Vehicle Loan Default Prediction

Shicheng Chen

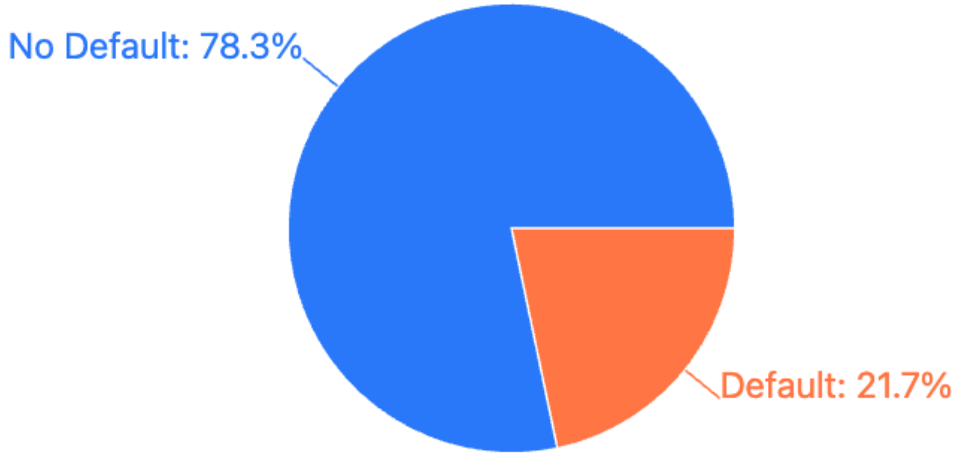
Problem Statement & Objective

- Problem: Financial institutions need to identify potential defaulters before loan approval
- Objective: Build a predictive model to classify good vs. bad borrowers
- Business value: Reduce default rates, optimize lending decisions

Dataset Overview

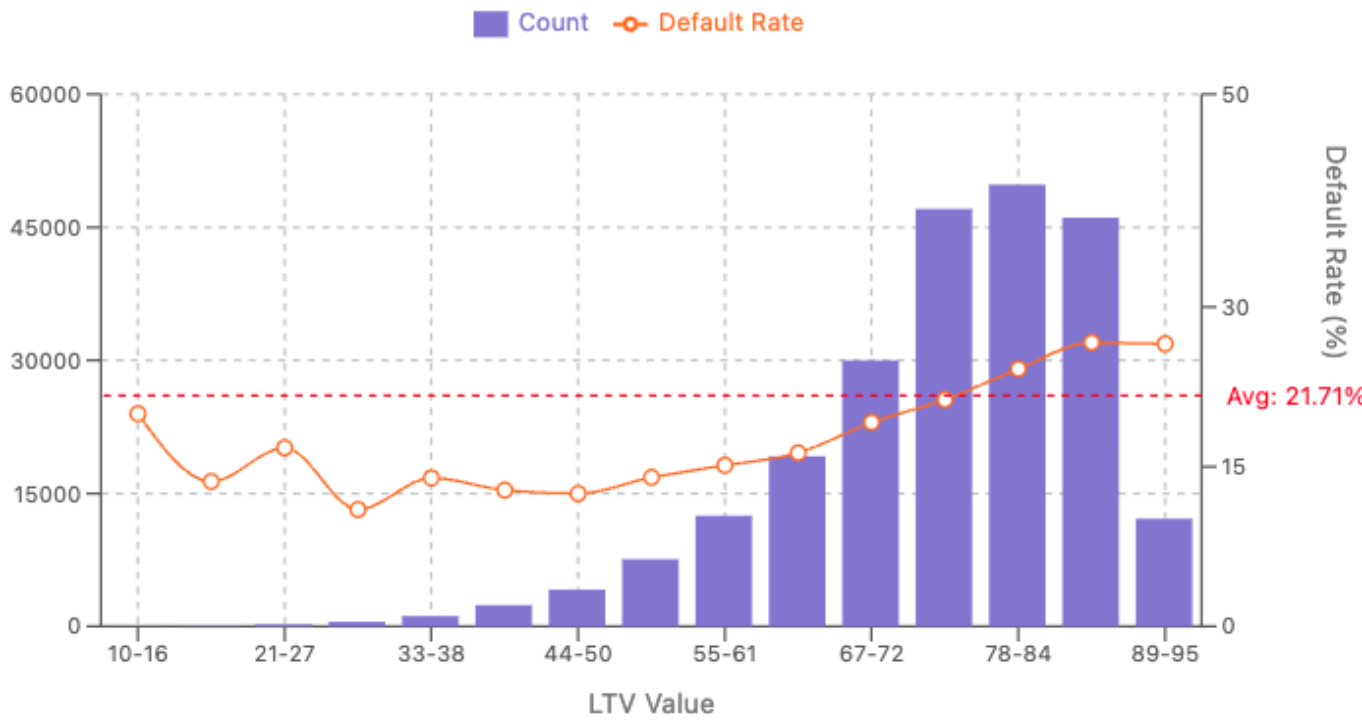
- Dataset size: 233,154 rows and 42 columns
- Data duration: 3 months (03-08-2018 to 31-10-2018)

- Key features: Loan details, borrower information, credit history
- Target variable: loan_default (binary classification problem)



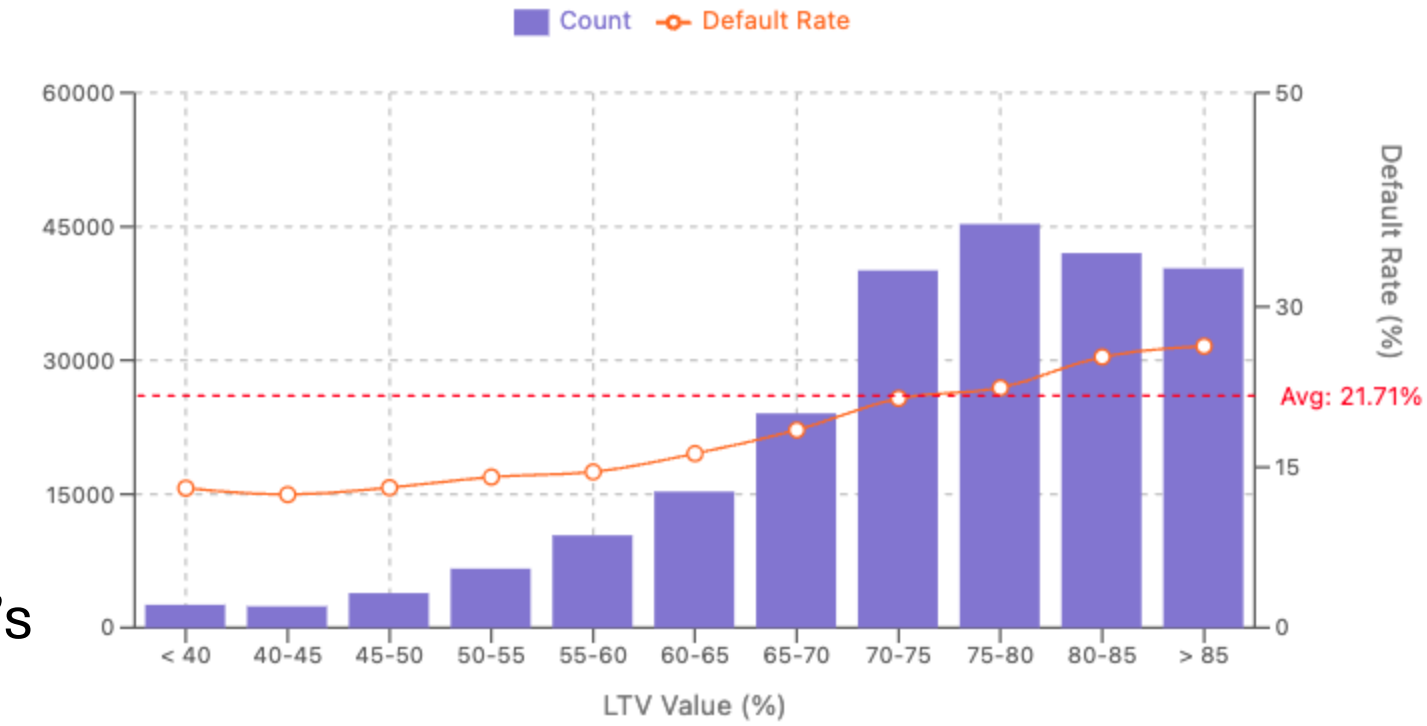
Feature Exploration: LTV

- Loan-to-Value (LTV): the amount of a loan compared to the vehicle's value
- Default rates show a positive correlation with LTV values.
- Correlation weakens at extreme LTV ranges (<40 and >85) due to limited sample sizes in these segments.



