scientific reports

OPEN

Predicting merchant future performance using privacy-safe network-based features

Mohsen Bahrami[1,5], Hasan Alp Boz[2,5], Yoshihiko Suhara[3], Selim Balcisoy[2], Burcin Bozkaya[4] & Alex Pentland[1,3]

Small and Medium-sized Enterprises play a significant role in most economies by contributing to job creation and economic growth. A majority of such merchants rely on business financing, and thus, financial institutions and investors need to assess their performance before making decisions on business loans. However, current methods of predicting merchants' future performance involve their private internal information, such as revenue and customer base, which cannot be shared without potentially exposing critical information. To address this problem, we first propose a novel approach to predicting merchants' future performance using credit card transaction data. Specifically, we construct a merchant network, regarding customers as bridges between merchants, and extract features from the constructed network structure for prediction purposes. Our study results demonstrate that the performance of machine learning models with features extracted from our proposed network is comparable to those with conventional revenue- and customer-based features, while maintaining higher privacy levels when shared with third-party organizations. Our approach offers a practical solution to privacy concerns over data and information required for merchants' performance prediction, enabling safe data-sharing among financial institutions and investors, helping them make more informed decisions on allocating their financial resources while ensuring that merchants' sensitive information is kept confidential.

Small and Medium-sized Enterprises (SMEs) constitute a significantly large portion of economies around the world and thus play a vital role in economic stability and growth. According to the Annual Report on European SMEs 2021/2022[1], 99.8% of European enterprises consist of SMEs, which account for 65% of total employment in the business economy. The U.S. Small Business Administration (SBA) reported that SMEs were responsible for employing 61.7 million people, which is equal to 46.4% of the total employed workforce in the private sector in 2022[2]. These statistics indicate that the success and failure of SMEs can directly affect the levels of (un) employment in countries.

On the other hand, according to the U.S. Bureau of Labor Statistics, between 1994 and 2019, about 33% of the newly opened SMEs failed within their first two years, and only roughly 50% of them survived after their first five years[3]. Their historical records indicate that the SME failure and default rates can drastically worsen during economic downturns and financial crises. Considering the importance of SMEs for economic stability, it is of crucial importance for governments, responsible organizations, and authorities to support and oversee the temporal performance and well-being of SMEs.

Moreover, most SMEs heavily rely on business financing from banks and/or other financial institutions and entities to run their businesses[4]. However, given the noticeable failure rates of SMEs, financial institutions and investors have to carefully assess SMEs to make decisions on business loans. Such institutions and entities seek methodologies that provide insights into the determinants of success, creditworthiness, and distress level of an SME. Therefore, an accurate well-being assessment and performance prediction is of critical importance for financial institutions as they determine the overall risk level and leverage position for these institutions, and potentially the entire financial ecosystem.

[1]MIT Connection Science, Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. [3]MIT Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Sabanci Business School, Sabanci University, Istanbul, Turkey. [5]These authors contributed equally: Mohsen Bahrami and Hasan Alp Boz. email: bahrami@mit.edu

However, current methods of predicting future performance of SMEs, involve their private internal information, such as revenue and customer base, which cannot be shared without potentially exposing critical information. On the other hand, SMEs are typically involved in resource-constrained operations that suffer from the scarcity of certified financial statements and publicly available information on debt, equity, or liquidity[5–7]. Additionally, unlike public companies that publish financial reports periodically, SMEs are not obligated to publish their financial activity reports, and thus, it is not feasible for financial institutions to monitor and track the financial activities and positions of every single SME.

Given the highly competitive environments that SMEs operate in, they always have privacy concerns about sharing their internal information with other firms, which could harm their position in case of leakage to their competitors. Therefore, the scarcity or unavailability of SMEs' internal information and financial records is a big hurdle in replicating the models built on such data. On the other hand, sharing data with third-party organizations has always been a delicate matter for financial institutions. In such cases, legally binding but also time demanding non-disclosure agreements (NDA) between entities help data to be secured to a certain extent.

This study endeavors to examine the SME performance evaluation and prediction (also referred to as "merchants" hereafter) while addressing concerns regarding data privacy and data sharing safety. We aim to achieve these objectives without directly accessing the internal financial metrics of the merchants, instead, by leveraging a comprehensive dataset of credit card transactions on a large scale.

First, in order to evaluate a merchant's performance level, we present a new assessment measure as a function of the relative changes in future transactions, revenue, and customers of a merchant. Subsequently, we propose to derive a feature set from credit card transaction data instead of the internal financial metrics of merchants. Thus, banks can easily incorporate the approach presented here into their decision support models. We refer to these features as *revenue-based features*. Then we extract customer information to answer the question of 'who visits the shop?'. The features extracted from customer information are referred to as *customer-based features*. The revenue-based and customer-based feature sets will serve as the baselines for our analyses.

Next, we propose a novel network model of merchants based on customer credit card transaction patterns and investigate merchants' positions in the network using network centrality metrics namely: node degree, betweenness, closeness, and eigenvector centrality. In addition to these centrality metrics, we obtain diversity variables using the ego-network of each merchant to capture its surroundings. These features are referred to as *network-based features*. We further use the proposed network structure for learning continuous numeric feature representations for nodes (i.e., merchants) in the network (node2vec)[8]. We refer to these features as *node2vec features*.

Finally, we extract a set of features based on merchants' available information for banks and financial institutions. These features include information about merchants' business categories, socio-demographic information about their neighborhood residents, and physical factors including the number and diversity of other businesses and amenities in their vicinity. These *merchant-based features* are utilized in all of our analyses in addition to the other four extracted feature sets (i.e., revenue-based, customer-based, network-based, and node2vec feature sets). The resulting feature sets are fed into different types of machine learning models to predict merchants' future performance levels.

