Loan Default Risk Analysis

A Data-Driven Approach to Financial Risk Assessment in the Lending Sector

SECTOR
Financial Services

TEAM ID
G-15

FACULTY MENTOR
Aayushi Vashishth

MEMBERS

Husain Khorakiwala 2401010192

Adnan Rizvi 2401010043

Mukul Kumar 2401010285

Tanishk Agrawal 2401020071

Anurag Pandey 2401010089

Shivansh Tiwari 2401020108

G-15 | Financial Services | 2026 Capstone

Context & Problem Statement

SECTOR CONTEXT

Banks provide loans to customers, but many customers fail to repay. Loan defaults cause huge financial losses to banks.

PROBLEM STATEMENT

How can banks identify risky customers before approving loans?

OBJECTIVE

To analyze customer and loan data to reduce default risk and support better loan approval decisions.

Data Engineering

SOURCE

Rows	**24,999**
Columns (raw)	**35**
Columns (clean)	**29**
Period	**2019 (Single Year)**
Domain	**Mortgage Lending**
Target Column	**defaulted (0/1)**

CLEANING HIGHLIGHTS

Missing Nulls

Rate of interest (24.3%), Income (6.2%), LTV (9.7%), DTI (15.8%) — filled with median for robustness

Outlier Treatment

Income outliers replaced with median to prevent skew in income-based risk segmentation

Column Drops

year (constant), security_type (constant), loan_type & loan_purpose (no interpretable meaning)

Categorical Fixes

Missing categoricals (gender, loan_limit, age, etc.) filled with mode; all switched to UPPER CASE

KEY COLUMNS

credit_score	*Creditworthiness proxy*
dtir1	*Debt-to-income ratio*
ltv	*Loan-to-value ratio*
income	*Annual income*
age	*Borrower age group*
region	*Geographic segment*
neg_ammortization	*Loan structure risk*
interest_only	*Payment type*
defaulted	*Target: 0 = No, 1 = Yes*

KPI & Metrics Framework

24,999
Total Loans Analysed

24.40%
Overall Default Rate

>120.0
Highest Risk LTV

50-90
Highest Risk DTI

KPI Metric	Definition	Why It Matters
Default Rate by Segment	% of loans defaulted in each group	Isolates highest-risk cohorts for targeted policy
Loan-to-Value (LTV)	Loan ÷ Property Value × 100	Higher LTV = less collateral buffer → greater loss on default
Debt-to-Income (DTI)	Total debt payments ÷ Monthly income	Measures borrower's repayment capacity
Income Band Default Rate	Default % by monthly income bracket	Links economic vulnerability to default likelihood
Regional Default Rate	Default % by geographic region	Highlights macro-economic or policy-driven risk pockets

Key Insights from Exploratory Analysis

01

LTV is the Strongest Default Driver

Loans with LTV > 120% default 100% of the time. Even the 60–80% LTV band defaults at 32.49% — more than 2× the rate of low-LTV loans. Collateral coverage is the single most predictive variable.

02

Low Income = High Default Risk

Borrowers earning under $2,000/month default at 43.79% — nearly double the average. Risk drops consistently as income rises, falling to ~17% for high earners. Income bands should anchor approval tiers.

03

DTI Extremes Are Dangerous

Both very high DTI (50–59: 44%, 60–69: 41%) and high-middle DTI (30–39: 31.8%) drive elevated defaults. The sweet spot is DTI 20–29 at only 12.96%. DTI alone can identify a third of the high-risk population.

04

Credit Score Provides No Discrimination

Default rates across credit score bands (500–900) vary by only ±2%, ranging from 23.5% to 25.4%. This signals that credit score alone is a poor underwriting tool in this portfolio.

05

North-East Region is the Highest-Risk Area

NORTH-EAST has a 33.17% default rate vs. 22.24% in NORTH. Regional economic conditions, property markets, and borrower demographics likely drive this disparity. Geographic risk must be priced in.

