Customer Loan Default Risk Analysis

A Data-Driven Approach to Lending Risk Assessment

Sector: Financial Services – Lending
Team ID: G-15

Team Members:
- Husain Khorakiwala
- Adnan Rizvi
- Anurag Pandey
- Mukul Kumar
- Shivansh Tiwari
- Tanishk Agrawal

Institute: Rishihood University, Newton School of Technology
Faculty Mentor: Aayushi Vashishth

---

1. Executive Summary

Problem

Financial institutions face substantial losses due to loan defaults, which negatively affect profitability, regulatory capital, and financial stability. Traditional underwriting often relies heavily on individual indicators such as credit score, which alone cannot accurately measure repayment capacity. With increasing credit demand and diverse borrower profiles, lenders require data-driven risk identification systems.

Approach

This study analyzed a structured lending dataset of **24,999 loans** to identify key drivers of default. Interactive dashboards and KPI frameworks were developed to examine relationships between defaults and borrower characteristics including income, DTI, LTV, age, region, and loan structure. Segment-wise default analysis and a risk heatmap were used to evaluate combined effects of multiple variables.

Key Insights

- Portfolio default rate is **24.40%**
- Credit score alone has weak predictive power
- High DTI (>50%) strongly predicts default
- LTV >120% results in near-certain default
- Low income borrowers are highest risk
- Young and elderly borrowers show elevated defaults
- Regional variation exists across loan performance

Key Recommendations

- Implement multi-factor underwriting models
- Set strict LTV and DTI thresholds
- Introduce income-based approval limits
- Apply regional risk policies
- Automate risk screening rules

Conclusion

Data-driven underwriting significantly improves loan decision accuracy and reduces financial risk exposure.

---

2. Sector & Business Context

2.1 Sector Overview

The lending sector plays a critical role in financial systems by providing credit to individuals and businesses. However, lending inherently involves credit risk, as borrowers may fail to repay obligations.

2.2 Current Challenges

- Rising loan defaults increase losses
- Static approval criteria fail to capture true borrower risk
- Credit score dependence oversimplifies decision making

2.3 Problem Rationale

Reducing default risk directly improves profitability, regulatory compliance, and portfolio stability. Data analytics enables more accurate borrower evaluation and informed lending decisions.

---

3. Problem Statement & Objectives

Problem Definition

Develop a data-driven framework capable of identifying high-risk borrowers prior to loan approval.

Project Scope

- Analyze borrower demographics and financial attributes
- Identify major predictors of default
- Provide actionable underwriting insights

