Building a Trust Engine for the Unbanked

The Problem: Millions of responsible people, like '**Alex**,' are **invisible** to traditional banks. No credit history often means no loan, forcing them toward **predatory lenders**.

The Mission: To build a **predictive engine** that looks **beyond traditional** credit scores, using alternative data to identify **trustworthy borrowers**.

The Goal: To **safely approve** loans for the **deserving**, **unlocking financial inclusion**.

AUC: 0.50 (Baseline)

The Detective's Toolkit: How We Measure Success

ROC-AUC - The Ranking Game

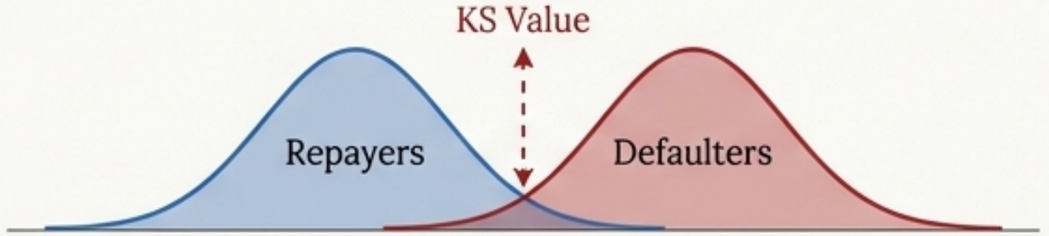
ROC-AUC doesn't care about a "Yes/No" answer; it measures how well the model ***ranks*** risky clients above safe ones.

"Imagine lining up photos of all customers from 'Most Risky' to 'Least Risky". A perfect model (AUC=1.0) places all defaulters at the front of the line. A random model (AUC=0.5) shuffles them randomly."

KS Statistic - The Separation

KS measures how well the model separates the scores of good and bad customers. It answers, "How confident is the model?"

"Imagine two mountains: a blue one for good guys' scores and a red one for bad guys' scores. High KS means the mountains are far apart, with a clear 'Valley of Separation" between them."



How to Judge KS Score

KS Score Range	Model Quality
< 0.20	Poor
0.20 – 0.40	Decent / Acceptable
0.40 – 0.60	Good
> 0.75	Suspicious

The Mission: Become Data Detectives

The Challenge

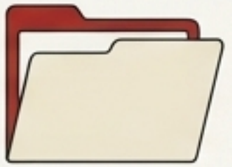
- Home Credit's mission is to serve the "unbanked" like Alex.
- Without traditional credit scores, we must look for alternative clues to assess trustworthiness.

The Evidence Locker

The Interview
"application_{train|test}.csv"
(What the applicant tells us).

The Gossip
"bureau.csv"
(What other banks say about them).

The History
"previous_application.csv"
(Their past dealings with us).

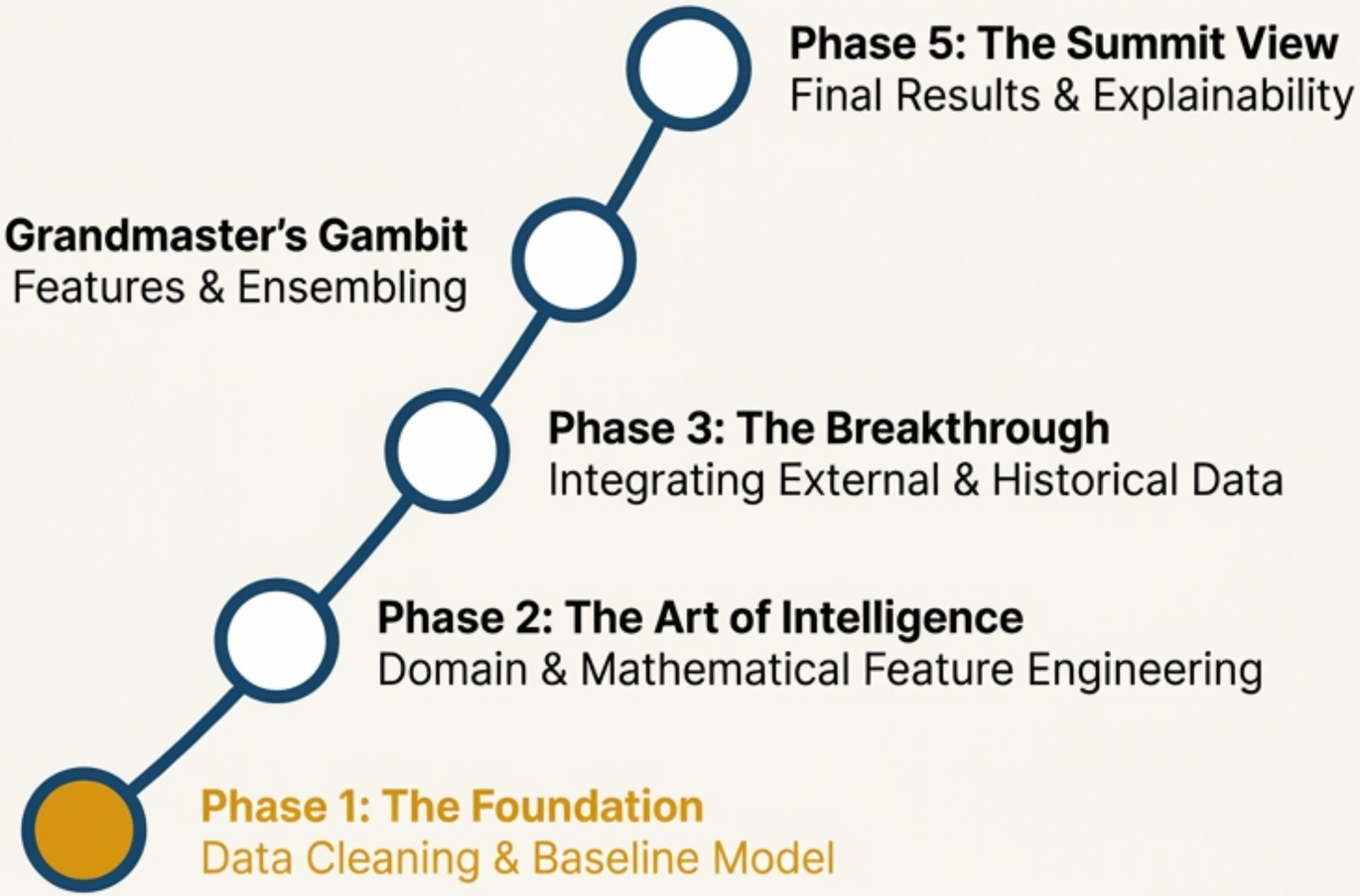


The Habits
"installments_payments.csv",
"credit_card_balance.csv",
"POS_CASH_balance.csv"
(Their month-to-month financial behavior).

Our goal: Build a 'Trust Engine' that can analyze these clues and predict a person's likelihood of repayment.

The Ascent: Our Roadmap from Raw Data to a High-Performance Engine

Our journey is a methodical climb, where each phase of work builds upon the last, systematically increasing the intelligence and predictive power of our model.

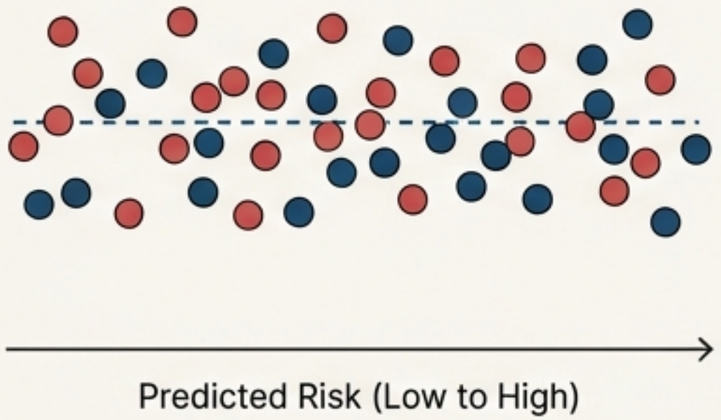


AUC: 0.50 (Baseline)

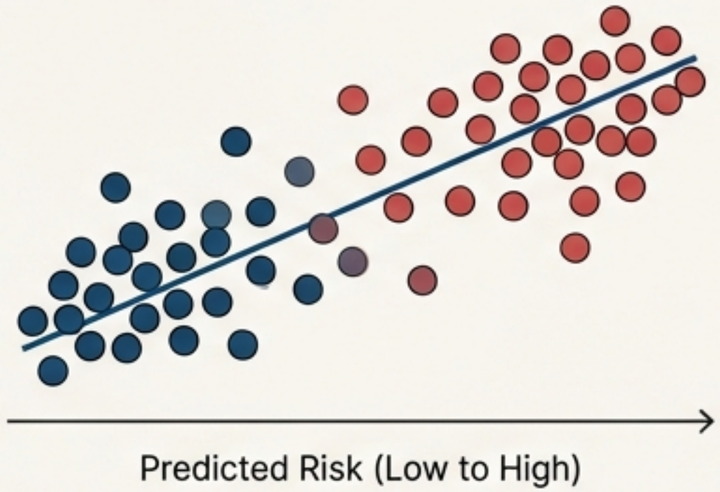
Our Compass: Why AUC is the Gold Standard for Ranking Risk.

ROC-AUC isn't about a simple 'Yes/No." It's about how well the model can **rank** customers from most risky to least risky. A higher AUC means a better, more confident ranking.