The results of these analyses show that the proposed network-based features' performance is comparable to those of conventional revenue-based and customer-based features, indicating that the proposed merchant network is effective in capturing the performance level of merchants and offering promising results for future studies. The proposed network-based and node2vec features possess a crucial attribute: not only do they offer valuable insights into the present and future performance of merchants, but they also ensure a significantly enhanced level of privacy protection compared to raw financial records. Therefore, these features can be safely shared with third-party organizations in a structured tabular format. This emphasis highlights the importance of these features in striking a balance between providing valuable insights and preserving merchants' privacy while facilitating secure data sharing with external entities.

To summarize, the contributions of this paper are as follows:

- We introduce a new approach to defining the performance level of merchants and verify that the labels do not exhibit any particular biases concerning the geographical locations, income levels of the region's residents, or the socio-economic status of the merchants' customers.
- We propose constructing a merchant network based on customer purchase transaction records. We then leverage the four commonly used network centrality scores (i.e., node degree, betweenness, closeness, and eigenvector centrality.) and diversity variables obtained from their ego networks as input features to predict their future performance.
- We conduct our analysis on a large-scale credit card transaction dataset and show that the predictive models with our proposed network-based and node2vec features yield comparable performance to those using conventional features sets (i.e., revenue- and customer-based) while ensuring privacy and safeguarding any sensitive information regarding merchants, and thus, enabling secure data sharing with third-party organizations.

The methods and approaches outlined in this research are poised to pave the way for the advancement of methodologies enabling businesses and financial entities to exchange the valuable insights within their data, while preserving the confidentiality of the raw data.

The remainder of this paper is outlined as follows. We commence by examining the current state-of-the-art literature, which serves as a foundation for the methodologies that have influenced our study. Next, we introduce our proposed methodology, describe the datasets employed in this research, and delineate the analytical setting.

Subsequently, we present a comprehensive analysis of our results. Finally, we engage in a discussion regarding the implications derived from our findings and offer concluding remarks to wrap up the paper.

Background

Machine learning constitutes the backbone of the majority of merchant well-being and performance prediction studies[9–12]. Dominant merchant performance assessment methodologies involve quantitative risk models that combine various financial metrics such as earnings per asset, equity per asset, and debt ratio[13–15]. Gallucci et al.[16] incorporate financial metrics, bank-firm interaction information, and corporate governance variables in SME loan default prediction using a Bayesian model. Features extracted from financial statements of corporations, such as debt ratio, total capital turnover and quick ratio, have been used by Son et al.[17] on a Gradient Boosting framework. To predict business failure, Kim et al.[12] make use of the tree-based majority voting ensemble method which includes, in addition to the financial features, the recession indicator computed by the National Bureau of Economic Research (NBER) as a macro-economic feature to account for the overall economic status.

Other research incorporates an SME's principal owner's credit information in the quantitative risk model in order to project the risk for the SME[18]. This approach seeks to tackle the fundamental lack-of-data issue underlying SME risk scoring; however, such methods still involve exclusively collected data. This data deficiency has led the researchers to explore potential substitute data and proxies. Another recent study[19] relies on web mining to extract proxy features from online resources, such as TripAdvisor and OpenStreetMap for predicting SME growth for restaurants in Switzerland where revenue information is known a priori.

To account for local economies' effect on SMEs, Fernandes and Artes[20] propose a new variable based on ordinary kriging to help assess the risk of credit default among SMEs. The proposed variable was able to enhance the predictive power of the logistic model on credit scoring. Yoon and Kwon[6] have shown that credit card data can be highly informative about the financial position of SMEs. They built a support vector machines (SVM) model to predict bankruptcies, using the variables such as sales fluctuation and sales patterns extracted from credit card transactions.

Inspired by social physics[21,22] and computational social science [23], which take data-driven approaches toward understanding human behavior, we take a similar approach to understand merchant performance. We go one step beyond the current stage of bankruptcy prediction by using a large-scale credit card transaction dataset and leveraging machine learning models to predict merchants' future performance without directly accessing their internal financial metrics.

Building upon the findings of a previous study[24], which highlighted the significance of social bridges in comprehending the purchasing behaviors across diverse communities, we leverage customers as bridges between merchants in order to gain deeper insights into the future performance of the merchants. This approach recognizes the influential role of social connections in shaping economic dynamics and allows us to explore the potential of network science in understanding social phenomena.

Furthermore, a substantial body of literature in social physics has employed network science to enhance our understanding of various social phenomena. Studies in this domain have consistently demonstrated that interactions and information exchange through networks can promote productivity[25–30]. It is widely recognized that networks serve as the fundamental framework for social and economic activities, underpinning the fabric of society[31]. By integrating these insights from network science into our study, we can shed light on the intricate dynamics of social and economic interactions, providing a comprehensive understanding of merchant performance.

As we live in a social network, merchants run their businesses in a merchant network as well; therefore, we contend that the structural properties of merchants in a network are essential to their performance prediction. To this end, we propose constructing a *merchant network*, using credit card transaction data, with the premise that customers can act as bridges between merchants. We use network centrality metrics (i.e., degree and strength, betweenness, closeness, and eigenvector centrality) as indicators of merchants' performance in the network. The merchant network should enable us to capture the positioning of merchants by incorporating their surrounding information. We hypothesize that the network centrality metrics in the merchant network can potentially provide holistic signals for better understanding and predicting merchants' future performance .