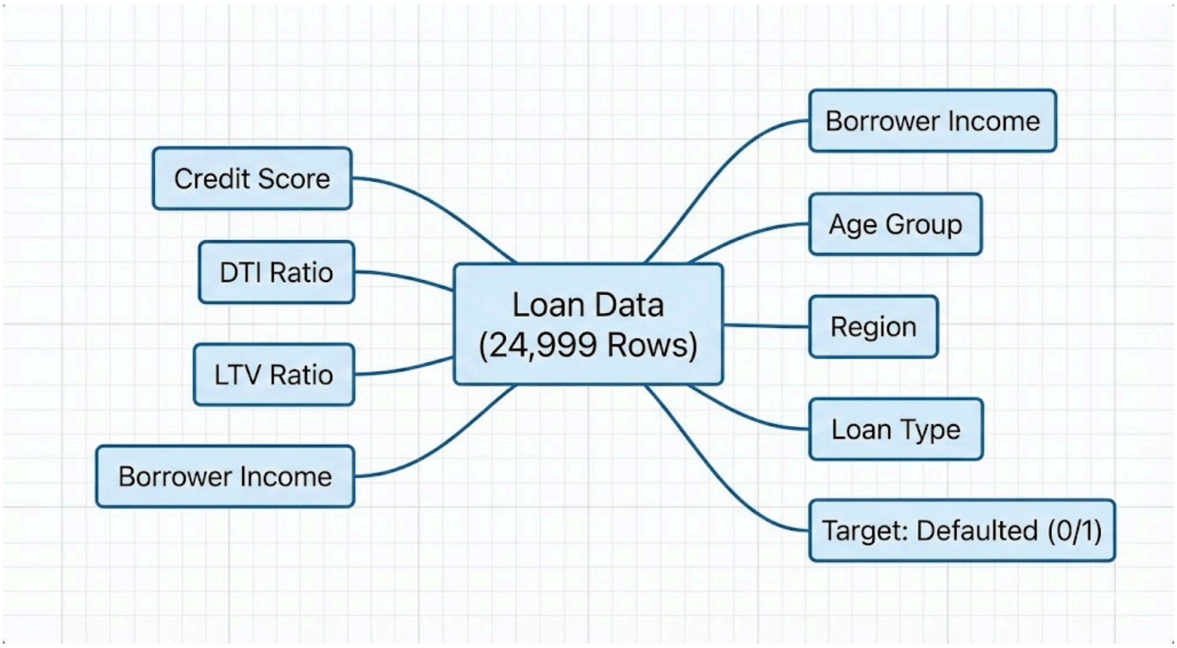
Success Criteria

- Accurate risk identification
- Measurable risk indicators
- Practical decision rules for lenders

---

4. Data Description

A statistically randomized sampling approach was used to select records from an initial dataset of 125,000 entries, producing a representative subset that was further refined through preprocessing to a final analytical dataset of 24,999 rows.



Dataset Source

Internal Mortgage Lending Dataset (2019 snapshot) containing 24,999 loan records.
https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data

4.1 Structure

Attribute	Value
Rows	24,999
Raw Columns	35
Clean Columns	30

Target Variable

Variable	Description
defaulted	Loan outcome (0 = No, 1 = Yes)

Key Features

Feature	Description
Credit Score	Creditworthiness proxy

Feature	Description
DTI	Debt-to-income ratio
LTV	Loan-to-value ratio
Income	Borrower income
Age	Age group
Region	Geographic segment
Interest Only	Loan type
Negative Amortization	Risk indicator

Limitations

- Single-year data
- Missing values
- No macroeconomic variables

---

5. Data Cleaning & Preparation

(All preprocessing conducted in Google Sheets.)

Initial Dataset Reduction

The original dataset obtained from Kaggle contained approximately **150,000 rows**. However, due to the technical limitations of Google Sheets — specifically the **10 million cell limit per spreadsheet** — it was not feasible to perform large pivot table operations on the full dataset.

Therefore, a representative subset of approximately **25,000 random rows** were selected for analysis. This sample size was sufficient to preserve overall data patterns while enabling efficient processing, dashboard creation, and KPI computation within the platform constraints.

The code for this can be found in the GitHub repository under `random-sampler/main.py`.

Missing Value Handling

Variable	Missing%	Action
Interest Rate	24.3	Median Imputation
DTI	15.8	Median Imputation
LTV	9.7	Median Imputation
Income	6.2	Median Imputation
upfront_charges	26	Median Imputation
term	0.032	Median Imputation
property_value	9.6	Median Imputation
income	6	Median Imputation
loan_limit	2.41	Mode
approv_in_adv	0.66	Mode
neg_ammortization	0.10	Mode
age	0.12	Mode
submission_of_application	0.12	Mode

Outlier Treatment

Median replaces the income outliers to prevent skew.

Transformations

Categorical values standardized to uppercase.

Dropped Columns

year, security_type, loan_type, loan_purpose, total_units
(removed due to low predictive value)

Credit Score Band Creation

A new column named `credit_score_band` was created to group individual credit scores into fixed numerical ranges for easier analysis and segmentation.

Why This Method Was Used

Financial risk analysis often focuses on score ranges rather than exact values. Grouping scores allows clearer interpretation of default behavior across different credit tiers and supports decision-making in lending policies.

Assumptions

Median best represents skewed financial distributions. Dropped variables assumed non-predictive.

