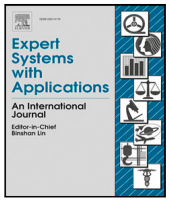
Expert Systems With Applications 252 (2024) 124100

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Leveraging network topology for credit risk assessment in P2P lending: A comparative study under the lens of machine learning

Yiting Liu [a,b], Lennart John Baals [a,b,*], Jörg Osterrieder [a,b], Branka Hadji-Misheva [a]

[a] Bern Business School, Bern University of Applied Sciences, Brückenstrasse 73, 3005 Bern, Switzerland
[b] Faculty of Behavioural, Management and Social Sciences, Department of High-Tech Business and Entrepreneurship, Section Industrial Engineering and Business Information Systems, University of Twente, Enschede, the Netherlands

ARTICLE INFO

JEL classification:
G00
G1
G12
G14
G02
G4

Keywords:
Peer-to-Peer-lending
Credit default prediction
Machine Learning
Network centrality

ABSTRACT

Peer-to-Peer (P2P) lending markets have witnessed remarkable growth, revolutionizing the way borrowers and lenders interact. Despite the increasing popularity of P2P lending, it poses significant challenges related to credit risk assessment and default prediction with meaningful implications for financial stability. Traditional credit risk models have been widely employed in the field of P2P lending; however, they may not be capable to capture latent factor information inherent to a loan network based on similarity distances. Thus, in this study we propose an enhanced two-step modeling approach for Machine Learning (ML) that utilizes insights from network analysis and subsequently combines derived network centrality metrics with traditional credit risk factors to improve the prediction accuracy in the credit default prediction process. Through a comparative analysis of three classical ML models with varying degrees of complexity, namely Elastic Net (EN), Random Forest (RF), and Multi-Layer Perceptron (MLP), we showcase novel evidence that the systematic inclusion of network topology features in the credit scoring process can significantly improve the prediction accuracy of the scoring models. Additional robustness tests via the inclusion of randomly shuffled centrality metrics in the analysis, and a further comparison of the graph-based models against a pertinent state-of-the-art credit scoring model in form of XGBoost, further confirm our results. The insights from this study bear valuable conclusions for P2P lending platforms to further improve their scoring systems with graph-enhanced metrics, thereby reducing default risk and facilitating greater access to credit.

1. Introduction

Personal Peer-to-Peer (P2P) lending, in its facet as profit-driven crowdfunding, has progressively emerged as a compelling alternative to the traditional banking system. While P2P lending platforms mirror certain characteristics of traditional banks, they deviate substantially in terms of the financial products offered and their operational mechanisms. For instance, these platforms are mainly characterized through an online business model and chiefly provide medium-term financial solutions, generally accommodating maturities of up to 5 years (Coakley & Huang, 2020).

A distinguishing aspect of P2P lending lies in its ability to democratize access to credit. The traditional banking system, characterized by stringent credit approval processes, often poses barriers for potential borrowers to quickly access funding. In contrast, P2P lending platforms commonly simplify the borrowing process, thereby improving credit access. This disparity can be seen as instrumental for the growing adoption of P2P lending platforms.

The platforms commonly promote the offer of a win–win scenario for both lenders and borrowers. Borrowers gain access to easy-accessible credit, while lenders obtain an opportunity to earn higher returns, establishing a platform for mutual profitability (Malekipirbazari & Aksakalli, 2015). This distinctive attribute, coupled with a rapidly scaling online business character for most P2P lending platforms, has spurred a substantial surge in the issuance of P2P loans in recent years.

However, limited access to traditional credit data for P2P lending platforms, compared to banks, further amplifies the degree of information asymmetry between lenders and borrowers (Duarte, Siegel, & Young, 2012), potentially escalating the default risk. Our study is motivated by an essential difference in the risk ownership of traditional bank lending compared to P2P lending. This difference manifests in the constellation that in traditional lending, banks provide credit scores for borrowers and also bear the risk of the loan defaulting. Hence, the

* Corresponding author at: Bern Business School, Bern University of Applied Sciences, Brückenstrasse 73, 3005 Bern, Switzerland.
E-mail addresses: yiting.liu@utwente.nl (Y. Liu), lennart.baals@bfh.ch (L.J. Baals), joerg.osterrieder@bfh.ch (J. Osterrieder), branka.hadjimisheva@bfh.ch (B. Hadji-Misheva).

https://doi.org/10.1016/j.eswa.2024.124100
Received 14 August 2023; Received in revised form 22 March 2024; Accepted 21 April 2024
Available online 7 May 2024

incentive to provide accurate credit scoring lies in the interest of the bank. Conversely, in P2P lending, the scoring of credit risk is conducted by the P2P lending platform but the default risk of the loan is fully born by the lender that issues the credit (Giudici, Hadji-Misheva, & Spelta, 2019; Havrylchyk & Verdier, 2018). This misalignment can lead to the flaw of inaccurate credit scoring as P2P lending platforms are primarily incentivized to expand credit volume whereas lenders hold primary interest in accurate credit scores. In this realm it becomes clear that P2P lending lacks the traditional mechanisms of credit risk control present in conventional banking systems. Thus, we strive to enrich the ongoing academic discourse around modeling credit scores in P2P lending markets (Chen, Chong, Giudici, & Huang, 2022; Chen, Leu, Huang, Wang, & Takada, 2021; Giudici et al., 2019; Giudici, Hadji-Misheva, & Spelta, 2020a; Lee, Lee, & Sohn, 2021; Liu, Zhang, & Fan, 2022; Lyocsa, Vasanicova, Hadji Misheva, & Vateha, 2022; Niu, Zhang, Liu, & Li, 2020). The majority of the literature on modeling credit default risk in P2P lending markets has focused on traditional credit risk factors without taking into account the intricate network structures of these platforms through network centrality metrics. In this context, it appears crucial to point scholarly attention to the usage of innovative credit risk modeling approaches that could improve the accurate prediction of loan defaults in P2P lending markets. Therefore, in this study we propose an enhanced two-step Machine Learning (ML) process that utilizes network analysis in the first step and consecutively combines network-based centrality metrics with conventional credit risk factors to improve loan default classification.

Our study leverages data from Bondora, a European P2P lending platform, which has been active since 2009 in Estonia, Finland, Spain, and Slovakia. The platform boasts funds from 225,837 individual lenders and has disbursed €867.5 million. in loans.[1] We focus on six network centrality measures: PageRank (Brin & Page, 1998), betweenness centrality (Freeman, 1977; Freeman et al., 2002a), authority score (Kleinberg, 1999), katz centrality (Katz, 1953), hub centrality (Kleinberg, 1999), and closeness centrality (Sabidussi, 1966), which we hypothesize to have a substantial impact on loan default prediction. To evaluate the utility of the centrality measures, we utilize three ML models: Elastic Net (EL) (Zou & Hastie, 2005), Random Forest (RF) (Breiman, 2001), and Multi-Layer Perceptron (MLP) (Goodfellow, Bengio, & Courville, 2016) with varying hyper-parameter and factor configurations to rigorously assess the influence of the network centrality measures on the model prediction accuracy. Subsequently, a feature importance analysis is conducted to determine the individual contribution of the six centrality features within the model performance and compare their relative importance across the three models.

By proposing this enhanced two-step approach to loan default prediction in personal P2P lending, our study aims to make a significant contribution to the field of credit risk modeling. Through the integration of graph theory into our modeling approach, we aim to capture a loan's position and similarity in contrast to other loans within the broader network. Such information can impact the credit risk assessment for several reasons: I. the similarities of various loans captured by the network centrality measures can capture latent features which we do not observe directly, thus adding additional information to the loan default classification. II. how similar a loan contract is with others, can affect the platform's ability to accurately score the respective loan. If we consider the context of a new borrower applying for a loan and the loan's profile is very dissimilar to any of the other loan contracts that the platform has seen, it may raise difficulties for the platform to accurately score the particular loan contract. III. The inclusion of network centrality measures may also provide a reflection of a loan's similarity to the entire loan pool of the P2P lending platform, where higher levels of loan similarity with solvent loans can imply a higher level of collateralized risk. A borrower with a high degree of similarity to previous non-defaulted loans may have a stronger likelihood to repay the loan, thereby potentially lowering the default risk for the lender. This nuanced information would not become available in traditional tools for credit risk assessment. Through our comparative study design we further aim to uniformly assess the effect of incorporating centrality measures into credit scoring models of varying complexity. By choosing three different ML models, belonging to the class of statistical, shallow ML models, and deep learning methods, we aim to provide holistic evidence on the suitability of graph-based credit risk modeling in P2P lending markets. Our findings indicate that the inclusion of network-based features, across all model types, significantly improves the loan default classification of the scoring models. The results remain robust under the utilization of randomly shuffled centrality measures and confirm our initial findings. Insights derived from this study should guide practitioners to improve their credit risk assessment and decision-making processes in P2P lending to grant more stable lending conditions.

The remainder of this paper is organized as follows. Section 2 outlines the existing literature on credit default prediction, the methodical process transition to ML and the role of graph-based network models in risk estimation. We also identify research gaps in the current literature and introduce the foundations of our methodological approach. The methodological framework, presented in Section 3, introduces our proposed two-step ML methodology, embeds it in the concept of graph theory, and outlines the process of network construction. It further describes the statistical and ML models applied in our study and outlines the metrics used for the model evaluation. In Section 4, we describe the dataset used for our analysis, detailing its source, variables, and pre-processing steps. Section 5 presents the empirical findings of our study, including the feature importance analysis and model performance comparison. It further constitutes of the robustness checks and provides implications for P2P lending platforms and borrowers alongside the limitations and potential biases of our study approach. Finally, our conclusion in Section 7 summarizes the main findings of this study and points to potential future research directions.