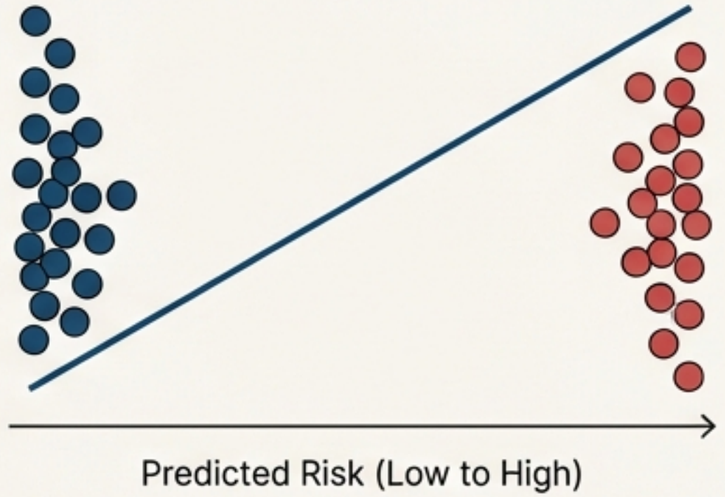
AUC = 0.5 (Random Guessing)



AUC = 0.75 (Good Model)



AUC = 1.0 (Perfect Model)

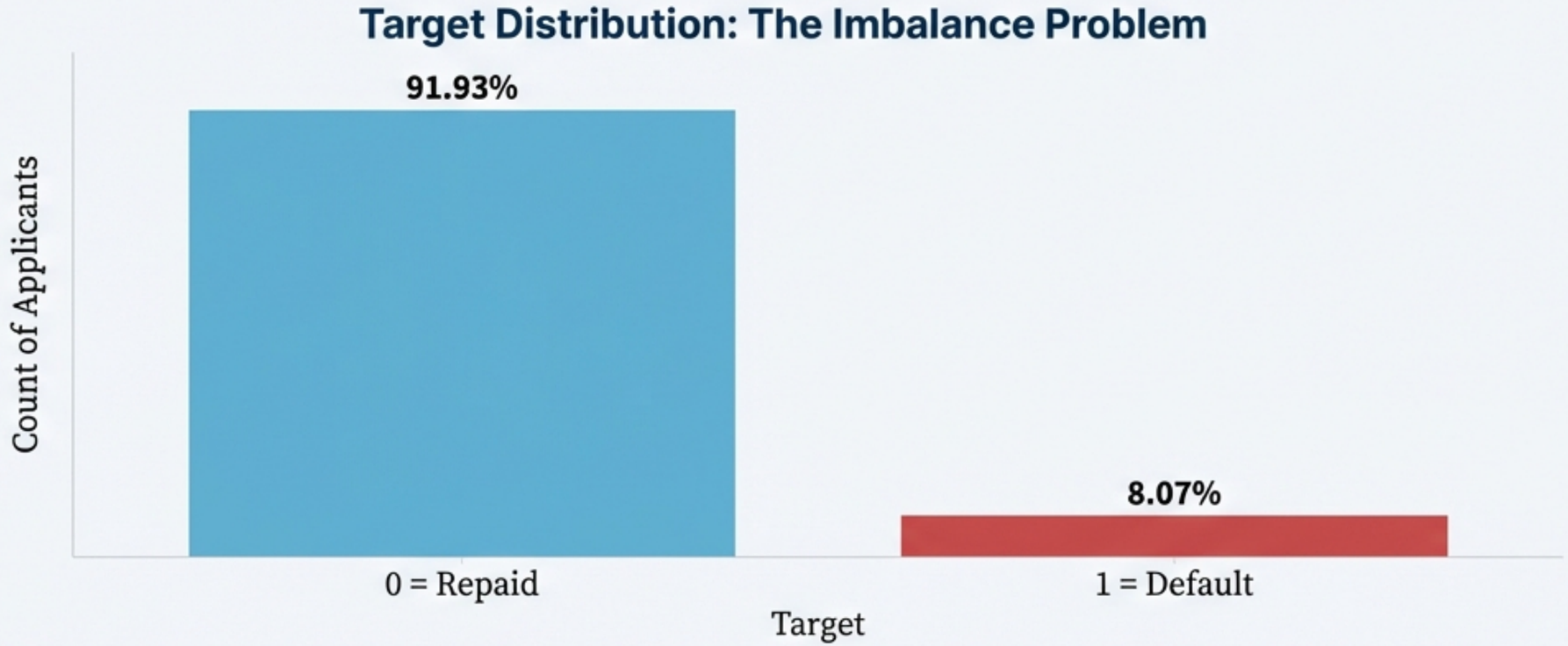


Unlike accuracy, AUC is independent of any specific cutoff threshold. It measures the quality of the model's ranking across all possibilities, making it stable, reliable, and the best metric for a credit risk model.

AUC: 0.50 (Baseline)

The Core Challenge: A Severe Class Imbalance

The goal is to predict loan default, but only 8% of applicants default. A model that predicts “no default” for everyone would be 92% accurate but commercially **useless.**



Our Strategic Response

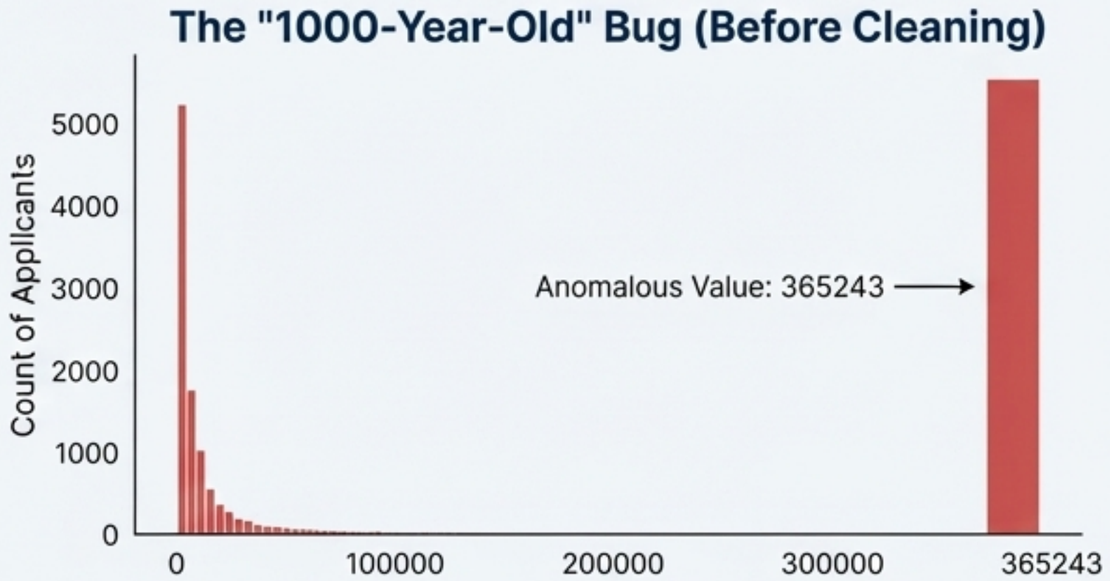
- Accuracy is a misleading metric. We must use **ROC-AUC**, which measures the model’s ability to distinguish between classes.
- Models will be forced to pay attention to the rare default cases using class weights (`scale_pos_weight = 11.3`).

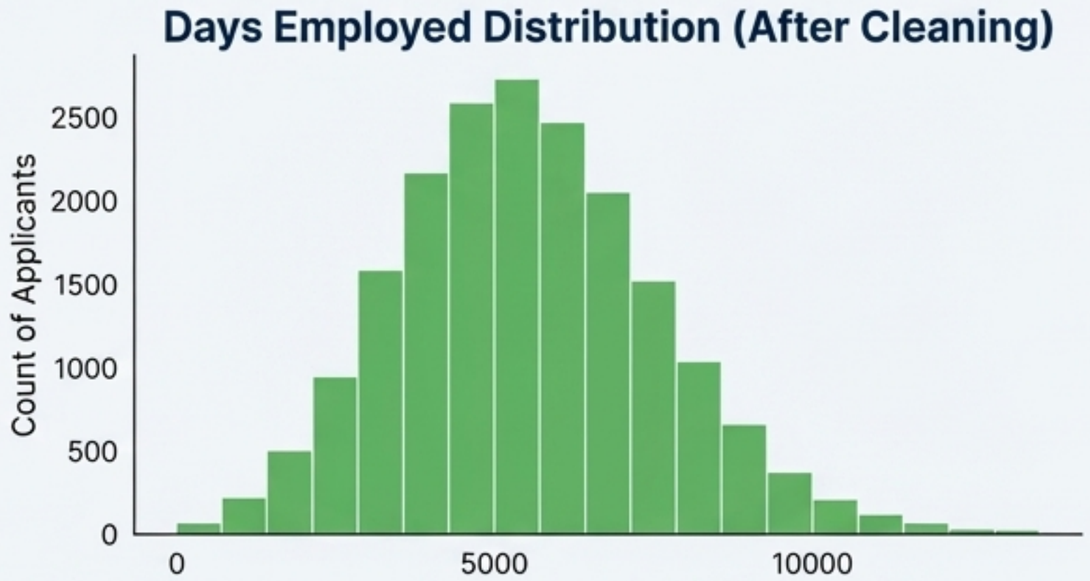
The First Ascent: Purging Anomalies from the Data

Challenge: The `DAYS_EMPLOYED` feature contained an anomalous value of `365243', equivalent to over 1000 years. This would skew any statistical calculations.

Solution: We flagged these entries with a new binary feature (`DAYS_EMPLOYED_ANOM`) and replaced the original value with NaN to prevent it from corrupting our models.

Before & After





Foundational data hygiene is the first step to reliable model performance.

