Predictive Modeling of Mortgage Loan Approval

Team 2: Aman Bansal, Sumant Chirde, Venkat Kalyan Duvvuru, Debal Saha

Business Problem

A mortgage loan is a long-term, property-secured loan for buying or refinancing homes.

HMDA requires transparent reporting of mortgage decisions; CFPB enforces consumer protection and fair-lending laws.

As one of the largest U.S. mortgage lenders, Bank of America faces high regulatory expectations.

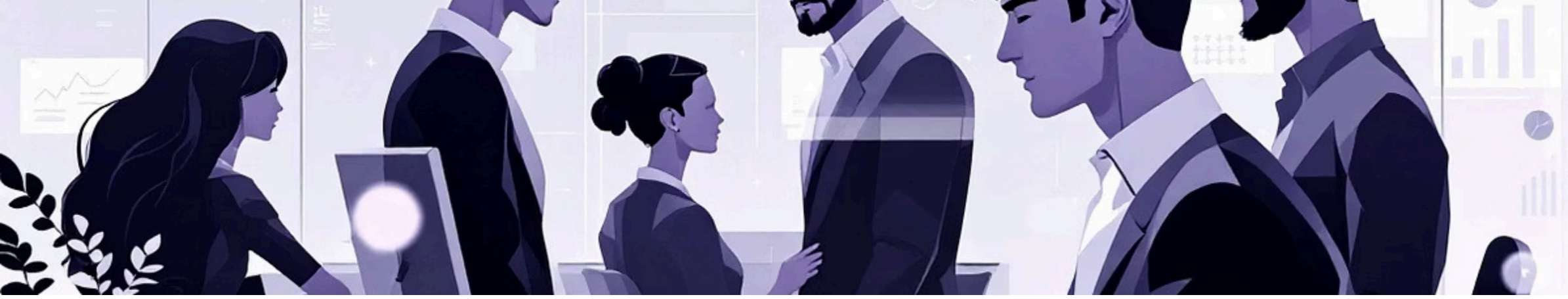
Problems Faced by Bank of America

- **Large application volumes** → slow, inconsistent underwriting
- **Misclassified loans** → default losses or unfair denials
- **Strong scrutiny** due to recent Bank of America regulatory scandals, increasing the need for accurate, explainable lending decisions
- **High risk** of CFPB or fair-lending actions if patterns are not understood and monitored

Why This Matters

Bank of America processes hundreds of thousands of home loan applications annually across diverse products and geographies. Each application must be evaluated accurately, consistently, and transparently—not only to protect the bank's balance sheet, but also to uphold CFPB regulations, meet HMDA reporting requirements, and reinforce the bank's commitment to fair lending.

The Goal: Build predictive models to improve underwriting accuracy, reduce false approvals/denials, and ensure regulatory compliance while supporting sustainable homeownership.



Addressing Mortgage Lending Challenges

Stakeholder

- Bank of America (BofA) Mortgage Lending Division: Underwriting, risk management, and compliance teams.
- BofA aims to enhance transparency in its lending processes to ensure full alignment with evolving regulatory requirements.

Challenge

- Manual underwriting inefficiencies (15-20 day average)
- Human bias in loan approvals
- Complex regulatory demands for explainable lending (HMDA, CFPB, HOEPA)
- High application volume

Opportunity

- Accelerate decisions in loan approvals
- Improve accuracy of approval models
- Ensure fair lending practices
- Optimise human resource allocation
- Achieve $15-20M annual savings.

Data Mining: Predictive & Explanatory

Problem Type

- Supervised Learning: Target variable (approval/denial) is known.
- Classification: Binary prediction of loan approval or denial.

Primary Goal: Predictive

Build a model to predict loan approval/denial for faster decisions and risk optimization.

Secondary Goal: Explanatory

Understand key factors driving approval/denial for policy insights.

Key Outcomes

Predicted approval probability, classification rules, and feature importance.

Dataset Overview

Dataset Overview

- Source: **HMDA* dataset for Bank of America**
- Volume: 163,529 loan applications from 2024
- Geography: Covers 51 states in the US
- Loan types: Conventional, FHA†, VA^, and RHS**

What is a Row?

Each row represents one mortgage loan application details

Input Variables

Applicant, loan, property, lender and underwriting details

Output Variable

action_taken: Decision for each mortgage loan

(Denied: 49.4%, Approved: 50.6%)

	state_code	county_code	conforming_loan_limit	derived_loan_product_type	derived_dwelling_category	action_taken	purchaser_type
2	CA	6087	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
3	UT	49043	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
4	NY	36061	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
5	CA	6059	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	3	0
6	MA	25017	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	3	0
7	NC	37097	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	1	0
8	CO	8069	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
9	CA	6075	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	1	0
10	WA	53033	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	1	0
11	HI	15007	C	Conventional:First Lien	Single Family (1-4 Units):Site-Built	3	0
12	FL	12109	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
13	PA	42091	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
14	FL	12099	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	3	0
15	CA	6037	C	Conventional:First Lien	Single Family (1-4 Units):Site-Built	3	0
16	NC	37183	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
17	CO	8045	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	3	0
18	NY	36103	C	Conventional:First Lien	Single Family (1-4 Units):Site-Built	3	0
19	NJ	34029	C	Conventional:First Lien	Single Family (1-4 Units):Site-Built	3	0
20	GA	13121	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	1	0
21	NJ	34025	C	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
22	MA	25021	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	1	0
23	CA	6041	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	3	0
24	CA	6081	C	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	0
25	CA	6013	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	3	0
26	AZ	4013	NC	Conventional:Subordinate Lien	Single Family (1-4 Units):Site-Built	3	0
27	GA	13121	NC	Conventional:First Lien	Single Family (1-4 Units):Site-Built	1	

*Home Mortgage Disclosure Act, †Federal Housing Association, ^Veteran Affairs, **Rural Housing Service

Analytical Approach: Two Complementary Models

LASSO Logistic Regression

Primary predictive model, handles sparse data, reduces variables, highly interpretable.

