PROJECT
December 2025

NATIONAL ECONOMICS UNIVERSITY, VIETNAM

The Power of Data Preparation: A Data Storytelling Approach to Credit Default Risk Prediction

Course ID: Data Preparation and Visualization

Student ID: 11230545
Full Name: Nguyen Khanh Huyen

Student ID: 11230532
Full Name: Nguyen Thi Huong Giang

Student ID: 11230569
Full Name: Nguyen Dang Minh

Student ID: 11230585
Full Name: Nguyen Thi Ha Phuong

Student ID: 11230593
Full Name: Duong Thi Huyen Trang

Supervisor: Dr. Nguyen Tuan Long

Contents

Part 1: Data Storytelling

1 Executive Summary

In the high-stakes world of microfinance, where every lending decision affects the lives of underserved communities, the difference between success and failure often does not lie in the complexity of machine learning algorithms but in the quality of the data that feeds them. While it is tempting and common for practitioners to chase the latest sophisticated models or hyperparameter optimizations, the reality is stark: without clean, well-prepared data, even the most advanced model produces little more than garbage output, a classic case of "Garbage In, Garbage Out (GIGO)."

Microfinance institutions (MFIs) serve as financial lifelines for millions of underserved individuals, yet they face a critical dilemma: how to identify borrowers at risk of defaulting without access to traditional credit histories or comprehensive financial records. This report tells the story of how rigorous data preparation transformed a struggling credit risk prediction model into a reliable decision-making tool for MFIs.

Armed with a dataset containing borrower profiles, transaction histories, and repayment records, we initially set out to build a predictive model capable of flagging high-risk borrowers. However, our first attempts quickly revealed a universal truth familiar to any data analyst: raw data, regardless of quantity, is often more obstacle than asset.

Our analysis demonstrates that data preparation is not a minor technical step but the very foundation of reliable machine learning predictions. Through systematic understanding, cleaning, transformation, and feature engineering of microfinance transaction data, we show how proper data preparation converts mediocre models into robust decision-making tools that meaningfully reduce default risk.

The story unfolds through three critical phases. First, we expose the hidden dangers lurking in raw data, including missing values, inconsistencies, duplicate records, and noise that obscure meaningful patterns and mislead algorithms. Second, we document the transformation journey, where each data preparation technique (cleaning, transformation, reduction) progressively reveals clearer insights and stronger predictive signals. Finally, we present irrefutable evidence of impact through side-by-side comparisons that show dramatic improvements in model accuracy, reliability, and business value.

The takeaway is unambiguous: investing time and rigor in data preparation yields returns far greater than chasing complex model architectures alone. For microfinance institutions, this translates into fewer missed defaults, more responsible lending decisions, and sustainable operations that continue to serve vulnerable communities. Through data-driven storytelling, this report proves that in machine learning for credit risk assessment, data preparation is not optional; moreover, it is everything.

2 Introduction and Context

2.1 The Real-World Problem

Millions of individuals around the world face a fundamental barrier to economic opportunity: they cannot access formal credit. The reason is straightforward yet devastating that they lack traditional credit histories. Without records of past loans, credit cards, or formal banking relationships, these individuals are invisible to conventional lending systems. Financial institutions, unable to assess their creditworthiness through standard metrics, simply reject their applications.

This credit invisibility creates a vicious cycle. Unable to secure loans from legitimate institutions, the unbanked population often turns to predatory lenders who charge exorbitant interest rates and impose exploitative terms. These untrustworthy lenders trap vulnerable borrowers in cycles of debt, undermining their financial stability, and perpetuating poverty. The consequence is not only merely individual hardship, but is also a systemic failure that locks entire communities out of economic participation.

Home Credit, a consumer finance provider committed to broadening financial inclusion, confronts this challenge head-on. Their mission is to provide positive and safe borrowing experiences for the unbanked population, ensuring that the lack of traditional credit history does not become a permanent barrier to financial opportunity.