Methods

In this section, we introduce the utilized credit card dataset and its basic statistics. We then describe our methodology and the analytical setting for evaluating the performance of our proposed methods.

Data and pre-processing.

In this study, we use a credit card transaction dataset from a major bank in an OECD country[32]. The sampled dataset consists of three tables, namely: customer, merchant, and transaction tables.

The customer table contains an anonymous customer ID, and customer demographic information including customer age, gender, marital status, education level, estimated income, and home and work district IDs. Districts are administrative areas within the metropolitan area of study, and each district is governed by a local municipality. Districts have an average size of 150 square kilometers and an average population of 380,000 residents.

The merchant table contains anonymous merchant ID, merchant ISO 18245 category code (MCC)[33], and merchant district ID. Merchants may belong to various categories such as grocery stores and supermarkets, restaurants, gas stations, clothing stores, and bookstores, which are described by their MCCs.

The transaction table contains transaction records of customers, including merchant ID, customer ID, the amount of the transaction, and the timestamp. The original dataset consists of 4,507 merchants, 62,194 customers, and 2,511,527 transactions over the span of one year from July 2014 to July 2015. More information about the tables described above is provided in "Supplementary Information".

Since our aim is to predict future performance level for private sector merchants, we remove non-discretionary MCCs such as government-owned places, parking lots, lodging, and other similar categories. Then, to ensure the robustness of the analyses, we filter out merchants that recorded less than 10 transactions per month in each period. A period can be equal to one or more months based on the amount of available data. More detailed information is provided in the section Analytical setting.

Finally, we only keep the MCCs that contain more than 10% of remaining merchants in the dataset, and remove the districts with less than 10 merchants remaining. As a result, the pre-processed data contains 1,977 merchants in 3 MCCs. Table 1 shows the number of remaining merchants in each MCC after data pre-processing.

The pre-processed dataset contains no information that could potentially be used to identify individual customers. A complete list of business categories and the related numbers of transactions and merchants are provided in "Supplementary Information".

Methodology. In this section, we initially establish the merchant network concept and provide a detailed explanation on its construction using customer credit card transactions. Then, we describe merchant network features that capture signals of merchants' performance in a holistic manner. Finally, we delineate the analytical setting for evaluating the method's performance in comparison to baseline methods.

Merchant network. In recent years, there has been a growing recognition of the ability of networks and their dynamics to capture distinctive characteristics of various phenomena. Generally, networks characterizing a certain phenomenon exhibit inherent suitability for making inferences about the future state of entities (nodes) in those networks. As previous studies have shown[34,35], structural properties of nodes in a network are capable of encoding the capacity for improvement for the nodes; hence, networks are particularly well-suited for applications that involve investigating and making inferences about the future of specific entities[36]. Inspired by those findings, using the credit card transaction dataset we define and form a merchant network as follows.

Let $v_i \in V$ denote a vertex for merchant i (shown by m_i) and $e_{ij} \in E$ is an edge between vertices v_i and v_j. A merchant network is an undirected graph $G = (V, E, W)$, where the weight $w_{ij} \in W$ is defined as:

$$w_{ij} = \left|\left\{c_k | \exists(c_k, m_i) \in \mathscr{D} \ \wedge \exists(c_k, m_j) \in \mathscr{D} \ , c_k \in \mathscr{C} \ \right\}\right| \tag{1}$$

where $\mathscr{D}$ is credit card transaction data and $\mathscr{C}$ is the set of all customers. The edge weight w_{ij} is defined by the number of distinct customers who made purchases at both merchants i and j, and we define neighbors (also referred to as alters) of merchant i as the set of merchants that are directly connected to merchant i in graph G and denoted by A_i. Figure 1 depicts an example of how a merchant network is created from credit card

Description	MCC	Merchant count
Grocery stores, supermarkets	5411	1118
Men's and women's clothing stores	5691	464
Service stations (with or without ancillary services)	5541	395

Table 1. Business categories ordered by their merchant counts after data pre-processing.

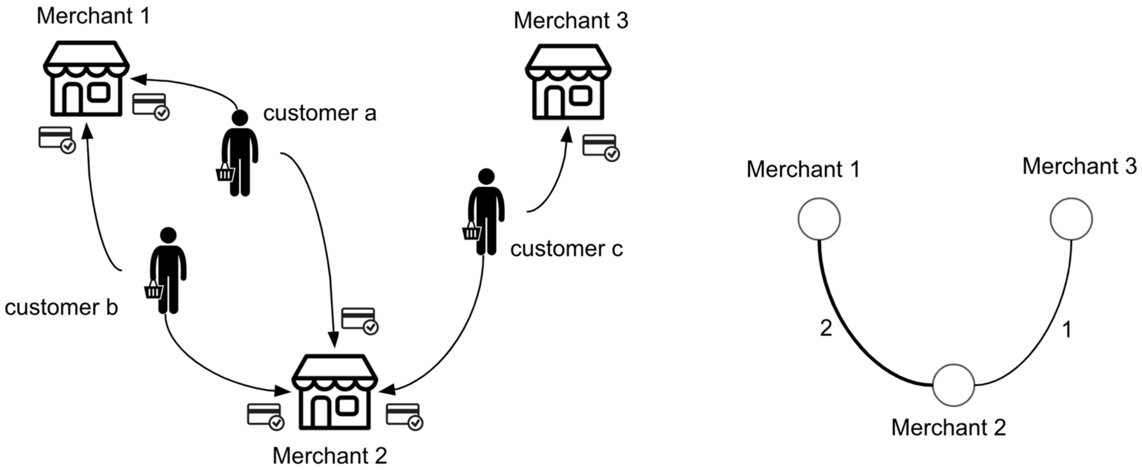

Figure 1. (Left) Example of customer transactions at merchants. (Right) Merchant network extracted from the transactions.

transactions. Intuitively, the weight of an edge between two merchants increases as more customers purchase at both merchants.