2. Literature review

2.1. Credit risk and default prediction

One of the major frontiers in modern finance is the quantification of credit risk (Kealhofer, 2003). Since the inception of the famous Black–Scholes model (Black & Scholes, 1973) for pricing equity derivatives, new approaches of modeling inherent default risk in conventional credit have continuously evolved. Following the foundational works of Altman (1968) and Merton (1974), various statistical methods such as Multiple Discriminant Analysis (MDA) (Altman, 1968), Binary Quantile Regression (BQR) (Li & Miu, 2010), probit analysis (Dierkes, Erner, Langer, & Norden, 2013) and logistic regression (Abdou, Pointon, & El-Masry, 2008; Crook, Edelman, & Thomas, 2007; Verbraken, Bravo, Weber, & Baesens, 2014) have been proposed to predict credit default risk. However, with advancements in computer technology these modeling concepts got rivaled by more complex analytical techniques originating from the field of ML, capable to handle large data sets with multiple dimensions. Models like Gradient Boosting Machines (GBM) or Extreme Gradient Boosting (XGBoost) (Li, Ding, Chen, & Yang, 2018; Tian, Xiao, Feng, & Wei, 2020), Adaptive Boosting algorithms (AdaBoost) (Onan, Korukoğlu, & Bulut, 2016), Support Vector Machines (SVM) (Barboza, Kimura, & Altman, 2017; Bellotti & Crook, 2009), and random forest models (Malekipirbazari & Aksakalli, 2015) are found to have yielded superior classification accuracy in estimating credit default over conventional statistical techniques. Similarly, studies have also applied deep learning models like multi-layer perceptrons (MLP) (Angelini, Di Tollo, & Roli, 2008; Battiston, Puliga, Kaushik, Tasca, & Caldarelli, 2012), long-short-term memory (LSTM) (Shen, Zhao, Kou, & Alsaadi,

[1] https://www.bondora.com/en

2021), and probabilistic neural networks (PNN) (Huang, Liu, & Ren, 2018), which have shown considerable accuracy over traditional statistical models in default prediction. Congruently, within the context of our study we place further focus on scholarly work that has specifically explored the application of ML in credit default prediction.

2.2. Credit default prediction using ML

In the realm of traditional banking and corporate lending, several studies have successfully employed ML models for default prediction (Barboza et al., 2017; Bellotti & Crook, 2009; García, Marques, & Sánchez, 2019; Ghatasheh, 2014; Huang, Chen, Hsu, Chen, & Wu, 2004; Lessmann, Baesens, Seow, & Thomas, 2015). Huang et al. (2004) made an early contribution by comparing the performance of different ML models, including ANNs, decision trees, and SVMs, in predicting bond defaults, thereby showcasing that the relative importance of financial input variables differed substantially across the different ML methods. Similarly, Lessmann et al. (2015) conducted an extensive comparison of 41 different classifiers, providing a comprehensive benchmark for credit scoring and modeling approaches. The authors showcased in their comparative analysis that several classifier, specifically heterogeneous ensemble classifiers, predict credit risk significantly more accurate than the industry standard logistic regression model. Bellotti and Crook (2009) investigated SVMs on a large credit dataset and found SVMs to be effective in classifying defaulting credit card customers, requiring a notable number of support vectors for optimal performance. Studies by Barboza et al. (2017) and García et al. (2019) specifically focused on the study of credit risk as a binary classification problem and applied ML classifier models to assess the default prediction accuracy in restrictive conditions such as high variable correlation, outliers, and missing values (Barboza et al., 2017) or a varying distribution of sample types from data (García et al., 2019). Both studies found that ML models can provide improvements in prediction accuracy of scoring models.

Within the literature a considerable strand further stressed the importance of network models in assessing credit risk particularly in the context of complex and interlinked modern financial systems (Allen, Babus, Kleindorfer, & Wind, 2009; Angelini et al., 2008; Battiston et al., 2012). The interconnected nature of these systems implies that an individual entity's failure can precipitate cascading effects on others, underscoring the utility of network models for capturing such dynamics. In the context of credit default prediction for private, corporate and financial entities, scholars applied neural network techniques to estimate credit default in various settings. In a seminal work, Galindo and Tamayo (2000) applied neural networks to predict the likelihood of default in consumer loans. By analyzing a set of socio-economic variables along with traditional financial metrics, their neural network model was able to identify complex interactions that significantly impact the risk of default. Subsequent studies on consumer loan default prediction focused on optimizing the application of neural network techniques for predicting default (Babaev, Savchenko, Tuzhilin, & Umerenkov, 2019; Dastile & Celik, 2021; Iwai, Akiyoshi, & Hamagami, 2020). Other studies as in Angelini et al. (2008) conceptualized feedforward neural networks for predicting loan default in corporate settings and validly classified corporate loan default with enhanced accuracy. Further scholarly work focused on the application of specific deep learning models on corporate loan default prediction such as artificial neural networks (ANN) (Leong, 2016), LSTM (Shen et al., 2021), or PNN (Huang et al., 2018). More recently, studies by Kou et al. (2021) and Poenaru-Olaru, Redi, Hovanesyan, and Wang (2022) considered the utilization of neural network techniques under the specific usage of network-based features to extract additional information from a network structure of financial performance data for small- and medium-sized enterprises (SMEs) to predict firm default.

While these studies have made significant strides, they have primarily been focused on specific alternative or traditional risk factors in the feature engineering process. Recently, there has been a growing recognition for the need to consider additional factors and techniques, specifically when aiming to capture complex nonlinear relationships among the credit features and credit risk (Huang & Kou, 2014). Hence, complex network-based modeling techniques may promise advanced feature engineering to fully mine dependency relationships among the variables, which requires further exploration.

2.3. Graph-based scoring models and their application in P2P lending

The field of network analysis and graph-based modeling has recently shifted its focus on the application of network analysis in the context of credit default prediction. Scholars have been concerned with clarifying the role of network importance and enriching the pool of informative features to more accurately predict credit default. Early studies have documented the important role of relational networks in lending markets. Garmaise and Moskowitz (2003) provide first evidence on the informative role that informal networks play in establishing access to credit finance in the U.S. commercial real estate market. Later studies have emphasized the positive influence of personal networks in the formation of interest rates for loan issuance (Engelberg, Gao, & Parsons, 2012), and decisions on credit allocation in corporate settings from banks to firms (Haselmann, Schoenherr, & Vig, 2018). Stanton, Walden, and Wallace (2018) are among the first to empirically proof the importance of a firm's network position, showcasing that the loan default rates of a firm and its comparative performance are related to the relative position of the respective company against its industry peers within the considered network. Further studies on systemic risk transmission and contagion among financial institutions (Constantin, Peltonen, & Sarlin, 2018; Torri, Giacometti, & Paterlini, 2018) highlight the benefit of structural information in graphs or networks that indicate latent factors irretrievable from conventional modeling approaches.

This evidence points to the influence of network effects within lending networks and consequently raises the case to consider them in the classification of default scenarios for financial entities (Kou et al., 2021; Shi, Qu, Chen, Mi, & Wang, 2024; Sukharev, Shumovskaia, Fedyanin, Panov, & Berestnev, 2020; Yıldırım, Okay, & Özdemir, 2021; Zhou et al., 2023). Kou et al. (2021) utilize network information derived from payment networks of SMEs, where nodes represent firms and edges indicate common payment transactions, to predict corporate default. The gathered topological information proves transactional data to improve SME bankruptcy prediction. In turn, Yıldırım et al. (2021) further advances this network modeling approach by deriving graph centrality metrics and comparatively assessing statistical and tree-based credit scoring models on a data set of Turkish companies. The findings reveal a uniform improvement across all graph-enhanced model ROCs in contrast to their conventional peers. Sukharev et al. (2020) coincide in this finding and report that graph-induced transactional information significantly improves the prediction accuracy of neural networks in classifying loan defaults of bank customers. Conversely, studies focusing on credit default prediction in consumer lending find similar evidence for the utility of graph-topological information in the credit risk modeling process. Zhou et al. (2023) apply graph-attention-based networks to model consumer credit risk under the influence of complex inter-relationships stemming from the credit providers' users. The approach results in superior prediction accuracy for the graph-based model over standard ML techniques. Shi et al. (2024) apply a hybrid graph neural network (GNN) approach combined with k-nearest-neighbor (KNN) to perform the graph transformation in an unsupervised way to enhance loan default prediction. Their superior classification accuracy over conventional ML models emphasis the benefit of constructing links between observed instances prior to the modeling process to further exploit latent information in the credit data.

In the context of P2P lending platforms, the application of advanced ML models under the utilization of network features has been a young

but increasingly recognized research field. Several studies have focused on the prediction of loan default in P2P lending (Ahelegbey, Giudici, & Hadji-Misheva, 2019a, 2019b; Chen et al., 2022, 2021; Giudici et al., 2019, 2020a; Lee et al., 2021; Liu et al., 2022; Lyocsa et al., 2022) but only some (Ahelegbey et al., 2019a, 2019b; Chen et al., 2022; Giudici et al., 2019, 2020a) have leveraged network effects measured by centrality measures to predict loan defaults. Giudici et al. (2019, 2020a) employ graph-based centrality features, namely PageRank and degree centrality in their effort to predict loan defaults in SME-focused P2P lending through similarity networks. The inclusion of the topological variables is found to increase the predictive performance of the methods employed. Ahelegbey et al. (2019a, 2019b) coincide in this matter and apply network-based modeling techniques on an SME loan sample to retrieve latent community information. By utilizing topological features, namely degree centrality in a single value decomposition (SVD) framework, the authors create a factor network-based approach to segment companies into homogeneous clusters using logistic-type models, thereby approving their enhanced default prediction accuracy. Conversely, Chen et al. (2022) are first to investigate the effect of including degree, betweenness, and eigenvector centrality in the credit default modeling process for personal P2P lending, using logit regression. The authors find that the position of a lender within the network, classified by the topological features does positively contribute to the classification of default risk. From this review, we determine the following gap in the literature on credit default prediction in personal P2P lending: (i) no scholarly effort so far has systematically investigated the effect of including network topological features in credit scoring models within a comparative setting to arrive at a uniform conclusion. (ii) The varying complexity of different scoring models has not been accounted for in previous studies, which leaves uncertainty about the concrete effectiveness of including network topological features in the credit default classification process. Hence, we strive out to investigate this conundrum in the following sections of this study.

