Identifying Mispriced Loans through Interest Rate-Based Network Analysis and Clustering in P2P Lending Markets

Lennart Baals

Bern University of Applied Science, University of Twente

December 2, 2025

Outline

Table of Contents

Introduction

Background:

Peer-to-peer (P2P) lending platforms facilitate direct lending between individuals.
Loans are assigned interest rates based on perceived borrower risk.

Problem Statement:

Mispricing occurs when interest rates do not accurately reflect the borrower's risk profile.
This can lead to unfair loan terms and increased default risk.

Objective:

The development of a methodology to identify mispriced loans by analyzing similarities between loans.

Approach Overview:

Construct a loan network based on interest rates and borrower risk factors.
Use network analysis and clustering techniques to detect mispricing.

Table of Contents

Data Preparation

Dataset Overview:

- **Source:** Bondora P2P lending platform.
- **Sample Size:** 4,000 loans (2,000 defaulted, 2,000 non-defaulted).
- **Features Include:**
  - Borrower characteristics: Age, gender, income, employment status.
  - Loan details: Interest rate, loan amount, monthly payment.
  - Loan performance: Default status.

Data Cleaning and Preprocessing:

- Addressed missing values and outliers.
- Normalized numerical variables for consistency.
- One-hot encoded categorical variables.
- Removed highly correlated features (correlation > 0.95).
- Balanced the dataset to prevent bias.

Feature Selection

Interest Rate:

- Treated as a key factor due to its direct link to loan pricing and potential mispricing.
- Included individually in the composite edge weight to capture pricing similarities.

Risk Factors Selected:

- Age
- Gender
- Debt-to-Income Ratio
- Monthly Payment
- Number of Previous Loans
- Amount of Previous Loans
- Total Income
- Log of Loan Amount

Rationale for Selection:

- These risk factors significantly influence a borrower's risk profile.
- Ensures a comprehensive assessment of similarities in borrower characteristics.

Data Transformation:

- Converted binary variables to factors where appropriate.
- Standardized variables to ensure compatibility in similarity computations.

Similarity Measures

Interest Rate Similarity:

Calculated pairwise differences between loans' interest rates.
Converted differences to similarities:

$$\text{Interest Rate Similarity}_{ij} = 1 - \left(\frac{|\text{Interest Rate}_i - \text{Interest Rate}_j|}{\text{Max Difference}} \right)$$

Risk Factor Similarity:

Employed Gower's distance to handle mixed data types.
Converted distances to similarities:

$$\text{Risk Factor Similarity}_{ij} = 1 - \text{Gower Distance}_{ij}$$

Composite Edge Weights:

Combined similarities to form edge weights:

$$\text{Edge Weight}_{ij} = 0.5 \times \text{Interest Rate Similarity}_{ij} + 0.5 \times \text{Risk Factor Similarity}_{ij}$$

Equal weighting to balance influence.

Graph Creation

Adjacency Matrix Formation:

- **Thresholding Approach:**
  - Applied a distance threshold (≤ 0.3) to retain only strong similarities between loans.
- **K-Nearest Neighbors (KNN) Approach:**
  - Connected each node to its 5 nearest neighbors based on similarity, ensuring each loan has sufficient connections.
  - Used KNN to handle isolated nodes and ensure full network connectivity.
- **Handling Isolated Nodes:**
  - For any nodes left isolated after thresholding, applied KNN to create additional connections, guaranteeing every loan node is reachable within the network.

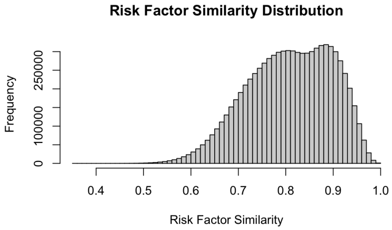
Graph Characteristics:

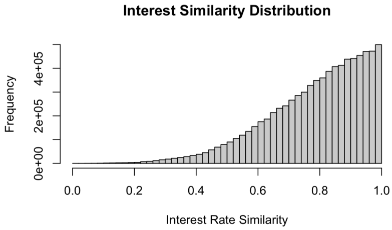
- Type: Undirected, weighted graph.
- Nodes represent individual loans.
- Edges represent similarities based on composite edge weights.

