Leveraging Network Topology for Credit Risk Assessment in P2P Lending: A Comparative Study under the Lens of Machine Learning

Yiting Liu, Lennart John Baals, Jörg Osterrieder, Branka Hadji-Misheva

July 26, 2023

Table of Contents I

Table of Contents

Introduction to P2P Lending - Overview

Definition of P2P Lending

Peer-to-peer (P2P) lending is a method of debt financing that enables individuals to borrow and lend money without the use of an financial institution as an intermediary.

Online Platform Character

P2P lending occurs through online platforms that pair lenders with potential borrowers.

These platforms used to offer more inclusiveness and better funding efficiency than banks.

Benefits for Lenders

Lenders can earn higher returns compared to traditional savings and investment products.

Lenders can choose which borrowers to invest in.

Benefits for Borrowers

Borrowers can access financing faster and often with less stringent credit checks.

This makes it beneficial for personal loan seeking.

Growing Importance

With the power of technology and the internet, P2P lending has grown in popularity since it started in the early 2000s. Today, P2P lending platforms have facilitated billions of dollars in loans *(Chen, Huang, and Shaban 2022)*.

Introduction to P2P Lending - Overview Part 2

Challenges in P2P Lending

In this context, P2P lending also imposes challenges to creditors as issued loans are usually unsecured which results in higher default rates, and a lower number of recovery options.

Need for Regulation

Given its nature, there's also a need for regulation to protect both borrowers and lenders from increased default risk.

Thus, governments and regulatory bodies are increasingly focusing on P2P lending platforms.

Need for Efficient Credit Scoring

Credit scoring is an important part of P2P lending.

It is used to assess the risk associated with a given borrower.

Integration of Machine Learning

Machine learning is now being used to enhance credit scoring processes.

The potential to improve loan default predictions is immense but yet sparsely investigated.

Summary

In summary, P2P lending is an innovative and growing field, but one that requires careful risk management. The use of machine learning could be a key to improving these risk assessments.

Table of Contents

Bondora: A Leading European Peer-to-Peer Lending Platform

Overview

Bondora (`https://www.bondora.com/en`) is a peer-to-peer (P2P) lending platform established in Estonia. The platform's operational design allows lenders and borrowers to transact directly between each other.

Disintermediation and Reputation

Acting as a prime example of financial sector disintermediation, Bondora boosts efficiency and democratizes credit access by eradicating the need for an intermediary, consequently reducing entry barriers for borrowers.

Diverse Participant Base and Rich Data

Bondora boasts a diverse user base, encompassing 221,357 individual lenders that lend funds to borrowers with varied demographic and credit backgrounds.

The platform has issued €815.4 Mio. in loans, generating rich data on loan listings, bidding records, and payment histories.

Table of Contents

Two-step Model: Step 1 - Build a Graph on Data

An Explanation

Graph theory offers a robust framework for analyzing complex relational data, including financial networks. These networks encapsulate financial entities as nodes and relationships as edges, providing valuable insights into risk dynamics *(Biggs, Lloyd, and Wilson 1986; Newman 2010; DeMarzo 2003)*.

Key Concepts

Graph: A set of nodes (vertices) and edges (links).

Degree: Number of edges connected to a node.

Path: Sequence of nodes in which each node is connected to the next by an edge.

Cycle: Path whose first and last nodes are the same.

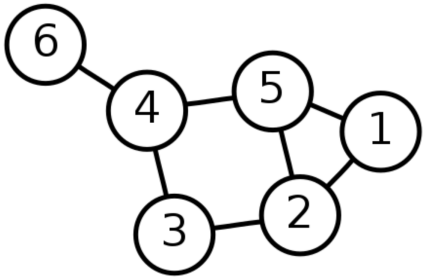


Figure: An example of graph

Two-Step Model: Step 1 - Build a Graph on Data

Overview

Centrality measures are important tools in network analysis, used to identify the most important nodes within a network. In the context of P2P lending, these nodes could represent key borrowers or lenders.