3. Methodological framework

We describe a two-step modeling approach designed to capture and analyze the complex interactions of loans to predict default probability based on initial credit features and centrality measures under the use of different ML models. Our first step consists of constructing a network based on loans. The development of this network involves systematically representing the loans as nodes and their interactions as edges. From this network, we then extract graph features that effectively summarize the structural characteristics of the loans and their similarities. The second step involves using these graph features and initial features as inputs to train the three ML models. Finally, to ensure that our results are robust and reliable, we make use of several model evaluation metrics. We begin with commonly used ML metrics such as accuracy, precision, recall, and the F1 score. To compare the performance of models based on dataset with and without graph features, we apply the DeLong Test.

3.1. Step 1: Network construction and centrality feature extraction

3.1.1. Process of network construction on loans

Our methodology to construct the network draws on previous work in social network analysis, particularly in the context of financial systems (Battiston, Caldarelli, May, Roukny, & Stiglitz, 2016; Glasserman & Young, 2015; Newman, 2003).

We opt for an undirected weighted network, as the relationship between two loans sharing common attributes, does not involve directionality, while we still need weights to express the distance or similarity among loan contracts. We use vector $X_i = (x_{i1}, x_{i2}, \dots, x_{ip}, \dots, x_{iP})'$ to denote node i, where x_{ip} is the pth feature among the total P features of this loan.

First, a fully connected graph is built. For the fully connected graph, there is an edge between any pair of nodes. We assign weights to every edge by calculating the Gower's distance (Gower, 1971) between two nodes. Gower's distance is a metric that effectively measures dissimilarity between observations for mixed data types, including continuous and categorical features. It normalizes the differences within each feature to a 0–1 scale, allowing meaningful comparison across disparate data ranges. This metric accommodates variables with different scales and is not influenced by the range of features. For each pair of loans i and j, we calculate the Gower's distance d_{ij} as follows:

$$d_{ij} = w_{ij} = \sum_{p=1}^{P} \frac{1}{P} \times \frac{d_{ij}^{p}}{\max(x_{\cdot p}) - \min(x_{\cdot p})},$$

$$\text{where } d_{ij}^{p} = \begin{cases} |x_{ip} - x_{jp}| & \text{if } x_p \text{ is a continuous variable,} \\ 1 & \text{if } x_p \text{ is a categorical variable and } x_{ip} \neq x_{jp}, \\ 0 & \text{if } x_p \text{ is a categorical variable and } x_{ip} = x_{jp}. \end{cases}$$

We then reduce this fully connected graph to its MST (Kruskal, 1956; Prim, 1957) and extract the graph. In the fully connected graph, all graph features will be the same across all loan contracts, while with the MST, we take the most connected sub-graph and extract information efficiently (Giudici, Hadji-Misheva, & Spelta, 2020b).

3.1.2. Network centrality features used in the analysis

Our analysis of loan similarity within the constructed network was based on several centrality measures, each contributing to a comprehensive understanding of borrower behavior and influence in the network. Here we assume there are N nodes in the graph and detail each centrality measure from a generic perspective:

3.1.2.1. Degree centrality ($C_D(X_i)$). This metric quantifies the number of direct connections a node has in the network. It encapsulates the immediate risk exposure or influence of a given node. A node with high degree centrality is typically heavily involved in the network's interactions, implying a more significant role or potential vulnerability. Mathematically, it is expressed as $C_D(X_i) = \sum_{j=1, j\neq i}^{N} a_{ij}$, where a_{ij} denotes the adjacency between node X_i and X_j (Freeman et al., 2002b).

3.1.2.2. Closeness centrality ($C(X_i)$). This metric calculates the average length of the shortest paths to reach all other nodes in the network from a given node. It captures how quickly information can propagate from a given node to others. Formally, it is represented as $C(X_i) = \frac{1}{\sum_{j=1, j\neq i}^{N} d(X_i, X_j)}$, where $d(X_i, X_j)$ represents the shortest-path distance from X_i to X_j (Freeman et al., 2002b).

3.1.2.3. Betweenness centrality ($C_B(X_i)$). This measure captures the influence of a node over the flow of information between other nodes in the network.

It is computed as $C_B(X_i) = \sum_{j,k\in\{1,2,\dots,N\}} \frac{\sigma(X_j, X_k|X_i)}{\sigma(X_j, X_k)}$, where $\sigma(X_j, X_k)$ is the total number of shortest paths from node X_j to node X_k, and $\sigma(X_j, X_k|X_i)$ is the number of those paths passing through node X_i (Freeman, 1977).

3.1.2.4. PageRank ($PR(X_i)$). Our PageRank variable evaluates the importance of nodes based on the quality of incoming links. It is defined as $PR(X_i) = (1-d) + d\sum_{X_j\in M(X_i)} \frac{PR(X_j)}{L(X_j)}$, where $M(X_i)$ is the set of pages that link to X_i, $L(X_j)$ is the number of outbound links on page X_j, and d is a damping factor, usually set to 0.85 (Brin & Page, 1998).

3.1.2.5. Katz centrality ($C_{Katz}(X_i)$). Katz Centrality considers both direct and indirect influence of a node's neighbors. It is represented as $C_{\text{Katz}}(X_i) = \sum_{j=1}^{N} \beta A_{ij} C_{\text{Katz}}(X_j) + \alpha$, where A_{ij} denotes the adjacency matrix element, β is a scaling factor, and α is a constant term representing the node's intrinsic centrality (Katz, 1953).

3.1.2.6. HITS algorithm (authority score $a(X_i)$ and hub score $h(X_i)$). The HITS (Hyperlink-Induced Topic Search) calculates the authority and hub scores for a node based on its incoming and outgoing links, respectively. The Authority Score $a(X_i)$ is computed as the sum of the hub scores of each node X_j that points to X_i, i.e., $a(X_i) = \sum_{X_j \in M(X_i)} h(X_j)$, where $M(X_i)$ is the set of nodes that point to X_i and $h(X_j)$ is the hub score of node X_j. Similarly, the Hub Score $h(X_i)$ of a node X_i is computed as the sum of the authority scores of each node X_j that X_i points to, i.e., $h(X_i) = \sum_{X_j \in N(X_i)} a(X_j)$, where $N(X_i)$ is the set of nodes that X_i points to and $a(X_j)$ is the authority score of node X_j (Kleinberg, 1999).

3.2. Step 2: ML models

For the second step of our methodology, we introduce several ML techniques. We employ EN, RF, and MLP models due to their reliant estimation and processing power as well as convenient interpretability. A grid search algorithm is further employed to fine-tune the hyperparameters of each model. In the following subsection, we will briefly introduce each model specification and focus on its hyperparameter settings. For the process of the model application, we use the H2O platform (H2O.ai, 2017) as fully integrated estimation tool to train the models.

3.2.1. Elastic Net (EN)

We apply the EN,[2] facilitated by a Generalized Linear Estimator (GLM), as initial ML model. The EN, originally proposed by Zou and Hastie (2005), is an advantageous regularization and variable selection method, integrating the strengths of both L1 and L2 penalties, synonymous with Lasso and Ridge regression respectively. We opt to implement the EN model over the standard logistic regression because the modeling technique is found to perform better in two-class classification tasks with any consistent loss function (Zou & Hastie, 2005). Furthermore, the EN conducts automatic feature selection with the ability to perform grouping, in contrast to penalized logistic regression that applies either univariate ranking or recursive feature elimination (Zhu & Hastie, 2004) to reduce the number of features in the final model. Hence, the EN provides an effective balance between bias and variance, addressing potential overfitting issues and ensuring model generalizability. Particularly for our dataset, it also aids in the recognition of key predictive features linked to loan default.

The α parameter governs the mixing proportion of the L1 and L2 penalties, with values ranging from 0 (only Ridge penalty) to 1 (only Lasso penalty). As such, an α hyperparameter space of $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ is defined. Concurrently, the λ controls the overall strength of the penalty, with a larger λ leading to more regularization and feature selection. We specify a diverse range of lambda values to cover different degrees of regularization, setting $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. This comprehensive grid search of hyperparameters aids in identifying the optimal model specification with regards to our credit risk data.

The model estimation is executed under a binomial family setting to suit the binary nature of our loan default outcome.

3.2.2. Random Forest (RF)

RFs, introduced by Breiman (2001), have emerged as a robust and versatile ML method with excellent performance across a diverse range of applications. The non-parametric nature of these models makes them exceptionally capable of handling high-dimensional spaces and intricate interactions. In the realm of credit risk modeling, the RF algorithm is particularly effective, given its ability to model complex non-linear relationships between predictors and the probability of loan default, thereby making it a potent tool for default prediction (Lessmann et al., 2015).

In this study, we employ a random forest model[3] using a distributed and paralleled implementation of the RF algorithm for maximum efficiency. The model is configured with varying hyperparameters - `ntrees` and `max.depth`. `ntrees` defines the number of trees in the forest, with range $\{50, 100, 150, 200, 250\}$. `max.depth` stipulates the maximum depth of each tree in the forest, with range $\{5, 10, 15, 20\}$.

Table 1
Confusion matrix.

Actual	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

3.2.3. Multi-Layer Perceptron (MLP)

To further account for the complexity and the latent interconnectedness of credit data, we employ a deep learning approach in the form of a MLP. Deep learning methods enable models to automatically learn representations of data through the use of neural networks with multiple layers (LeCun, Bengio, & Hinton, 2015). MLP, a specific type of deep feed-forward ANN, consists of multiple layers of nodes: an input layer, one or multiple hidden layers, and an output layer. In line with the predictive power of deep learning models reported in the literature (Sadhwani, Giesecke, & Sirignano, 2021), we anticipate that the MLP model, through its ability to model intricate structures in high-dimensional data, could offer meaningful insights and improved prediction accuracy in the context of credit risk modeling.