06

Young & Elderly Borrowers Default More Often

Borrowers under 25 (29.1%) and over 74 (29.2%) default at rates well above the 24.4% average. Prime working-age groups (35–54) are the lowest-risk cohort, suggesting lifecycle-based risk adjustment is warranted.

Advanced Analysis — Risk Heatmap & Segmentation

Credit Score	< 30%	30-34%	35-39%	40-44%	45-50%	> 50%
500-549	16.80%	13.39%	40.37%	12.54%	19.64%	44.04%
550-599	15.01%	11.38%	38.21%	14.19%	16.22%	42.37%
600-649	18.49%	13.77%	37.23%	14.60%	17.06%	38.38%
650-699	14.08%	12.61%	37.02%	14.07%	17.37%	44.37%
700-749	16.57%	12.43%	38.05%	11.88%	16.12%	47.94%
750-799	17.82%	13.13%	41.06%	13.45%	16.22%	48.26%
800-849	15.53%	12.28%	37.06%	14.56%	18.27%	41.92%
850-900	16.18%	9.07%	39.09%	14.60%	22.86%	39.10%

Risk Legend: < 13% — Low | 13–37% — Medium | 38–44% — High | ≥ 45% — Critical

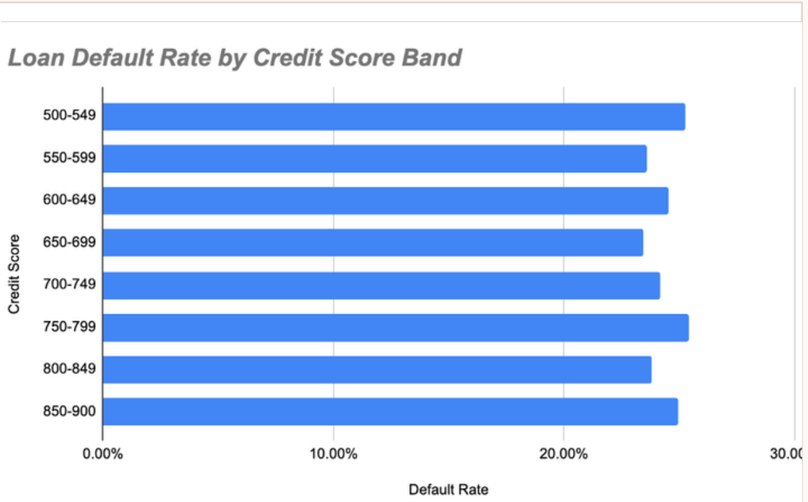
KEY FINDING

DTI > 50% is a universal danger zone — regardless of credit score.

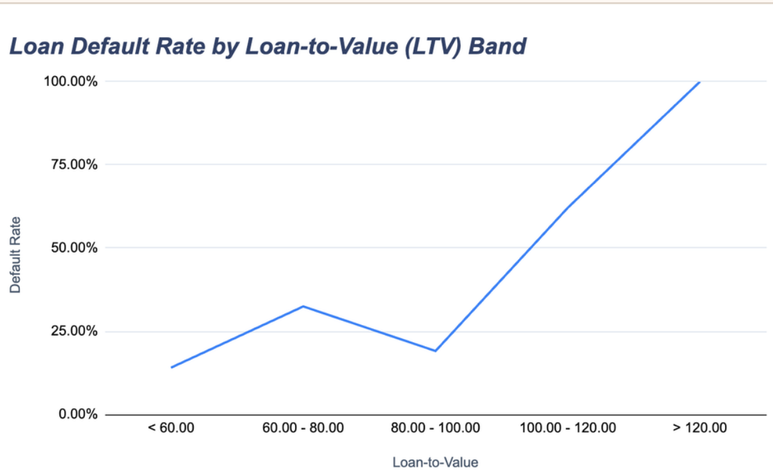
Even 850–900 score borrowers default at 39–48% when DTI exceeds 50%.