Key Measures

Degree Centrality: Number of connections a node has.

Closeness Centrality: How fast information can spread from a given node to other reachable nodes in the network.

Betweenness Centrality: A node's centrality.

Eigenvector Centrality: A node is considered important if it is connected to other important nodes.

Mathematical Formulation

Formula for Degree Centrality:
$C_D(v) = \deg(v)$

Formula for Closeness Centrality:
$C(x) = \frac{1}{\sum_y d(y,x)}$

Formula for Betweenness Centrality:
$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$

Formula for Eigenvector Centrality:
$PR(p) = (1-d) + d \sum_{i \in M(p)} \frac{PR(i)}{L(i)}$

Two-Step Model: Step 1 - Build a Graph on Data (cont.)

Overview

Other centrality measures like Katz Centrality, Hub Centrality, and PageRank could provide valuable insights about key participants in P2P lending.

Additional Measures

PageRank: More important websites are likely to receive more links from other websites.

Katz Centrality: Takes into account the total number of walks between a node and all others.

Hub Centrality (HITS): The HITS (Hyperlink-Induced-Topic-Search) algorithm is an analysis tool to rate links between nodes.

Mathematical Formulation

Formula for PageRank:

$$PR(p) = (1 - d) + d \sum_{i \in M(p)} \frac{PR(i)}{L(i)}$$

Formula for Katz Centrality:

$$C_{\text{Katz}}(i) = \sum_{j=1}^{n} \beta A_{ij} C_{\text{Katz}}(j) + \alpha$$

Formulas for Hub Centrality (HITS):

Authority Score: $a(i) = \sum_{j \in M(i)} h(j)$

Hub Score: $h(i) = \sum_{j \in N(i)} a(j)$

Significance

Centrality measures can identify potential hotspots of credit risk in P2P lending.

Two-Step Model: Step 2 - Supervised Models based on Graph Features

Introduction to Elastic Net

The Elastic Net Logistic Regression model integrates the strengths of both L1 (Lasso) and L2 (Ridge) penalties, providing an effective balance between bias and variance to ensure model generalizability.

Model Implementation

$$\beta = \arg\min_{\beta} \left(\frac{1}{2}||y - X\beta||_2^2 + \lambda((1-\alpha)||\beta||_2^2 + \alpha||\beta||_1) \right)$$

Here, $||\cdot||_2$ is the L2 norm, $||\cdot||_1$ is the L1 norm, λ is the regularization parameter, and α is the mixing parameter that ranges between 0 and 1.

Model Configuration

The alpha and lambda parameters are varied, with alpha ranging from 0 (Ridge penalty) to 1 (Lasso penalty) and $\lambda \in [1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100]$, enabling different degrees of regularization and feature selection.

Two-Step Model: Step 2 - Supervised Models based on Graph Features

Introduction to Random Forest

Random Forests, due to their nonparametric nature, are effective in handling high-dimensional spaces and complex interactions, making them a strong tool for default prediction in credit risk modeling.

Model Implementation

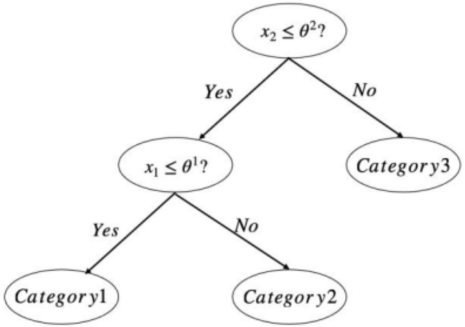


Figure: A decision tree

Model Configuration

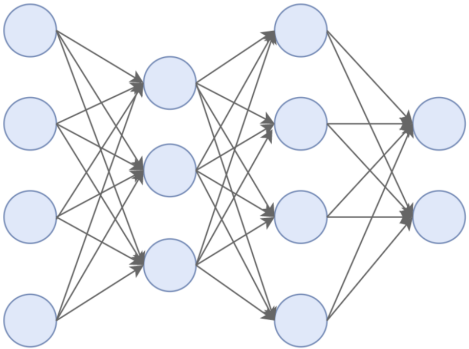
The model is configured with varying 'ntrees' (50-250) and 'max-depth' (5-20), to explore and determine the most suitable model configuration for the dataset in use.