However, the challenge is nuanced. The goal is not simply to approve more loans; it is to make the right lending decisions. Home Credit must identify clients who are capable of repayment but would be rejected by traditional systems, while simultaneously protecting vulnerable borrowers from taking on debt they cannot manage. Every false negative (rejecting a creditworthy applicant) denies someone an opportunity for economic advancement. Every false positive (approving a risky borrower) can lead to default, financial distress, and harm to both the institution and the individual.

2.2 The Dataset

The dataset provided contains comprehensive information about borrowers, their loan applications, repayment histories, and related transactional data. It includes demographic attributes, financial indicators, and behavioral information that collectively reflect client profiles. Key features include:

- **Client Information:** Gender, age, family status, education level, housing situation, income, number of children, employment details, ownership of assets such as cars or real estate.
- **Loan Details:** Loan type (cash or revolving), loan amount, annuity, goods price, application timing, and credit bureau inquiries.
- **Behavioral Indicators:** Previous payment patterns, social network influence on repayment (30 and 60 days past due), and document submissions.
- **External Sources:** Normalized scores from external data sources, regional ratings, and address consistency flags.

Each observation in the dataset corresponds to a loan application, with the target variable indicating whether the client experienced payment difficulties during the initial installments.

Column	Description
ID	ID of loan in our sample

Column	Description
TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
CNT_CHILDREN	Number of children the client has
AMT_INCOME_TOTAL	Income of the client
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	Client's income type (businessman, working, maternity leave, ...)
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means more populated region)
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	How many days before the application the person started current employment
DAYS_REGISTRATION	How many days before the application did client change his registration
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
OWN_CAR_AGE	Age of client's car
FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_EMAIL	Did client provide email (1=YES, 0=NO)

Column	Description
OCCUPATION_TYPE	What kind of occupation does the client have
CNT_FAM_MEMBERS	How many family members does client have
REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives taking city into account (1,2,3)
WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
ORGANIZATION_TYPE	Type of organization where client works
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source
OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 DPD
DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
FLAG_DOCUMENT_2	Did client provide document 2

Column	**Description**
FLAG_DOCUMENT_3	Did client provide document 3
FLAG_DOCUMENT_4	Did client provide document 4
FLAG_DOCUMENT_5	Did client provide document 5
FLAG_DOCUMENT_6	Did client provide document 6
FLAG_DOCUMENT_7	Did client provide document 7
FLAG_DOCUMENT_8	Did client provide document 8
FLAG_DOCUMENT_9	Did client provide document 9
FLAG_DOCUMENT_10	Did client provide document 10
FLAG_DOCUMENT_11	Did client provide document 11
FLAG_DOCUMENT_12	Did client provide document 12
FLAG_DOCUMENT_13	Did client provide document 13
FLAG_DOCUMENT_14	Did client provide document 14
FLAG_DOCUMENT_15	Did client provide document 15
FLAG_DOCUMENT_16	Did client provide document 16
FLAG_DOCUMENT_17	Did client provide document 17
FLAG_DOCUMENT_18	Did client provide document 18
FLAG_DOCUMENT_19	Did client provide document 19
FLAG_DOCUMENT_20	Did client provide document 20
FLAG_DOCUMENT_21	Did client provide document 21
AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 months before application (excluding one month before application)
AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one year before application (excluding last 3 months)

Table 1: Description of dataset variables

3 The Problem: Raw Data and Associated Challenges

3.1 Exploring Raw Data

3.2 Experiments with Raw Data

This section evaluates the performance of three baseline machine learning models trained directly on the raw, unprocessed credit dataset. The dataset contained missing values, outliers, inconsistent feature scales, categorical variables with 50+ levels, and strong right-skewness. No cleaning, transformation, encoding, or feature engineering was applied.

Three models were tested:

- **Logistic Regression:** linear-based model
- **Random Forest:** tree-based model
- **XGBoost:** tree-based gradient boosting model

3.2.1 Logistic Regression (Linear-Based Model)

The Logistic Regression baseline demonstrates how unscaled numeric variables, high missingness, and extreme outliers disrupt linear separability. As a result, the model produces unstable and inconsistent decision boundaries.

