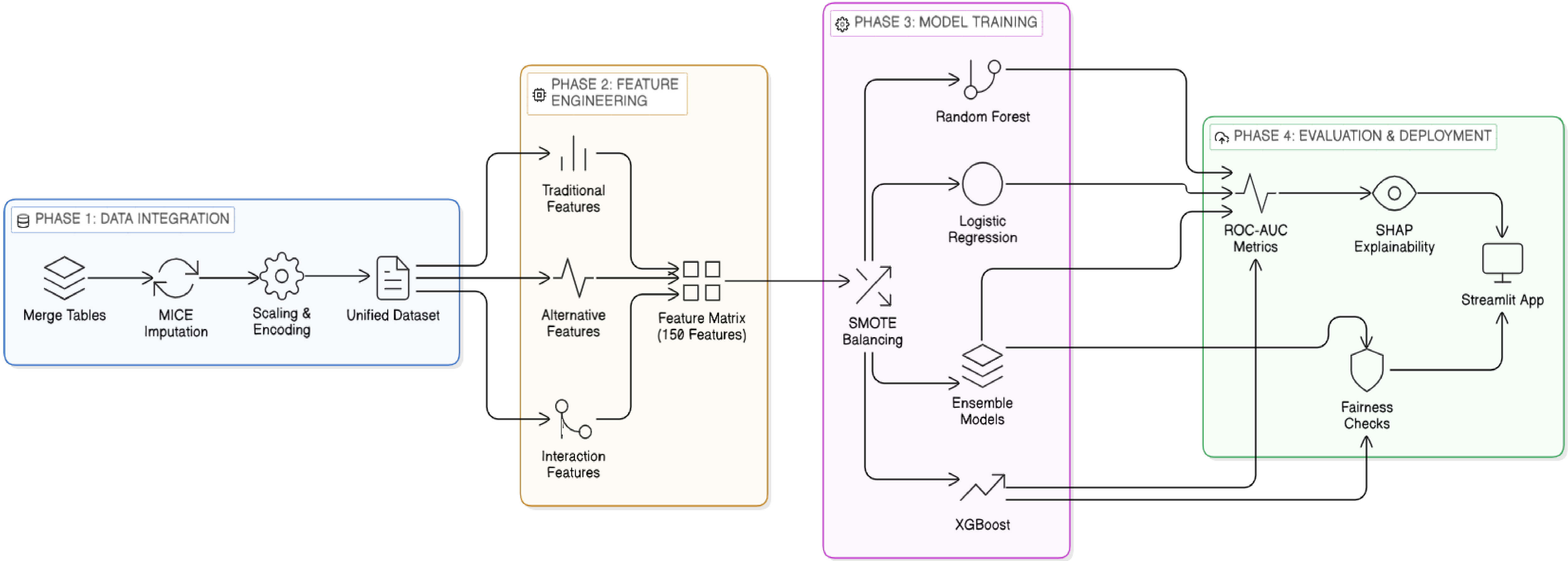
- People who repaid: EXT_SOURCE scores cluster around 0.5-0.7
- People who defaulted: EXT_SOURCE scores cluster around 0.2-0.4

The Strategy: No single feature dominates, so we'll combine MANY signals - exactly what machine learning is built for

Inference 3 – Top Predictors



Methodology



Summary

What We've Shown You Today:

✅ **The Challenge**: 1.7B people excluded from financial services
✅ **Our Solution**: Alternative credit scoring using payment behavior, social data
✅ **The Data**: 307K applications + 57M alternative data records
✅ **The Proof**: Alternative data reduces default risk by 21%
✅ **The Innovation**: 10M payment records as utility proxy (68% pay early!)
✅ **The Plan**: 4-phase approach to deployed model in 4 weeks

Key Numbers to Remember:

- **8.07%** default rate (manageable)
- **21%** risk reduction with alternative data
- **68.4%** early payment rate (strong discipline)
- **0.179** top correlation (external alternative scores)
- **ROC-AUC > 0.75** target for next week

What's Next: Week 2 - Build baseline model with 150 engineered features