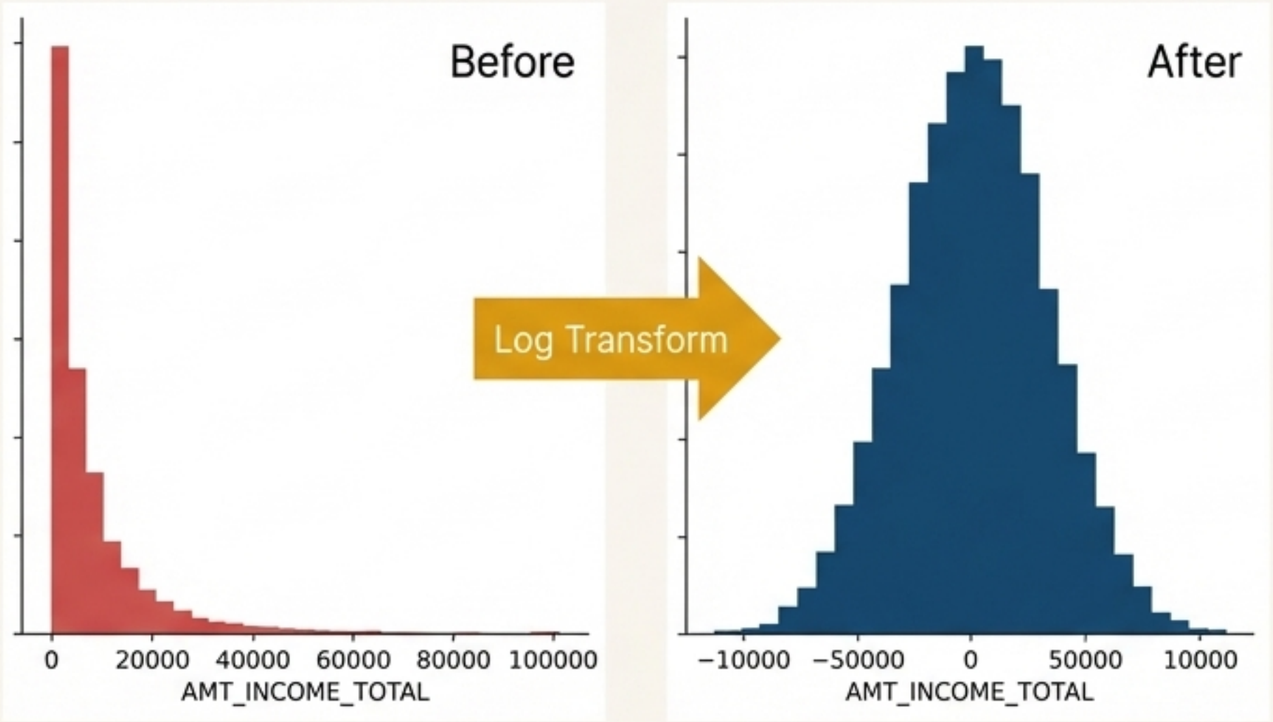
Phase 2: Taming the “Billionaire” Skew with Log Transformation

The Problem

The “Bill Gates” Effect

One outlier with a $1B income can skew the average so much that a model considers a normal $50k earner “extremely poor.”

This blinds the model to nuances among average applicants.



The Solution

Log Transformation squashes extreme values without changing the rank order.

A $1B income is pulled from “outer space” back down to “Earth,” allowing the model to see all applicants clearly.

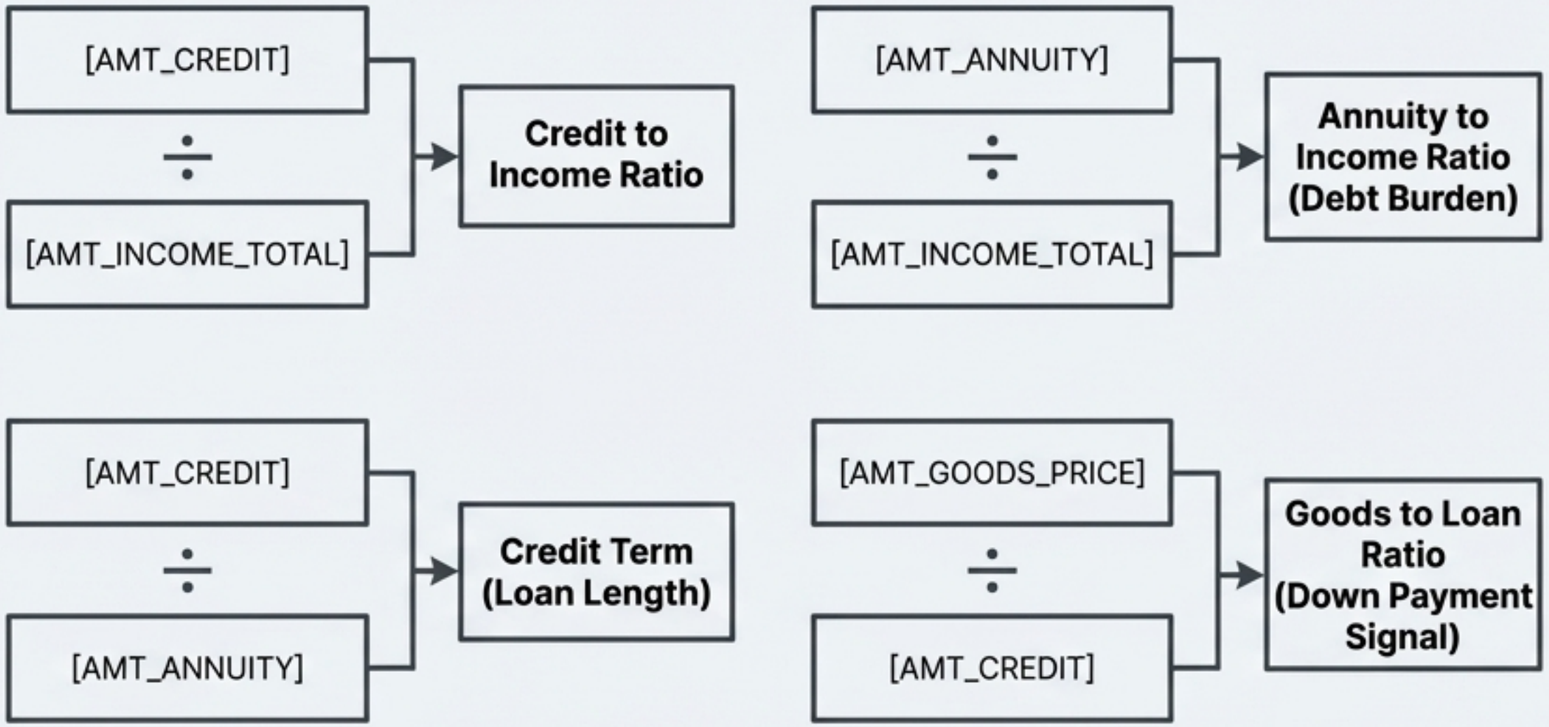
Why Not Other Methods?

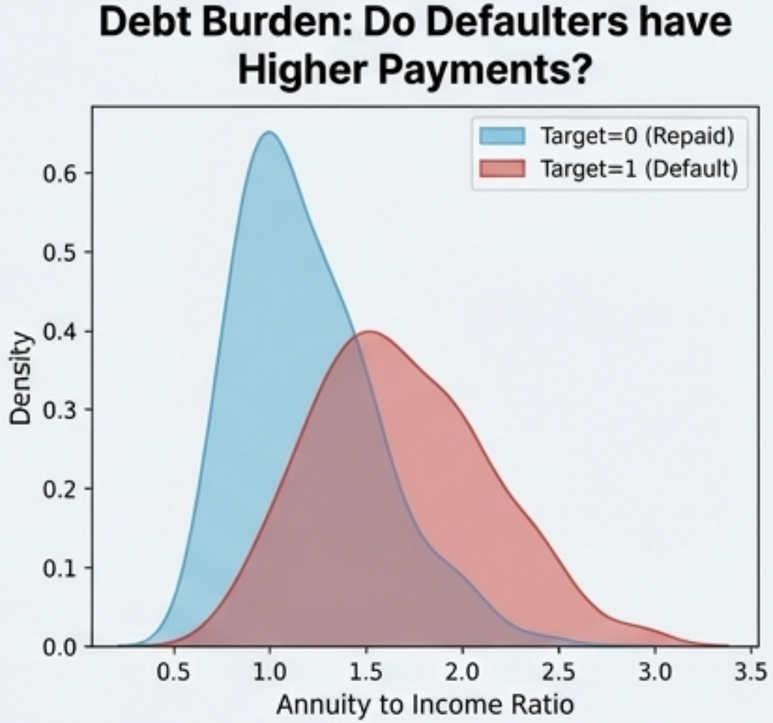
- ****Clipping****: Falsely equates a millionaire with an upper-class manager.
- ****Min-Max Scaling****: Crushes all normal applicants into a value near zero.

AUC Lift: +0.015

Engineering Financial Intuition: Domain Features

We translated business logic into four powerful **financial ratios** that the model can easily interpret.



