Customer Credit Risk Analysis

A comprehensive data-driven approach to predicting customer delinquency using machine learning and behavioral patterns.

DATASET OVERVIEW

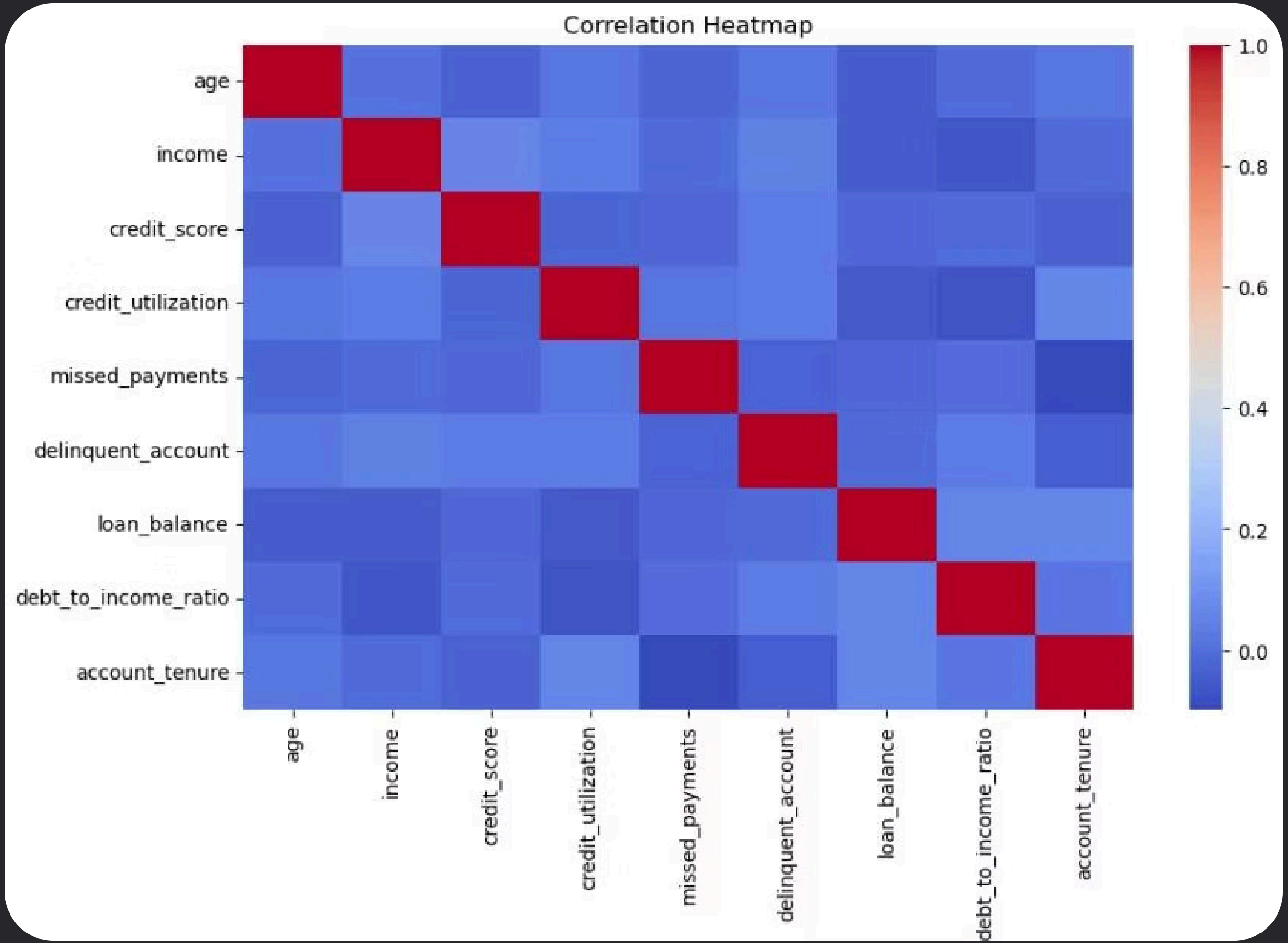
Understanding Our Data

500	19	0
Total Customers	Variables	Missing Values
Complete dataset analyzed	Comprehensive attributes tracked	Clean, complete data

Key Variable Categories

- **9 numerical columns:** age, income, credit score, utilization, missed payments, loan balance, debt ratio, account tenure
- **10 categorical columns:** employment status, card type, location, 6-month payment history
- **Target variable:** delinquent account status

Correlation Insights



Delinquency Correlations

Analysis reveals weak correlations with delinquent accounts:

No single variable strongly predicts delinquency, suggesting multivariate modeling is essential.