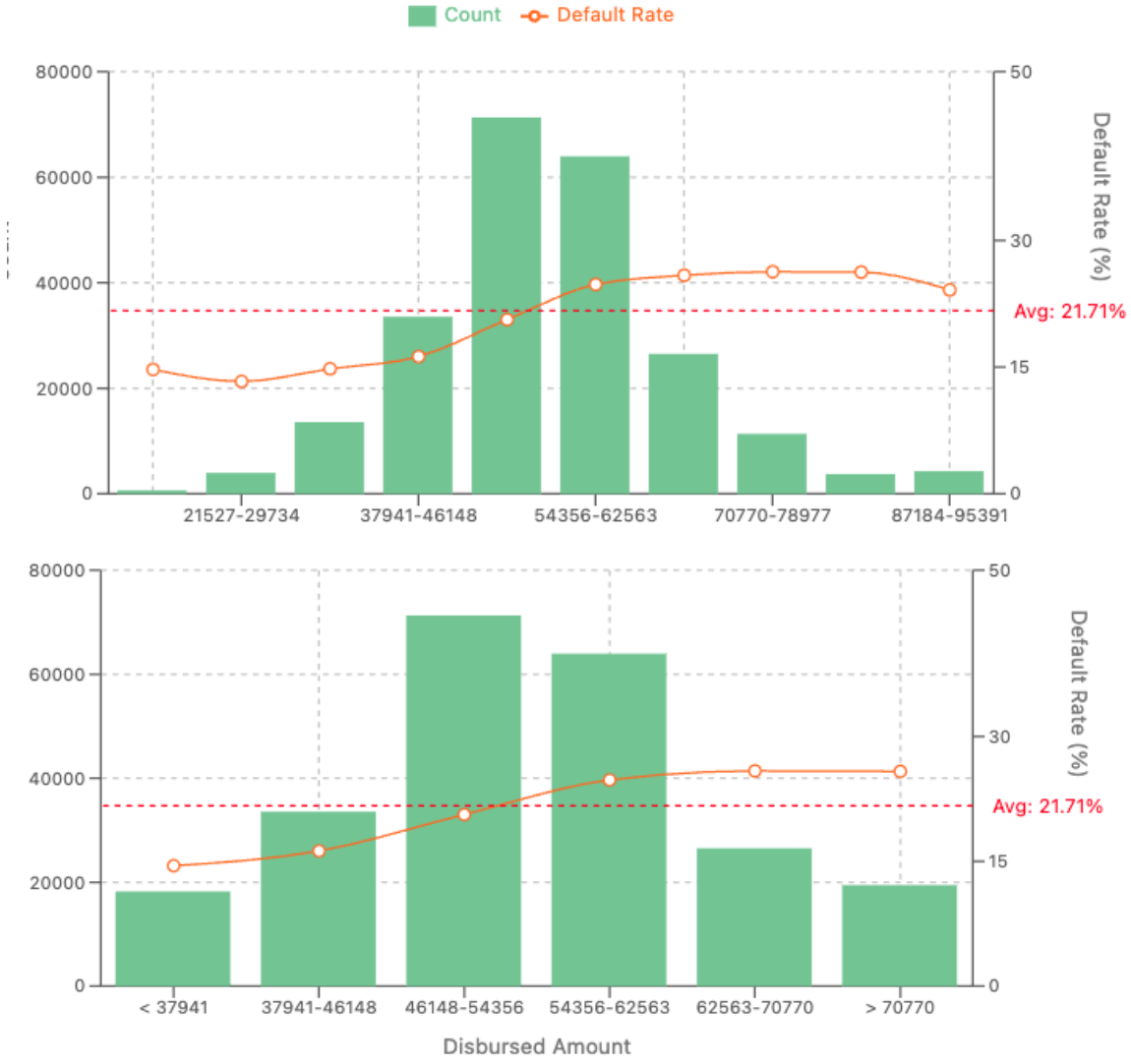
LTV vs Default Rate: Risk Prediction (Improved Visualization)

- Observations:
  - Strong positive correlation with default rate
  - Lowest default: <40% LTV
  - Highest default: >85% LTV
- Reasons:
  - Higher LTV increases payment burden
  - Lower equity reduces borrower's commitment



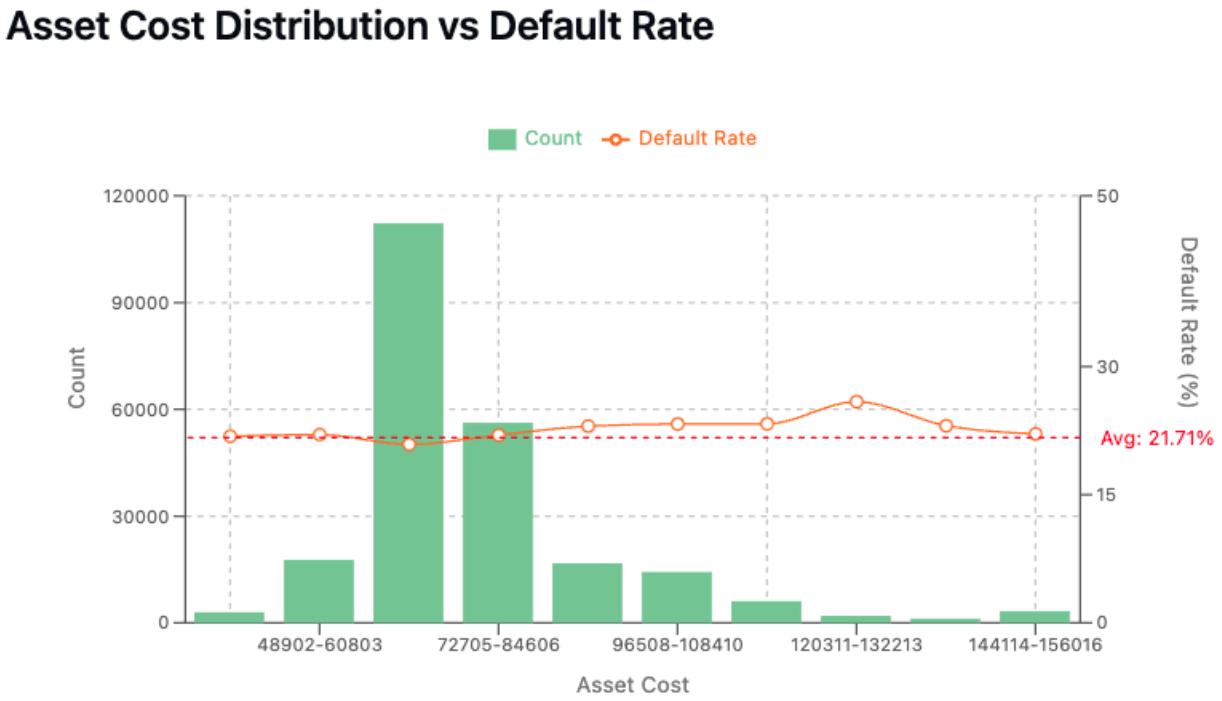
Loan Amount vs. Default Risk Analysis

- Feature:
  - Disbursed Loan Amount
- Observation:
  - Highest loan volume: $46K-$63K
  - Default rate grows: small (15%) → large loans (25%)
- Reasons:
  - Small loans: affordable repayments
  - Large loans: heavy repayment burdens; higher financial stress



Asset Cost vs. Default Rate – Feature Effectiveness Evaluation

- Observation:
  - Predictive value limited due to uneven distribution.
- Why This Happens:
  - Limited samples in extreme asset ranges can skew predictability.