Model Performance (Raw Data)

- Test ROC–AUC: 0.6203
- Test Accuracy: 0.6097
- Class 1 Precision: 0.0946
- Class 1 Recall: 0.5700
- Class 1 F1-score: 0.1622

Although recall for the minority class (default) appears deceptively high, precision is extremely low, indicating large numbers of false alarms.

Visual Analysis of Logistic Regression

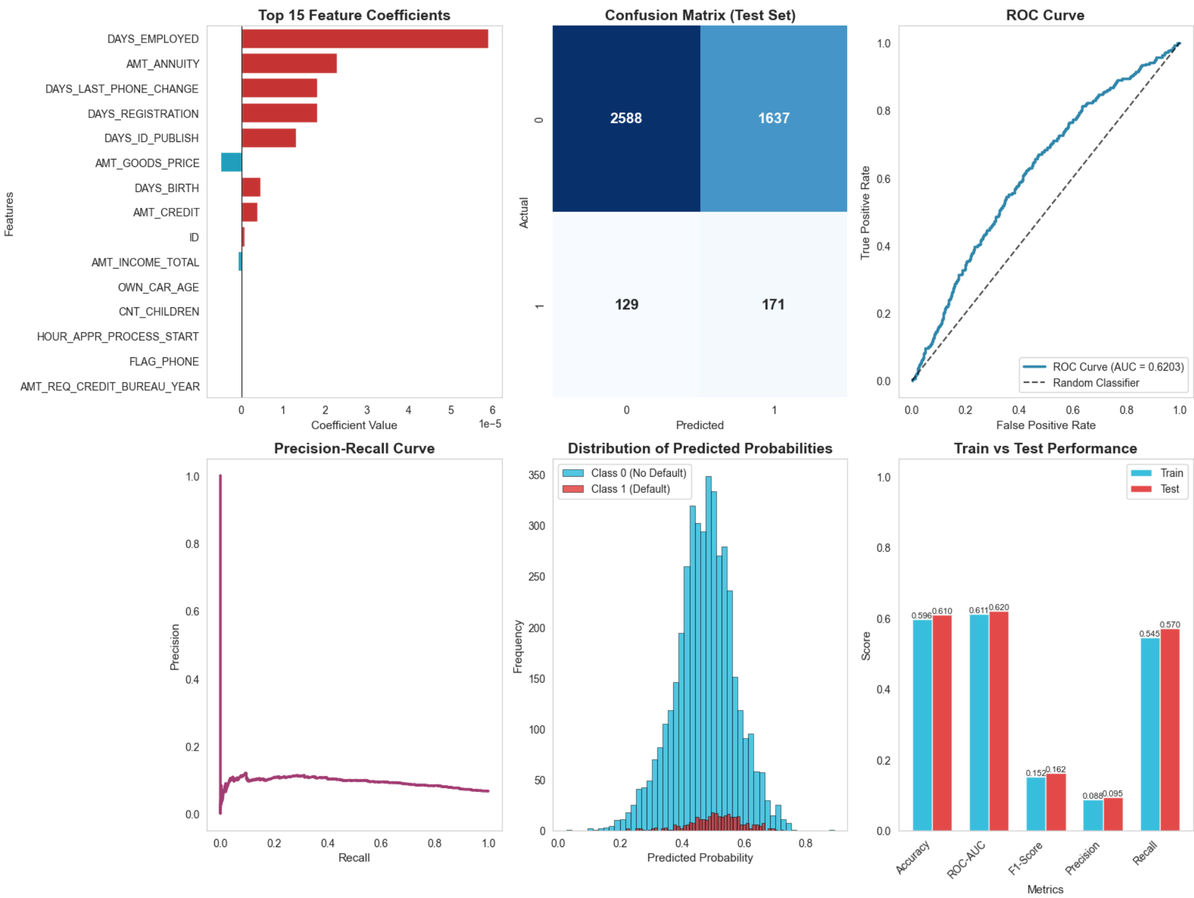

Figure 1: The visualization of Logistic Regression on Raw Data

Key Observations

- **Feature Coefficients Lack Meaning**
  The coefficient plot shows extremely small magnitudes and inconsistent signs. This indicates multicollinearity, scale distortion, suppressed signal due to noise. Even domain-relevant features like EXT_SOURCE variables do not appear among top predictors.

