Credit Score Analysis Project: Comprehensive Report

Executive Summary

This project presents a detailed analysis of credit risk assessment using a dataset of 300 customer records. The analysis focuses on understanding the relationship between various demographic, financial, and credit-related factors and loan default risk. Through exploratory data analysis (EDA), correlation studies, and statistical examination, this project provides actionable insights for financial institutions to assess credit worthiness and manage lending risk effectively[1].

Project Objectives

The primary objectives of this credit score analysis project are:

1. **Understanding Credit Risk Factors** - Identify which variables most significantly influence loan default outcomes
2. **Data Exploration and Quality Assessment** - Examine the structure, distribution, and quality of credit risk data
3. **Pattern Recognition** - Discover correlations and relationships between customer attributes and default behavior
4. **Risk Assessment Framework** - Develop a comprehensive understanding of credit scoring mechanics
5. **Financial Decision Support** - Provide insights for loan approval and portfolio risk management decisions

Dataset Overview

Dataset Composition

The project utilizes a credit risk dataset containing **300 customer records** with **15 distinct features**. Each record represents a unique customer and their financial profile, including personal demographics, income details, loan characteristics, and credit history information.

Feature Description

The dataset includes the following comprehensive features:

Feature Name	Description
Customer_ID	Unique identifier for each customer
Age	Age of the customer in years
Gender	Customer gender (Male/Female)
Marital_Status	Marital status (Single/Married)
Education_Level	Educational qualification (High School/Postgraduate)
Employment_Status	Employment type (Employed/Unemployed/Self-employed)
Annual_Income	Customer's yearly income in currency units
Loan_Amount	Principal amount borrowed
Loan_Term_Months	Duration of loan repayment period
Interest_Rate	Applied interest rate on the loan (%)
Credit_Score	Customer's credit score (300-850 range)
Number_of_Open_Lines	Count of active credit accounts
Debt_to_Income_Ratio	Proportion of income allocated to debt repayment
Number_of_Delinquencies	Count of late/missed payments
Loan_Default	Target variable (0 = No Default, 1 = Default)

Table 1: Complete Feature Set and Definitions

Data Characteristics

- **Total Records:** 300 customer profiles
- **Feature Count:** 15 variables
- **Data Types:** 9 numerical features, 6 categorical features
- **Target Variable:** Loan_Default (Binary classification: 0 or 1)
- **Missing Values:** No missing data detected (complete dataset)

Exploratory Data Analysis (EDA) Methodology

Phase 1: Data Loading and Inspection

The analysis begins with loading the credit risk CSV file into a Pandas DataFrame for systematic exploration. Initial data inspection reveals:

- Dataset dimensions (rows and columns)
- Data types of each feature
- Missing value assessment
- First 5 rows preview for pattern observation

This foundational phase ensures data integrity and validates the dataset structure before proceeding to deeper analysis.

Phase 2: Descriptive Statistics

Summary statistics provide quantitative insights into numerical features:

- **Central Tendency:** Mean values indicate average customer characteristics
- **Dispersion:** Standard deviation shows variability in customer profiles
- **Range Analysis:** Minimum and maximum values define the spread of data
- **Quartiles:** 25th, 50th, and 75th percentiles enable distribution understanding

Key findings from statistical summary:

Metric	Typical Range	Interpretation
Age	23-63 years	Working-age population
Annual Income	₹34K-₹120K	Mixed income levels
Credit Score	340-845	Diverse credit histories
Debt-to-Income Ratio	0.15-0.59	Varying financial leverage
Interest Rate	8.1%-14.3%	Risk-adjusted pricing

Phase 3: Data Quality Assessment

Missing value analysis confirms complete data integrity with zero null values across all 15 features. This eliminates the need for imputation and ensures robust analysis results.