---

6. KPI & Metric Framework

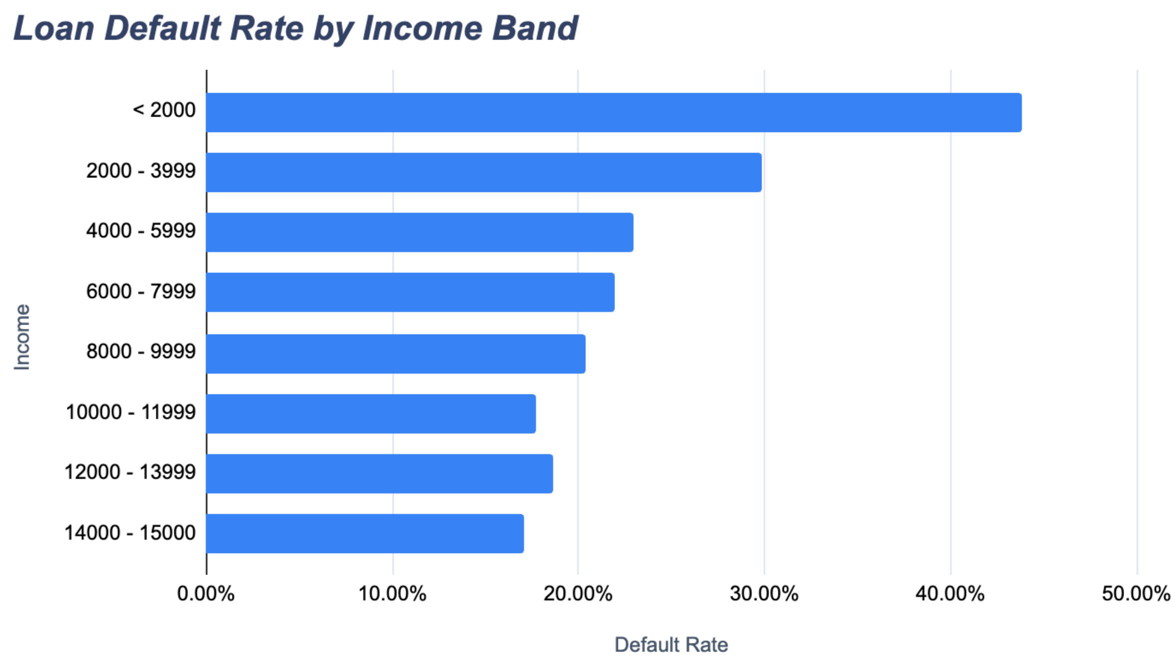
KPI	Formula	Value	Meaning
Default Rate	Defaults ÷ Total Loans ×100	24.40%	Portfolio risk level
LTV	Loan ÷ Property Value ×100	>120 highest risk	Collateral buffer
DTI	Debt ÷ Income	50–59 highest risk	Repayment stress
Income Risk	Defaults(income band)/Loans(band)	<2000 highest	Financial vulnerability
Regional Risk	Defaults(region)/Loans(region)	North-East highest	Geographic risk