- **Confusion Matrix: High False Positives**
  The model misclassifies 1,637 non-defaulters as high-risk. This demonstrates the instability of raw linear separation.

- **ROC Curve Near the Diagonal**
  The ROC curve lies close to the random classifier line, confirming low discriminative power.

- **Precision–Recall Curve Shows Almost No Predictive Value**
  Precision rapidly collapses → the model cannot reliably identify default cases at any threshold.

- **Predicted Probabilities Cluster Around 0.45–0.55**
  The distribution of predicted probabilities shows: lack of separation between classes, no clear decision boundary, heavy noise overshadowing signal.

Conclusion Logistic Regression confirms that linear models cannot detect meaningful patterns when raw financial data is unscaled, inconsistent, and dominated by noise.

3.2.2 Random Forest (Tree-Based Model)

Random Forest is more flexible than linear models, yet it still failed to extract meaningful relationships from the raw dataset due to noisy splits, missing values, and irrelevant high-cardinality features.

Model Performance (Raw Data)

- Test ROC–AUC: 0.7007
- Test Accuracy: 0.9193
- Class 1 Precision: 0.1905
- Class 1 Recall: 0.0667
- Class 1 F1-score: 0.0988

The model overwhelmingly predicts the majority class, yielding floor-level recall for defaulters.