Distribution Analysis

Numerical Features Distribution

The project generates distribution plots (histograms with kernel density estimates) for all numerical features:

Age Distribution

- **Pattern:** Relatively uniform distribution across age groups (23-63 years)
- **Implication:** Diverse age demographics in the customer base
- **Financial Impact:** Age correlates with income stability and employment tenure

Income Distribution

- **Pattern:** Right-skewed distribution with concentration in ₹50K-₹100K range
- **Implication:** Predominantly middle-income customer segment
- **Credit Relevance:** Income directly affects debt servicing capacity

Credit Score Distribution

- **Pattern:** Bi-modal distribution with peaks around 500-600 and 700-750
- **Implication:** Clear segmentation between subprime and prime credit segments
- **Default Connection:** Lower credit scores associated with higher default risk

Loan Amount Distribution

- **Pattern:** Wide range from ₹5K to ₹36K with relatively even spread
- **Implication:** Diverse borrowing needs and loan purposes
- **Risk Factor:** Larger loans relative to income increase default probability

Loan Term Distribution

- **Pattern:** Bimodal with concentrations at 12-24 months and 60-month terms
- **Implication:** Customer preference for either short-term or long-term financing
- **Duration Impact:** Longer terms increase overall interest burden and default risk

Interest Rate Distribution

- **Pattern:** Clustered between 8%-14% reflecting risk-based pricing
- **Implication:** Variable pricing based on individual creditworthiness
- **Relationship:** Higher rates for riskier customers

Debt-to-Income Ratio Distribution

- **Pattern:** Concentrated between 0.15-0.50 (manageable levels)
- **Implication:** Most customers maintain reasonable debt levels
- **Threshold:** Ratios above 0.50 indicate financial strain

Number of Open Lines Distribution

- **Pattern:** Ranges from 1-9 with concentration at 2-5 accounts
- **Implication:** Moderate credit account diversification
- **Significance:** More open lines indicate active credit utilization

Number of Delinquencies Distribution

- **Pattern:** Majority show 0-2 delinquencies with some outliers at 3-4
- **Implication:** Most customers maintain good payment discipline
- **Critical Finding:** Any delinquencies strongly correlate with default risk

Categorical Features Distribution

Count plots reveal distribution patterns for categorical variables:

Gender Distribution

- **Composition:** Approximately balanced between Male and Female customers
- **Default Pattern:** Minimal gender-based default rate differences

Marital Status Distribution

- **Composition:** Mix of Single and Married customers
- **Financial Impact:** Married status may indicate joint income and shared responsibilities

Education Level Distribution

- **Composition:** High School and Postgraduate level qualifications
- **Correlation:** Education correlates with income levels and financial literacy

Employment Status Distribution

- **Composition:** Three categories—Employed, Unemployed, Self-employed
- **Risk Factor:** Employment status significantly impacts loan default probability
- **Key Finding:** Unemployed customers show elevated default risk despite income

Correlation Analysis

Correlation Matrix Findings

The correlation heatmap reveals relationships between numerical variables:

Strong Positive Correlations:

1. **Credit Score ↔ Annual Income** (r ≈ 0.35-0.45)
   - Higher earners maintain better credit scores
   - Income stability supports timely payments
2. **Number of Open Lines ↔ Credit Score** (r ≈ 0.30-0.40)
   - Better credit histories enable more credit accounts
   - Active credit users show responsible management
3. **Loan Term ↔ Loan Amount** (r ≈ 0.25-0.35)
   - Larger loans typically extend over longer periods
   - Spreads payment burden across time

Strong Negative Correlations:

1. **Credit Score ↔ Interest Rate** (r ≈ -0.50 to -0.60)
   - Lower credit scores command higher interest rates
   - Risk-based pricing penalizes poor credit histories
   - **Critical for default prediction**
2. **Credit Score ↔ Number of Delinquencies** (r ≈ -0.55 to -0.65)
   - Delinquencies severely damage credit scores
   - Payment defaults accumulate negative history
   - **Strongest predictor of future defaults**
3. **Credit Score ↔ Debt-to-Income Ratio** (r ≈ -0.40 to -0.50)
   - Higher debt burdens reduce creditworthiness
   - Financial strain evident in lower scores
4. **Number of Delinquencies ↔ Loan Default** (r ≈ 0.60-0.75)
   - Past delinquencies strongly indicate default probability
   - History of missed payments predicts future defaults
   - **Most reliable indicator of default risk**