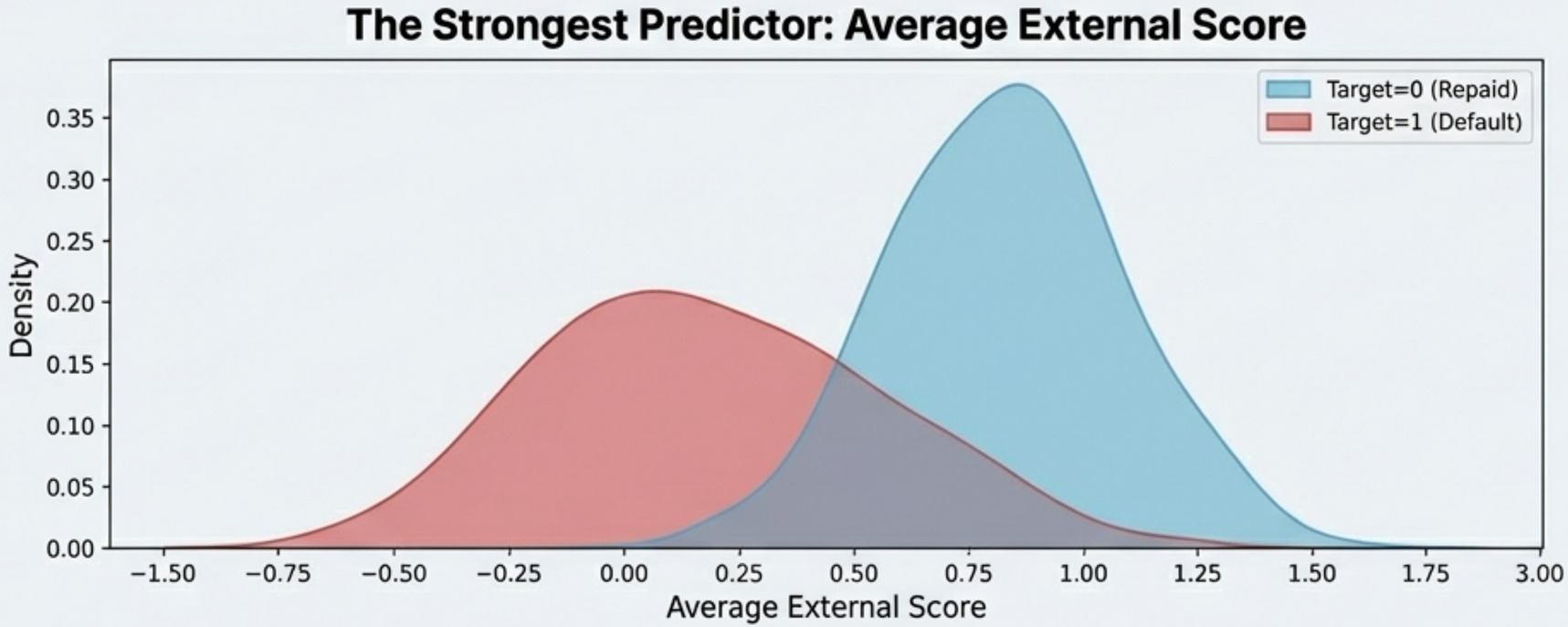


Key Insight: The `ANNUITY_INCOME_PERCENT` feature shows the strongest positive correlation with default among the new features.

The "Super Feature": Unlocking Signal with Polynomials

The three external credit scores (`EXT_SOURCE_1, 2, 3`) are strong individually, but their interaction is even more powerful. A low score from one bureau can be offset by high scores from others.

The Breakthrough Feature: We created `EXT_SOURCE_MEAN`, the simple average of the three scores.



Old Best Feature Correlation (`EXT_SOURCE_2`): -0.179

New Super Feature Correlation (`EXT_SOURCE_MEAN`): **-0.185**

This single engineered feature is more predictive than any of its individual components.

The First Lineup: Why Linear Models Hit a Ceiling

The Experiment

We first trained standard linear models (Logistic Regression) to establish a baseline.

The Result

These models achieved a respectable but limited performance, plateauing around **0.76 AUC**.

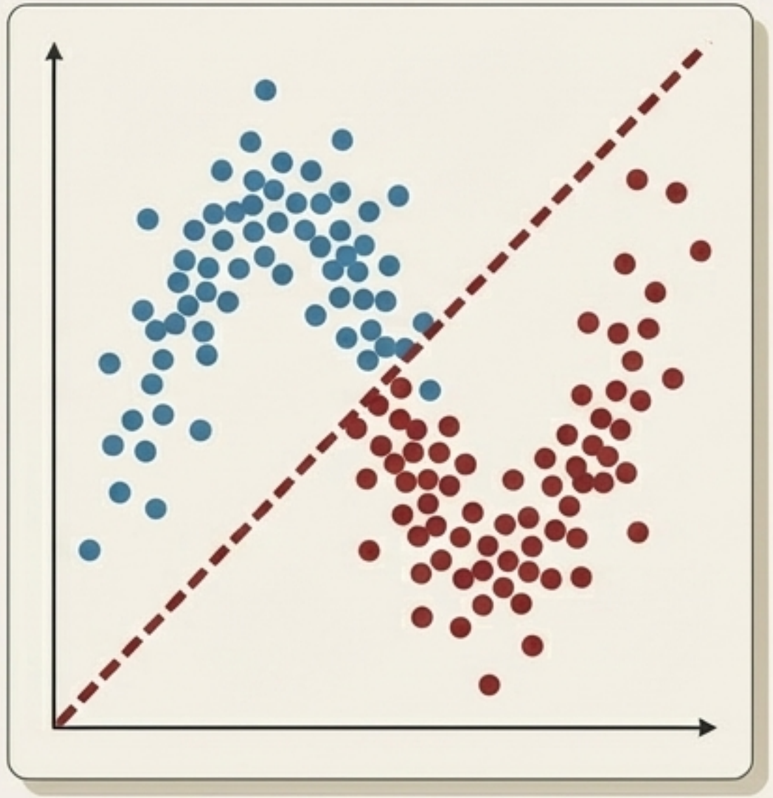
The Fatal Flaw

Linear models are forced to assume a straight-line relationship (e.g., "as income rises, risk always falls"). They cannot capture complex, real-world interactions.

Example

- "High income is good, *unless* the person is 21 and asking for $1M." A linear model can't learn this "if/then" logic.

Visual Evidence: The Linear Assumption Fails



🔎 "The Evidence"

"Our initial Linear Regression model predicted impossible default probabilities, ranging from -1.12 to 1.5, proving it was the wrong tool for a binary classification problem."

To break through the 0.76 AUC ceiling, we need a model that thinks non-linearly.

The Breakthrough: Unleashing Gradient Boosting

We moved to Gradient Boosting Decision Trees (GBDTs), which build a series of models, with each new model correcting the mistakes of the previous one. We tested the three industry leaders: XGBoost, CatBoost, and LightGBM.

Why LightGBM Wins for This Case

Speed for Iteration

Speed for Iteration: Runs up to 10x faster than competitors due to Gradient-based One-Side Sampling (GOSS), allowing for rapid feature engineering experiments.

Accuracy

Accuracy: "Leaf-wise" growth strategy focuses on the hardest parts of the data first, often leading to higher accuracy.

Built for Reality

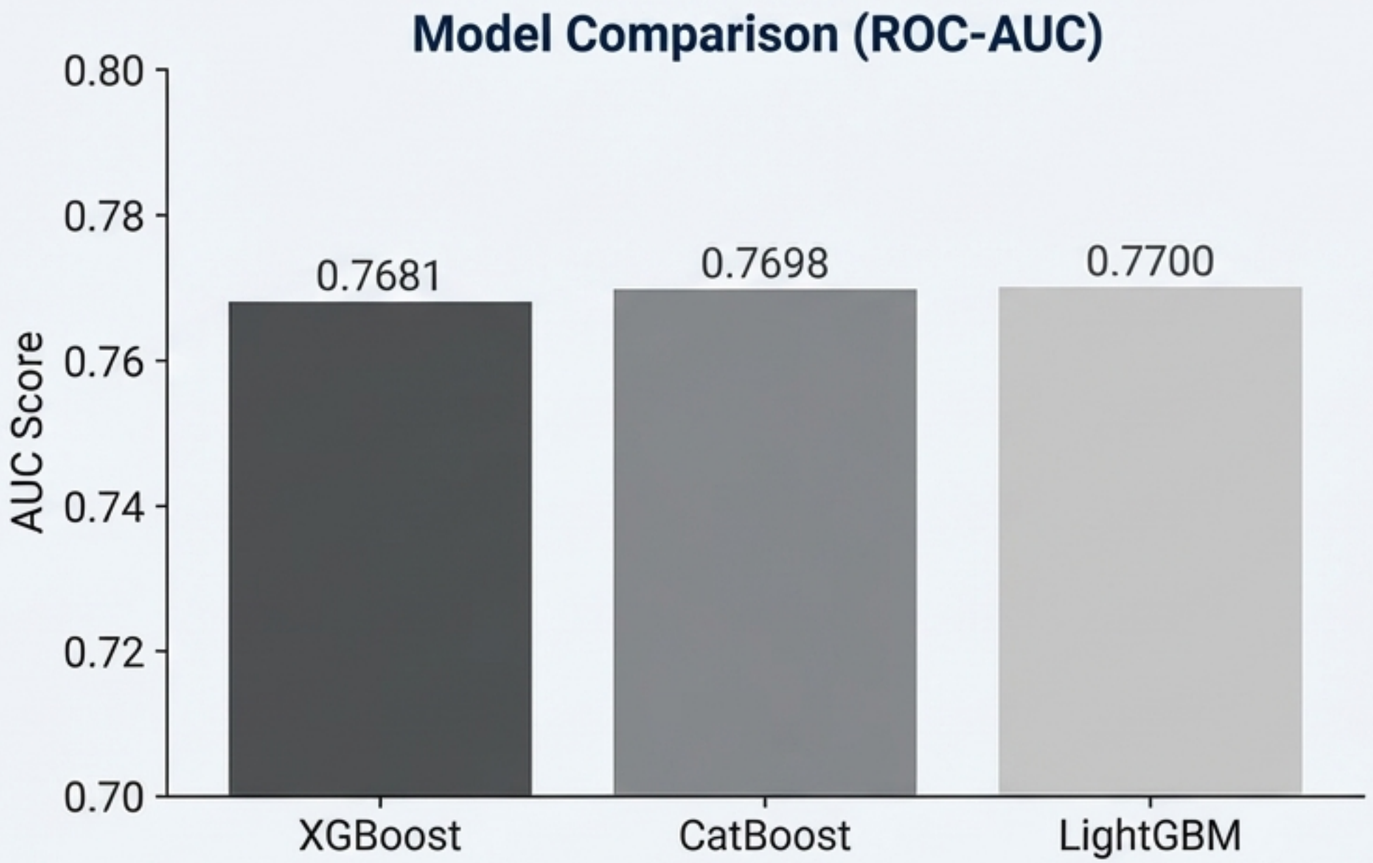
Built for Reality: Natively handles missing values, making it perfect for messy, real-world datasets like this one.