For our MLP model,[4] the input layer is adjusted according to the number of features (P) of each observation to reach at a different number of neurons. The output can only have one neuron with softmax (Rumelhart, Hinton, Williams, et al., 1985) as activation function, outputting a value between 0 and 1 as the probability of default. We focus on `hidden` layers with range $\{[50, 50], [100, 100], [200, 200], \}$ $\{[100, 200, 100], [200, 100, 200]\}$. The count of individual numbers in each square bracket represent the number of hidden layers, and the value of each number represents the amount of neurons in this layer. For instance, $[100, 200, 100]$ indicates that this MLP contains an input layer, three hidden layer and an output layer. There are $100, 200, 100$ neurons for the 1st, 2nd and 3rd hidden layer, respectively. The hyperparameter `epochs` controls the time of the weights to be updated to minimize the loss function, with range $\{10, 50, 100\}$.

3.3. Model evaluation

3.3.1. General machine learning performance metrics

In this section we outline the main evaluation metrics used in this study to measure the performance of the EN, RF and MLP. In this binary classification task, we have the Table 1 confusion matrix (Gupta, Kose, Khanna, & Balas, 2022):

Several additional metrics at the forefront accuracy, precision, recall, and F1-score are employed. Each method is explained below:

We utilize accuracy to assess the performance of a classification model. It measures the proportion of total correct predictions (both positive and negative) out of all predictions made. Mathematically, accuracy is calculated as the sum of True Positives (TP) and True Negatives (TN) divided by the sum of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$.

[2] For detailed elaborations on the technical notation of the EN please refer to Appendix A.

[3] For detailed elaborations on the technical notation of the RF please refer to Appendix B.

[4] For detailed elaborations on the technical notation of the MLP model please refer to Appendix C.

Precision (also known as positive predictive value) measures the proportion of true positive predictions among all positive predictions. It is calculated as: Precision = $\frac{TP}{TP + FP}$.

Recall or sensitivity measures the proportion of true positive predictions among all actual positives and is calculated as: Recall = $\frac{TP}{TP + FN}$.

The F1-score is the harmonic mean of precision and recall, acting as a balance of these two metrics. It is particularly useful when dealing with imbalanced data and conversely computed as: F1-score = $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

These metrics can help assess each model's performance more accurately, particularly in cases where the original class distribution is uneven (Powers, 2020). Each of these metrics offers a distinct perspective on model performance and allows us to draw a collective assessment of the models' predictive capabilities in this study.

In addition, we use the Area Under the Receiver Operating Characteristic curve (AUC-ROC) (Fawcett, 2006) as additional measure to visually compare the model performances. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, while the AUC quantifies the overall performance of the classifier. An exemplary AUC of 0.5 suggests no discrimination, i.e., the model has no ability to distinguish between the positive and negative classes. Conversely, an AUC of 1.0 signifies perfect discrimination.

3.3.2. Statistical performance metrics

Furthermore, we utilize two statistical metrics, namely Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to measure the average magnitude of errors from our models. We define the MSE as: MSE = $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$, where y_i are the observed values and $\hat{y}_i$ are the predicted values. Conversely, we define the RMSE as the square root of the mean squared error indicated as: RMSE = $\sqrt{\text{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$.

3.3.3. DeLong test

In this subsection we also introduce the DeLong Test (DeLong, DeLong, & Clarke-Pearson, 1988; Sun & Xu, 2014), which serves as a statistical method for comparing the areas under two or more correlated ROC curves. We will utilize this test to compare the prediction performance of the different ML models in our study. This approach recognizes and accounts for the correlated nature of the data, specifically when tests are performed on identical individuals.

In the DeLong test, let AUC_1 and AUC_2 represent the areas under the ROC curves of the two predictive models under comparison. The difference between these areas is denoted by $\Delta = AUC_1 - AUC_2$. We apply the two-side hypothesis test:

$$H_0: \quad \Delta = AUC_1 - AUC_2 = 0,$$
$$H_1: \quad \Delta = AUC_1 - AUC_2 \neq 0.$$

The null hypothesis, H_0, suggests no difference in the AUCs of the two models. The alternative hypothesis proposes a significant difference. Under the null hypothesis H_0, DeLong et al. (1988) showed that the test statistic follows a standard normal distribution, or Z-distribution, allowing us to employ a Z-test to ascertain the statistical significance of the observed difference Δ. The test statistic is calculated as $Z = \frac{\Delta}{SE(\Delta)}$, where $SE(\Delta)$ represents the standard error of the difference in AUCs. The standard error $SE(\Delta)$ is estimated using the formula:

$$SE(\Delta) = \sqrt{SE_1^2 + SE_2^2 - 2 \cdot COV},$$

where SE_1 and SE_2 represent the standard errors of AUC_1 and AUC_2, respectively, and COV is their covariance. These quantities are derived from the estimated variance–covariance matrix of the AUCs, as detailed by DeLong et al. (1988). With the $SE(\Delta)$ estimated, it can be used in the Z-test to determine the statistical significance of the observed difference Δ.

3.3.4. Feature importance

Subsequently, we introduce the assessment method how we evaluate the feature importance for each model based on the training data set.

For the EN model, feature importance is calculated based on the magnitude of the estimated coefficients. The larger the absolute value of the coefficient, the more important the corresponding feature is considered. Features with zero coefficient, which were eliminated by the L1 penalty, are considered least important.

For the RF model, feature importance is computed using the Mean Decrease Impurity (MDI) method (Breiman, 2001). This method calculates the total decrease in node impurities from splitting on the variable, averaged over all trees. The impurities are measured by the Gini index for classification or variance for regression. Variables that are used at the top of the tree contribute to the final prediction of a larger fraction of the input samples, and will therefore have a higher importance value.

As for the MLP model, assessing feature importance can be quite intricate due to the complexity and non-linearity of deep neural networks. For this study, we utilize the Gedeon method (Gedeon, 1997), which measures the importance of an input by calculating the sum of the products of the weights and the derivative of the activation function, taking into account both the input-hidden layer and hidden-output layer connections. Thus, this method considers the overall network architecture and the interaction effects between different features. It offers an insightful way to comprehend the contribution of each feature within the context of a neural network model.

To make the values of the feature importance comparable within and among models, we calculate and apply the relative feature importance to $[0, 1]$ by dividing the importance of each feature by the importance of the most important feature.

3.4. Overview of the data modeling workflow

In this subsection, we provide an overview of the data modeling workflow employed in our study. The data modeling workflow can be summarized by Fig. 1.

Initial data cleaning serves as the first step of our workflow. At this stage, we remove columns identified as irrelevant to the default status of a loan, such as *DateOfBirth*. We adjust variables to correct formats, that is, continues values, categorical values, dates and strings. We delete a list of forward-looking biased variables that cannot be known prior to the target variable and drop these alongside duplicates. We exclude rows containing missing values. Furthermore, columns exhibiting high correlation with other variables, as indicated by the Variance Inflation Factor (VIF), are dropped to reduce multicollinearity. Most importantly, we create the binary label, *default*, which each model will predict. The definition of *default* is:

$$y = default = \begin{cases} 1 & \text{if } DefaultDate \text{ is not null,} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

which means we regard a loan as defaulted if an interest is not paid on time before the data download date. Otherwise, we regard the loan as non-defaulted.

Afterwards we perform a sampling procedure to handle the class imbalance in our binary classification task. In the cleaned dataset, there are 12, 228 default loans and 20, 241 non-default loans. As supervised models usually perform better on a balanced dataset (Jing et al., 2019), we randomly sample 12, 000 defaulted loans and 12, 000 non-defaulted loans. We then construct a network, based on the balanced sample of 24, 000 loans, and calculate graph features by applying the methodology introduced in Section 3.1.1.

Our dataset is then split into training, validation, and testing sets at a ratio of $0.6 : 0.2 : 0.2$. A feature selection procedure is carried out on the training set, removing features where most observations share the same value, as well as graph features that cannot be calculated on the graph in the context of our study.

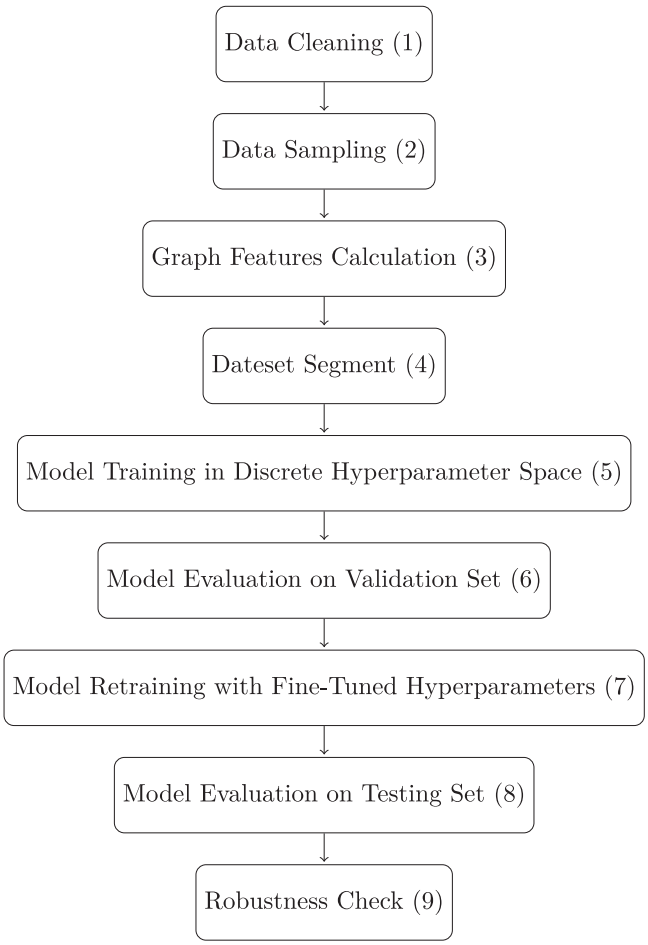


Fig. 1. The data modeling workflow.

Models of three types (EN, RF, MLP) under all respective possible combinations of hyperparameters introduced in Section 3.2 are subsequently trained on the training set. This is done for two groups of models: the first group is trained only on the initial features, and the second group is trained on both the initial and graph features.