Credit score does NOT compensate for overleverage.

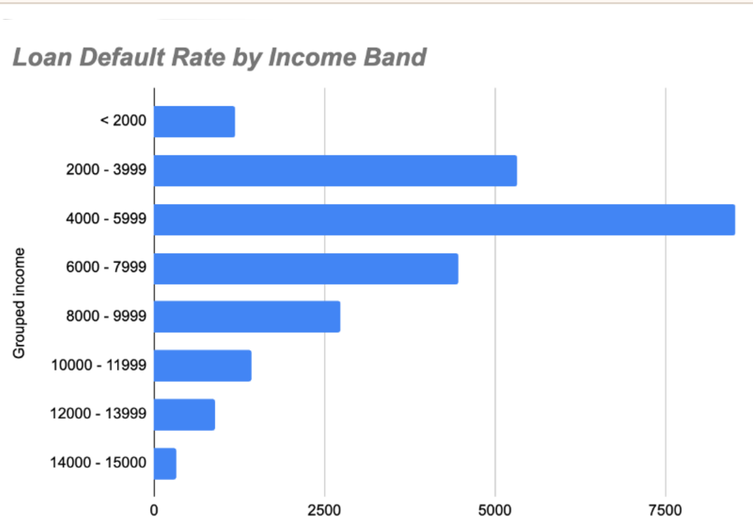
Dashboard OVERVIEW



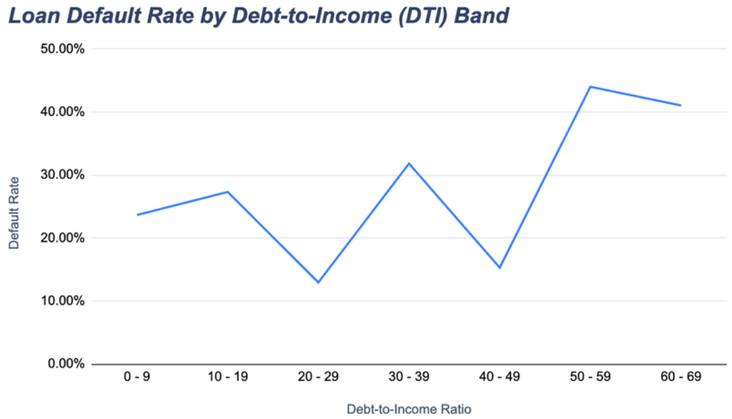
Credit Score vs. Default: Rates are nearly flat (±2%) across all bands — credit score alone is a poor predictor.



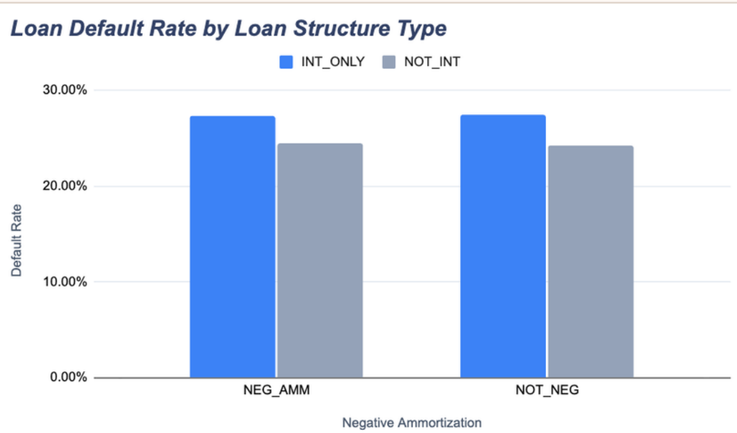
LTV vs. Default: LTV > 120% means 100% default. Even the 60–80% band defaults at 32.49%.