Visual Analysis of Random Forest

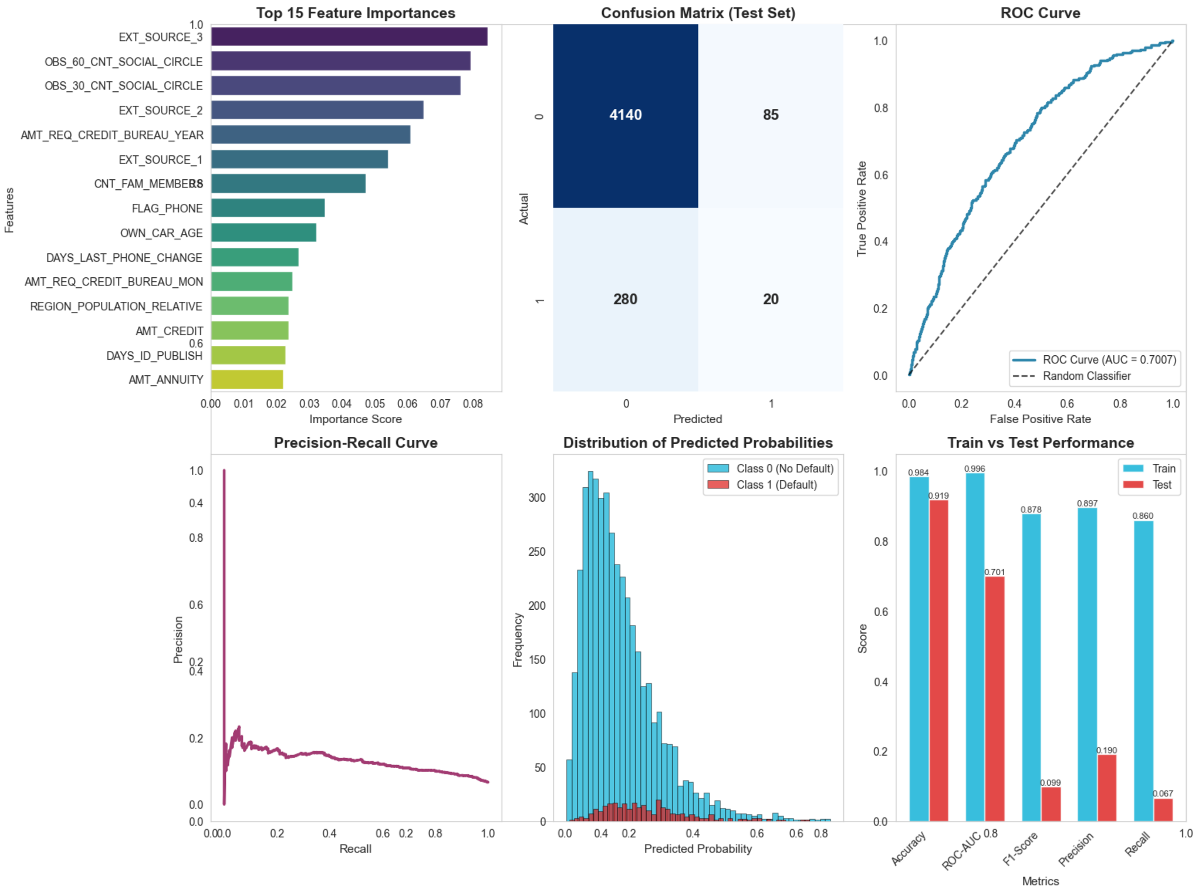

Figure 2: The visualization of Random Forest on Raw Data

Key Observations

- **Feature Importance Dominated by Noise:** Although EXT_SOURCE_3 appears as the top feature, the rest of the importance distribution is unstable and inconsistent with actual credit-risk behavior.
- **Confusion Matrix: Misses 93% of Defaults:** Only 20 out of 300 true default cases are detected. This shows the model overfits majority patterns while failing on minority detection.
- **Precision–Recall Curve Approaches Zero:** No meaningful trade-off exists; the curve indicates almost no ability to identify defaults.
- **Predicted Probabilities Extremely Skewed:** Most predictions cluster below 0.2, showing that the model saturates toward predicting non-default and that splitting rules fail due to missing values and inconsistent numeric scales.

Conclusion Random Forest is severely misled by raw noise, overfitting the majority class while missing critical minority patterns.

3.2.3 XGBoost (Tree-Based Gradient Boosting Model))

XGBoost is typically powerful at extracting nonlinear relationships. However, without imputation, encoding, or cleaning, it becomes biased toward noisy splits and missing-value default directions.

Model Performance (Raw Data)

- **Test ROC–AUC:** 0.7253
- **Test Accuracy:** 0.9335
- **Class 1 Precision:** 0.4737
- **Class 1 Recall:** 0.0300
- **Class 1 F1-score:** 0.0564

The model predicts extremely few defaulters, producing high precision but near-zero recall.