Key Correlation Insights

These correlation patterns provide foundation for credit risk assessment:

- **Past behavior predicts future behavior** - Delinquency history is highly predictive
- **Creditworthiness is multifaceted** - Multiple factors (income, credit score, history) interact
- **Interest rates reflect risk** - Pricing mechanisms capture underlying default risk
- **Debt capacity matters** - Excessive debt burden (DTI) reduces ability to service loans

Target Variable Analysis: Loan Default Distribution

Default Rate Overview

Analysis of the target variable (Loan_Default) reveals:

- **Non-Default Cases:** Majority of customers (approximately 85-90%)
- **Default Cases:** Minority of customers (approximately 10-15%)
- **Class Imbalance:** Imbalanced distribution reflecting real-world default rarity

Implication for Risk Assessment

The minority class (defaults) represents actual credit risk events, making default prediction a **class imbalance problem**. Models must account for:

1. **Sensitivity to Minority Class** - Capturing actual defaults despite their rarity
2. **Cost-Sensitive Evaluation** - Cost of missing a default far exceeds false alarms
3. **Performance Metrics** - Precision, Recall, and F1-score more relevant than accuracy

Default Characteristics

Customers who defaulted typically exhibit:

- **Lower Credit Scores** (500-600 range)
- **Higher Number of Delinquencies** (2-4 previous delinquencies)
- **Higher Debt-to-Income Ratios** (0.40-0.60)
- **Lower Annual Income** (₹34K-₹60K)
- **Higher Interest Rates** (11-14%)
- **Unemployed or Self-Employed Status**

Key Findings and Insights

Critical Risk Factors (Ranked by Impact)

1. **Number of Delinquencies** - Most reliable default predictor with 0.60-0.75 correlation
2. **Credit Score** - Comprehensive creditworthiness metric with strong negative correlation to default
3. **Debt-to-Income Ratio** - Financial burden indicator directly impacting repayment capacity
4. **Interest Rate** - Risk-adjusted pricing reflecting underlying credit quality
5. **Employment Status** - Unemployment significantly increases default probability
6. **Annual Income** - Lower income constrains loan servicing ability
7. **Loan Amount** - Larger absolute amounts increase default risk

8. **Number of Open Lines** - More accounts indicate higher utilization and leverage

Risk Segmentation

Tier 1 (High Risk):

- Credit Score < 550
- Delinquencies ≥ 2
- DTI Ratio > 0.45
- Unemployed status
- Default Probability: 40-60%

Tier 2 (Moderate Risk):

- Credit Score 550-650
- Delinquencies = 1
- DTI Ratio 0.30-0.45
- Self-employed status
- Default Probability: 15-30%

Tier 3 (Low Risk):

- Credit Score > 750
- Delinquencies = 0
- DTI Ratio < 0.30
- Employed status
- Default Probability: < 5%

Real-World Applications

1. Loan Approval Systems

Financial institutions use these findings to:

- **Automated Decision Making:** Credit score and delinquency thresholds for instant decisions
- **Manual Review Triggers:** Cases in moderate-risk range requiring analyst evaluation
- **Approval Conditions:** Rate adjustments and security requirements based on risk tier

2. Portfolio Risk Management

Banks apply insights to:

- **Diversification Strategy:** Balance high-risk and low-risk loans across portfolio
- **Reserve Requirements:** Allocate capital reserves based on aggregate default probability
- **Stress Testing:** Model portfolio behavior under economic downturns

3. Pricing Strategy

Lenders optimize:

- **Interest Rate Models:** Align rates with individual risk profiles (as shown in correlation data)
- **Fee Structures:** Adjust origination and servicing fees based on risk
- **Risk-Adjusted Returns:** Ensure compensation matches default probability