We then separately perform hyperparameter tuning on both model groups using the validation set. The best performing combination of hyperparameters for each model type in each group is afterwards selected based on their performance on the corresponding validation set. According to the model performance on the validation set, we also delete unimportant features in the dataset. Afterwards we train the models in both groups under the best performing combination of hyperparameters. This is also done on the training set excluding unimportant features. Then we test the models in both groups on the testing set and report the performance.

Finally, we conduct a robustness check by shuffling the graph features and repeating the steps outlined above. If the performance difference between the two model groups disappears, we conclude that this signifies the observed performance improvement to be indeed attributable to the network centrality features.

4. Data

In this analysis, we employ a holistic dataset acquired from Bondora, a prominent European P2P lending platform based in Estonia, in order to explore the dynamics of loan default. This comprehensive dataset encompasses 32,469 individual borrowers, each of which is detailed through a combination of 155 categorical and continuous variables. This set of variables provides a multi-dimensional perspective into the complex factors that could influence loan default risk. Lending on Bondora is characterized by a diversity of borrowers and lenders, a trait that is inherently captured in our dataset. Borrowers on the platform are largely individuals seeking personal loans, and they encompass a wide array of demographics, credit histories, and borrowing needs.

4.1. Dataset characteristics

All credit features provide valuable context about the borrower's past interactions with the credit market, and by extension, her potential risk profile.[5]

Based on an initial screening process via histograms of each credit feature within our dataset, we are able to filter and classify a range of variables that we assume to have a stronger influence on the default classification of an individual loan.[6] Table 2 provides the summary statistics that disclose the lower moments of these variables in the original dataset. In total, the dataset contains 12,228 defaulted and 20,241 non-defaulted loans.

From Table 2, we observe that most of the variables lie within similar ranges.

Nonetheless, individual borrowers in our sample, are displaying a wide range of financial backgrounds and borrowing behaviors. For instance, the total income ('inc.total') varies considerably, as shown by its standard deviation of 0.53 around the mean of 6.98. Similarly, there is significant variation in the number of previous loans ('No. Prev. Loans'), with some borrowers having as many as 26 previous loans while others have none.

Furthermore, it is noteworthy that borrower groupings appear diverse, not limited to any particular age group or credit background. However, the dataset is unbalanced, specifically in the context of our target variable, the individual loan default status. This can pose challenges for statistical learning and prediction accuracy, which we will address in Section 3 of this article.

Finally, the average interest rate across all loans is 17.76%, with a standard deviation of 5.39%, indicating a considerable degree of variation in the cost of borrowing. These characteristics underscore the heterogeneity of our sample and sheds light on the complexity of predicting loan outcomes.

In order to arrive at a balanced dataset, containing equal numbers of defaulted versus non-defaulted loans, we subsequently draw a randomized sub-sample of 24,000 observations from the original dataset. Table 3 shows the summary statistics for this particular sub-sample.

4.2. Description of the six network-based centrality features

In addition to the standard variables typically found in P2P lending datasets, our dataset incorporates network-related measures, reflecting our objective to integrate graph theory into credit risk modeling. In Table 4 we present a detailed table of the summary statistics for the computed centrality measures:

Upon analyzing the computed network centrality features, we find in Table 4 that the mean values for most of these measures are close to zero, thereby indicating that the average loan contract exhibits relatively low centrality within the network. This finding underscores a lower influence and prominence of individual borrowers within the broader loan network, thus showcasing the rather decentralized nature of the P2P loan network in our sample.

[5] A detailed description of the informative features used for the analysis after the pre-processing of the data sample can be found in Appendix F.

[6] The detailed histograms of the most informative loan features of the Bondora dataset can be found in Appendix E.

Table 2
Descriptive statistics for the top 9 informative loan features on the cleaned and unbalanced dataset.

	Count	Mean	Std. Dev.	Min	Max
liab.l	32 469	5.11	2.08	0.00	10.48
inc.total	32 469	6.98	0.53	0.79	13.09
Monthly Payment	32 469	4.03	0.97	0.00	7.55
log. amount	32 469	7.28	0.92	4.66	9.27
time	32 469	3.55	0.64	1.84	4.80
Interest	32 469	17.76	5.39	7.27	38.00
Amt. Prev. Loans Bef. Loan	32 469	5317.28	6034.19	0.00	74 740.00
No. Prev. Loans	32 469	2.71	3.02	0.00	26.00
Age	32 469	39.63	11.46	18.00	70.00

Table 3
Descriptive statistics for the top 9 most informative loan features.

	Count	Mean	Std. Dev.	Min	Max
liab.l	24 000	5.13	2.05	0.00	10.48
inc.total	24 000	6.96	0.52	0.79	13.09
Monthly Payment	24 000	4.01	0.98	0.00	7.55
log. amount	24 000	7.29	0.92	4.66	9.27
time	24 000	3.53	0.63	1.84	4.80
Interest	24 000	17.72	5.37	7.29	38.00
Amt. Prev. Loans Bef. Loan	24 000	5347.45	5914.89	0.00	74 740.00
No. Prev. Loans	24 000	2.75	3.03	0.00	26.00
Age	24 000	39.79	11.59	18.00	70.00

Table 4
Descriptive statistics for the network centrality measures.

	Count	Mean	Std. Dev.	Min	Max
PageRank	24 000	0.000042	0.000026	0.000023	0.000359
betweenness	24 000	2.589516e−09	8.100456e−09	0.000000	2.153047e−07
closeness	15 221	26.07	15.44	3.76	663.01
katz	24 000	0.01	0.00	0.01	0.01
authority	24 000	0.00	0.01	0.00	0.93
hub	24 000	0.00	0.01	0.00	0.33

5. Results

5.1. Best performing combination of hyperparameters for both groups

In this subsection, we elaborate on the hyperparameter combination that results in the optimal performance of our models on the validation set, as delineated in Step 6 of Fig. 1. The resulting fine-tuned hyperparameter set, displayed in 5, has been subsequently employed for the retraining of the model, as outlined in Step 7 of the same figure. We report the value for all three model types in two groups: Group 1 - represents models trained on the training set without centrality measures; Group 2 - represents models trained on the training set with centrality measures.

Table 5
Best performing combination of hyperparameters.

Model	Hyperparameter	Group 1	Group 2
EN	Family	binomial	binomial
	Link	logit	logit
	α	0.2	0
	λ	0.0001	0.0001
RF	`ntrees`	250	250
	Min Depth	20	20
	`max_depth`	20	20
	Min Leaves	430	659
	Max Leaves	2609	2463
	Mean Leaves	2066.14	2090.48
MLP	`hidden`	[200, 200]	[200, 100, 200]
	`epoch`	20	20
	`max_depth`	10	10

5.2. Model performance comparison

To thoroughly assess the efficacy of our three models, EN, RF, and MLP, we employ a diverse set of evaluation metrics, as detailed in Section 3.3. This comprehensive evaluation is conducted on the testing set, corresponding to Step 8 in Fig. 1, for both model group 1 and group 2.

5.2.1. Model performance metrics

Tables 6 and 7 encapsulate the comparative performance metrics of the fine-tuned EN, RF, and MLP models on the test dataset. These tables are structured to present a clear differentiation of the model effectiveness with respect to standard metrics including the confusion matrix, accuracy, precision, recall, F1-score, AUC, MSE, and RMSE. We should notice here that the prediction of all models is the probability of default, a continues value between 0 and 1. Thus, the calculation of confusion matrix, accuracy, precision, recall and F1-score is affected by the choice of the threshold, while the calculation of AUC, MSE, and RMSE is not. In this subsection, we report the confusion matrix, accuracy, precision, recall and F1-score on the testing set when the threshold to maximize the F1-score on training set is applied.

In an intra-model comparison, the RF model distinguishes itself with the highest recall values in both groups, suggesting its robustness in identifying positive cases. The EN model, despite not outperforming the other two models, achieves a reliant level of predictive power. This is particularly notable given its simple and linear modeling approach. The MLP model presents a more balanced profile between precision and recall, leading to competitive F1-scores.

The inter-group comparison highlights the impact of incorporating centrality measures into the training process. All models in Group 2, which include these measures, demonstrate enhancements in accuracy, precision, and F1-scores compared to their counterparts in Group 1.

Table 6
Confusion matrix for three types of models in two groups.

Model		Group 1				Group 2			
		0	**1**	**Error**	**Rate**	**0**	**1**	**Error**	**Rate**
EN	**0**	1004	1364	0.576	(1364.0/2368.0)	1586	782	0.3302	(782.0/2368.0)
	1	311	2121	0.1279	(311.0/2432.0)	423	2009	0.1739	(423.0/2432.0)
	Total	1315	3485	0.349	(1675.0/4800.0))	2009	2791	0.251	(1205.0/4800.0)
RF	**0**	978	1390	0.587	(1390.0/2368.0)	1839	529	0.2234	(529.0/2368.0)
	1	212	2220	0.0872	(212.0/2432.0)	492	1940	0.2023	(492.0/2432.0)
	Total	1190	3610	0.3337	(1602.0/4800.0)	2331	2469	0.2127	(1021.0/4800.0)
MLP	**0**	1151	1217	0.5139	(1217.0/2368.0)	1731	637	0.269	(637.0/2368.0)
	1	332	2100	0.1365	(332.0/2432.0))	427	2005	0.1756	(427.0/2432.0)
	Total	1483	3317	0.3227	(1549.0/4800.0)	2158	2642	0.2217	(1064.0/4800.0)

Table 7
Model performance metrics.