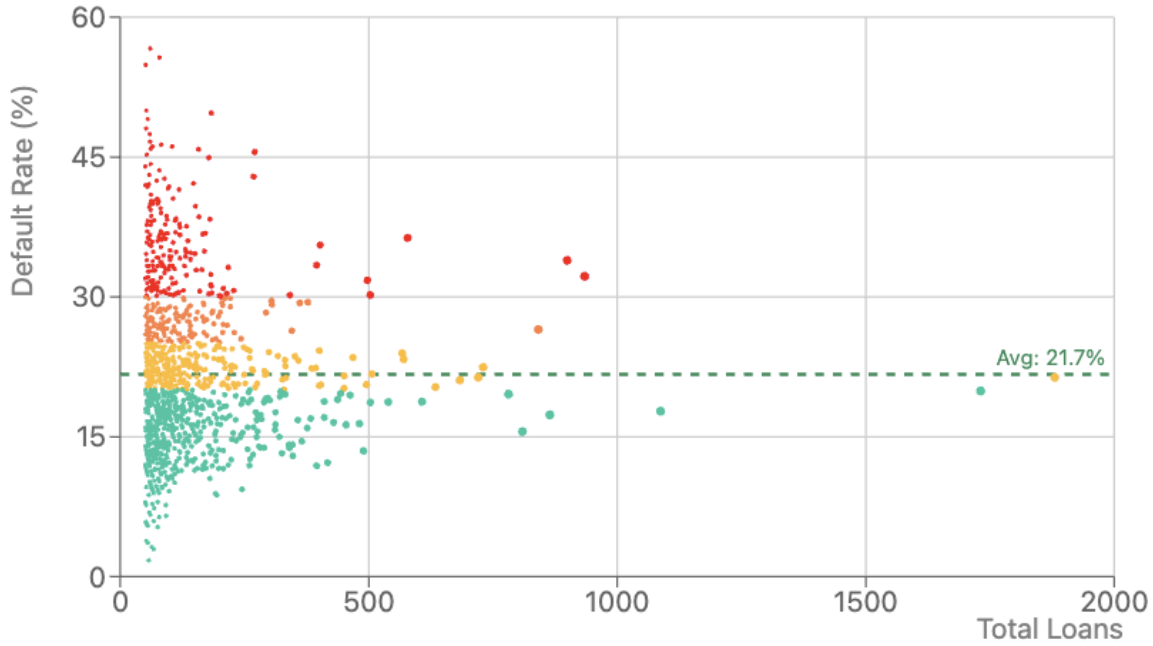
The Importance of ID Features

- Decision Pattern Analysis:
  - Branch/Employee IDs reveal location patterns and decision tendencies
- Geographic Factors:
  - State IDs and Pincodes capture economic conditions, regulatory frameworks, and cultural variations
- Product Quality:
  - Manufacturer and Supplier IDs identify reliability variations

Geographic Loan Default Analysis: Pincode Impact

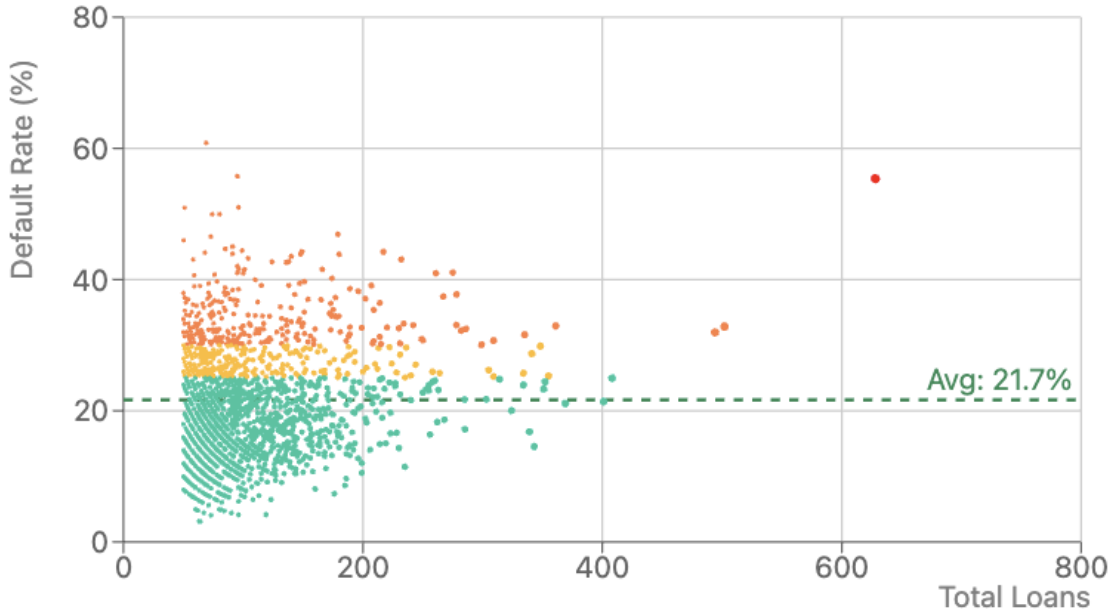
- Most pincodes having fewer than 50 loans
- Only 21 pincodes have more than 500 loans
- Default rates vary significantly by pincode, even for areas with sufficient data

- Pincode 3000 has the highest default rate (36.3%) among high-volume areas
- Pincode 1794 has the lowest default rate (15.6%) among high-volume areas



Human Factor in Loan Defaults: Employee ID Performance Metrics

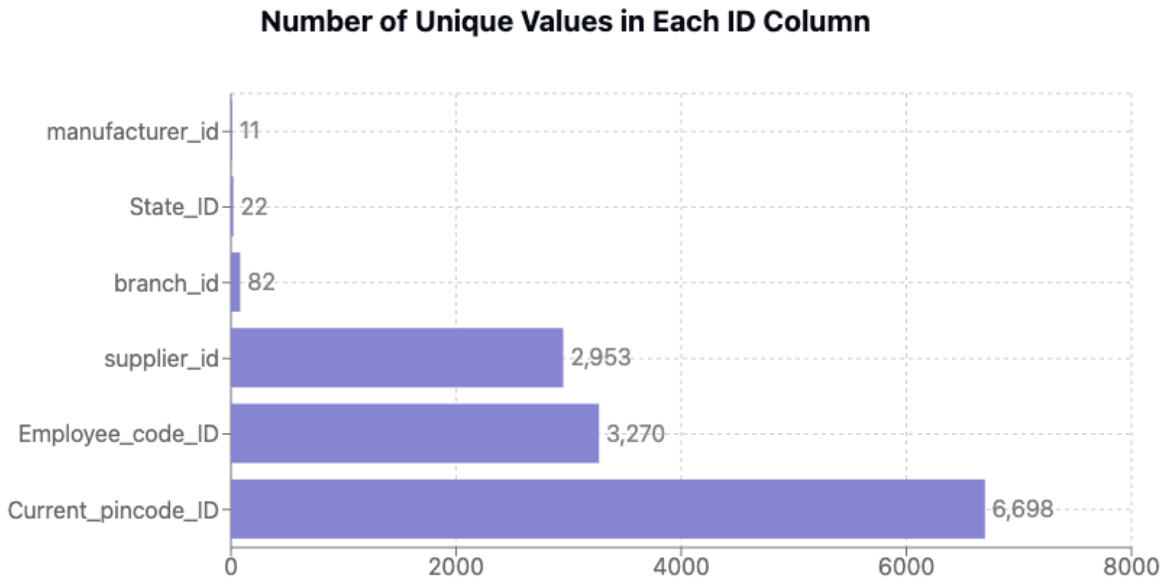
- Employee performance varies dramatically
- Employee 2546 processed a high volume (628 loans) but has an alarmingly high default rate (55.4%)
- Some employees consistently maintain low default rates despite high volumes