---

7. Exploratory Data Analysis

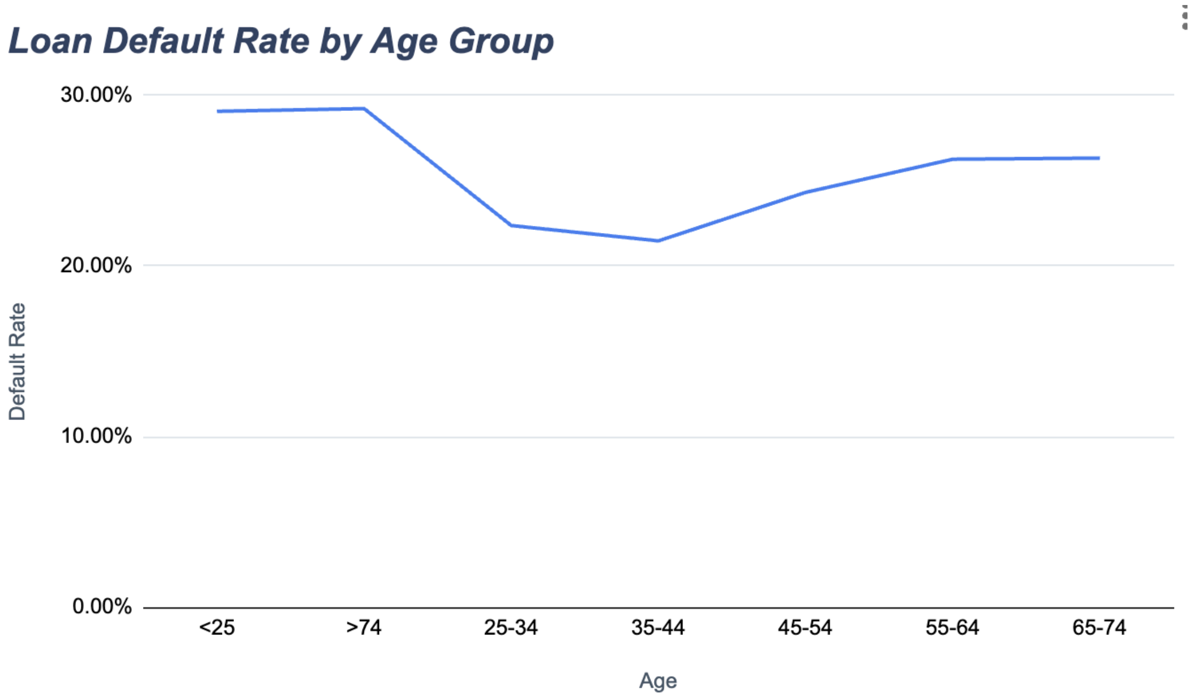
Comparison Analysis

Low-income borrowers default nearly twice as often as high-income borrowers.



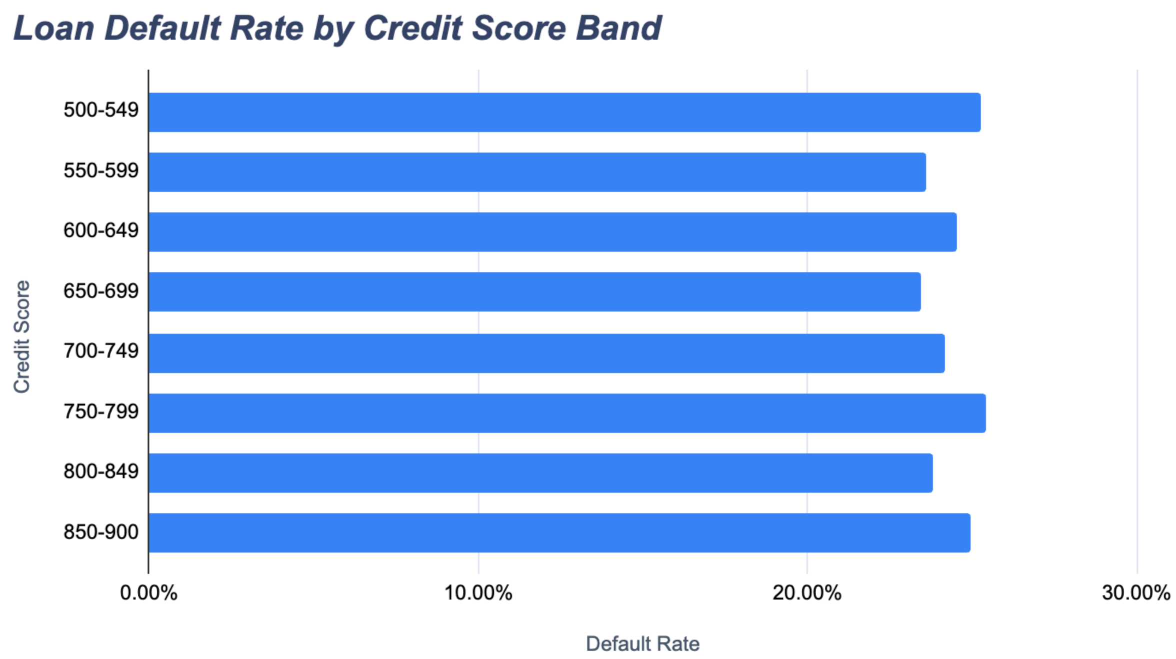
Distribution Analysis

Prime working-age borrowers show lowest defaults.