Merchant network-based features. In order to extract signals specific to each merchant from a constructed merchant network, it is necessary to aggregate information within each node of the network. In this paper, we consider using the four most commonly employed centrality metrics (i.e., degree and strength, betweenness, closeness, and eigenvector centrality) and two diversity metrics as the signals of merchant current and future performance level. Each merchant's importance is measured by four centrality metrics based on its position in the network. Those metrics are known to capture different perspectives on the importance of a node[37]. In what follows, we introduce those centrality measures and interpret what each of those measures may mean in our proposed merchant network. We also derive two diversity measures from the proposed network structure to understand a merchant's capability in attracting customers from different areas and with different preferences and needs.

Degree and strength[38]. are computed based on a node's direct ties with other nodes in a network. Degree of a node is equal to the number of its direct ties with other nodes, which is also equal to the number of alters of merchant i as defined to be members of the set A_i. Node strength is equal to the sum of the weights of edges that the node shares with others. Here, the strength of a node (i.e., merchant) is identical to the number of distinct customers in common with the other merchants in the network. Merchants' degrees and strengths are well correlated with their revenues during each corresponding period of time.

Betweenness centrality[38]. is a measure of how important a node is based on the shortest paths through the network. It can also be seen as the extent to which a node lies on the shortest path between other nodes in the network. This measure takes into account the connectivity of the node's neighbors, giving a higher value for nodes which bridge clusters of nodes. In our case, it can capture the number of merchants to which a shop is connecting indirectly through their direct links[39].

Eigenvector centrality[40]. is an extension of a node's degree centrality. This centrality metric measures a node's importance while giving consideration to the importance of its neighbors[41]. In our network, it roughly encodes the probability that a random customer will visit a particular merchant.

In addition to the centrality measures, we also consider two ego network diversity metrics using the constructed merchant network to capture the dynamics of a node's interactions with different geographies and business categories.

Closeness centrality[38]. is the extent that a node is close to other nodes in a network. It is a proxy of measuring the ability of a merchant to access information of other merchants in its network through its customers.

Geographical diversity. measures the geo-spatial diversity of neighbors of the target merchant i. We denote this variable with D_g^i and calculate the geographical diversity based on the geographical distributions of the neighbor merchants of a merchant. Specifically, we use the district information to compute the probabilistic distribution of the districts. To quantify the diversity, we utilize Shannon entropy [42], a widely adopted metric in this context [43]. The formalization of Shannon entropy for computing the geographical diversity is depicted by Eq. (2).

$$D_g^i = \sum_{h \in I_H^i} -p_h^i \log p_h^i \tag{2}$$

where p_h^i is the fraction of the edge weights of merchant i connected to other merchants from district h over all districts having edges with merchant i denoted by I_H^i.

Business-category diversity. measures the diversity of the business categories of neighbors of merchant i. Similar to the geographical diversity, we consider the business-category diversity of a merchant's ego network with respect to the business category of its neighbors. We formalize this metric with D_c^i and use the following equation to calculate it.

$$D_c^i = \sum_{b \in I_B^i} -p_b^i \log p_b^i \tag{3}$$

where p_b^i shows the fraction of the edge weights of merchant i connected to other merchants from business type b over all business types having edges with merchant i denoted by I_B^i.

Label definition. The main objective of this section is to establish a performance evaluation metric that can effectively gauge a merchant's performance in a competitive environment, and at the same time can be derived from its transactional data. Prior research has employed a range of objective indicators (e.g., financial metrics) and subjective indicators (e.g., managers' perceptions) to propose various performance measurements for SME merchants[44]. However, accessing a firm's internal goals and its management's perception of performance may not always be feasible, and even if attainable, there is no direct method for translating such information into a metric that can accurately indicate the firm's market position and relative standing among competitors. Therefore, in this study, we leverage objective indicators, encompassing merchants' sales, attractiveness to passers-by, and customer relationship information, to define a new performance metric.

Considering the magnitude and temporal scope of our dataset, along with the outcomes of preliminary analyses, we establish a period duration of 6 months, dividing the data records into two equal-length periods. Subsequently, we juxtapose a merchant's revenues, number of unique customers, and number of transactions during the initial 6-month observation period with corresponding measures in the subsequent 6 months. We then calculate the rate of change for each metric using Eqs. (4–6).

$$\Delta R^i_{t+1,t} =(R^i_{t+1} - R^i_t)/R^i_t \tag{4}$$

$$\Delta C^i_{t+1,t} =(C^i_{t+1} - C^i_t)/C^i_t \tag{5}$$

$$\Delta N^i_{t+1,t} =(N^i_{t+1} - N^i_t)/N^i_t \tag{6}$$

where R^i_t, C^i_t, and N^i_t denote the revenue, number of unique customers, and number of transactions of merchant i in period t (here we use the first 6 months) and R^i_{t+1}, C^i_{t+1}, and N^i_{t+1} denote the revenue, number of unique customers, and number of transactions of merchant i in period $t+1$ (the remaining 6 months). It is important to note that the revenue of a merchant is calculated by aggregating the spending amount of all transactions made at the merchant in the corresponding period. Then we compare the rates resulting from Eqs. (4–6) for merchant i with the median value of the same indicators' rates over all merchants in the same business category (MCC) as merchant i's. By contrasting the rate of change in these metrics among merchants within the same MCC, one can significantly mitigate the impact of seasonality. This becomes particularly crucial when the available historical data is insufficient for capturing the full extent of seasonality effects.