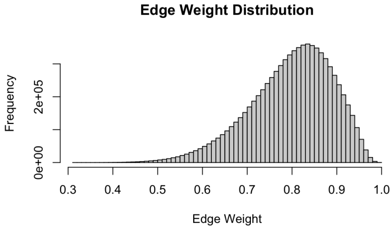
Connectivity Check:

- Ensured the graph is connected.

Similarity and Edge Weight Distributions of the Network







Network Attributes and Metrics

Node Attributes Added:

- Interest rate
- Loan amount
- Selected risk factors (e.g., age, gender, debt-to-income)

Computed Network Metrics:

- **Degree Centrality:** Number of direct connections.
- **Betweenness Centrality:** Importance in connecting different parts of the network.
- **Clustering Coefficient:** Degree to which nodes tend to cluster together.

Community Detection:

- Applied the **Louvain Method**.
- Loans grouped into communities based on similarity and information is added to the loan sample.

Comments on Network Metrics Computation

Computed Network Metrics:

- **Degree Centrality (k_i):** Number of direct connections of loan i.
- **Betweenness Centrality (b_i):** Measure of how often loan i lies on the shortest paths between other loans.
- **Clustering Coefficient (C_i):** Degree to which loans connected to loan i are interconnected.

Purpose:

- Identify influential loans within the network.
- Detect patterns that might be associated with mispricing.

Implementation:

- Used `igraph` package in R for efficient computation.

Gaussian Mixture Models (GMM) Clustering

Objective:

- Identify underlying patterns and group loans into clusters.
- Complement network communities with statistical clustering.

Features Used:

- Interest Rate, Degree Centrality, Betweenness Centrality, Clustering Coefficient.

Methodology:

- Standardized features for consistency.
- Applied GMM to capture data complexity and clusters.

Results:

- Determined optimal number of clusters using BIC scores.
- Assigned cluster memberships to loans.

Cluster Analysis

Cluster Distribution: Identified 9 clusters with varying sizes, shown in a bar chart or table.

Cluster Characteristics: Weighted averages of Interest Rate and Default Rate.

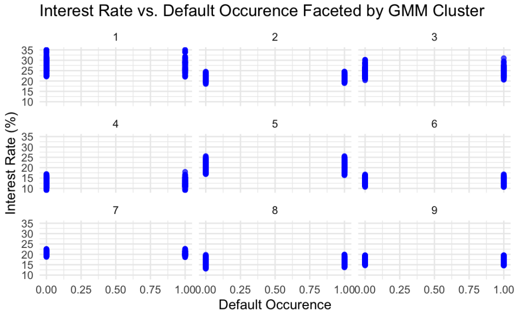


Figure 1: Interest Rate vs. Default Rate

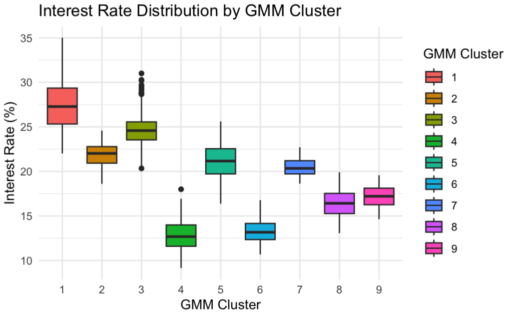


Figure 2: Interest Rate Distribution by Cluster

Comparative Analysis

ANOVA Tests:

- Conducted ANOVA to assess differences in:
  - Interest Rates across clusters.
  - Betweenness Centrality across clusters.
  - Degree Centrality across clusters.
- Significant differences found ($p < 0.001$).

Post-hoc Analysis:

- Performed Tukey's HSD tests to identify specific cluster differences.
- Detected clusters with significantly higher or lower metrics.

Implications:

- Validates the heterogeneity among clusters.
- Supports the need for further investigation of targeted pricing strategies.