On the main `application` data alone, a tuned LightGBM model achieved a score of 0.7705 AUC, successfully breaking the linear model ceiling.

The Grand Tournament: Gradient Boosting Champions Compete

Current AUC: 0.7700

After extensive feature engineering, we ran a tournament with three state-of-the-art gradient boosting models to find the highest-performing algorithm on our dataset. All models used `scale_pos_weight=11.3`.



And the winner is... LightGBM! Its speed Its speed and performance make it our champion model for the next phase.

Fine-Tuning the Engine: Hyperparameter Optimization

Using Optuna, an automated optimization framework, we systematically searched for the best combination of LightGBM settings (`n_estimators`, `learning_rate`, `num_leaves`, etc.) to maximize the cross-validated AUC score.

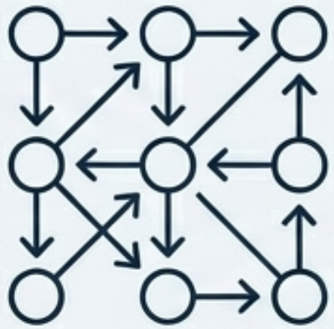
Parameter Space

learning_rate: [0.005, 0.1]

num_leaves: [20, 300]

n_estimators: [800, 1500]

Optuna: 50 Trials



Best Configuration Found

learning_rate: 0.0235

num_leaves: 147

n_estimators: 1120

Baseline LightGBM Score: 0.7700

Optimized LightGBM Score: **0.7737**

Key Insight: Automation discovered a superior configuration, yielding a **+0.0037** AUC gain. We are now hitting a performance ceiling with the current data.

Phase 3: The Breakthrough by Integrating External & Historical Data.

The Insight: We could never break the 0.78 ceiling with only what applicants *'told us'*. The real truth lies in what they've *'done'*.

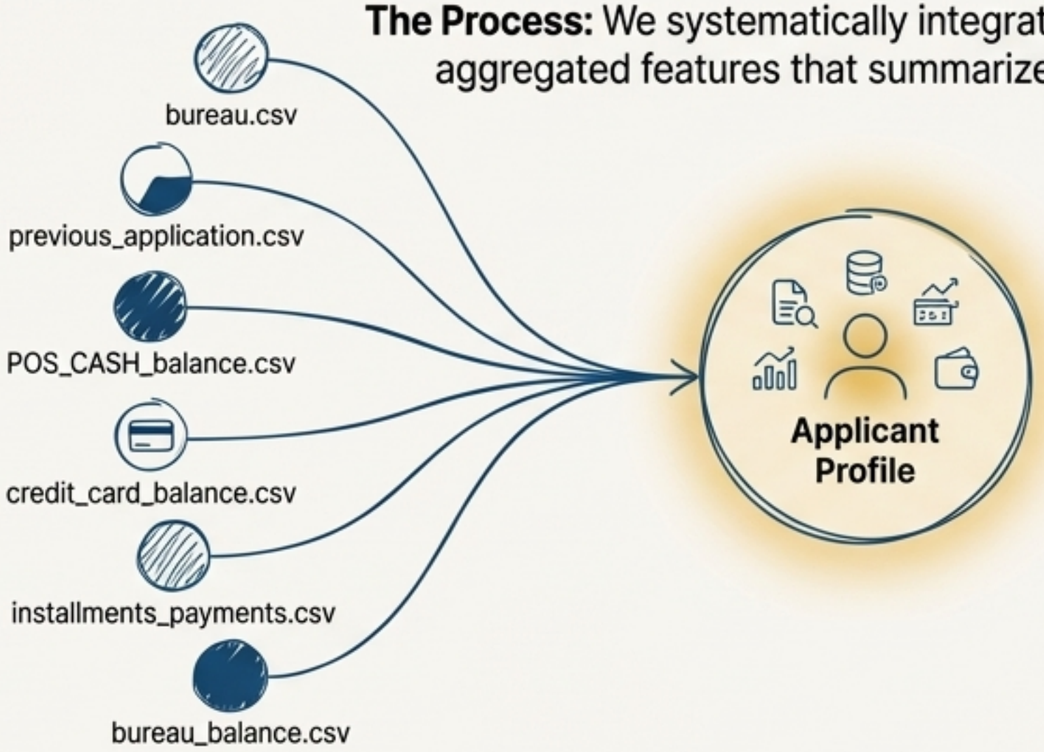
The Process: We systematically integrated data from 6 additional files, creating aggregated features that summarized an applicant's entire financial history.

The 'Bureau Attack'

Merged `bureau.csv` data to see credit history with **other** banks.

The 'Previous History' Attack

Merged `previous_application.csv` to analyze past behavior with Home Credit.



The 'Lost Data' Recovery

Merged `POS_CASH_balance.csv` and `credit_card_balance.csv` to capture spending habits.

Result: This fusion of data sources provided the single largest performance gain in the project, validating that behavioral history is the most powerful predictor of future risk.

AUC: 0.7850
(All Source Data Integrated)

Major Breakthrough #I: The Bureau Attack

The Problem

Our 0.77 AUC score was the ceiling for the `application` data. We couldn't judge a person's trustworthiness without knowing their history with *other banks*.

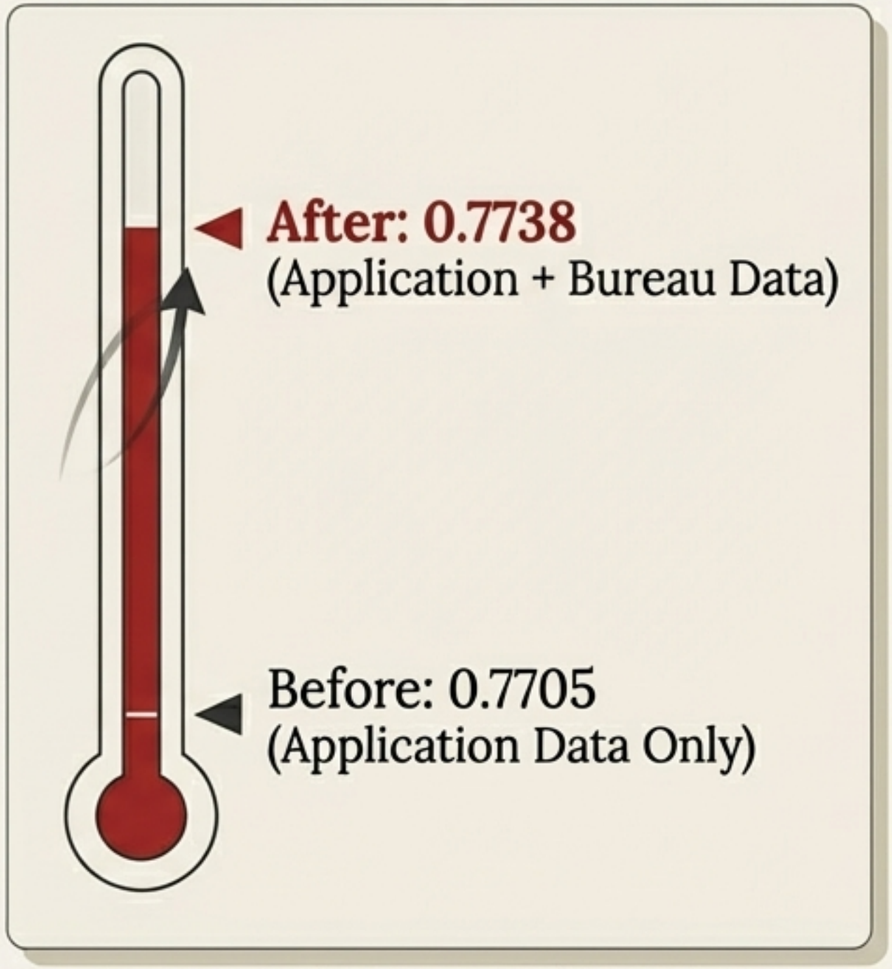
The Action

1. Processed `bureau.csv` and `bureau_balance.csv` (27 million rows of transaction history).
2. Aggregated this history into 25 new high-power features for each applicant.
3. Examples: `AVG_DAYS_OVERDUE_OTHER_BANKS`, `ACTIVE_LOAN_COUNT`, `TOTAL_DEBT_ACROSS_BANKS`.

Conclusion

Adding external credit history provided the first significant performance lift.

The Impact



Major Breakthrough #2: The Previous History Attack

The Next Question

How did this applicant behave on previous loans *with us*?