Conditional Inference Trees

Interpretable validation model, generates business-friendly rules, captures interactions.

Model Evaluation

LASSO Logistic Regression Performance

- Accuracy: 83.0%
- AUC: 0.90
- F1 Score: 0.84

Benchmark: Traditional Logistic Regression

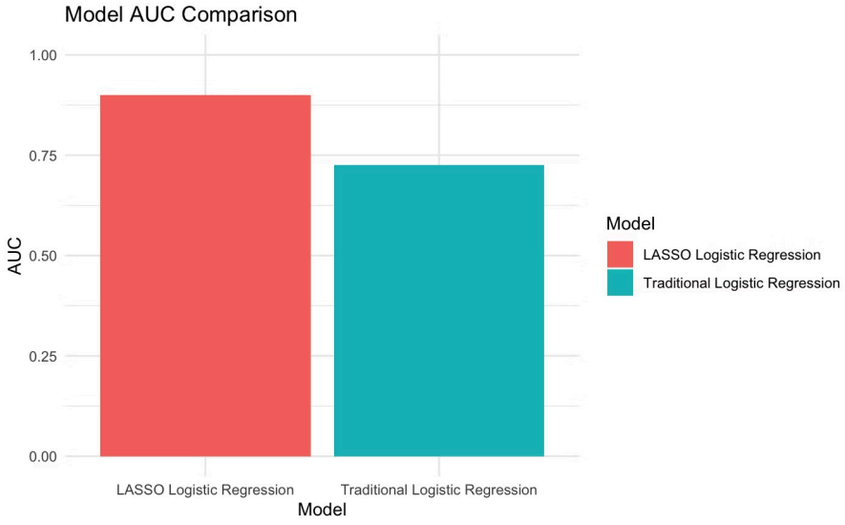
- **Accuracy: ~72–75%**
- **AUC: ~0.70–0.75**
- **F1 Score: ~0.70**

Improvement vs baseline:

- +8–10 percentage points in accuracy
- +0.15–0.20 improvement in AUC

Business impact:

- Reduced false approvals → lower default risk
- Reduced false denials → fewer lost customers



Our model shows significant improvement, reducing false approvals and denials, and complying with regulations.

Key Findings from LASSO Model

The LASSO model selected 69 significant features, revealing the following strongest drivers of loan approval and denial:

Automated Underwriting System (AUS)

- Loans processed by AUS such as **DU, LP, and TOTAL** are much more likely to be denied
- When AUS was bypassed, denial probability is lower

Loan Product Structure

- Refinance purpose loans, open-end line of credit (LOC), commercial purpose loans, interest-only loans, and balloon payment loans are **features of risky loan products** that consistently increase denial probability

Property & Neighborhood

- Higher minority percentages areas have slightly lower denial risk, indicating **fair lending practices**
- Higher loan to value (LTV) ratio slightly lowers denial risk

Submission Channels

- **Direct submissions** show lower denial probability
- Loans submitted directly to the bank appear less risky or more effectively underwritten

Borrower Affordability Indicators

- Higher **income** and lower **debt to income (DTI)** ratio improve approval probability

Geographic Characteristics

- West region has slightly higher denial rate
- South/Midwest region has slightly lower denial

Key Findings from Decision Tree

The decision tree reveals interpretable approval rules and clarifies how specific combinations of features lead to outcomes:

HOEPA Status (Primary Split)

- Loans under HOEPA which are not high-cost mortgages frequently appear in the dataset with characteristics that strongly correlate with denial
- For loans where HOEPA does not apply, the approval rate is very high for primary residences

Occupancy Type

- Owner-occupied homes have higher approval probability
- Second homes/investment properties have lower approval probability
- Influences deeper splits involving DTI, LOC, and rate spread

High-Risk Product Combinations

- Multiple risk factors stack: open-end line of credit + interest-only payment + DTI >42% + rate spread >0.94%
- These combinations lead to near-universal denials
- Indicates strong product-level underwriting restrictions

Pricing Thresholds

- Rate spread thresholds: ≤0.74 vs >0.74
- Interest rate cutoffs: ≤/≥6.875% or 6.99%
- Pricing interacts with borrower risk to determine approvals

Applicant Financial Strength

- Low income thresholds (e.g., income ≤1,191)
- DTI cutoffs (25%, 42%)
- Weak affordability metrics consistently lead to denials

Geographic Interaction

- Tract minority population percentage and tract family homes data
- Not direct drivers but interact with product risks

Top 5 Must-Implement Recommendations

Based on our analytical findings, these are the critical actions to enhance your mortgage loan approval process:

1

Strengthen Affordability Screening (DTI & Income)

Business Impact: Reduces early defaults and strengthens overall portfolio quality.

2

Control High-Risk Loan Products (LOC, IO, Balloon Loans)

Business Impact: Cuts processing time and cost while avoiding loss-prone loans.

3

Recalibrate AUS Decisioning (DU, LP, TOTAL)

Business Impact: Converts false denials into profitable approvals; increases interest-earning volume.

4

Implement Pricing Threshold Controls

Business Impact: Protects margins, improves risk-adjusted returns, and stabilizes revenue.

5

Build an Explainable-AI Compliance Layer

Business Impact: Lowers regulatory risk, prevents costly penalties, and improves decision consistency.