Correlation Analysis

Credit score shows minimal variation across default rates.



Trend Analysis

Default probability increases sharply as leverage rises. Loans with LTV >120% show near-certain default.

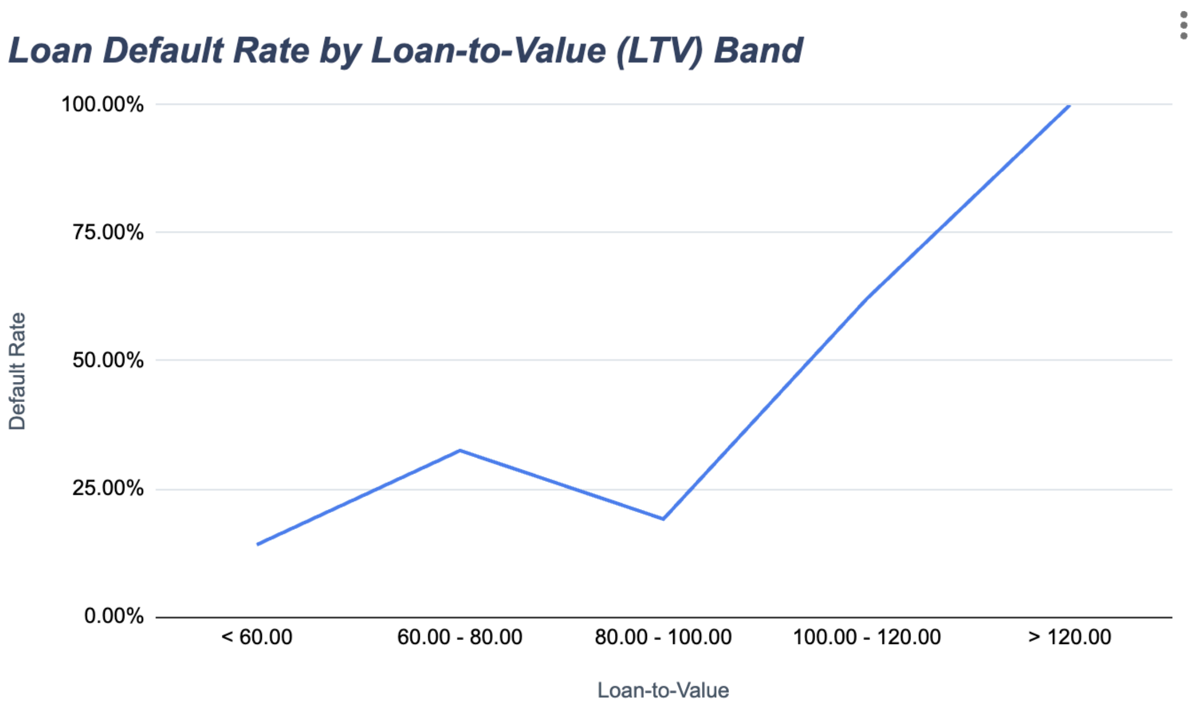


Chart Interpretation

Visualizations included:

- Bar charts → income risk
- Line charts → DTI trend
- Heatmap → credit score vs DTI
- Column charts → regional comparison