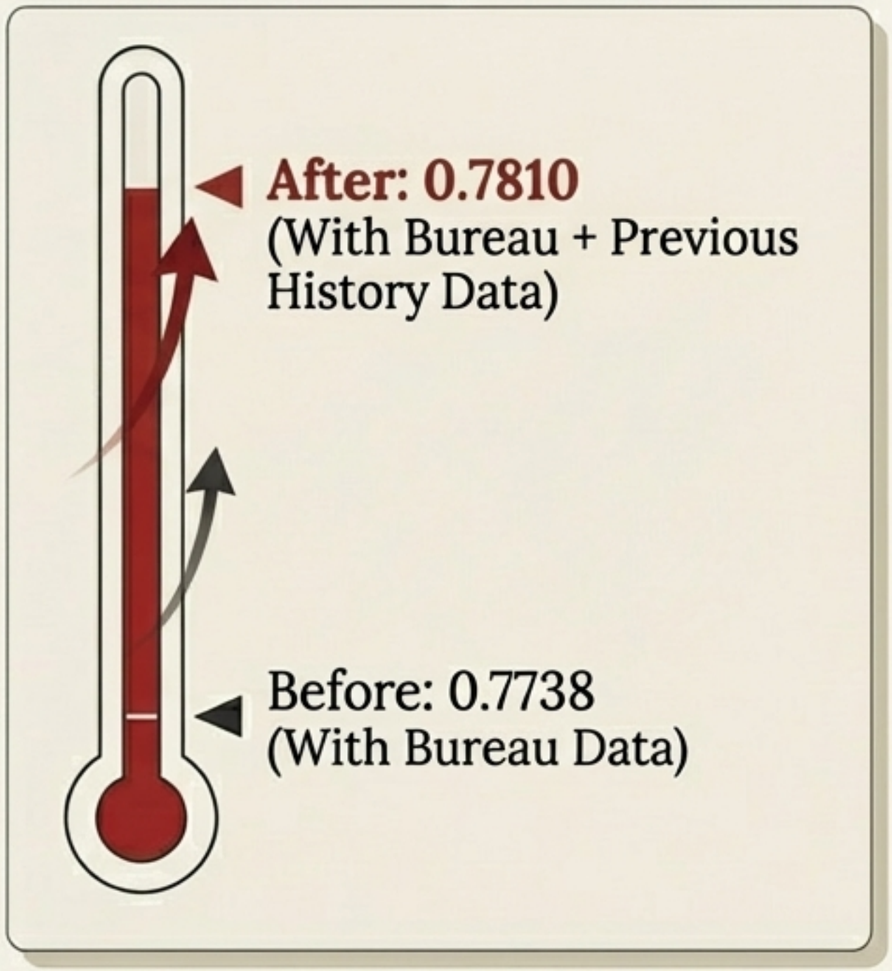
The Action

1. Processed `previous_application.csv` and `installments_payments.csv`.
2. Engineered features based on past behavior.
3. Examples:
   `AVG_LATE_PAYMENT_DAYS`,
   `REFUSED_CONTRACT_RATIO`,
   `DAYS_SINCE_LAST_COMPLETED_LOAN`.

Conclusion

Understanding an applicant's direct history with the lender is a powerful predictor of future behavior.

The Impact



The Final Clue: Following the Money

The Final Frontier

The last unexamined evidence was the applicant's day-to-day transaction behavior.

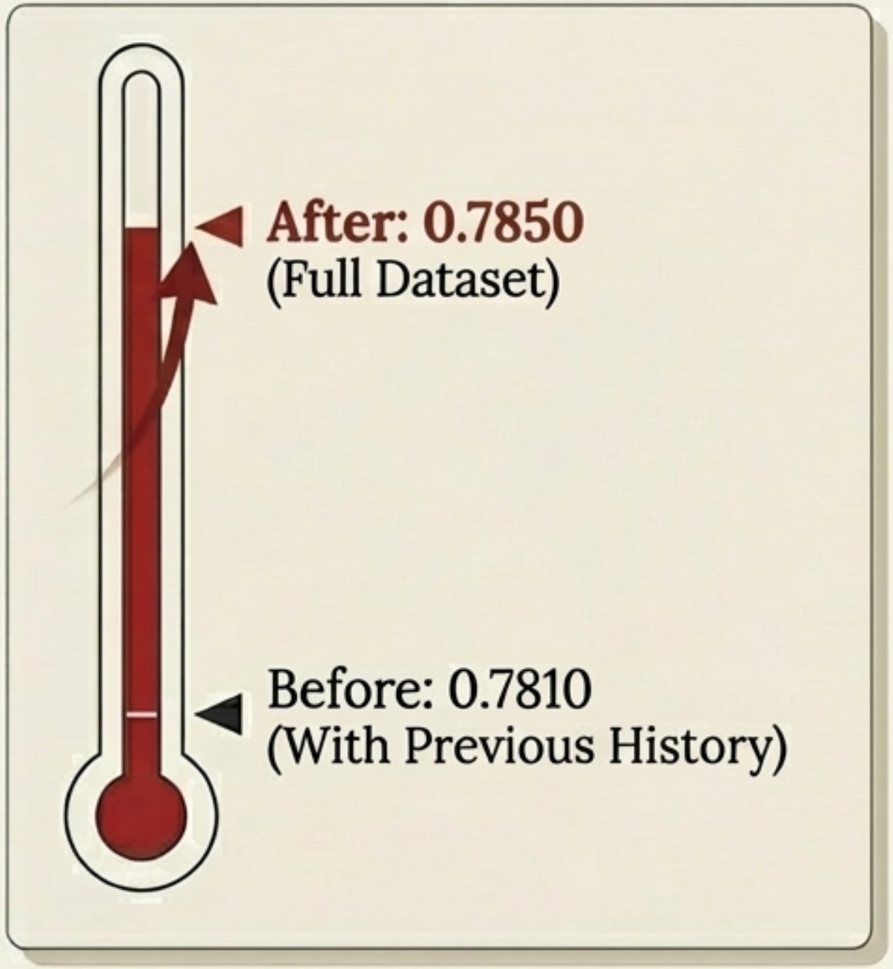
The Action

1. Integrated `POS_CASH_balance.csv` and `credit_card_balance.csv`.
2. Created features capturing spending habits, credit utilization, and payment patterns on revolving credit lines.

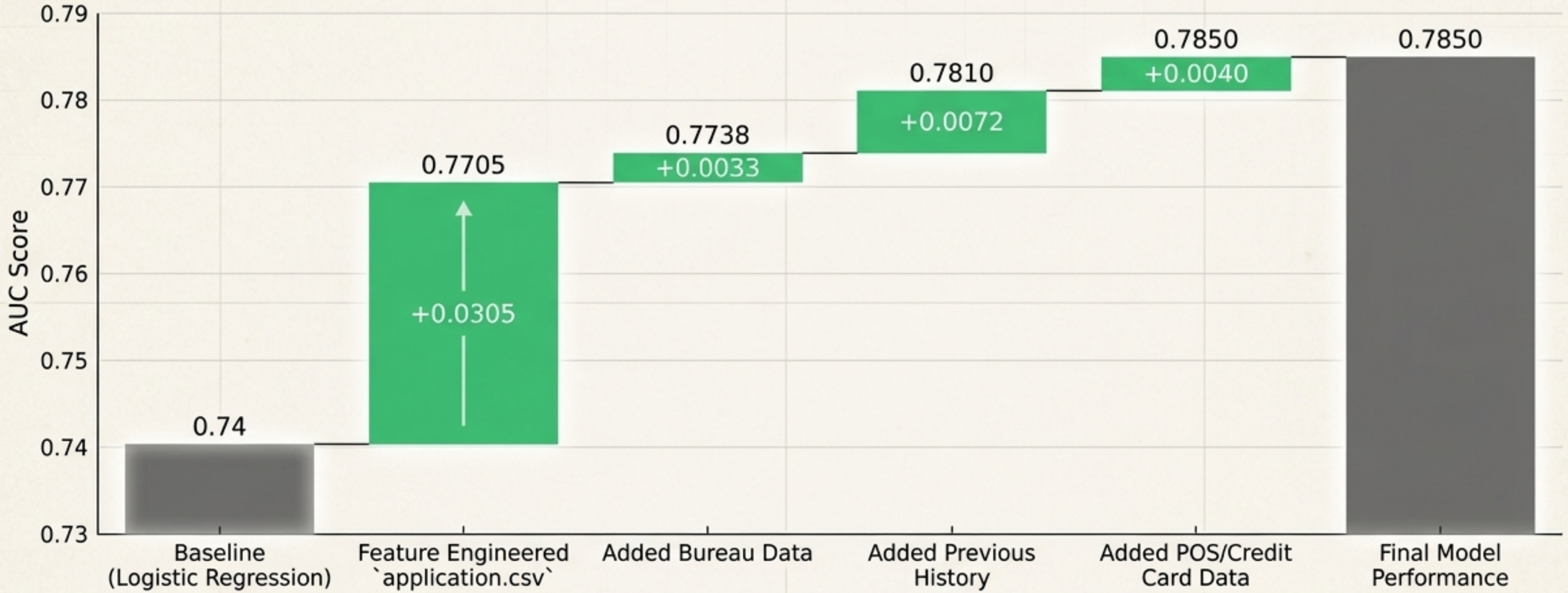
Conclusion

A complete 360-degree view of an applicant, combining application data, external history, internal history, and spending habits, is required for maximum predictive power.

The Impact



The Path to 0.785: A Story of Incremental Gains



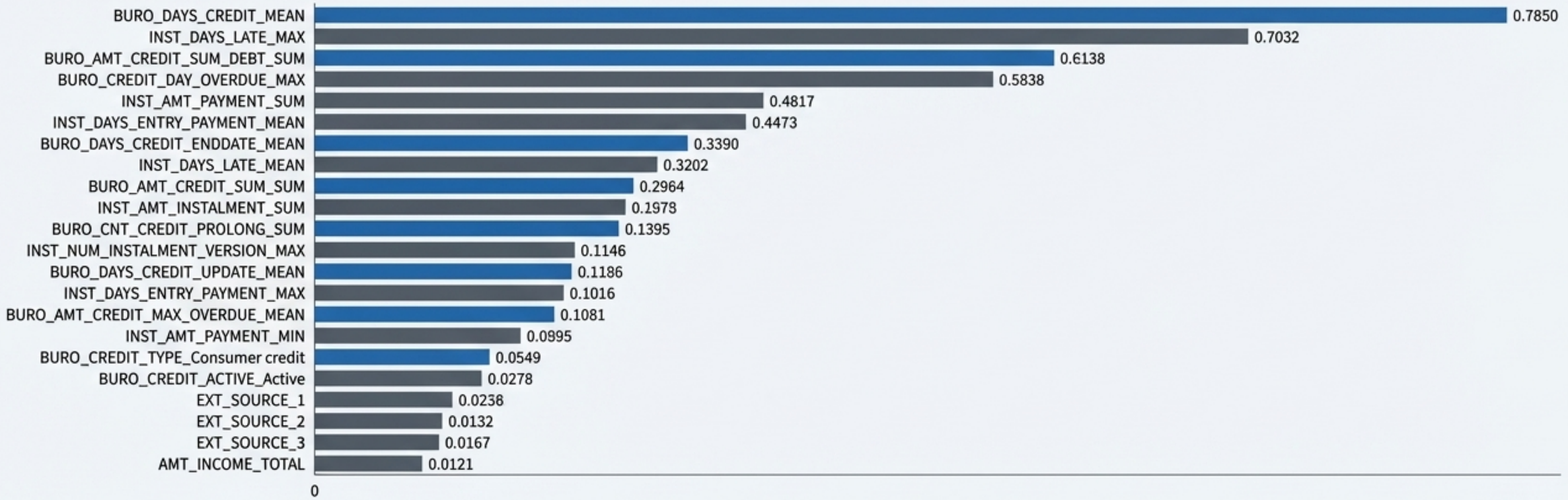
Hyperparameter tuning yielded minimal gains (<0.003 lift). The vast majority of the model's performance—a lift of over 0.04 AUC—came directly from intelligent feature engineering and comprehensive data integration. Information is more valuable than algorithms.