- **Income: +0.044**

  Slight positive correlation

- **Credit Score: +0.035**

  Minimal relationship

- **Account Tenure: -0.040**

  Weak negative correlation

Outlier Detection & Removal

Outlier Impact

Before removal: 500 customers

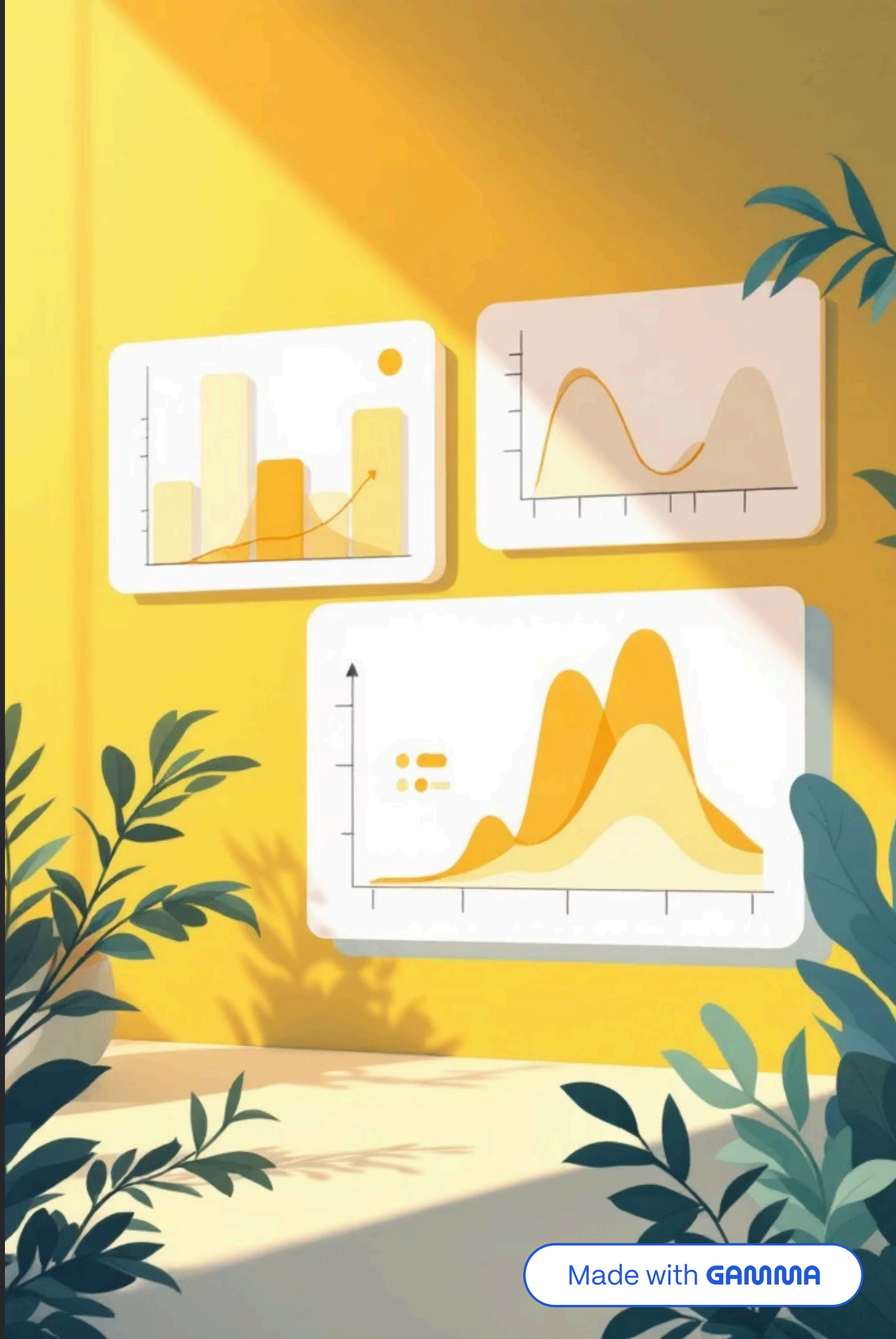
After removal: 420 customers

Outliers removed: 80 (16%)

Detection Method

Applied IQR (Interquartile Range) method across all numerical variables. Outliers identified as values beyond 1.5 × IQR from Q1/Q3.

Key outliers found in age, income, credit score, and loan balance distributions.