High Cardinality Features

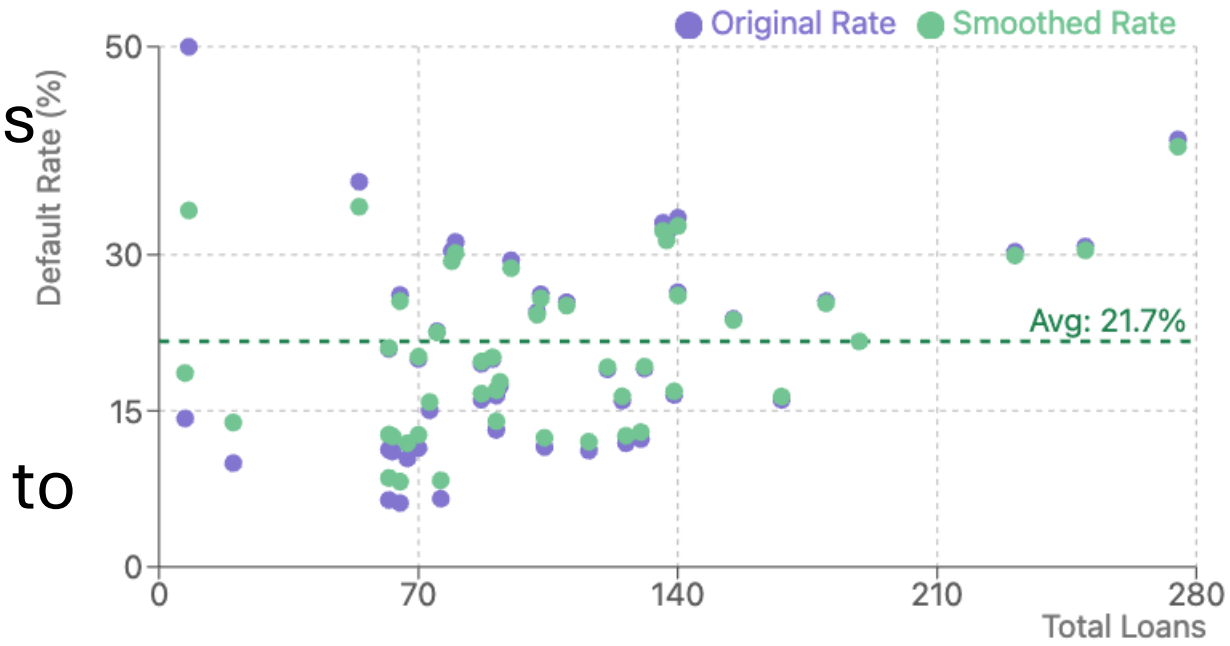
- Problem: One-hot encoding becomes infeasible
  - too many columns, memory issues.

- Solutions:
  - Smoothed target encoding
    - replaces categories with regularized target statistics
  - CatBoost
    - directly handles categorical features



Smoothed Target Encoding

- Smoothed target encoding
  - replaces categorical variables with a blend of the target variable's mean for that category and the global mean
- Original and smoothed values converge as sample size increases
  - Minimal smoothing effect on larger samples
- Use only training set default rates to prevent data leakage



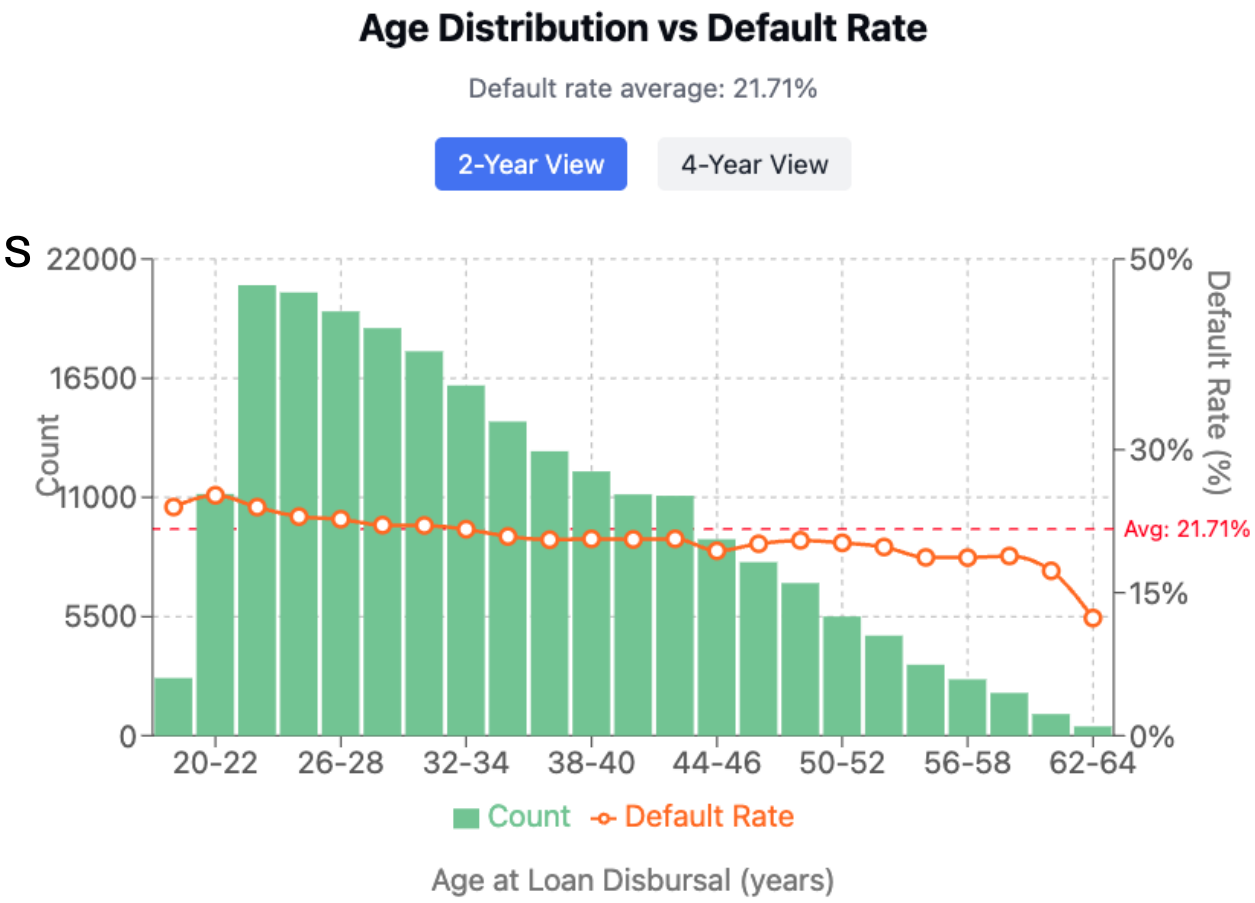
Feature engineering

- Handling Missing values:
  - Employment.Type: filled missing values with Unknown
- Created features
  - Age at Disbursal:
    - Captures customer age at loan disbursement
  - Disbursal Day:
    - Identifies potential cyclical patterns in loan disbursement timing

- *Note: These are select examples from a total of* 105 *engineered features created for the model*.