Two-Step Model: Step 2 - Supervised Models based on Graph Features

Introduction to Deep Neural Network

Deep learning methods such as Multi-Layer Perceptron (MLP) enable models to automatically learn representations of data through neural networks with multiple layers, accommodating the complexity of credit risk data *(LeCun 2015)*.

Model Implementation



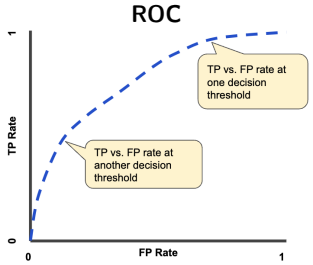
Model Configuration

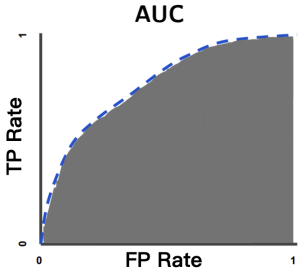
The model consists of various configurations of hidden layers, using the rectified linear unit (ReLU) activation function. The hyperparameter 'epochs' denotes the number of complete passes through the entire training dataset.

ROC and AUC: The performance of one model

Introduction to ROC and AUC

The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The Area Under the Curve (AUC) is literally the area under the ROC curve.





Why do we care about ROC and AUC?

A model with a higher AUC is generally better because it has a higher probability of ranking a randomly chosen positive instance higher than a randomly chosen negative instance. This helps us to quantify how well the model distinguishes between classes and to compare different models' performance.

DeLong Test: Compare the prediction performance of two models

Introduction to DeLong Test

The DeLong test is a statistical test used to compare the areas under two or more correlated ROC curves. This method takes into account the correlated nature of the data when the tests are performed on the same individuals.

Statistics in DeLong Test

Let AUC_1 and AUC_2 be the areas under the ROC curves of the two models.
Calculate the difference: $\Delta = AUC_1 - AUC_2$.
null hypothesis: $\Delta = AUC_1 - AUC_2 = 0$ vs. alternative hypothesis: $\Delta = AUC_1 - AUC_2 \neq 0$
Estimate the standard error of Δ: $SE(\Delta)$.
The DeLong statistic is $Z = \frac{\Delta}{SE(\Delta)}$.

Interpreting DeLong Test Results

The DeLong test provides a p-value, p. We interpret the test results as follows:

$$\text{Decision} = \begin{cases} \text{Reject } H_0 & \text{if } p < \alpha \text{ (Significant difference)} \\ \text{Fail to reject } H_0 & \text{if } p \geq \alpha \text{ (No significant difference)} \end{cases}$$

where H_0 is the null hypothesis that the two AUCs are the same, and α is the significance level (commonly 0.05).

Variable Importance

Introduction to Variable Importance

Variable importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. These scores can be used to interpret the data, understand the impact of specific features on the model's predictions, and to perform feature selection.

GLM: In Generalized Linear Models (GLMs), variable importance is represented by the magnitude of the coefficients. These coefficients serve as predictor weights of the standardized data and are mainly used for comparison of the relative variable importance.

Tree-Based Models: Tree-based models such as Gradient Boosting Machine (GBM), Distributed Random Forest (DRF), and XGBoost determine variable importance by calculating the relative influence of each variable during the tree building process. This is done by considering whether that variable was selected for splitting, and how much the squared error improved (decreased) as a result. The variable importances are computed from the gains of their respective loss functions during tree

Neural Network: For neural networks, variable importance is computed using the Gedeon method. This method is used to calculate the contribution of each input variable to the output of the neural network.