---

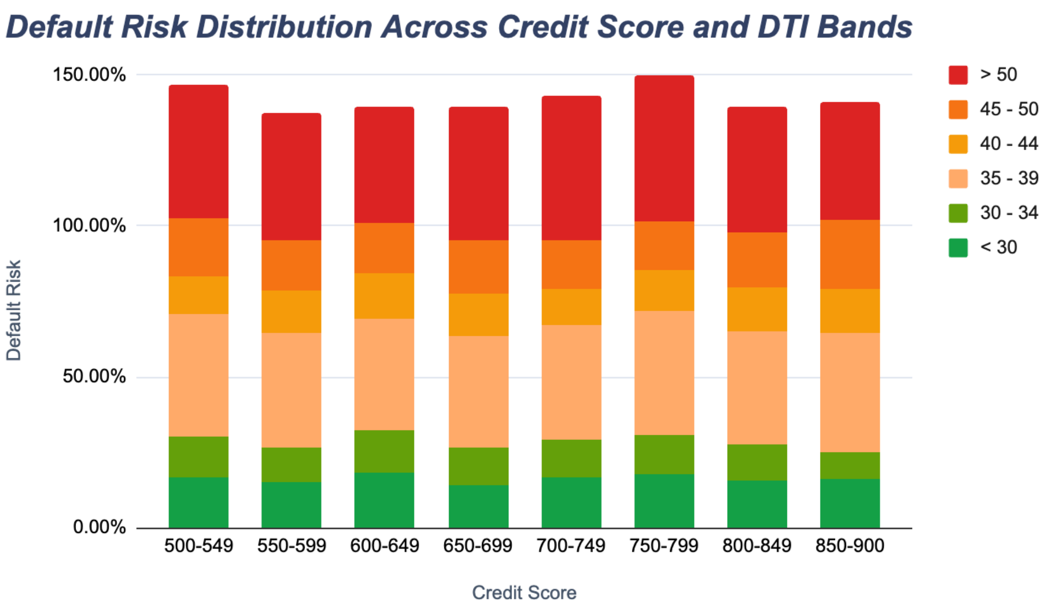
8. Advanced Analysis

Method: Segmentation + Interaction Analysis

Technique Used: Credit Score × DTI Heatmap

Finding:
DTI levels above 50% represent a universal high-risk threshold across all credit score categories. Even borrowers with strong credit profiles show elevated default probabilities once their debt burden crosses this level.

Interpretation:
Debt burden is a more reliable predictor of loan default than credit history alone, indicating that current financial capacity outweighs past credit behavior in risk assessment.