Subsequently, for each rate of change, we label the merchants with rates above the median with 1 and those with rates below the median with 0 as binary-class labels using Eqs. (7–9). This kind of binary labeling is a common practice in the literature[45].

$$I^i_R(t+1) \rightarrow \begin{cases} 1 & if \quad \Delta R^i_{t+1,t} \ \geq \ median(\Delta R^{b_i}_{t+1,t}) \\ 0 & otherwise \end{cases} \tag{7}$$

$$I^i_C(t+1) \rightarrow \begin{cases} 1 & if \quad \Delta C^i_{t+1,t} \ \geq \ median(\Delta C^{b_i}_{t+1,t}) \\ 0 & otherwise \end{cases} \tag{8}$$

$$I^i_N(t+1) \rightarrow \begin{cases} 1 & if \quad \Delta N^i_{t+1,t} \ \geq \ median(\Delta N^{b_i}_{t+1,t}) \\ 0 & otherwise \end{cases} \tag{9}$$

Then for each merchant, we sum those three binary indicators (Eq. 10). If they sum up to 3–that is, the merchant's performance is better than the median in all three indicators (i.e., change in revenue, unique customer count, and transaction count)–we label the merchant as a 'well-performing' merchant. These merchants can be considered as low-risk merchants since they have performed better than at least half of their counterparts. On the other hand, if all indicators for a merchant are equal to 0, it means that the merchant is under-performing in all indicators considering its competitors in the same MCC. Those merchants can be considered as high-risk merchants for investment and we label such merchants as 'poorly-performing' merchants. The remaining merchants are labeled as 'medium-performing' with medium risk level for financial institutions, and thus, in total, we have three performance classes. Equation (11) formalizes the last step in the labeling task.

$$I^i_S(t+1) = I^i_R(t+1) + I^i_C(t+1) + I^i_N(t+1) \tag{10}$$

$$L^i(t+1) \rightarrow \begin{cases} \text{`well-performing'} & if \quad I^i_S(t+1) = 3 \\ \text{`medium-performing'} & if \quad I^i_S(t+1) = 2 \ \text{ or } \ I^i_S(t+1) = 1 \\ \text{`poorly-performing'} & if \quad I^i_S(t+1) = 0 \end{cases} \tag{11}$$

Analytical setting. In this study, we use machine learning techniques to evaluate the effectiveness of the signals extracted from our proposed merchant network. In such evaluation frameworks that are based on supervised learning methods, one needs to derive input features and define the output labels. Our analytical setting is explained in what follows.

Feature extraction. In order to assess the effectiveness of the network-based feature set (Network), we conduct a comparative analysis with two baseline feature sets: revenue-based features (Revenue) and customer-based features (Customer). Previous research by Yoon and Kwon[6] has demonstrated the viability of revenue-based features as a substitute for internal financial data in predicting merchant bankruptcy. Furthermore, studies such as those by Anderson et al.[46] and Simester et al.[47] have utilized customer-based features to forecast the success and failure of new products.

In addition to the features extracted from the credit card transaction data, we include the features that pertain to a merchant's business category (i.e., MCC), its physical surrounding, and neighborhood attributes. Based on merchants' locations, we extract the population and average household income of the districts they are located in. It is important to note that as shown by previous studies[29,48,49], the number and diversity of points-of-interest (POIs) and amenities in a merchant's proximity can increase its attractiveness for transient customers and improve its market potential. To this end, we use a dataset of POIs, provided by Here.com (a digital map production company), in which the POIs are grouped into twelve categories, namely: business centers, community service centers, financial institutes, educational institutes, entertainment places, shopping places, restaurants, hospitals, parks, travel destinations, auto services, and transportation hubs.

By leveraging the POI dataset and the geographic information of merchants, we examine the POIs situated within a 200-meter radius of each merchant. Subsequently, we calculate the quantity and category diversity of these nearby POIs. We refer to these features as merchant-based features (Merchant), as presented in Table 2. These features are incorporated as a fixed subset of input features into all models throughout the analyses.

The customer-based features are derived from the data related to individuals who have conducted transactions with the merchants included in our study. Some merchants target a specific demographic group of customers and some attract a wide range of demographic groups. By adding such features, we aim to take into account the attractiveness of each merchant with respect to the variety of customers. These features help us understand whether the merchants of interest are successful in attracting their target customers from different demographic groups and regions. The set of customer-based features is shown in Table 2. On the other hand, an existing study has used revenue-based features for bankruptcy prediction[6]. The proposed model uses historical records of a merchant's sales, revenue, and customer profitability statistics. Here, we use the revenue-based features presented in Table 2.

The Network features are the main contribution of this study. We compute the network features based on a merchant network constructed for the first 6 months. As explained in the previous section, we extract the features shown in Table 2, including 4 centrality metrics and 2 diversity metrics based on a node's position in the network structure. It is worth noting that although there are alternative centrality metrics available (e.g., PageRank), we chose not to incorporate them into the network-based feature set as they did not contribute to our analysis results.

Lastly, as a privacy-enhanced alternative to Network features, we employ features generated through representations of nodes utilizing the proposed network structure and topology. This is accomplished by using the node2vec algorithm[8]. We generate a feature vector of 128 dimensions for each merchant with return parameter p and in-out parameter q set to 1.3 and 1.2 respectively. Although it is possible to adjust all dimensions and hyperparameters of this technique, and several methods have been proposed to identify the optimal dimensions[50], this study focuses on a different aspect. Instead of exploring and implementing optimization approaches, we adopt a commonly used practice of comparing the performance of various dimension choices (e.g., 50, 100, 128, 200, and 300) in the downstream task of predicting merchants' future performance.