Table of Contents

Models Description and Representation

In our study, we utilized six different models. These models were differentiated based on the types of features they incorporated:

Model 1: Initial features (informative and uninformative). Python representation: ——.

Model 2: Initial features (informative and uninformative), and graph features. Python representation: - - - - - - - - - -.

Model 3: Only initial features (informative). Python representation: ——.

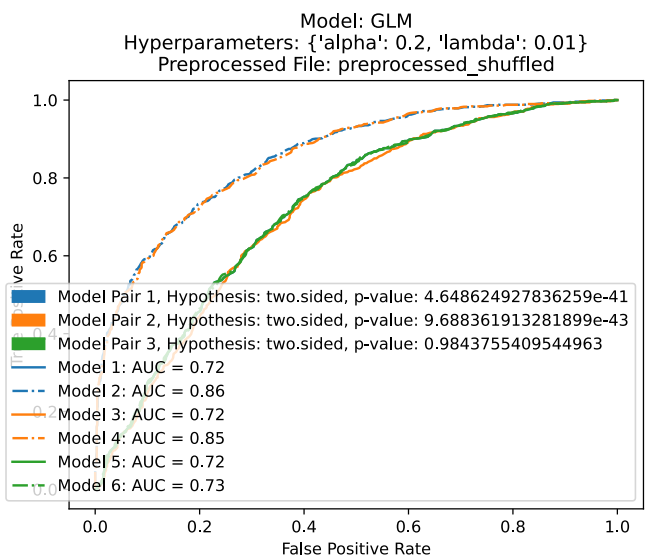
Model 4: Initial features (informative), and graph features. Python representation: - - - - - - - - - -.

Model 5: Initial features (informative and uninformative). Python representation: ——.

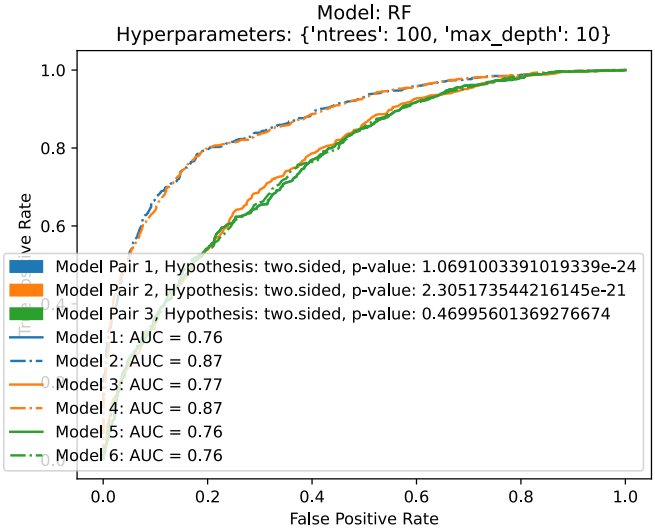
Model 6: Initial features (informative and uninformative), and shuffled graph features. Python representation: - - - - - - - - - -.

Each model is represented in Python using a specific color and linestyle.