Age Distribution vs Default Rate

- Feature: Borrower age (20-64 years)
- Key Points:
  - Strong inverse relationship: default rates decrease with age
  - Youngest borrowers show highest risk (25%)
- Factors:
  - Financial stability increases with age
  - Younger borrowers face income uncertainty
  - Older borrowers have more debt management experience

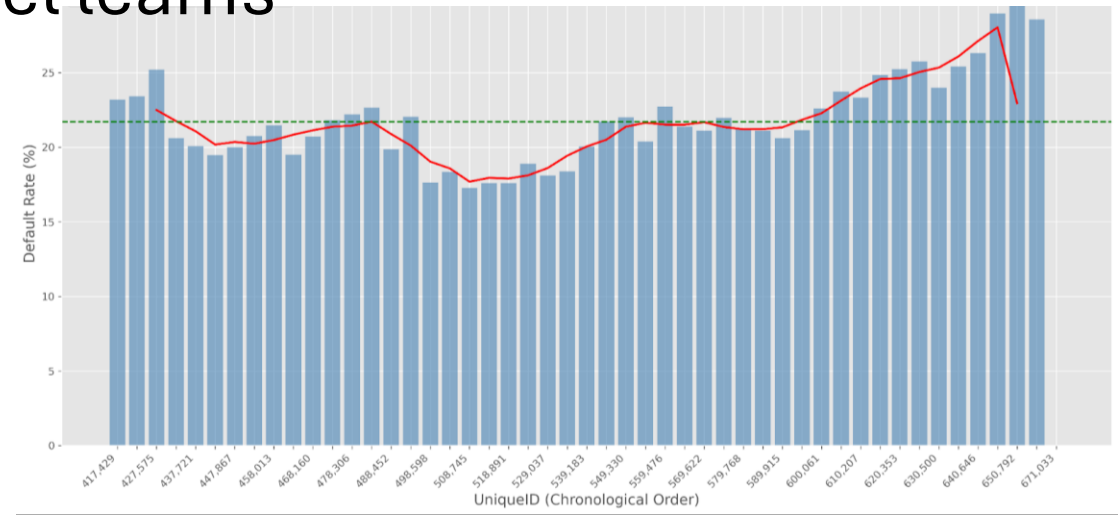


Modeling Approach

- Data split:
  - First 80% for training, last 20% for testing
- Model:
  - CatBoost
- Evaluation:
  - AUC-ROC metric on held-out test data to handle class imbalance
- Ensemble:
  - CV Ensemble
  - Mean Ensemble
  - Stacking Ensemble

Avoiding Data Leakage by Using Time-Based Train-Test Split

- Initial approach: 80/20 stratified split
- Discovery: UniqueID ranked as top 5 important feature
- Investigation: UniqueID appears time-ordered with visible patterns
- Problem: Information leakage as model learns temporal patterns
- Solution: Changed to chronological split (first 80% for training, last 20% for testing)



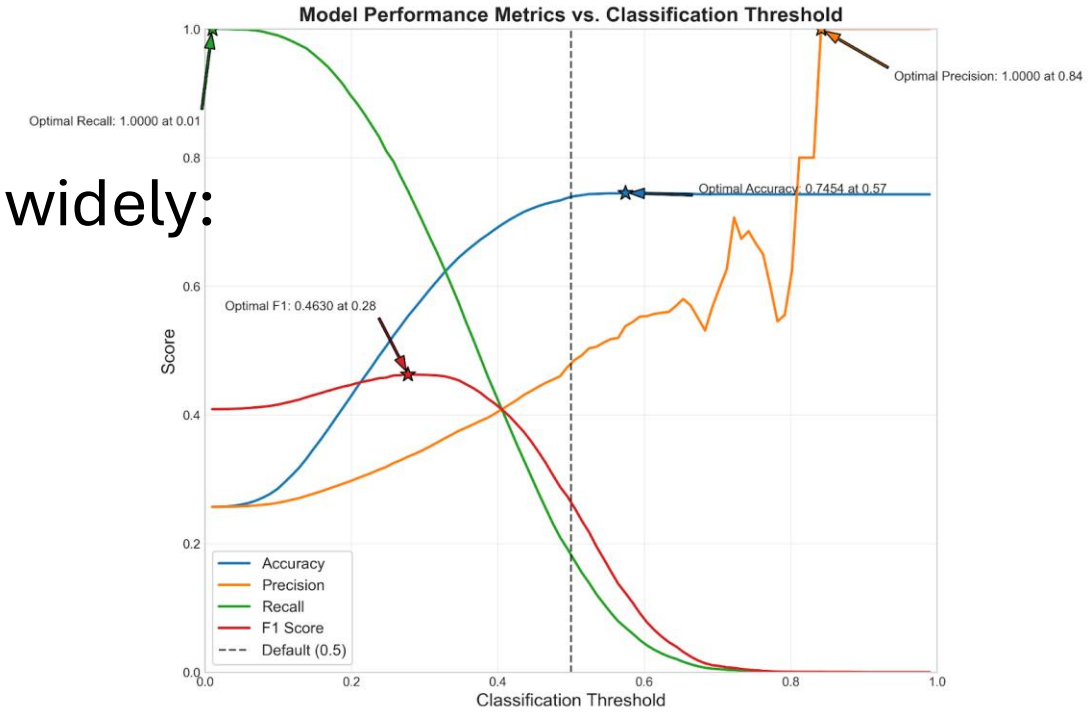
CatBoost Superiority for High Cardinality Categorical Features

- CatBoost outperforms other models (LightGBM, XGBoost) when processing high-cardinality categorical features
- Achieves superior AUC (0.6592) compared to competitors
- Remains effective even when others use target encoding or native category handling
- Recommended as the optimal choice for datasets with important categorical variables

Model	AUC
LightGBM (native category handling)	0.5635
XGBoost (native category handling)	0.5499
XGBoost (CatBoost-style Encoding)	0.6443
CatBoost	0.6592

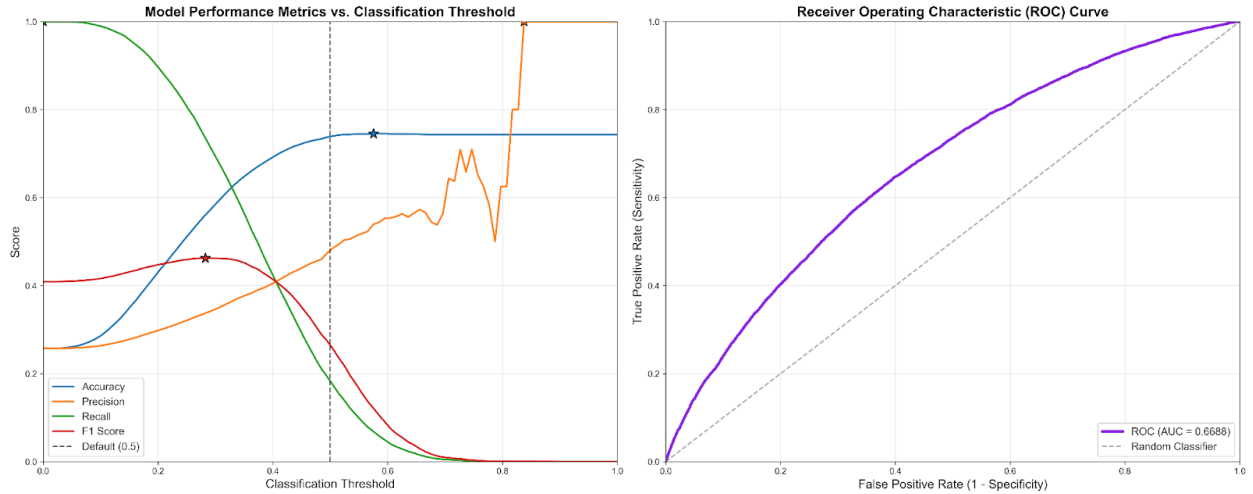
Why Common Metrics Mislead in Imbalanced Classification