Model	Performance measure	Group 1	Group 2
EN	Accuracy	0.65	0.75
	Precision	0.61	0.72
	Recall	0.87	0.83
	F1-score	0.72	0.77
	AUC	0.72	0.84
	MSE	0.21	0.16
	RMSE	0.46	0.40
RF	Accuracy	0.67	0.79
	Precision	0.61	0.79
	Recall	0.91	0.80
	F1-score	0.73	0.79
	AUC	0.76	0.88
	MSE	0.20	0.15
	RMSE	0.45	0.39
MLP	Accuracy	0.68	0.78
	Precision	0.63	0.76
	Recall	0.86	0.82
	F1-score	0.73	0.79
	AUC	0.75	0.86
	MSE	0.22	0.16
	RMSE	0.47	0.40

This suggests that the integration of centrality measures bolsters model prediction capabilities. Notably, the RF model benefits substantially in precision and AUC. Additionally, reductions in MSE and RMSE across all models in Group 2 affirm the positive influence of centrality measures on the model performance. On the other hand, while observing a decline in recall for all models in Group 2, it is important to recognize that this is a consequence of the threshold selection strategy aimed at optimizing the overall model performance. Subsequent sections will demonstrate that, despite the lower recall, Group 2 models exhibit superior ROC and AUC results compared to Group 1. The advanced ROC and AUC outcomes provide Group 2 models with the latitude to adjust the threshold, facilitating a tailored balance between false positives and false negatives. This flexibility ensures that the performance enhancements are not merely nominal but translate into pragmatic gains in predictive accuracy.

5.3. ROC, AUC and DeLong test

Fig. 2 shows the ROC curve and results of the DeLong test for all three model types in two groups.

Across all three models within Group 2, we find an elevation in performance that is consistently observed across the entire spectrum of threshold values. Correspondingly, the AUC metrics for these models surpass those of their counterparts in Group 1. The DeLong test corroborates the enhancements, confirming their statistical significance.

5.4. A comprehensive comparison of model performance

To further corroborate on these findings, Fig. 3 presents a detailed comparison of model performances, illustrating the significance of centrality features. The ROC curves in the upper-left corner represent models trained exclusively on initial credit features, whereas the ROC curves in the upper-right corner depict models trained solely on centrality features. It is observed that models relying only on centrality features achieve a similar level of predictive capability as those based solely on initial credit features. Combining both feature groups significantly enhances the performance across all model types. The empirical analysis demonstrates that models trained on a unified dataset comprising of both initial credit and centrality features outperform those confined to individual feature sets. This finding substantiates the assertion that the integration of initial credit and centrality features is instrumental for enhancing predictive performance.

Additionally, the final ROC curve plot includes the performance of an XGBoost model with fine-tuned hyperparameters on the testing set, demonstrating that all three model types, incorporating centrality measures, outperform a contingent ML model. This thorough comparison confirms that the introduced centrality measures contribute positively to predictive accuracy. The improvement of model prediction caused by the centrality features is common to all model types.

5.5. Feature importance analysis

Our investigation is centered on the evaluation of feature importance, particularly focusing on network centrality measures such as PageRank, betweenness, authoriy- and hub-centrality, katz, and closeness. We employ the methodology discussed in Section 3 to elaborate

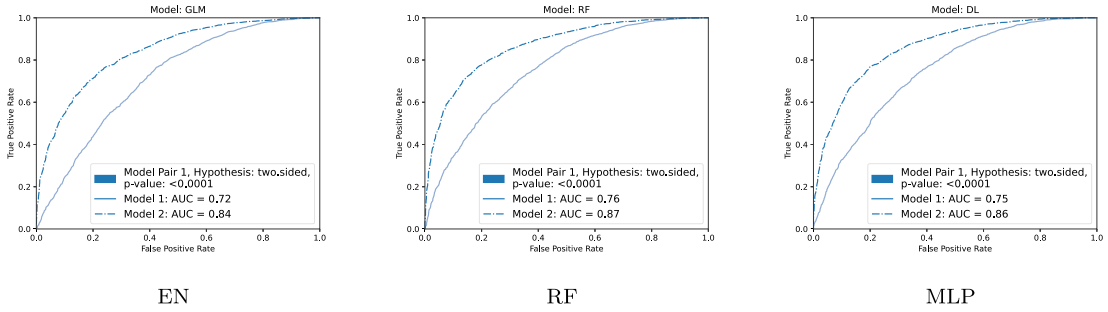


Fig. 2. ROC curves for all models.

Models trained on only initial credit features

Model	Features in Training Set	AUC
EN	Initial credit features	0.72
RF	Initial credit features	0.76
MLP	Initial credit features	0.75

Models trained on only centrality features

Model	Features in Training Set	AUC
EN	Centrality features	0.77
RF	Centrality features	0.78
MLP	Centrality features	0.78

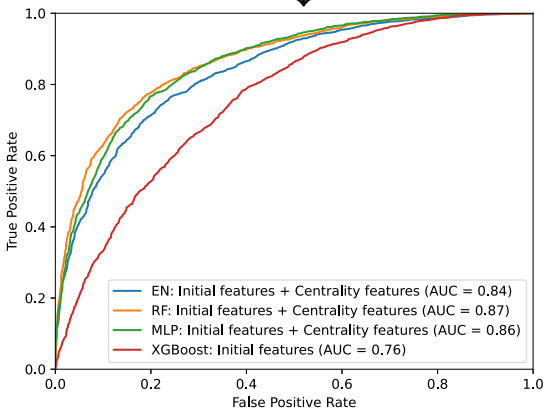


Fig. 3. A comprehensive comparison.

on the significance of the network-based features in credit risk prediction. The following sections will discuss the role of the network centrality features in influencing the performance of our EN, RF, and MLP models, especially from the feature importance perspective. Fig. 2 presents the feature importance for all three model types across two groups. Models within Group 1 lack values for centrality features. Nonetheless, for comparative convenience, we continue to enumerate all significant features along the x-axis.

The comparison across these three models yields several important observations. While the PageRank feature consistently emerges as an influential predictor across all models, other features such as betweenness show varying levels of importance depending on the specific model type. This variability in feature importance across different models indicates the complex interplay between the data characteristics and model architectures (see Fig. 4). It is imperative to reconcile that while these findings shed light on the relative contributions of the various network measures to the model's prediction capability, they do not imply causality (Hastie, Tibshirani, & Friedman, 2009). They merely reflect the relationships within the given dataset and the specific models used (James, Witten, Hastie, & Tibshirani, 2013).

The contrasting ranking of the network-based features across the three ML models underscores the inherent differences in how these algorithms capture and interpret the structure of the data. The EN model, being a linear model, tends to emphasize the direct linear effects of variables on the outcome. It assigns importance based on the degree to which a feature contributes to reducing the prediction error, under the constraints imposed by its regularization term (Zou & Hastie, 2005). Hence, features like PageRank and closeness centrality, which may exhibit a more prominent linear relationship with the outcome, are assigned a higher importance.

In contrast, tree-based methods like RF inherently account for higher-order interactions and non-linear relationships between variables (Breiman, 2001). This allows them to highlight importance not only based on direct effects but also due to the structural influence a feature might have in relation to other features. Therefore, it is immanent to see a broader set of centrality measures like PageRank, closeness, and katz receiving high importance.

Similarly to non-parametric tree-based models, neural networks like the MLP model are known for their capability to capture complex non-linear interactions and high-dimensional relationships, attributing importance through the weights learned across multiple layers of the network (Hastie et al., 2009). Consequently, we observe a wider set of network-based features like betweenness, PageRank, katz, and closeness being emphasized.

5.6. Robustness checks

In order to control for potential deficiencies in our model specifications and provide robust model estimates, we added six shuffled centrality features to the data sample and model training process that consisted of randomly rearranged observations of the actually computed network centrality measures. The structure of the shuffled

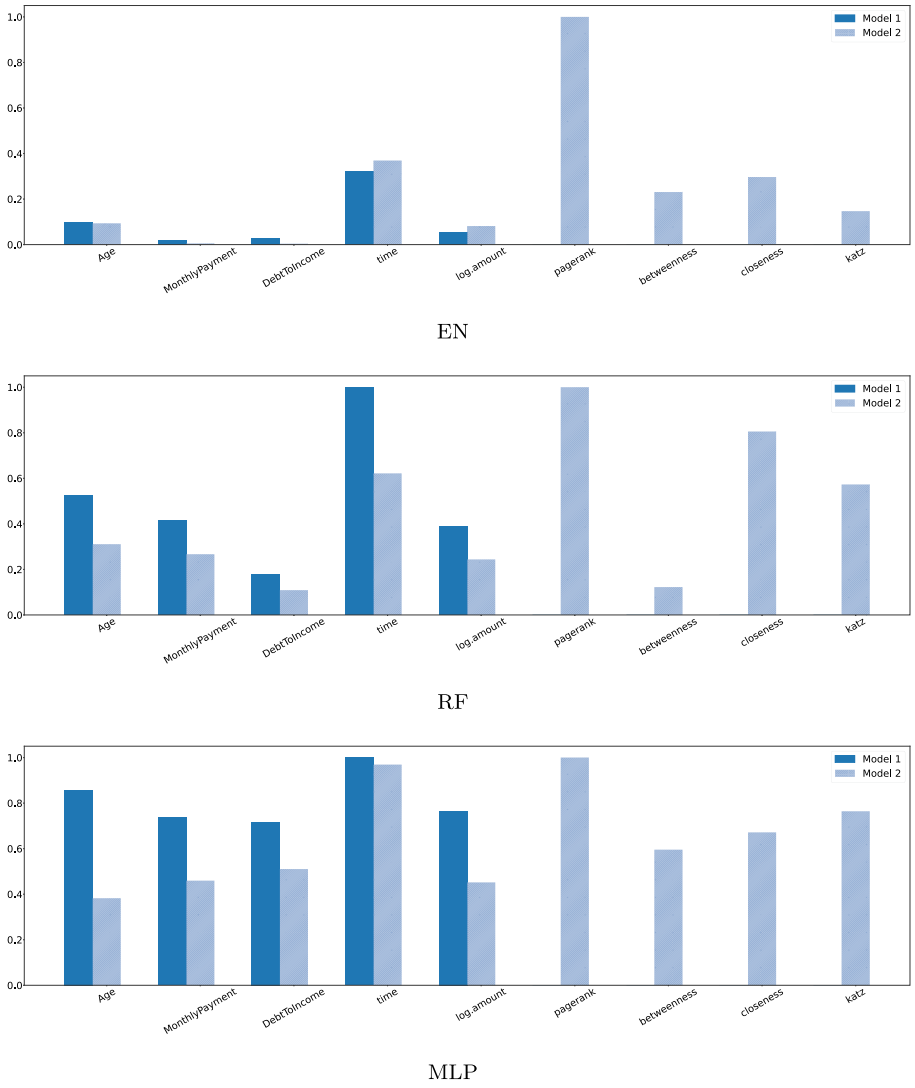


Fig. 4. Feature importance for all models with and without graph-based features.

features was randomly rearranged to eliminate any dependencies and should consecutively represent white noise in the estimation of the model predictions. Consequently, the inclusion of the shuffled centrality features should not be anymore helpful in predicting loan defaults than a randomly drawn i.i.d sequence (Dimpfl & Peter, 2018). As introduced in Section 3 we additionally trained each model on datasets that only included the conventional credit features in combination with the randomized network centrality features to assess the model performance under the corresponding feature selection. Hitherto, by *apriori* assumption none of the models, trained under the inclusion of the randomized centrality features should detect any meaningful feature importance in the default classification process.