It is important to highlight that all feature sets are exclusively derived from credit card transactions within the initial 6 months (i.e., observation period). This strict selection ensures that no information from the prediction period is incorporated into the predictive models, thus maintaining their integrity.

Machine learning algorithms. We use four machine learning models [51] built on four different types of algorithms in the analyses, namely: Multi-class Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Naive Bayes (NB). These algorithms are among the common choices for linear and non-linear models and major machine learning algorithms for a wide variety of domains. These methods are widely used in the bankruptcy prediction literature.

Evaluation metric. Following the standard machine learning evaluation framework , we conduct ten-fold cross-validation for classification and report area under the receiver operating characteristic curve (AU-ROC)[51] as an evaluation metric for model performance. AU-ROC is a measurement of how well the classification models can distinguish between different classes. In our study case, since the labels are not severely imbalanced, AU-

Feature sets			
Merchant	**Customer**	**Revenue**	**Network**
- Merchant MCC	- Number of unique customers	- Period total revenue	- Degree
- District population	- Age mean	- Number of transactions	- Closeness centrality
- District's average per month household income	- Income mean and median	- Number of unique customers	- Eigenvector centrality
	- Education diversity		- Betweenness centrality
- Surrounding POI count	- Gender diversity		- Ego network geo diversity
- Category diversity of surrounding POIs	- Marital status diversity		- Ego network MCC diversity
	- Employment type diversity		
	- Home district diversity		
	- Work district diversity		
	- Distinct home district count		
	- Distinct work district count		

Table 2. Feature sets employed as predictive variables and input for machine learning models in this study.

ROC is considered to be a suitable evaluation metric. In this setting, to account for multi-class labels, we use the One versus Rest (OVR) evaluation metric where AU-ROC is calculated for each class against the rest, and resulting scores are averaged. During cross-validation, we compute AU-ROC on test-fold and obtain mean and standard deviation across folds.

Results

For the purpose of conducting our analyses, we construct a merchant network by utilizing customer co-purchase (edge) information extracted from the credit card transaction dataset. Subsequently, we compute the previously introduced centrality and diversity features from the derived network for each individual merchant (node). Next, we use our proposed performance evaluation metric for labeling the merchants according to their rank among their counterparts in the same MCC considering the rate of change in their revenue, number of transactions, and number of unique customers from one period to another. Finally, utilizing the extracted feature sets (e.g., network-based features), along with the performance labels, we leverage different machine learning methods to evaluate the performance of different features and feature sets in predicting the future performance of merchants.

Label analysis. Among the studied 1,977 merchants, 590 (29.84%) are labeled as well-performing, 818 (41.37%) are labeled as medium-performing, and 569 (28.78%) are labeled as poorly-performing merchants. To ascertain the reliability of the labels in providing insights into the future performance of merchants and to assess the potential influence of merchant location and customers' socio-demographic factors, such as income and wealth, on the assigned labels, we conduct the following analyses.

Label indication. To investigate if the defined labels are able to distinguish between well-performing and poorly-performing merchants, and provide insights into merchants' performance in the longer terms, we compare the merchants that possess similar magnitudes of revenue, number of customers, and transaction counts in the first period, but are labeled oppositely (i.e., poorly-performing vs. well-performing) bringing into account their performance in the second period.

To this end, we first convert the revenue, transaction count, and number of unique customers of merchants during the first period (first 6 months) into quartiles (i.e., Q1, Q2, Q3, and Q4). Table 3 illustrates an example of the resulting data table structure by providing a random sample of rows. Then we choose the merchant pairs from the same MCC and the same quartiles of revenue, transaction count, and the number of unique customers in the first period.

We only keep the pairs that are labeled oppositely (i.e., well-performing and poorly-performing) based on their second-period performance indicators. Those merchant pairs are the ones that: (1) fall into the same quartiles of merchant revenue, transaction count, and distinct customer count, and (2) one of them is labeled as well-performing and the other is labeled as poorly-performing. There are 11,813 pairs of merchants in our dataset that satisfy both conditions. Table 4 provides an example of the data table resulting from using the information presented by Table 3 as input.

Next, using the ordinary least squares method, for each merchant we compute three fitted line slopes taking into account their monthly revenue, transaction count, and number of unique customers as dependent variables

$\mathbf{ID^m}$	$\mathbf{L^m(t+1)}$	$\mathbf{Q^m_R(t)}$	$\mathbf{Q^m_C(t)}$	$\mathbf{Q^m_N(t)}$
119811014	well-performing	Q4	Q3	Q3
119811011	poorly-performing	Q4	Q3	Q3
119811067	medium-performing	Q2	Q1	Q1
119811051	medium-performing	Q2	Q1	Q1
119811017	well-performing	Q1	Q3	Q3
119811002	poorly-performing	Q1	Q3	Q3
119811051	well-performing	Q1	Q2	Q3
119811018	poorly-performing	Q1	Q2	Q3

Table 3. Example table of merchants' revenue, transaction count, and unique customers count, presented in quartiles.