- Accuracy
  - Misleads with imbalanced data
  - Example: With 79% positive cases, predicting "always positive" yields 79% accuracy despite zero insight
- Recall, Precision, F1
  - Threshold-dependent values
- The graph shows optimal thresholds vary widely:
  - F1 optimal at 0.28
  - Accuracy optimal at 0.57
  - Precision optimal at 0.84
  - Recall optimal at 0.01



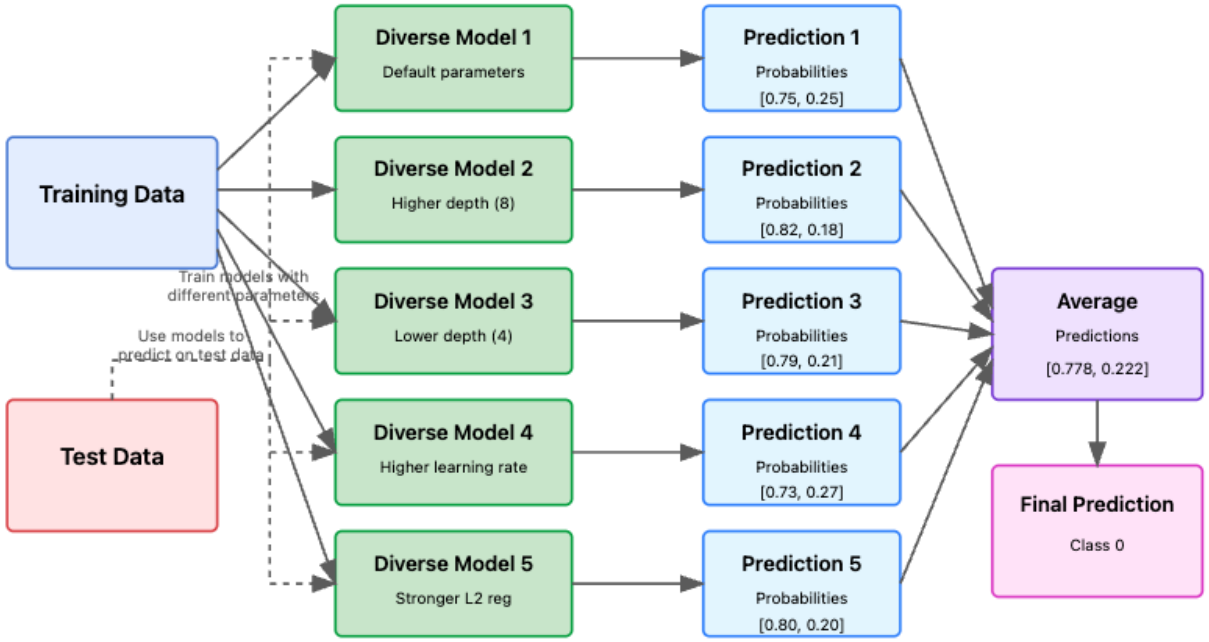
Why AUC-ROC Is Superior for Imbalanced Classification

- AUC-ROC Advantages
  - Threshold-independent evaluation
    - Measures performance across all possible thresholds
  - Not affected by class imbalance unlike accuracy, precision, recall
- Left Graph Problem
  - Shows how threshold choice drastically affects traditional metrics
  - No single optimal threshold satisfies all metrics simultaneously



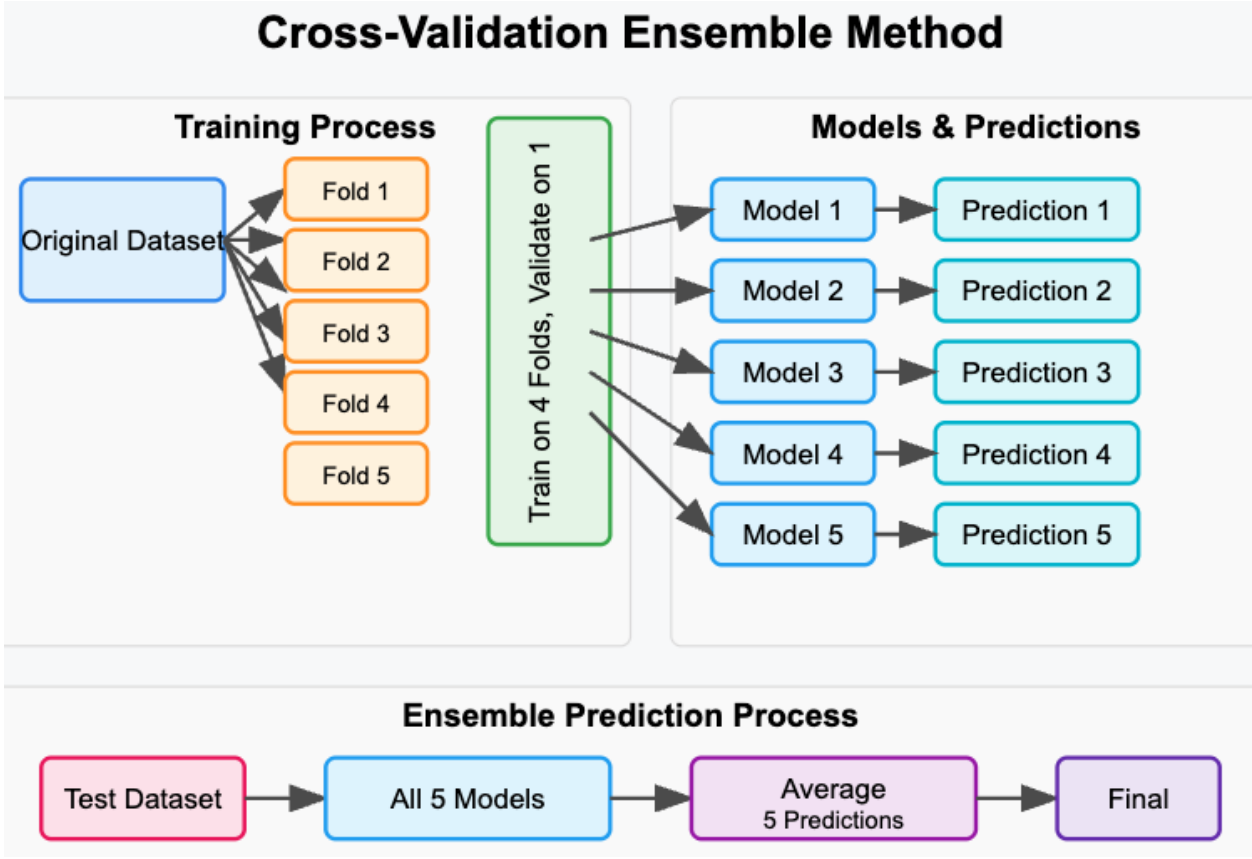
Diverse Model Ensemble for Improved Prediction Accuracy

- Multiple diverse models with different parameters trained on the same data, then averaged to produce more robust final predictions.