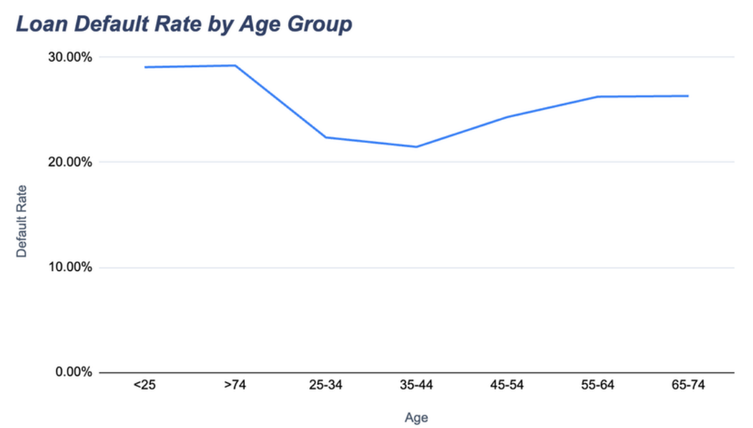
Income vs. Default: Under $2k/mo earners default at 43.79%. Risk drops as income rises.



DTI vs. Default: DTI 50–59 band shows 44% default; DTI 20–29 is the safe zone at 12.96%.

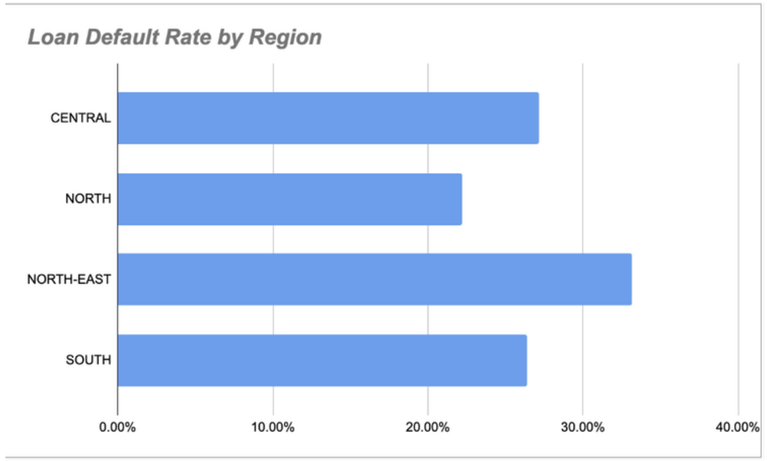


Loan Structure: INT_ONLY loans default at ~27% vs. 24% for standard; neg-amm adds further risk.

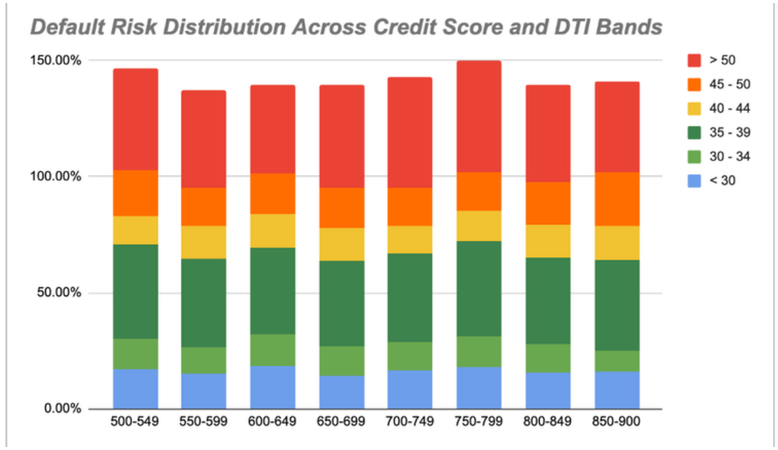


Age vs. Default: Under-25 (29.1%) and over-74 (29.2%) borrowers default above average. Ages 35–54 are safest.