$\mathbf{ID^{m1}}$	$\mathbf{ID^{m2}}$	$\mathbf{L^{m1}(t+1)}$	$\mathbf{L^{m2}(t+1)}$	$\mathbf{Q^{m1,m2}_R(t)}$	$\mathbf{Q^{m1,m2}_C(t)}$	$\mathbf{Q^{m1,m2}_N(t)}$
119811014	119811011	well-performing	poorly-performing	Q4	Q3	Q3
119811017	119811002	well-performing	poorly-performing	Q1	Q3	Q3
119811051	119811018	well-performing	poorly-performing	Q1	Q2	Q3

Table 4. Merchant pairs with the same quartiles of revenue, transaction count, and unique customers count but opposite labels.

and month numbers as the independent variable. Those slopes can provide an overall indication of a merchant's performance considering each variable over 12 consecutive months. Equation (12) shows the closed-form expression of the regression model where β_1 denotes the value we use as fitted line slope in our analysis.

$$\begin{aligned} &Y = \beta_0 + \beta_1 X + \epsilon \\ &Y = (Y_1, ..., Y_{12})^\intercal,\ X = (1, ..., 12)^\intercal,\ \text{and}\ \epsilon = (\epsilon_1, ..., \epsilon_{12})^\intercal \end{aligned} \tag{12}$$

Figure 2 shows the time series plots for three instances provided in Table 4. Each sub-figure (i.e., 2a, b, and c) depicts time series plots of a merchant pair's monthly revenues (i.e., 2a$_R$, 2b$_R$, and 2c$_R$), monthly transaction counts (i.e., 2a$_N$, 2b$_N$, and 2c$_N$), and numbers of their monthly distinct customers (i.e., 2a$_C$, 2b$_C$, and 2c$_C$) over 12 months of historical credit card transaction records. The vertical blue line in each plot splits the time-frame into two periods of 6 months as we used for labeling. All pairs are chosen from the same quartile of customer count, transaction count, and revenue in the first period and may have similar trends during the first 6 months. In contrast, each merchant pair have opposite labels (i.e., well-performing vs. poorly-performing) according to their performance indicators values in the subsequent period. As illustrated in Figure 2, it is evident that each merchant pair have a comparable performance during the first period, but their performance is significantly different during the second period. For visualization purposes, the fitted lines representing merchants labeled as well-performing are depicted in green, while the fitted lines for those labeled as poorly-performing are displayed using red color. The lines in each chart are the OLS fitted lines and their slope is equal to β_1 in Eq. (12). Those lines provide insights into merchants' overall performance trend considering each of the three factors, namely: revenue, transaction count, and distinct customer count over the course of 12 months.

In the subsequent step, we compare the β_1 values for each pair of merchants. For each slope pair, if the merchant labeled as well-performing has a higher slope, we assign 1 to the slope indicator and otherwise assign 0, and then sum the three resulting indicators. If the well-performing merchant has higher β_1 values in all three indicators, the sum will be equal to 3. Conversely, if the well-performing merchant has lower values in all three indicators, the sum will be equal to 0. The closed-form expressions of these calculations are presented in Eqs. (13–16).

$$I^R_{\beta_1} \rightarrow \begin{cases} 1 & if \quad \beta_1^{R_{well-performing}} \geq \beta_1^{R_{poorly-performing}} \\ 0 & otherwise \end{cases} \tag{13}$$

$$I^C_{\beta_1} \rightarrow \begin{cases} 1 & if \quad \beta_1^{C_{well-performing}} \geq \beta_1^{C_{poorly-performing}} \\ 0 & otherwise \end{cases} \tag{14}$$