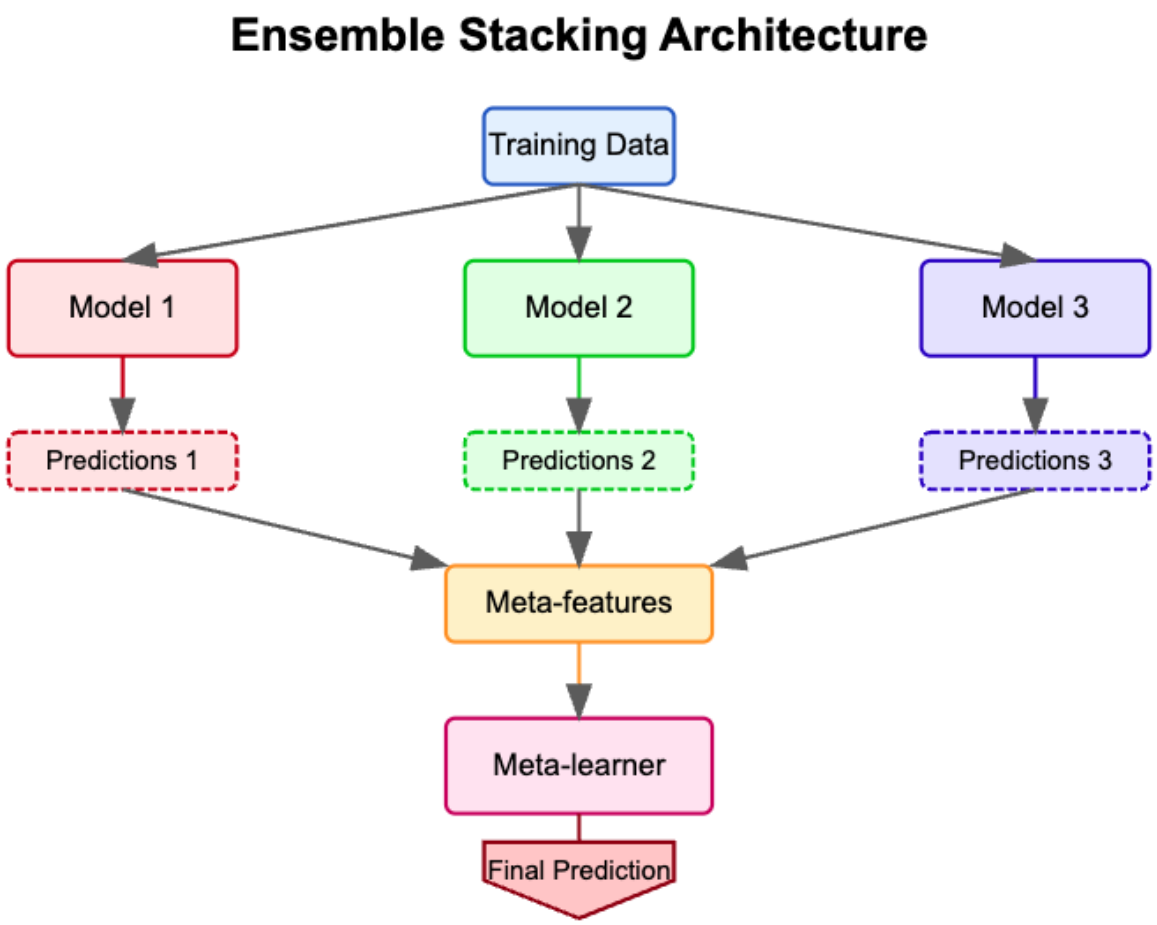
Cross-Validation Ensemble

- A technique combining predictions from multiple models trained on different data splits to improve performance and reliability.



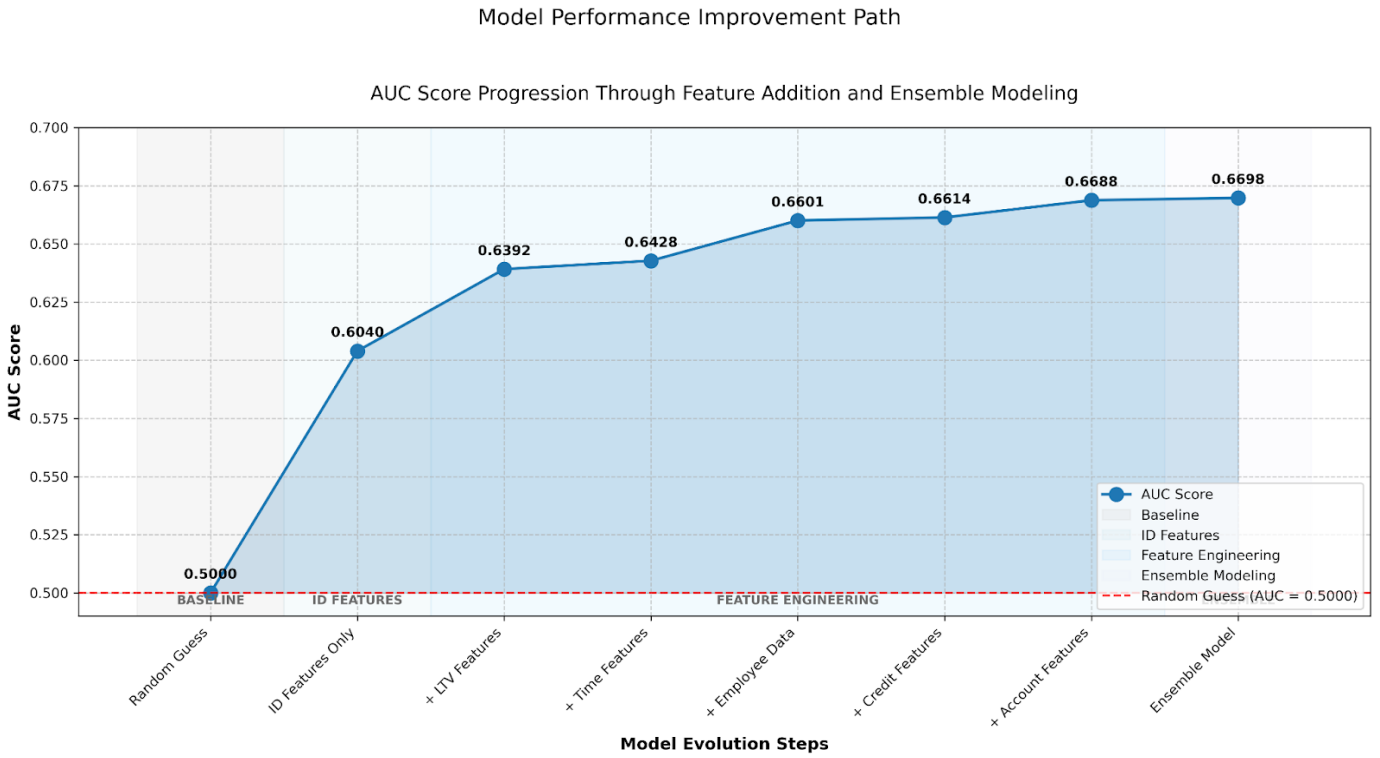
Ensemble Stacking Architecture

- Multiple base models make predictions that become meta-features for a meta-learner, which produces the final prediction.



Model Performance Improvement Path

- ID features provide foundation (+10.4%)
- LTV features offer significant value
- Diminishing returns after employee data
- Ensemble modeling provides modest final lift
  - CV ensemble perform best
  - Harnesses model diversity to create a robust predictor with lower overall variance.



Note: This visualization shows the progressive improvement in model performance from baseline through feature engineering to ensemble modeling.

Top Features Explained

- LTV: Higher ratios indicate less borrower equity, increasing default risk during property value declines.
- Employee Code: Loan officers vary in risk assessment skills and customer segments served.
- Supplier ID: Different origination channels attract varying customer risk profiles.
- Employment LTV Risk: Self-employed borrowers face income volatility unlike salaried employees.
- Branch ID: Regional economic conditions and lending practices create geographical default patterns.