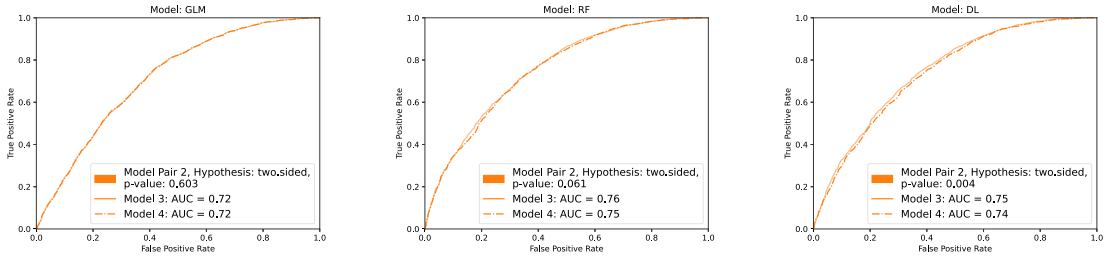
As can be seen in the respective plots (Fig. 5) of the ROC and AUC metrics of each ML model type in the corresponding feature configuration with and without randomized network centrality features, the inclusion of the latter to the credit features does not improve the predictive accuracy of any of the used credits scoring models.

Although the respective plot for the linear EN model does not appear to be statistically significant, we can still infer that the randomized network centrality features provide no additional explanatory power for the classification of loan defaults in any of the ML models.

6. Discussion

6.1. Result analysis and discussion of the centrality measures

From our analysis in Section 5 it is possible to observe that the inclusion of the graph-based centrality measures lead to a uniform improvement in prediction accuracy across all tested scoring models. We outline the rationale for such findings in the following: (I) Path-based and eigenvector-based measures like betweenness, Katz, and PageRank, measuring global node importance, perform exceptionally well due to their ability in detecting similar loan clusters within the network. PageRank expresses the highest uniform importance across all three model-types and is by definition a measure of node importance that is capable to capture loans that group in clusters of similar risk profiles. Similarly, a high betweenness centrality is signaling that a respective loan acts as a bridge to multiple loan clusters with dissimilar risk, thus yielding information about transitional risk profiles. This information is benefiting, particularly complex non-linear models like the MLP, in being able to better classify defaulted loans from non-defaulted counterparts. Contrarily, Katz centrality allows through its conceptualization to classify the indirect influence of a loan in contrast to the entirety of loans in the similarity network. In addition



EN RF MLP

Fig. 5. ROC curves for all models with shuffled centrality features.

to capturing cluster information on loans with similar credit profiles, Katz centrality also measures the immediate connection of an individual loan to all other loans in the network with diminishing weights added to loans with more distant connections. This information can signify common risk patterns on broader structural risk within the network for specific credit profiles, from which the scoring models, particularly models with increasing non-linear complexity benefit in the loan default classification process. (II) The direct consideration of the network structure and its latent information is inherently captured by closeness centrality which reflects the degree to how related a loan contrast with others within the similarity network. This information conveys details about a loans relative risk positioning, which empowers all tested scoring models in better predicting loan default outcomes. The lower feature importance in the EL model could be traced back to the model's linear character that might not capture more complex dependencies considered in the RF and MLP. In a more general sense, it is our premise that the centrality measures help to capture latent loan similarity factors which remain unobserved in a traditional model specification, accounting only for the original input features. Thus, by retrieving such latent information located in the network positioning of the individual loans respectively to other loans originated by the platform, we can further improve the classification accuracy of credit scoring models.

6.2. Implications for P2P lending platforms and borrowers

The findings of this study hold profound implications for P2P lending platforms and their borrowers. Our comparative study sheds light on the complex relationships in credit risk data and demonstrate the effectiveness of combining ML models with network-based feature extraction to develop more nuanced credit risk assessment tools.

For P2P lending platforms, enhancing the application of basic ML models like the EN, RF, and MLP with graph-based features can help to facilitate more accurate and robust credit risk assessments. In this study we also introduced a modeling approach that is easy to replicate for P2P lending platforms in that they can simply utilize network analysis prior to the model training process, and apply the network structure to any current risk frame work implemented without being obliged to systematically retrain underlying scoring models. The outcome of this study can foster understanding for P2P lending platforms in how the different scoring models perform under the influence of network analysis. Platforms can select or combine ML models that best suit their specific needs and risk tolerance to ultimately provide more accurate and stable credit scores. For borrowers, particularly those with limited credit history or unconventional credit profiles, these advancements in credit risk modeling could mean greater access to credit. As ML models can capture complex patterns and utilize a wider range of data types, borrowers may be evaluated more holistically. This can potentially lower the barrier to credit access, especially for underserved segments, without necessarily compromising the risk management standards of the platform.

6.3. Limitations and potential bias

While this study provides significant insights into the application of ML and network analysis in credit risk modeling, certain limitations and potential biases still deserve attention.

Our analysis, based on a specific dataset, brings to fore the potential limitations concerning generalizability to other contexts or problem applications. Despite our rigorous data cleaning process and the inclusion of a diverse set of features as well as randomized variants of our network centrality variables, we acknowledge that the specific dataset characteristics could still have affected our results. While we cannot completely rule out data-induced influences to have an effect on our modeling approach, we still committed best scholarly practices to rigorously ensure valid results within the scope of our study. Another crucial aspect is the inherent bias–variance trade-off. While tuning the models, we made conscious efforts to balance overfitting and underfitting, yet the risk of over- or under-optimization on the training data at the expense of general performance remains. Nevertheless, we also want to advocate that model optimization was not the primary goal of this study, hence we acknowledge that the evidence presented may not reflect the true optimal state for each ML model in the given parameter constellation as we rather aimed to demonstrate the significance and importance of considering the intricate network structure for the loan default classification process in P2P lending markets.

7. Conclusion and future research

In this study, we conducted a thorough analysis of credit default classification in personal P2P lending markets by utilizing an advanced network analysis approach in combination with three state-of-the-art ML models: Elastic Net (EN), Random Forest (RF), and Multi-layer Perceptron (MLP). We apply a two-step modeling approach to a comprehensive dataset from the Bondora P2P lending platform composed of various borrower characteristics and credit-related attributes. We further explore the usability of network-derived centrality measures such as PageRank, betweenness, closeness, katz, authority, and hub as prediction-enhancing tools in the credit default classification of P2P loans. By computing the Gower's distance between our initial data points, representing individual loans, we first study the intricate network structure of our sample and subsequently derive the network-based centrality features.

Our study findings reveal the crucial role that network centrality measures can fulfill in improving the predictive accuracy of credit scoring models. Across all ML models, we find that centrality features emerge as consistently influential, with PageRank proving to be a key attribute in credit default prediction. In addition, closeness, betweeness and Katz centrality demonstrate consistent importance across all models, simultaneously underlining their relevance as important feature for classifying defaulted loans. By systematically comparing the different graph-enhanced and conventional model types, we find a uniform improvement in the prediction accuracy of all scoring models. In additionally comparing the graph-enhanced models with a reputable

ML technique in form of XGBoost we find further evidence of the model superiority of the graph-based models in accurately predicting credit default. Additional tests with randomly shuffled centrality features confirm the robustness of our findings.

A notable observation from our study is the need for model-specific feature selection. Some features, including katz, betweennness, and closeness centrality, showed varying levels of importance across different models. This emphasizes the complex relationship between data characteristics and model architectures, suggesting further exploration into optimal feature selection and engineering strategies tailored for each model type. Here we see several promising directions for future research. Firstly, while we have demonstrated the efficacy of elastic net, random forest, and deep learning models in predicting credit defaults, there are several other ML techniques that are yet to be fully explored in this context. For instance, the application of support vector machines, gradient boosting algorithms, and newer deep learning architectures could be investigated. We also encourage future studies to consider the integration of even more diverse data types in credit risk modeling. In addition to financial and non-financial indicators, potential data sources could include alternative data such as text from social media, news sentiment, and other behavioral or psychological indicators. Given the increasing availability of such data, there is substantial scope for researchers to explore how these can be harnessed to improve the accuracy and comprehensiveness of credit risk assessments.

CRediT authorship contribution statement

Yiting Liu: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualisation, Project administration, Writing – review & editing. **Lennart John Baals:** Conceptualisation, Investigation, Validation, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Jörg Osterrieder:** Supervision, Project administration, Methodology, Funding acquisition, Resources, Writing – review & editing. **Branka Hadji-Misheva:** Supervision, Project administration, Methodology, Resources, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declares that they have no competing interests.

Data availability

Data will be made available on request.

Acknowledgments

The authors want to thank Štefan Lyócsa and Tomáš Plíhal from Masaryk University for their invaluable help and comments on the data pre-processing and design of the research methods — specifically related to the model estimations.

Funding

This work has been supported by several institutions, each of which has provided vital resources and expertise to the research project.

Firstly, we acknowledge the COST Action CA19130 and COST Action CA21163, under the auspices of the European Cooperation in Science and Technology (COST). COST Actions provide networking opportunities for researchers across Europe, fostering scientific exchange and innovation. This has been particularly beneficial for this research project on financial econometrics.