Identifying Mispriced Loans

Ex-Post Identification:

- Defined mispriced loans as:
  - High interest rate but no default (*overpriced*).
  - Low interest rate but defaulted (*underpriced*).
- Used quartiles to determine thresholds.

Analysis:

- Identified loans meeting mispricing criteria.
- Analyzed their distribution across clusters.

Findings:

- Certain clusters have higher proportions of mispriced loans.
- Indicates potential areas for pricing adjustments.

Visualization of Loan Amount per Cluster

Scatter Plot:

Plotted clusters based on Average Interest Rate & Average Default Rate, sized by total loan amount:

Figure:

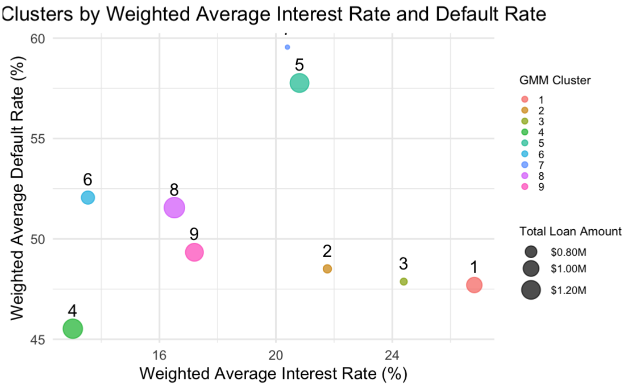


Figure 3: Clusters with Total Loan Amount

Visualization of Mispriced Loans per Cluster

Scatter Plot:

Plotted clusters based on Average Interest Rate & Average Default Rate, sized by total loan amount and proportion of mispriced loans per cluster:

Figure:

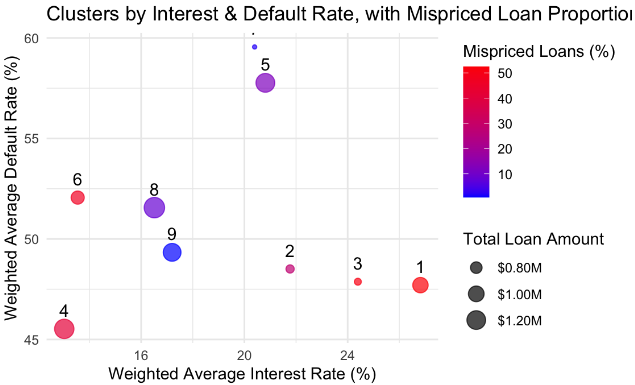


Figure 4: Clusters with Mispriced Loan Proportion

Ex-Ante Mispricing Detection

Objective:

Predict mispricing before loan issuance.

Utilize only ex-ante variables (available prior to loan approval).

Methodology:

Built a Multinomial Logistic Regression model to predict Bondora's credit ratings.

Predictor variables included borrower characteristics and network metrics.

Results:

Model achieved an accuracy of 83.57% on the test set.

Confusion matrix indicated good classification performance across ratings.

Comparing Expected and Actual Interest Rates

Mapping Predicted Ratings:

Mapped predicted credit ratings to expected interest rate intervals.
Used Bondora's official rating scale.

Interest Rate Difference:

Calculated difference between expected and actual interest rates.
Defined a mispricing threshold (e.g., 2%).

Mispricing Classification:

Loans with interest differences exceeding the threshold were labeled as *Mispriced*.
Others were labeled as *Properly Priced*.

Analysis:

Examined mispricing distribution across clusters.
Identified clusters with high mispricing rates.

Insights from the Analysis

Key Findings:

Network metrics are significant in identifying mispriced loans.
Clusters with high mispricing align with higher default rates.

Practical Implications:

P2P platforms can leverage network analysis for better pricing strategies.
Investors can assess loan portfolios for hidden risks.

Limitations:

Data limited to a specific platform and timeframe.
Model assumptions may not hold universally.

Future Work:

Incorporate more advanced machine learning models.
Validate the approach on larger and more diverse datasets.