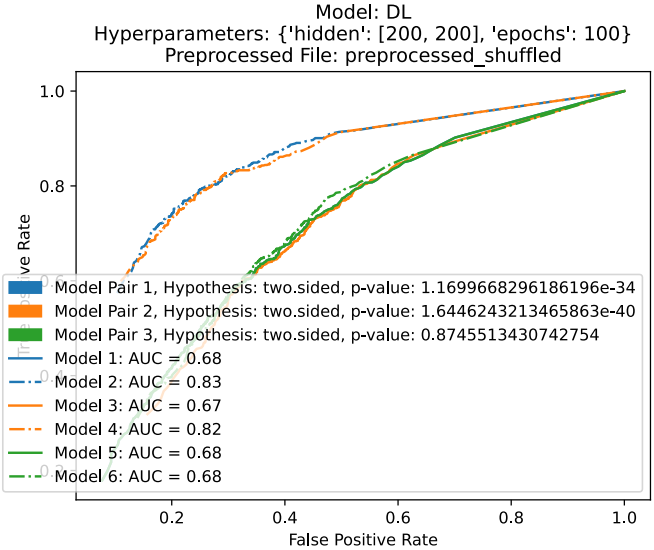
ROC and AUC: GLM



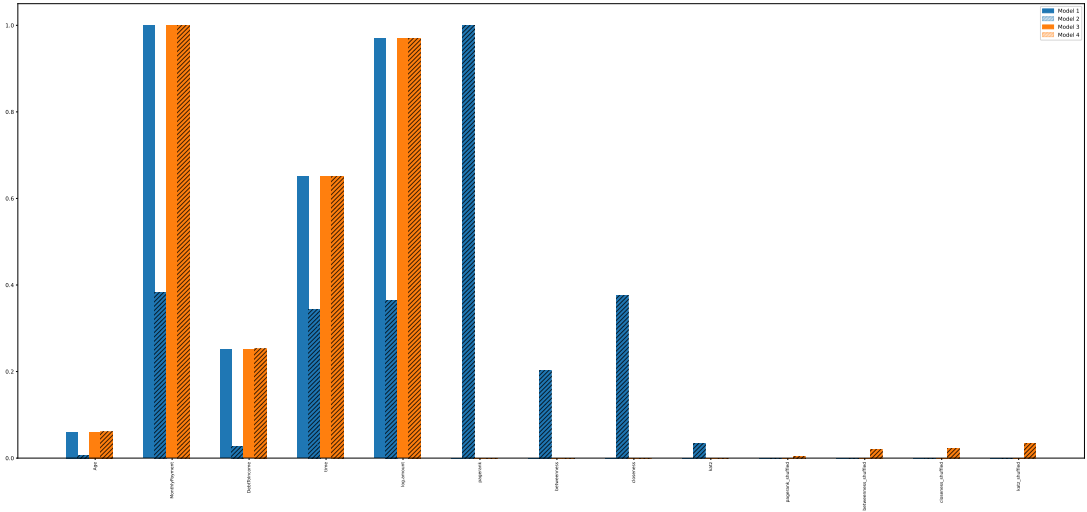
ROC and AUC: RF



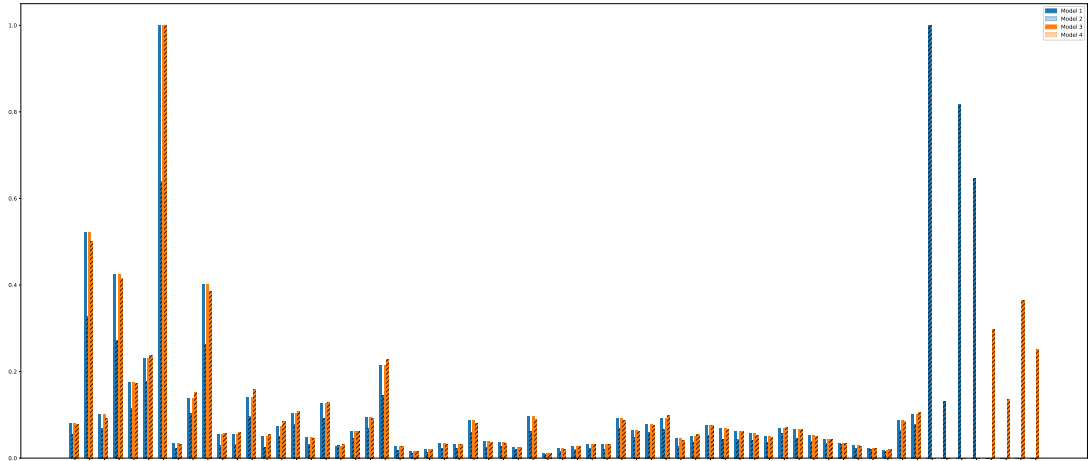
ROC and AUC: DL



Feature Importance: GLM



Feature Importance: RF



Feature Importance: DL