We would like to express our gratitude to the Swiss National Science Foundation for its financial support across multiple projects. This includes the project on Mathematics and Fintech (IZCNZ0-174853), which focuses on the digital transformation of the Finance industry. We also appreciate the funding for the project on Anomaly and Fraud Detection in Blockchain Networks (IZSEZ0-211195), and for the project on Narrative Digital Finance: a tale of structural breaks, bubbles & market narratives (IZCOZ0-213370).

Most notably, we are grateful for financial support from the Swiss National Science Foundation under the project Network-based credit risk models in P2P lending markets ($100019E - -205487$).

Furthermore, we gratefully acknowledge the support of the Marie Skodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym: DIGITAL, Project No. 101119635.

In addition, our research has benefited from funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA). This grant was provided for the FIN-TECH project, a training program aimed at promoting compliance with financial supervision and technology.

Lastly, we acknowledge the cooperative relationship between the ING Group and the University of Twente. This partnership, centered on advancing Artificial Intelligence in Finance in the Netherlands and beyond, has been of great value to our research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2024.124100.

References

Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, *35*(3), 1275–1292.

Ahelegbey, D. F., Giudici, P., & Hadji-Misheva, B. (2019a). Factorial network models to improve P2P credit risk management. *Frontiers in Artificial Intelligence*, *2*, 8.

Ahelegbey, D. F., Giudici, P., & Hadji-Misheva, B. (2019b). Latent factor models for credit scoring in P2P systems. *Physica A. Statistical Mechanics and its Applications*, *522*, 112–121.

Allen, F., Babus, A., Kleindorfer, P. R., & Wind, Y. (2009). In P. R. Kleindorfer, Y. Wind, & R. E. Gunther (Eds.), *The network challenge: strategy, profit, and risk in an interlinked world* (pp. 367–382).

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Angelini, E., Di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, *48*(4), 733–755.

Babaev, D., Savchenko, M., Tuzhilin, A., & Umerenkov, D. (2019). E.T.-RNN: Applying deep learning to credit loan applications. CoRR, abs/1911.02496.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417.

Battiston, S., Caldarelli, G., May, R., Roukny, T., & Stiglitz, J. (2016). The price of complexity in financial networks. *Proceedings of the National Academy of Sciences*, *113*(36), 10031–10036.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P., & Caldarelli, G. (2012). Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific Reports*, *2*(1), 1–6.

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, *36*(2), 3302–3308.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.

Chen, X., Chong, Z., Giudici, P., & Huang, B. (2022). Network centrality effects in peer to peer lending. *Physica A. Statistical Mechanics and its Applications*, *600*, Article 127546.

Chen, Y.-R., Leu, J.-S., Huang, S.-A., Wang, J.-T., & Takada, J.-I. (2021). Predicting default risk on peer-to-peer lending imbalanced datasets. *IEEE Access : Practical Innovations, Open Solutions*, *9*, 73103–73109.

Coakley, J., & Huang, W. (2020). P2P lending and outside entrepreneurial finance. *The European Journal of Finance*, 1–18, Publisher: Taylor & Francis.

Constantin, A., Peltonen, T. A., & Sarlin, P. (2018). Network linkages to predict bank distress. *Journal of Financial Stability*, *35*, 226–241.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465.

Dastile, X., & Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. *IEEE Access : Practical Innovations, Open Solutions*, *9*, 50426–50440.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 837–845.

Dierkes, M., Erner, C., Langer, T., & Norden, L. (2013). Business credit information sharing and default risk of private firms. *Journal of Banking & Finance*, *37*(8), 2867–2878.

Dimpfl, T., & Peter, F. J. (2018). Analyzing volatility transmission using group transfer entropy. *Energy Economics*, *75*, 368–376.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: the role of appearance in peer-to-peer lending. *The Review of Financial Studies*, *25*(8), 2455–2483.

Engelberg, J., Gao, P., & Parsons, C. A. (2012). Friends with money. *Journal of Financial Economics*, *103*(1), 169–188.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Freeman, L. C., et al. (2002a). Centrality in social networks: Conceptual clarification. *Social Network: Critical Concepts in Sociology. Londres: Routledge*, *1*, 238–263.

Freeman, L. C., et al. (2002b). Centrality in social networks: Conceptual clarification. *Social Network: Critical Concepts in Sociology. Londres: Routledge*, *1*, 238–263.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, *15*, 107–143.

García, V., Marques, A. I., & Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*, *47*, 88–101.

Garmaise, M. J., & Moskowitz, T. J. (2003). Informal financial networks: Theory and evidence. *The Review of Financial Studies*, *16*(4), 1007–1040.

Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, *8*(02), 209–218.

Ghatasheh, N. (2014). Business analytics using random forest trees for credit risk prediction: A comparison study. *International Journal of Advanced Science and Technology*, *72*(2014), 19–30.

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2019). Network based scoring models to improve credit risk management in peer to peer lending platforms. *Frontiers in Artificial Intelligence*, *2*, 3.

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020a). Network based credit risk models. *Quality Engineering*, *32*(2), 199–211.

Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020b). Network based credit risk models. *Quality Engineering*, *32*(2), 199–211.

Glasserman, P., & Young, H. P. (2015). How likely is contagion in financial networks? *Journal of Banking & Finance*, *50*, 383–399.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics. Journal of the International Biometric Society*, 857–871.

Gupta, D., Kose, U., Khanna, A., & Balas, V. E. (2022). *Deep learning for medical applications with unique data*. Academic Press.

H2O. ai (2017). H2O: Scalable machine learning platform. URL https://www.h2o.ai/.

Haselmann, R., Schoenherr, D., & Vig, V. (2018). Rent seeking in elite networks. *Journal of Political Economy*, *126*(4), 1638–1690.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Havrylchyk, O., & Verdier, M. (2018). The financial intermediation role of the P2P lending platforms. *Comparative Economic Studies*, *60*, 115–130.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, *37*(4), 543–558.

Huang, Y., & Kou, G. (2014). A kernel entropy manifold learning approach for financial data analysis. *Decision Support Systems*, *64*, 31–42.

Huang, X., Liu, X., & Ren, Y. (2018). Enterprise credit risk evaluation based on neural network algorithm. *Cognitive Systems Research*, *52*, 317–324.

Iwai, K., Akiyoshi, M., & Hamagami, T. (2020). Structured feature derivation for transfer learning on credit scoring. In *2020 IEEE international conference on systems, man, and cybernetics* (pp. 818–823). IEEE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: vol. 112*, Springer.

Jing, X.-Y., Zhang, X., Zhu, X., Wu, F., You, X., Gao, Y., et al. (2019). Multiset feature learning for highly imbalanced data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 139–156.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika.*, *18*(1), 39–43.

Kealhofer, S. (2003). Quantifying credit risk I: default prediction. *Financial Analysts Journal*, *59*(1), 30–44.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys (CSUR)*, *31*(4es), 5–es.

Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., et al. (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems*, *140*, Article 113429.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, *7*(1), 48–50.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, J. W., Lee, W. K., & Sohn, S. Y. (2021). Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. *Expert Systems with Applications*, *168*, Article 114411.

Leong, C. K. (2016). Credit risk scoring with bayesian network models. *Computational Economics*, *47*(3), 423–446.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136.

Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *Ieee Access*, *6*, 54396–54406.

Li, M.-Y. L., & Miu, P. (2010). A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach. *Journal of Empirical Finance*, *17*(4), 818–833.

Liu, J., Zhang, S., & Fan, H. (2022). A two-stage hybrid credit risk prediction model based on xgboost and graph-based deep neural network. *Expert Systems with Applications*, *195*, Article 116624.

Lyocsa, S., Vasanicova, P., Hadji Misheva, B., & Vateha, M. D. (2022). Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innovation*, *8*(1), 1–21.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*(10), 4621–4631.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*(2), 449–470.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256.

Niu, K., Zhang, Z., Liu, Y., & Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information Sciences*, *536*, 120–134.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, *62*, 1–16.

Poenaru-Olaru, L., Redi, J., Hovanesyan, A., & Wang, H. (2022). Default prediction using network based features. In *Complex networks & their applications x: volume 1, proceedings of the tenth international conference on complex networks and their applications COMPLEx NETWORkS 2021 10* (pp. 732–743). Springer.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, *36*(6), 1389–1401.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1985). Learning internal representations by error propagation.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika.*, *31*(4), 581–603.

Sadhwani, A., Giesecke, K., & Sirignano, J. (2021). Deep learning for mortgage risk. *Journal of Financial Econometrics*, *19*(2), 313–368.

Shen, F., Zhao, X., Kou, G., & Alsaadi, F. E. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, *98*, Article 106852.

Shi, Y., Qu, Y., Chen, Z., Mi, Y., & Wang, Y. (2024). Improved credit risk prediction based on an integrated graph representation learning approach with graph transformation. *European Journal of Operational Research*, *315*(2), 786–801.

Stanton, R., Walden, J., & Wallace, N. (2018). Mortgage loan flow networks and financial norms. *The Review of Financial Studies*, *31*(9), 3595–3642.

Sukharev, I., Shumovskaia, V., Fedyanin, K., Panov, M., & Berestnev, D. (2020). Ews-gcn: Edge weight-shared graph convolutional network for transactional banking data. In *2020 IEEE international conference on data mining* (pp. 1268–1273). IEEE.

Sun, X., & Xu, W. (2014). Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, *21*(11), 1389–1393.

Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science*, *174*, 150–160.

Torri, G., Giacometti, R., & Paterlini, S. (2018). Robust and sparse banking network estimation. *European Journal of Operational Research*, *270*(1), 51–65.

Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505–513.

Yıldırım, M., Okay, F. Y., & Özdemir, S. (2021). Big data analytics for default prediction using graph theory. *Expert Systems with Applications*, *176*, Article 114840.

Zhou, B., Jin, J., Zhou, H., Zhou, X., Shi, L., Ma, J., et al. (2023). Forecasting credit default risk with graph attention networks. *Electronic Commerce Research and Applications*, *62*, Article 101332.

Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, *5*(3), 427–443.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.