Visual Analysis of XGBoost

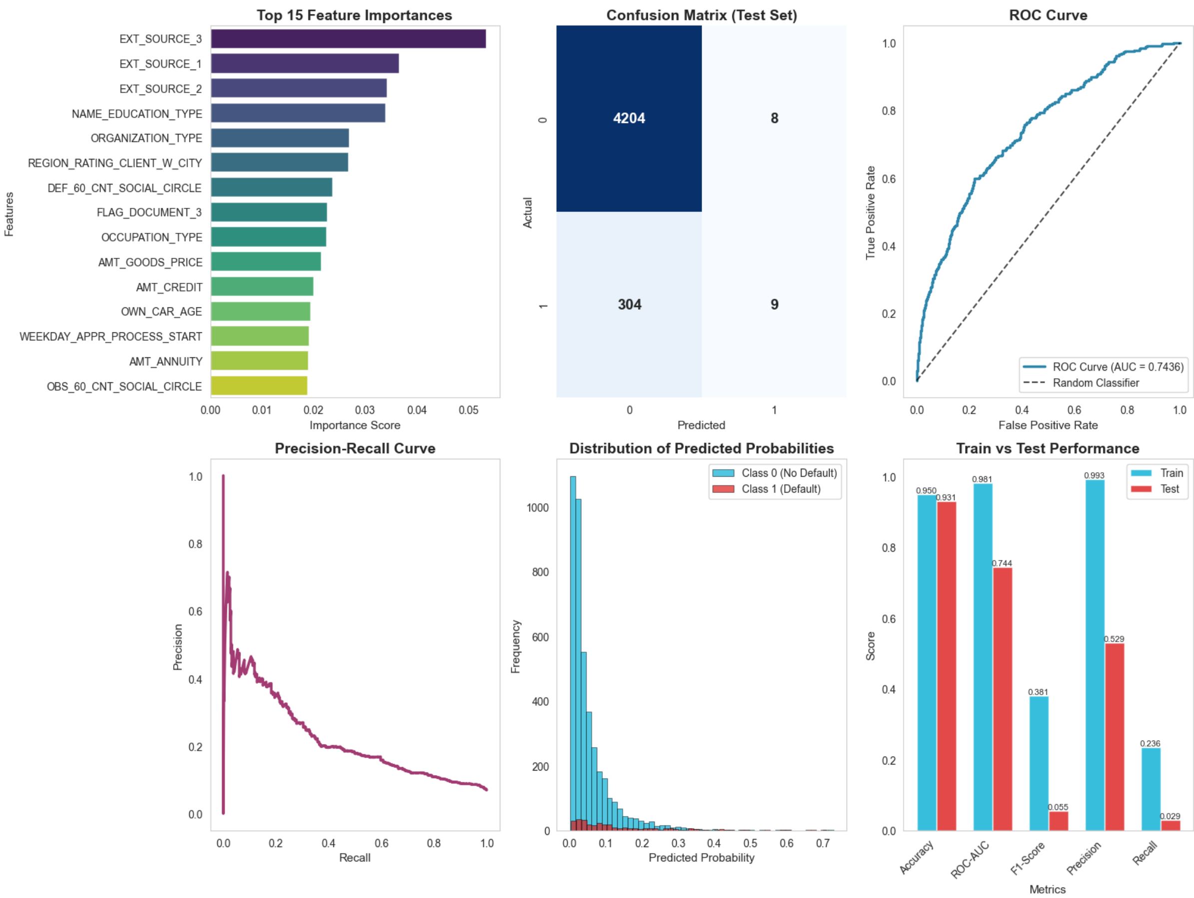

Figure 3: The visualization of on Raw Data

Key Observations

- **Feature Importance Shows Partial but Unstable Signals:** EXT_SOURCE features appear, but several high-noise variables also receive high importance, indicating brittle split logic.
- **Confusion Matrix: Almost Total Failure on Class 1:** Only 9 out of 304 default cases are correctly identified; the model collapses into predicting the majority class.
- **ROC Curve Slightly Higher Than Other Models:** Although AUC reaches 0.74, the performance does not translate into meaningful detection due to poor threshold calibration.
- **Predicted Probability Distribution Highly Skewed:** Most predicted probabilities are below 0.1, meaning the model views nearly all applicants as "very low risk" and misses critical signals.

Conclusion Even XGBoost cannot extract meaningful structure from unprocessed data—the underlying patterns are buried beneath missingness and noise.

3.2.4 Cross-Model Insights: Why Raw Data Fails

Across all three models, consistent failure patterns appear:

- **Hidden Insights:** Important financial behaviors—such as credit bureau history, late payments, and debt ratios—do not manifest in the models.
- **Visual Evidence of Noise:** Prediction probability histograms, PR curves, and ROC curves all reveal inconsistent distributions, class overlap, and missing structure.
- **Majority-Class Overlearning:** Tree models collapse into predicting nearly all observations as non-default.
- **Severe Distortion from Missing Values:** Missing EXT_SOURCE features alone cause poor splits in tree models and unstable coefficients in linear models.
- **Outliers Break Linear Relationships:** Unscaled income, loan amount, and annuity variables introduce extreme leverage effects.

The baseline results collectively show that raw data cannot support reliable credit-risk modeling: Logistic Regression misclassifies thousands of customers and fails to capture any meaningful trend. Random Forest and XGBoost severely overfit and fail to detect minority risk, despite high apparent accuracy. Visualizations clearly reveal noise-dominated distributions, unstable patterns, and poor class separation.

Therefore, data preparation is not optional. It is the prerequisite for uncovering meaningful insights and building effective ML models.