The Bureau Breakthrough: A New Performance Plateau

Current AUC:
0.7850

Adding the aggregated bureau features and re-running our optimized LightGBM model caused a significant jump in performance.

The “Killer” Features (Top 20)



Key Insight: An applicant’s payment history and debt level with *other lenders* are among the most powerful predictors of default. This confirms the value of data enrichment.

Phase 4: The Grandmaster's Gambit with 'Time-Window' Features

A single average of past behavior is not enough. We must capture the *story* of the borrower over time.

The Strategy

We engineered features that compared behavior across different time windows (e.g., last 6, 12, and 24 months) for all transactional data. This generated over 100 new, high-impact features.

Insights Unlocked

Short Term (6M)
Is the applicant in a financial crisis *right now*?

Medium Term (12M)
Are they generally stable?

Long Term (24M)
Do they have a solid, proven history?

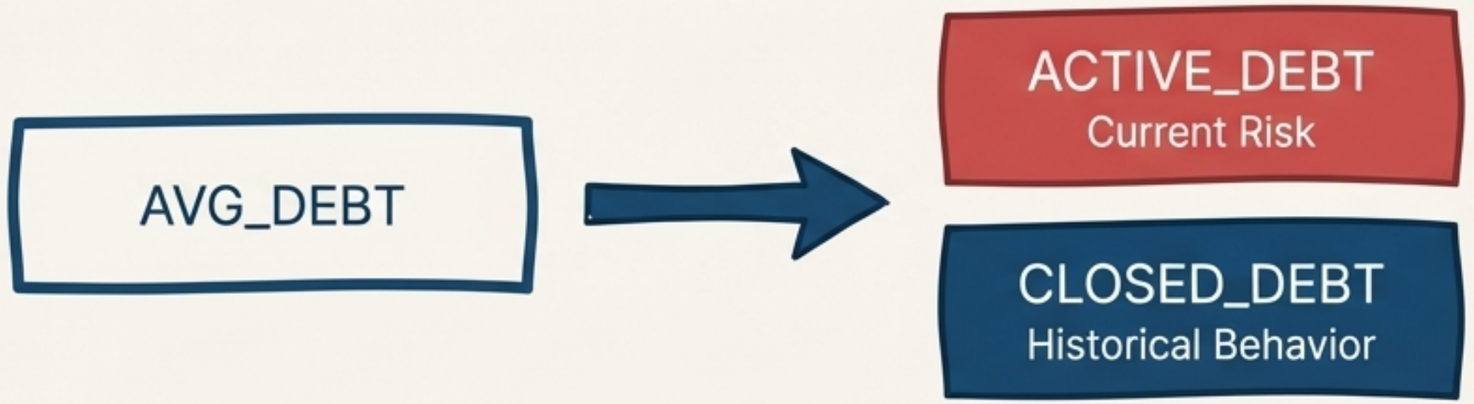
The Result

This temporal context created a paradigm shift, giving the model a dynamic view of the applicant's financial health and leading to our biggest single leap in model intelligence.

AUC: **0.7896**
(Time-Window Features Added)

Phase 4: Engineering the “Megaset” by Splitting by Context

The Insight: Not all history is equal. A current, active loan is a measure of present risk, while a closed loan from 5 years ago is a measure of past reliability. We needed to teach the model this distinction.



The Winning Techniques

1. ****Active vs. Closed Split****: Instead of `AVG_DEBT`, we created `ACTIVE_DEBT` (current risk) and `CLOSED_DEBT` (historical behavior).
2. ****Approval vs. Refused Split****: We calculated separate statistics for previously `Approved` applications versus `Refused` ones.

The Result

This final layer of context pushed our single LightGBM model to its peak performance, creating a 311-column "Megaset" that captured deep behavioral nuances.

AUC: 0.7921

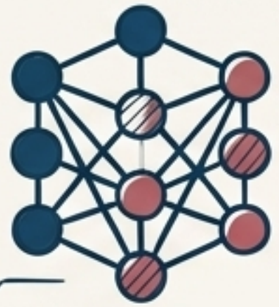
(Single LGBM on Megaset)

Phase 4: The Final Polish with a Dual-Model Ensemble

The Logic: No single model is perfect. By combining two world-class models, we allow them to correct each other's blind spots.

0.7921

LightGBM (The Aggressive Specialist): Incredibly fast and accurate, grows trees 'leaf-wise' to quickly minimize error. Our star player, scoring 0.7921."



CatBoost (The Robust Cousin): Handles data differently, using 'Ordered Boosting.' It often finds patterns the others miss. We upgraded it with our Megaset features to be a strong partner."

AUC: 0.7933 (Final Ensemble)

The Blend

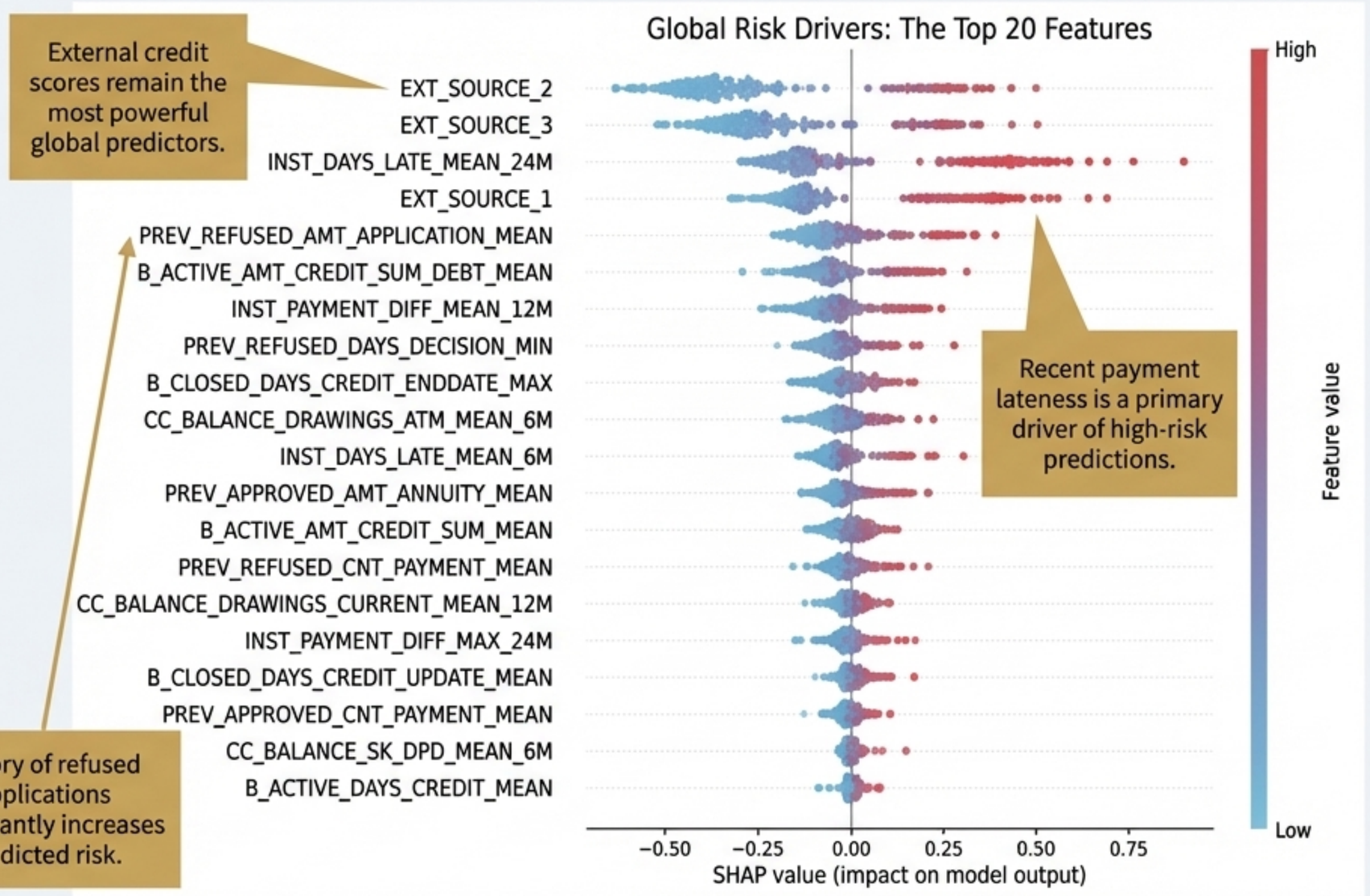
We created a weighted average of the predictions from both models (60% LightGBM / 40% CatBoost). This simple blend cancelled out individual errors and provided the final boost to our score.