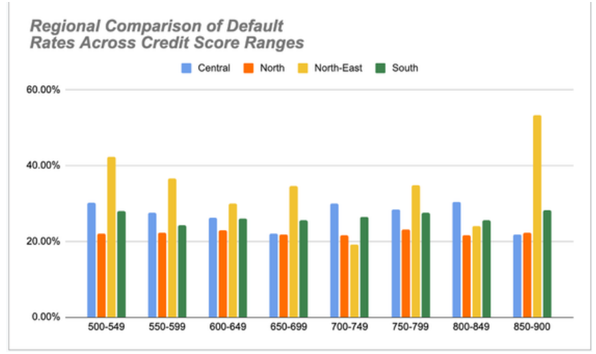
Dashboard OVERVIEW



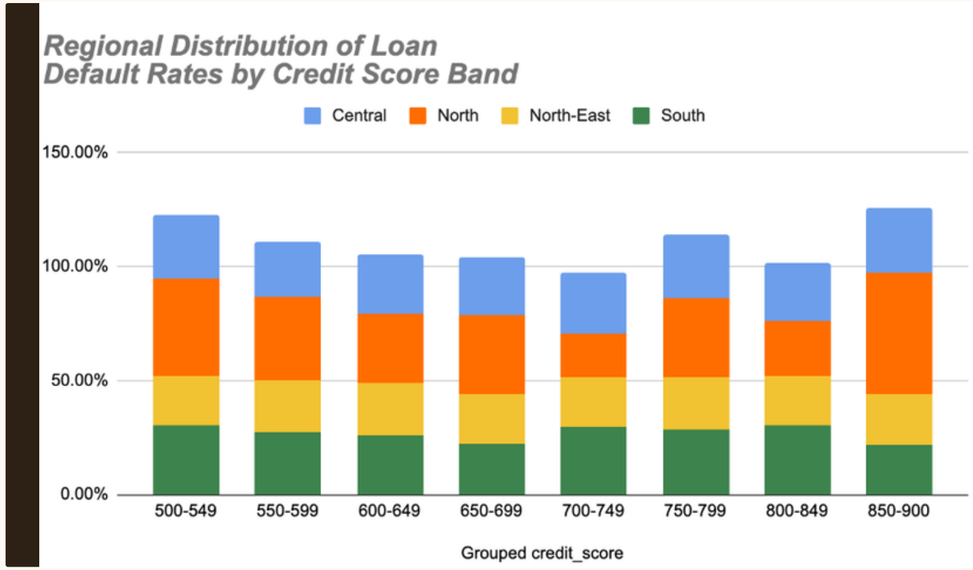
Region vs. Default: NORTH-EAST leads at 33.17%. All regions exceed 22%, signalling a portfolio-wide concern.



Risk Heatmap: DTI > 50% is a universal danger zone regardless of credit score band.



Default Rates Across Regions by Credit Score Band: North-East consistently shows elevated default risk across every credit tier — confirming that geography, not just borrower quality, drives risk. A regional risk premium of 50-75bps is warranted.



Stacked Default Distribution by Region & Credit Score: The composition shows that even within the same credit score band, regional exposure varies widely. This supports targeted underwriting and pricing strategies by geography.

Dashboard Walkthrough

EXECUTIVE VIEW

24.40%
Default Rate

33.17%
NE Region

100%
LTV > 120

44%
DTI 50-59

Headline view for CRO — four critical risk signals at a glance, enabling rapid escalation and board-level communication.

OPERATIONAL DRILL-DOWN

Region × Credit Score
Identify high-risk pockets (e.g. NORTHEAST, low credit)

LTV Band Analysis
Flag > 80% LTV loans for enhanced scrutiny

Income Bracket Drill
Segment <$2k/mo for referral to specialist underwriting

DTI × Credit Heatmap
Cross-filter borrowers in danger zones simultaneously

Neg-Amm / Int-Only
Compare loan structure impact on default (27% vs 24%)

Used by underwriters for daily application review — filters, slicers, and cross-dimensional drills support precise risk profiling.