FEATURE ENGINEERING

Creating the Risk Label

Missed Payments

More than 2 missed payments

Delinquent Account

Active delinquency status

Credit Score

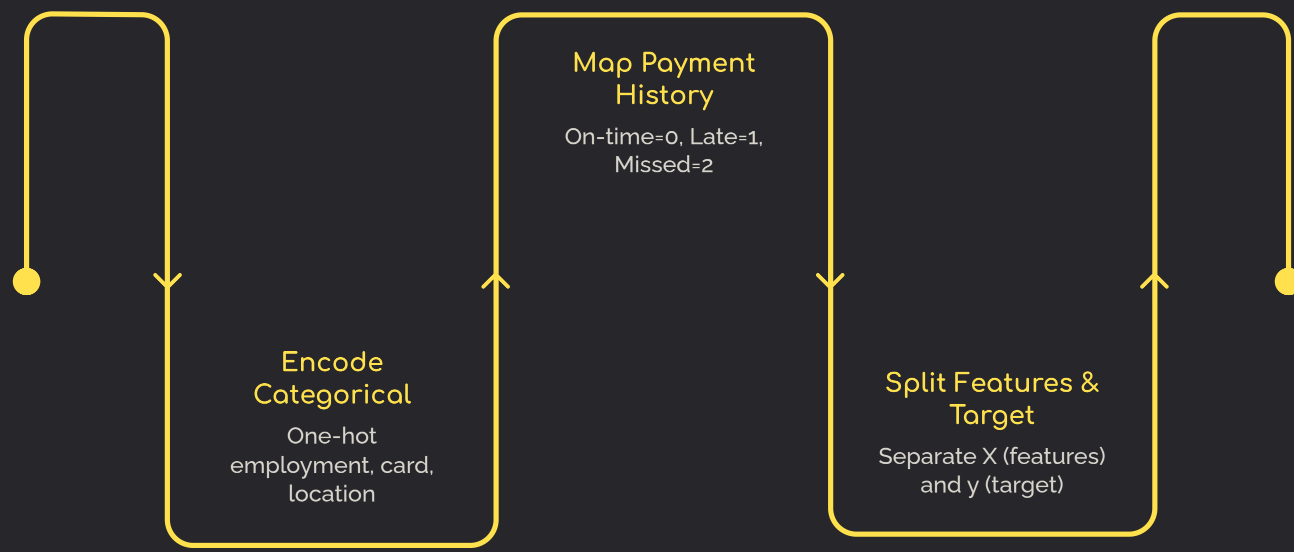
Score below 600

Debt Ratio

Debt-to-income above 40%

Customers meeting any of these criteria were labeled as high-risk (1), creating a binary classification target for predictive modeling.

Data Transformation Pipeline



Systematic transformation prepared data for machine learning modeling.

Key Transformations

01

One-Hot Encoding

Converted employment status, card type, and location into binary features

02

Payment Mapping

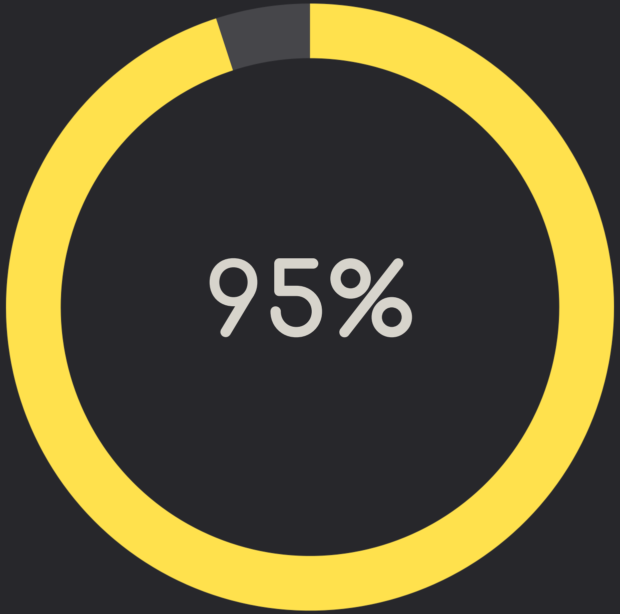
Transformed 6-month payment history into ordinal scale (0-2)

03

Feature Selection

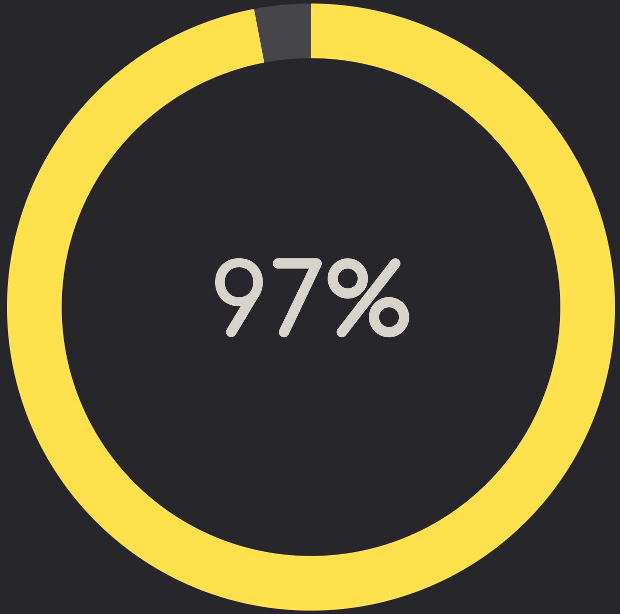
Separated predictors (X) from risk label target (y)