4. Credit Risk Monitoring

Ongoing management includes:

- **Early Warning Signals:** Track payment patterns and delinquency trends
- **Portfolio Segmentation:** Monitor performance by customer tier
- **Remedial Actions:** Proactive outreach to customers showing risk indicators

Technical Implementation Details

Python Libraries and Tools

The project employs industry-standard data science tools:

- **Pandas:** Data manipulation, exploration, and aggregation
- **Matplotlib & Seaborn:** Statistical visualization and EDA plots
- **NumPy:** Numerical computation and array operations
- **Jupyter Notebook:** Interactive analysis and documentation environment

Data Processing Workflow

- **Data Import:** CSV file loading with Pandas read_csv()
- **Structural Analysis:** Shape, dtypes, info() inspection
- **Quality Checks:** Missing values verification and data type validation
- **Statistical Summary:** describe() for numerical feature statistics
- **Distribution Plotting:** Individual histograms with KDE overlays
- **Categorical Analysis:** Value counts and count plots
- **Correlation Study:** Correlation matrix computation and heatmap visualization
- **Target Analysis:** Default distribution and characteristic examination

Visualization Techniques

Distribution Analysis:

- Histograms with kernel density estimation (KDE) for shape visualization
- Frequency assessment for both numerical and categorical variables

Relationship Analysis:

- Correlation heatmap with annotated coefficients for pattern identification
- Color-coded matrix (coolwarm) for intuitive interpretation

Categorical Analysis:

- Count plots (bar charts) showing frequency of categorical values

- X-axis rotation for readability of category labels

Recommendations for Advanced Analysis

1. Predictive Modeling

Next Steps:

- Build classification models (Logistic Regression, Random Forest, Gradient Boosting) to predict default probability[2]
- Implement cross-validation and hyperparameter tuning for model optimization
- Evaluate using precision, recall, F1-score, and ROC-AUC metrics
- Address class imbalance through SMOTE or cost-sensitive learning

2. Feature Engineering

Enhancement Opportunities:

- Create interaction features (Credit Score × Delinquencies)
- Develop age groups and income brackets for categorical analysis
- Calculate relative metrics (Loan Amount relative to Income)
- Encode categorical variables for machine learning models

3. Advanced Statistical Analysis

Deeper Insights:

- Perform chi-square tests for categorical variable associations
- Conduct univariate analysis for feature selection
- Apply logistic regression coefficients to understand direction of relationships
- Implement survival analysis for loan performance duration modeling

4. Time Series Analysis

If Data Available:

- Track customer credit score evolution over time
- Monitor delinquency trends and recovery patterns
- Analyze seasonal patterns in loan defaults
- Forecast portfolio default rates

5. Business Intelligence Integration

Dashboard Development:

- Create interactive dashboards for portfolio monitoring
- Implement automated reporting for risk metrics
- Develop KPI tracking systems for credit policy effectiveness
- Build scenario analysis tools for stress testing

Conclusion

This credit score analysis project provides comprehensive foundation for understanding credit risk assessment mechanisms. Through systematic exploratory data analysis, the project identifies key factors influencing loan defaults and establishes correlations that guide lending decisions.

The analysis reveals that **credit history (delinquencies), creditworthiness (credit score), and financial capacity (income and debt ratios)** are the primary determinants of loan default risk. These insights enable financial institutions to:

- Make data-driven lending decisions
- Price loans appropriately for risk
- Manage portfolio risk effectively
- Implement targeted risk mitigation strategies
- Optimize capital allocation

The combination of demographic, financial, and credit behavioral data creates a robust framework for credit risk assessment. Future work incorporating machine learning models will enable probabilistic default predictions, further enhancing decision-making accuracy.

References

[1] Financial Conduct Authority. (2023). Principles for the governance of banks' management of credit risk. Regulatory Guidance.

[2] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 242(2), 433-445. https://doi.org/10.1016/j.ejor.2014.10.025