Strategic Recommendations

R1

Implement Hard LTV Caps

Refuse or require mortgage insurance for LTV > 100%. Impose enhanced underwriting for 80–100% LTV. Loans above 120% LTV have a 100% default rate — this is a zero-tolerance threshold.

↑ *Insight: LTV is the #1 predictor*

R2

Introduce Income-Tiered Loan Limits

Cap maximum loan amounts relative to verified monthly income. Borrowers earning < $2,000/month should have a maximum LTV of 70% and DTI of 28%. This directly targets the 43.79% high-default cohort.

↑ *Insight: Low income = highest risk*

R3

Hard DTI Cut-Off at 50%

Decline or escalate for specialist review any application with DTI ≥ 50%. The data shows 44% default rate in the 50–59 band. Reward borrowers in the 20–29 DTI band with preferred rates (only 12.96% default).

↑ *Insight: DTI 50%+ is universal danger*

R4

North-East Regional Risk Premium

Apply a 50–75bps rate premium or stricter approval criteria in the NORTH- EAST region (33.17% default vs. 22.24% in NORTH). Investigate local economic or property market drivers to calibrate further.

↑ *Insight: Regional gaps are significant*

R5

Age-Adjusted Borrower Support

Require additional income verification or a co-borrower for applicants under 25 or over 74, who both default at ~29% — nearly 5% above the 24-54 age cohort. Consider targeted financial counselling programmes.

↑ *Insight: Age cohorts show lifecycle risk*

Impact & Business Value

The 'So What?'

Revenue Protection

Applying LTV caps and DTI thresholds could reduce the default pool by an estimated 30–40%, protecting millions in principal. Even a 5% reduction in defaults (24.4% → 19.4%) saves ~$12M per $1B in originations.

Underwriting Efficiency

A rules-based flagging system using LTV, DTI, and income thresholds can auto-escalate high-risk files, reducing underwriter review time by ~25% for borderline cases. Fast-track approvals for low-risk profiles.

Regulatory Capital Relief

Reducing high-LTV exposures directly lowers risk-weighted assets (RWA), improving capital adequacy ratios. Compliance teams gain defensible, data-driven documentation for audit and stress-testing submissions.

Early Warning System

The dashboard provides a real-time view of portfolio drift — enabling the risk team to intervene with workout programs, loan modifications, or targeted collections before delinquencies crystallise into charge-offs.

Implementation requires CRO sign-off, IT system integration (3–6 months), and a parallel-run validation period before full deployment.

Limitations & Next Steps

⚠ LIMITATIONS & DATA GAPS

Single Year Snapshot
Data covers only 2019. No multi-year trend analysis, no recession cycle, no COVID-19 shock. Findings may not generalise.

High Missing Rates
Rate of interest (24.3%) and DTI (15.8%) had significant nulls. Median imputation may mask patterns in the missing segment.

No Macroeconomic Context
Unemployment, interest rate cycles, and house price indices are absent. Economic environment heavily influences default behaviour.

Anonymous Loan Types
loan_type (type1/2/3) and loan_purpose (p1–p4) were dropped. These may contain significant predictive signal if decoded.

Credit Score Blind Spot
All credit scores are remarkably uniform in default rate. This may indicate data quality issues or portfolio selection bias in origination.

NEXT STEPS

1

Phase 1 (0-3 months)
Deploy predictive model (Logistic Regression or Random Forest) using LTV + DTI + income as primary features. Establish a model scorecard for real-time scoring at origination.

2

Phase 2 (3-6 months)
Integrate multi-year data (2015–2024) for time-series risk trending. Add macroeconomic covariates (unemployment rate, housing index) to the model.

3

Phase 3 (6-12 months)
Build an automated early-warning system — flag performing loans with rising DTI or declining property values for proactive workout outreach.

∞

Ongoing
Conduct quarterly model validation and recalibration. Establish a regional risk committee to monitor NORTH-EAST and CENTRAL portfolios.