4 The Solution: Data Preparation Journey

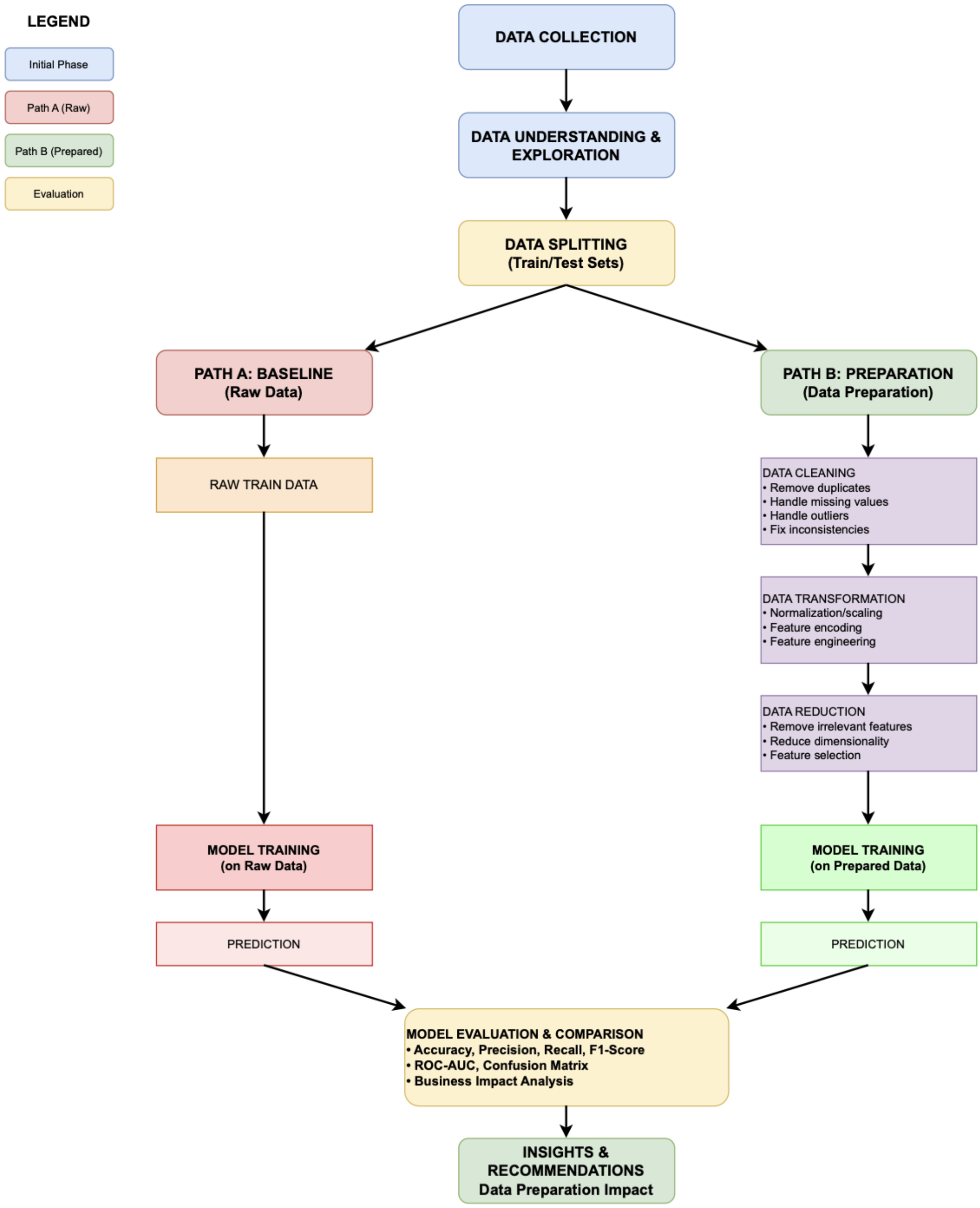

Figure 4: The flow diagram of this project

5 The Impact: Clear Comparison of Results

6 Call to Action and Recommendations

Part 2: Technical Analysis

1 Storytelling Strategy Evaluation

bvvnmhbjvgvvvhk

2 Principles of Data Visualization

bbbvvbvbjhbjhbjbjhb

3 Technical Implementation Details

nhdbskfjdsnfklsnfk

4 Design Decisions and Rationale

ghhjgkgkjgkjjvhvj

5 Lessons Learned and Best Practices

gjhgjghjgjhgjk