'The View from the Summit: Explaining the "Why" with SHAP'

The Need for Interpretability

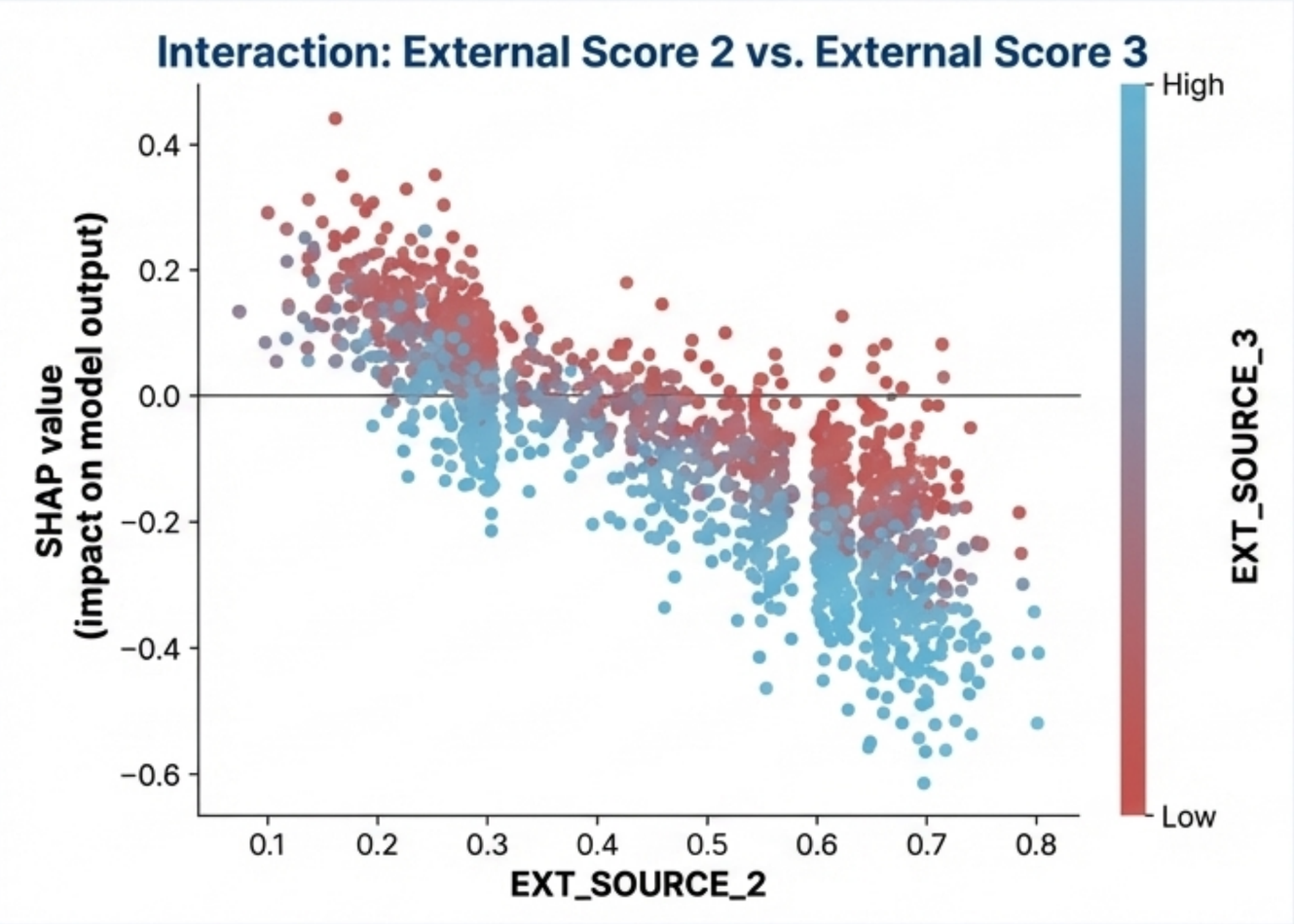
A high AUC score is not enough. We need to understand the key drivers of risk to trust the model and generate actionable insights.

Our Tool: We use **SHAP (SHapley Additive exPlanations)**, a game theory-based method to calculate the precise contribution of each feature to a specific prediction.



Uncovering Nuance: How Features Interact

SHAP allows us to visualize how the impact of one feature changes based on the value of another. This reveals complex, non-linear relationships learned by the model.

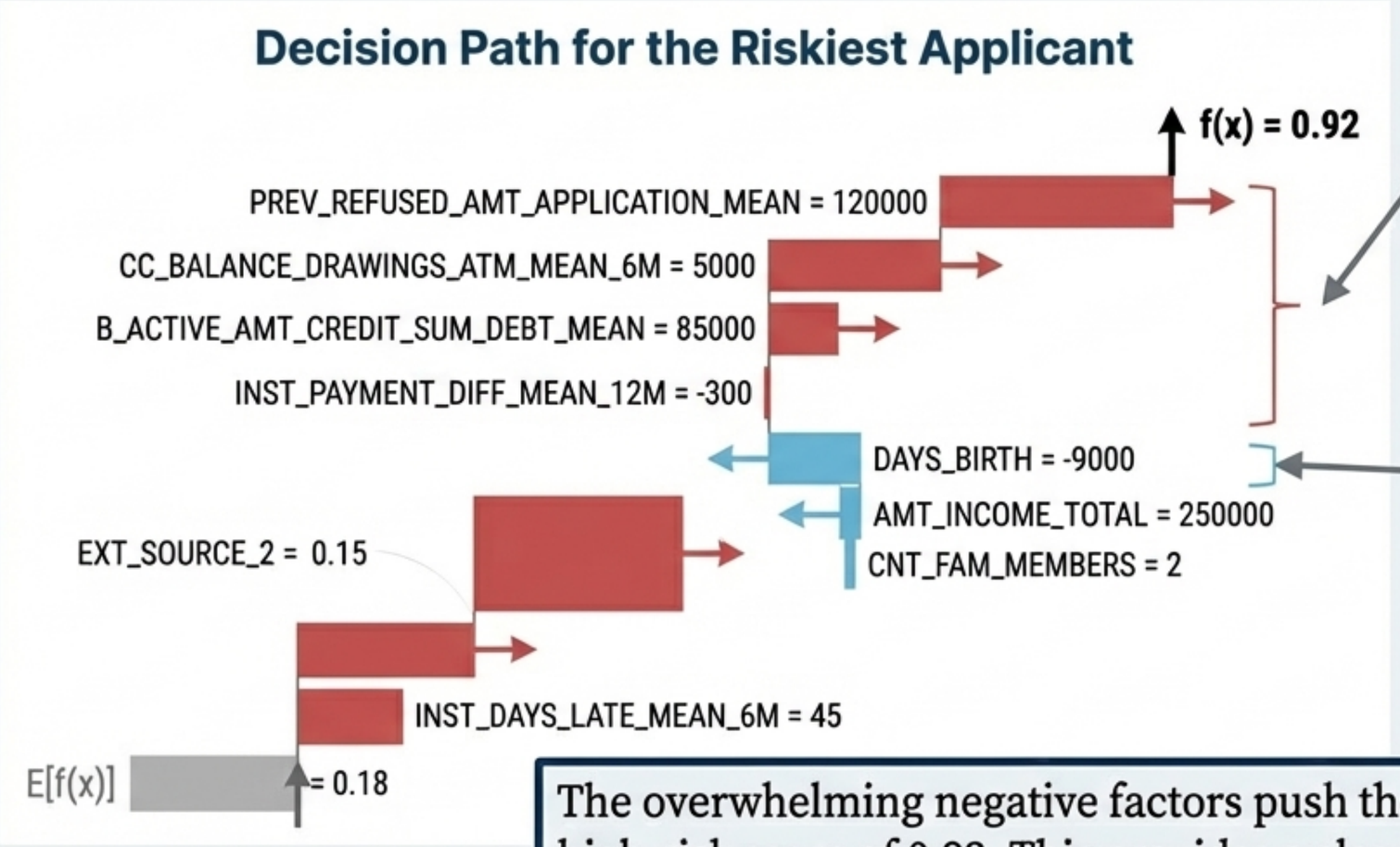


As EXT_SOURCE_2 increases (gets better), its SHAP value decreases, lowering the risk prediction.
The vertical coloring shows a powerful powerful interaction: for any given EXT_SOURCE_2 value, having a high EXT_SOURCE_3 (blue dots) further lowers the risk, while a low EXT_SOURCE_3 (red dots) negates the benefit.

Conclusion: The model learned that being highly rated by *both* credit bureaus is a multiplicative sign of safety.

From 300,000 Applicants to One: A Single Prediction Explained

The Scenario: Let's analyze the riskiest applicant in our validation sample, who had a **92% predicted probability of default.** Why did the model flag them?



Key Risk Drivers
Low external credit scores (EXT_SOURCE_2), significant late payments in the last 6 months, and a history of refused applications and high current debt significantly drive the risk score up.

Mitigating Factors
Higher age (DAYS_BIRTH), substantial income, and a stable family size slightly reduce the risk, but are heavily outweighed by the risk drivers.

The overwhelming negative factors push the prediction from the baseline to the final high-risk score of 0.92. This provides a clear, auditable trail for the decision.

The Summit View: A Journey from Baseline to Top 5% Performance.

Phase	Key Strategy	AUC Score
Phase 0	Logistic Regression Baseline	0.7471
Phase 1	LightGBM + Domain Features	0.7705
Phase 2	All Source Data Integrated	0.7850
Phase 3	Time-Window “Nuclear” Features	0.7896
Phase 4	“Megaset” Contextual Features	0.7921
Phase 5	**Final LGBM + CatBoost Ensemble**	🏆 **0.7933**

Developed a credit scoring ensemble achieving **0.7933 AUC**, placing in the **Top 5%** of global benchmarks by engineering 311 advanced features from over 27 million transactional records.