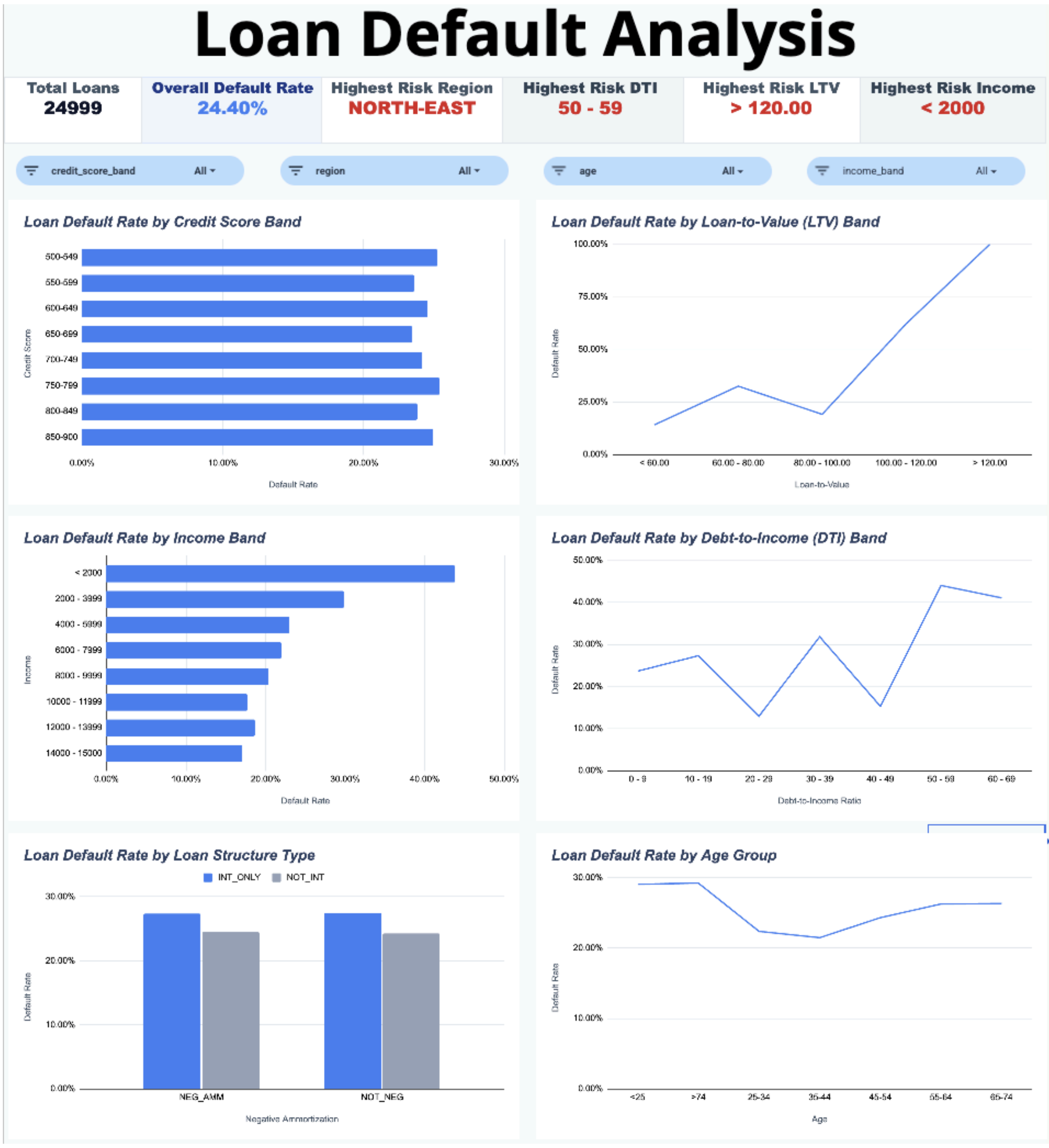
---

9. Dashboard Design

The dashboard was developed in Google Sheets using pivot tables, formulas, interactive filters, KPI panels, and dynamic charts to enable real-time monitoring, visualization, and analysis of borrower risk metrics.

Objective

Enable real-time risk monitoring for executives and underwriters.



Filters

Region, Credit score, LTV band, Income bracket, Loan structure.

The layout supports quick executive insights with deeper analytical drilldowns.

---

10. Insights Summary

1. Loan Default rate is high (24.40%), requiring stricter policies.
2. LTV is the strongest predictor.
3. Low income significantly increases risk.
4. DTI above 50% is a critical threshold.
5. Credit score alone is unreliable.
6. Regional disparities exist.
7. Age shows a U-shaped risk pattern.
8. Loan structure influences repayment.
9. Combined high DTI + LTV drives defaults.
10. Automated rules can detect risk instantly.

---

11. Recommendations

Recommendation	Insight	Impact
Implement LTV caps	LTV strongest predictor	Prevent extreme defaults
Income-based limits	Low income high risk	Reduce risky approvals
DTI cutoff 50%	High DTI danger	Filter unstable borrowers
Regional pricing	Regional variation	Adjust risk exposure
Age verification	Age risk	Reduce lifecycle risk
Prioritize DTI/LTV	Credit score weak	Better accuracy
Flag risky loans	Loan structure impact	Improve loan quality
Automated rules	Threshold detection	25% efficiency gain

---

12. Impact Estimation

Metric	Estimate
Cost savings	~$12M per $1B loans if defaults drop 5%
Efficiency	underwriting time ↓ 25%
Service	earlier intervention reduces delinquencies
Risk	default pool could fall 30–40%

---

13. Limitations

- Single year dataset
- Imputation may hide patterns
- No macroeconomic data
- Removed loan variables may contain signal
- Credit score uniformity may indicate bias

---

14. Future Scope

Further Analysis

- Logistic Regression model
- Random Forest model
- Time-series forecasting

Additional Data Needed

- Multi-year loan history
- Housing price index

---

15. Conclusion

This analysis demonstrates that structural financial indicators such as LTV and DTI outperform traditional credit score metrics in predicting loan defaults. Implementing a data-driven underwriting framework can significantly reduce financial losses, improve operational efficiency, and strengthen long-term portfolio stability.