Model Performance



Overall Accuracy

Model correctly classified 95 out of 100 test cases



High-Risk F1-Score

Excellent balance of precision and recall for high-risk customers

Low-Risk Recall

Perfect identification of all low-risk customers

DECISION TREE

Model Architecture

Hyperparameters

Max Depth: 5

Prevents overfitting

Min Split: 20

Ensures robust splits

Min Leaf: 10

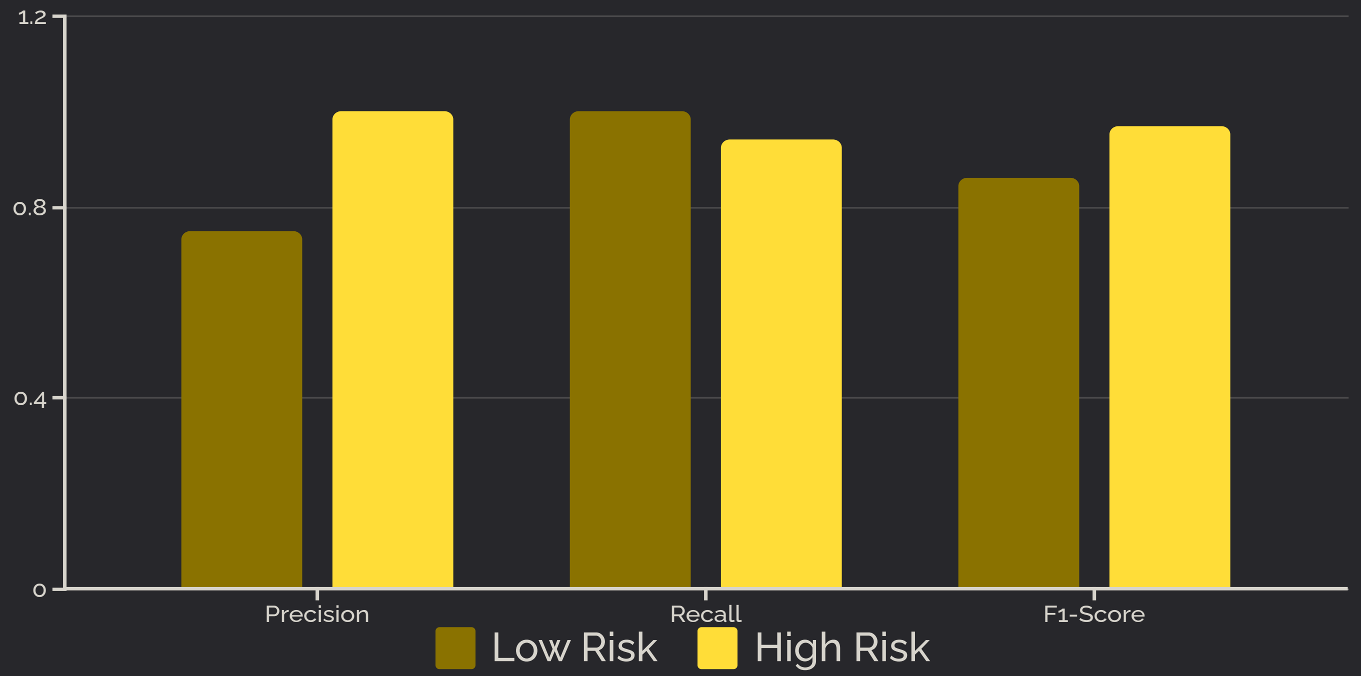
Stable predictions

Why Decision Trees?

- **Interpretability:** Clear visualization of decision logic
- **Non-linear relationships:** Captures complex patterns
- **Feature importance:** Identifies key risk drivers
- **No scaling required:** Works with raw features

80/20 train-test split with stratification ensured balanced evaluation.

Classification Results Breakdown



Performance Analysis

High-Risk Detection: Model excels at identifying risky customers with 100% precision and 94% recall.

Low-Risk Classification: Perfect recall (100%) ensures no low-risk customers are misclassified as high-risk.

Test Set: 85 high-risk and 15 low-risk customers evaluated.

Key Takeaways

Robust Prediction

95% accuracy enables confident risk assessment and proactive intervention

Actionable Insights

Decision tree reveals which factors drive delinquency for targeted strategies

Business Impact

Early identification of high-risk customers reduces losses and improves portfolio health

This analysis demonstrates how machine learning transforms raw customer data into strategic risk management capabilities.