Employee Code ID: Hidden Predictive Power

- Key Observation
  - EmployeecodeID is unexpectedly predictive of vehicle loan defaults
  - Outperforms traditional predictors (branch_id, supplier_id, geographic data)
  - Only used for logging disbursements, not approval decisions
- Why This Happens
  - Employee Specialization: Certain employees may handle specific risk segments
  - Hidden Correlations: Employee IDs indirectly map to customer profiles or loan types
  - Training Differences: Newer employees vs. experienced ones might handle different loan types or have different levels of scrutiny
  - Fraud Indicator: Strong correlation with particular employees could indicate potential fraud or policy violations

Conclusion

- ID features unexpectedly delivered strongest predictive power
- LTV ratio confirmed as fundamental default risk indicator
- CatBoost model outperformed competitors
- Employee Code ID revealed surprising correlation with defaults
- Ensemble approach maximized model performance
- AUC-ROC proved most reliable metric for imbalanced classification

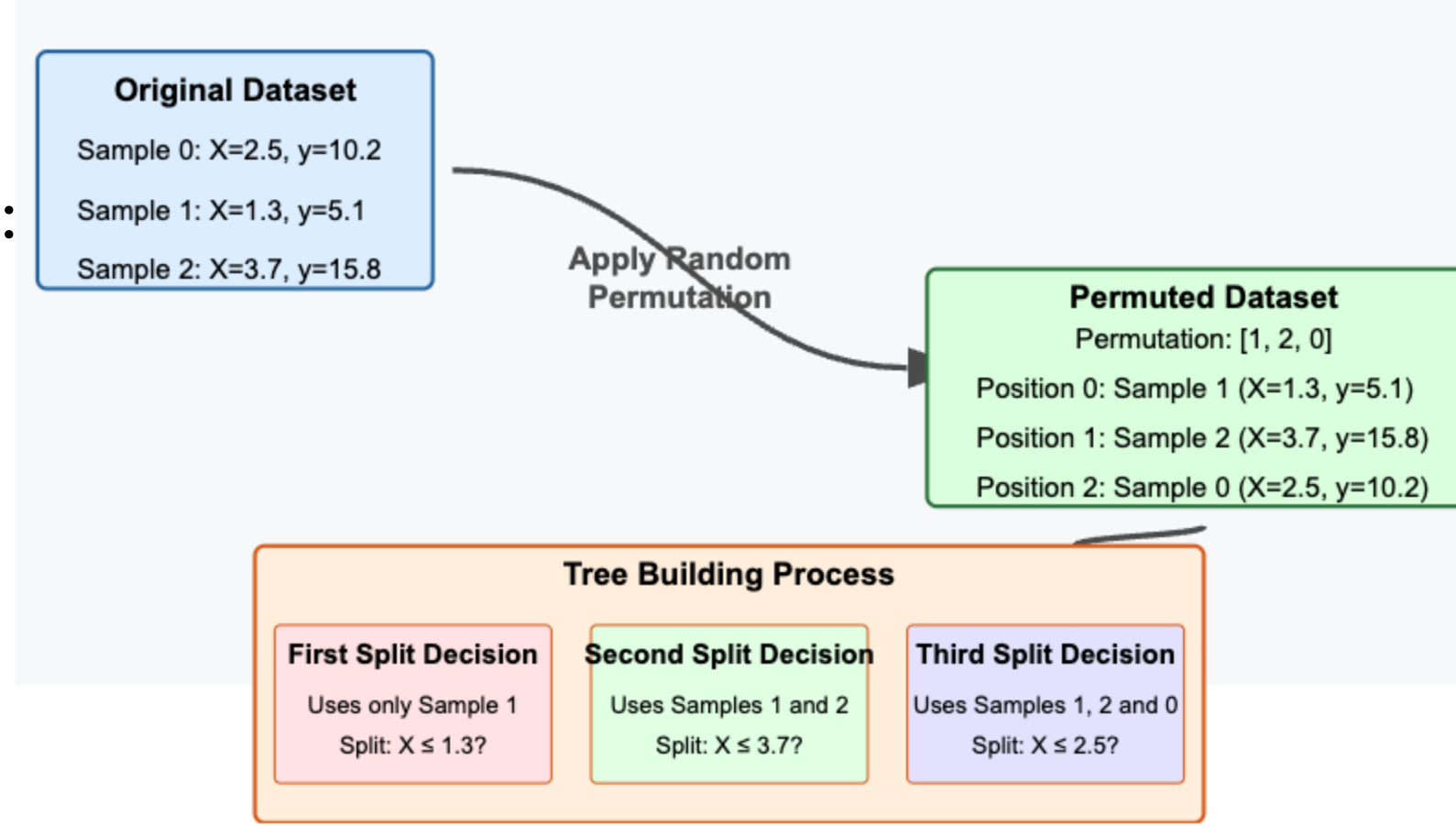
Thank You

CatBoost & Outliers

- Why CatBoost Handles Outliers Well
  - Ordered Boosting: Uses permutation-driven approach where sample influence depends on position
  - Gradient Dilution: Outlier impact varies across trees as position changes in each permutation
  - Tree-Based Splits: Makes decisions based on thresholds, not distances or absolute values
- When to Filter Outliers Anyway
  - Data errors (truly invalid values)
  - Regulatory/business requirements
- Key Insight
  - "CatBoost's permutation-based learning inherently dilutes outlier influence by randomizing their position and impact across multiple trees."

CatBoost Tree Building with Permutation

- **Original Dataset Box**
- **Random Permutatio**
- **Tree Building Process:**
- First split only uses Sample 1
- Second split uses Samples 1 and 2
- Third split uses all three samples: 1, 2, and 0



UniqueID

- UniqueID serves as customer identifier
- IDs correlate with chronological order
- Lower ID numbers indicate earlier customers
- Correlation confirmed by DisbursalDate data

loan_default

- Failure to pay first monthly loan installment by due date
- EMI = Equated Monthly Installment (fixed payment amount)
- Contains both principal and interest components
- Early default signals potential repayment issues

Flags

- Aadhar_flag
  - Tracks if customer provided Aadhar
  - Aadhar card is India's unique biometric identification document issued by the Unique Identification Authority of India (UIDAI)
- PAN_flag
  - The PAN card is an important tax identification document in India

sanctioned amount vs disbursed amount

- DISBURSED.AMOUNT
  - This refers to the total amount that was actually paid out or transferred to you across all loans.
- SANCTIONED.AMOUNT
  - This is the total amount that the lender has approved or agreed to lend across all your loans at the time of approval.
- Imagine you're approved for a home renovation loan:
  - Sanctioned Amount: $50,000 (the bank approves this total amount)
  - Disbursed Amount: $48,500 (after deducting $1,500 in processing fees)

CURRENT.BALANCE

- Total Principal outstanding amount of the active loans at the time of disbursement
  - this represents the total principal amount that remains unpaid on all active loans at a specific point in time.
  - Where principal amount = sum of money borrowed in a loan or invested
- A negative CURRENT.BALANCE typically indicates that the borrower has paid back more than what was required at that point in time

INSTAL.AMT

- PRIMARY.INSTAL.AMT:
  - Customer is the main borrower
  - Example
    - If you have Loan A with a monthly payment of $300
    - And Loan B with a monthly payment of $500
    - And both are loans where you are the primary borrower
    - Then your PRIMARY.INSTAL.AMT would be $800
- SEC.INSTAL.AMT:
  - Customer is a co-applicant or guarantor

NEW.ACCTS.IN.LAST.SIX.MONTHS

- This counts how many new loans the customer has taken in the 6 months before this loan was disbursed. Opening multiple new credit accounts in a short period can be a risk indicator, as it might suggest financial distress or poor planning.

DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS

- This counts how many loans the customer has defaulted on (missed payments) in the 6 months before disbursement. Recent delinquencies are strong negative indicators of creditworthiness.

AVERAGE.ACCT.AGE

- This measures the average age or tenure of all loans the customer has had. Longer average account age generally indicates more experience managing credit and is typically viewed positively by lenders.

CREDIT.HISTORY.LENGTH

- This measures how long the customer has been using credit, calculated from when they took their first loan. A longer credit history provides more data for lenders to assess reliability and is generally viewed positively.

NO.OF_INQUIRIES

- This counts how many times the customer has applied for loans (resulting in credit checks). Multiple inquiries in a short period can suggest that the applicant is desperately seeking credit and may be a higher-risk borrower.