---

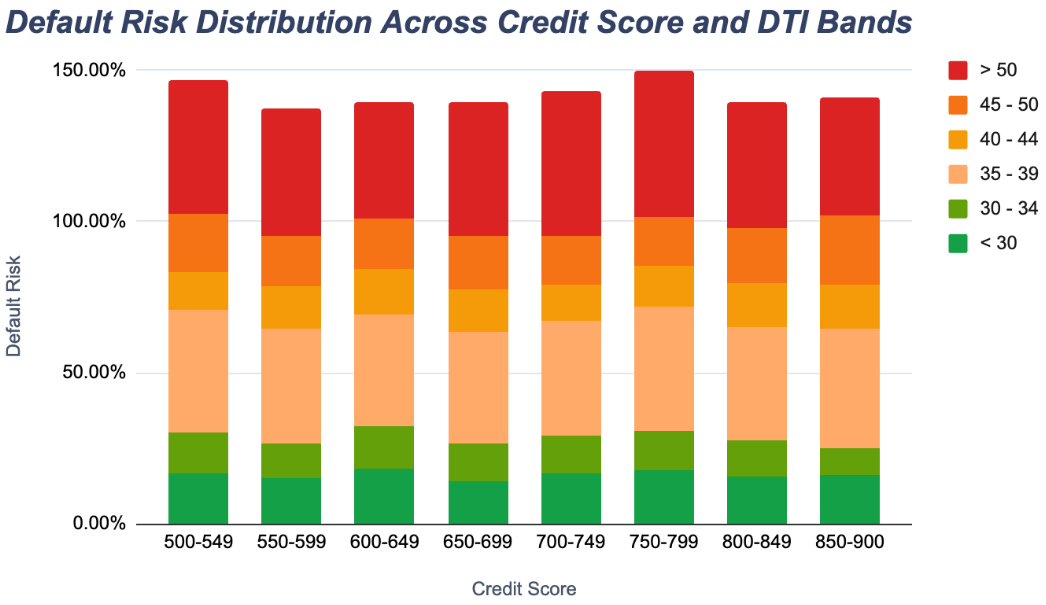
16. Appendix

A. Data Dictionary

Defined in the Data Description section.

B. Additional Analysis

- Risk heatmap



- KPI dashboard

C. Suggested Models

- Logistic Regression
- Random Forest

D. Sample Risk Logic

- IF LTV > 100 → Reject
- IF DTI ≥ 50 → High Risk Review
- IF Income < 2000 → Manual Review
- ELSE → Approve

---

17. Contribution Matrix

Member	Dataset	Cleaning	Analysis (Pivot Tables)	Dashboard	Report	PPT	Role
Husain Khorakiwala	✔	✔	✔	✔		✔	Project Lead
Adnan Rizvi	✔	✔			✔	✔	Documentation Lead
Anurag Pandey	✔					✔	Presentation Lead
Mukul Kumar		✔	✔			✔	Data Cleaning Lead
Shivansh Tiwari	✔	✔		✔		✔	Dashboard Lead
Tanishk Agrawal	✔		✔		✔		Analysis Lead

Declaration

We confirm that the above contribution details are accurate and verifiable.

Team Signatures

- HUSAIN KHURAKIWALA :
  ADNAN RIZVI : Adnan
  ANURAG PANDEY : Anurag
  MUKUL KUMAR : Mukul
  SHIVANSH TIWARI : shivansh
  TANISHK AGRAWAL : Tanishk