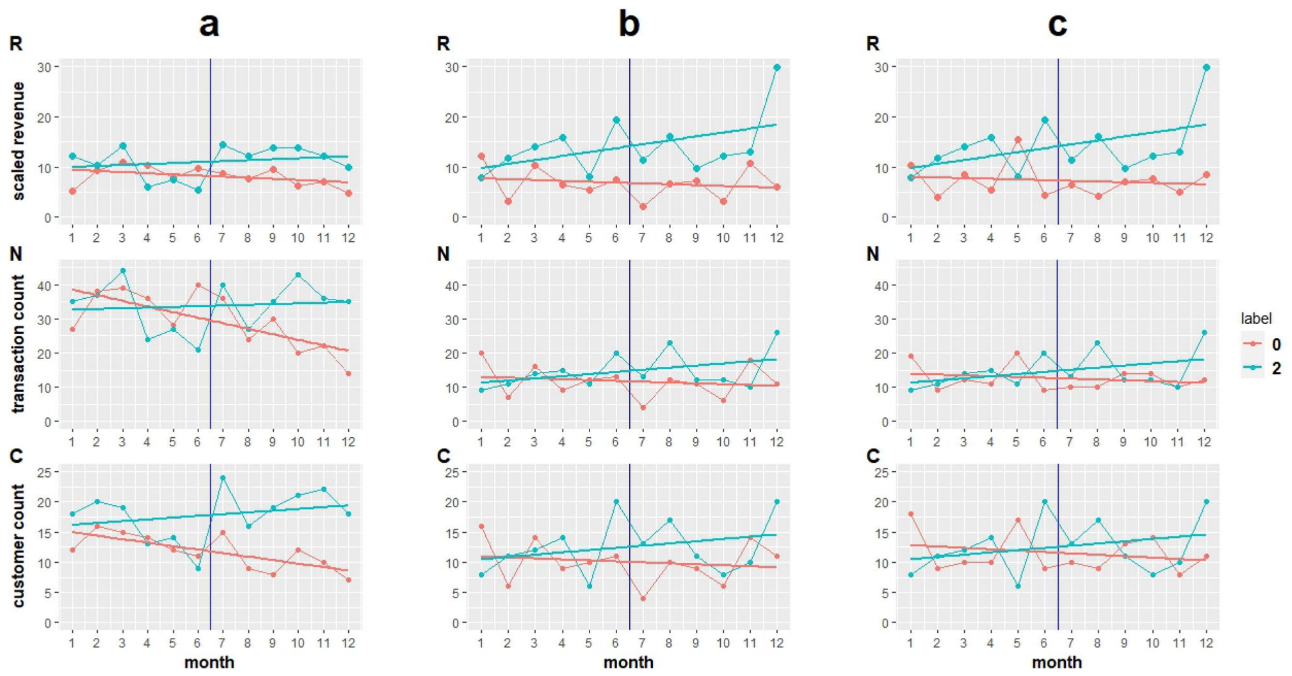

Figure 2. Each sub-figure (i.e., a, b, and c) depicts three time series plots for different merchant pairs with the same quartiles of revenue, transaction count, and unique customers count in the first period, but opposite labels. In each sub-figure the time series plots show monthly revenue (top: R), monthly transaction count (middle: N), and monthly unique customer count (bottom: C) including the OLS regression line for each merchant over the course of 1 year. In the figure legend, 0 corresponds to merchants labeled as poorly-performing, and 2 corresponds to those labeled as well-performing.

$$I^N_{\beta_1} \rightarrow \begin{cases} 1 & if \quad \beta_1^{N_{well-performing}} \geq \beta_1^{N_{poorly-performing}} \\ 0 & otherwise \end{cases} \tag{15}$$

$$I^S_{\beta_1} = I^R_{\beta_1} + I^C_{\beta_1} + I^N_{\beta_1} \tag{16}$$

Table 5 summarizes the result of this analysis. It is evident that in more than 90% of the pairs, the merchant labeled as well-performing demonstrates superior long-term and overall performance across all three factors: revenue, transaction count, and the number of unique customers. These results affirm the robustness and high accuracy of our proposed labeling approach in effectively distinguishing merchants based on their comprehensive performance measured by three distinct objective criteria.

District level analyses. In this section, we conduct three statistical analyses to explore potential relationships between merchants' performance and their geographical location (i.e., district).

Correlation analysis. Initially, we calculate the proportion or share of each performance group (i.e., well-performing, medium-performing, and poorly-performing) within each district based on their respective labels. These shares represent the probability of a merchant belonging to each performance class within a particular district. Furthermore, we compute the relative ratios of performance class pairs for each district. Subsequently, we conduct a correlation analysis to examine the potential associations between the probabilities and relative ratios of performance classes with the population and average household income of their corresponding districts. The results reveal no correlations between the performance class of merchants and the population size or income level of the residents in the districts where they are located (correlation table is provided in "Supplementary Information").

Chi-squared test. This test is performed between district ids as categorical variable (identification numbers) and performance class probabilities converted into categorical variables. This type of test is valid here as the sample size is small (33 districts) and the contingency table test is reasonable. Based on the results ($\chi^2(2560, n = 99) = 2629, p = 0.167$), there are no observed dependencies between the performance class probability distributions and relative ratios with the district ids indicating that within our dataset and based on the defined labels, a merchant's performance is independent of the district in which it is situated.

While based on Eq. (17) the inequality value can range from 0 (perfectly equal) to 1 (fully unequal), the histogram and the distribution range presented in Table 6 reveal that, in district-level distributions, the inequality scores tend to concentrate towards lower values, suggesting a trend towards reduced inequality. It is important to note that the maximum inequality index value obtained here (i.e., 0.388), is even less than 40% of the maximum possible value. Furthermore, the other statistical measures, such as median, mean, and interquartile range, further support the observed trend of the distribution favoring low inequality values.

In conclusion, the findings obtained from the three aforementioned analyses validate that there is no discernible dependency or substantial association between the defined performance labels for merchants and the districts in which they are situated.

Label distribution inequality analysis. For this analysis, we use the formulation presented in Eq. (17) to measure the inequality[52] in performance class labels' distributions within and among districts. This equation formally represents the inequality of labels' distribution in district i (denoted by $Inequality^i_L$) as a function of shares of the three labels in that district (denoted by $p^i_{L_k}$).

Sum Indicator ($I^S_{\beta_1}$)	Merchant pair count	Merchant pair percentage
0	160	1.35%
1	239	2.02%
2	673	5.69%
3	10,741	90.92%

Table 5. Number and percentage of merchant pairs by their sum indicators.

Range	Mean	Median	Standard deviation	Inter-quartile range
[0.042 , 0.388]	0.188	0.164	0.094	0.11

Table 6. Basic statistics of the label inequality measures at districts level.

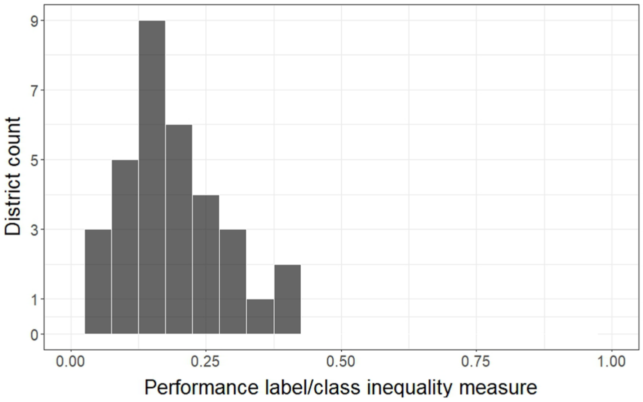

Figure 3. District level label inequality histogram.

$$Inequality_L^i = \frac{3}{4} \times \sum_{k=1}^{3} |p_{L_k}^i - \frac{1}{3}| \tag{17}$$

$$(Inequality_L^i \in [0, 1])$$

Since by definition, there are 3 performance labels ($k \in 1, 2, 3$), then the classes in a totally equal distribution should have the same share of the merchants in a district, that is equal to one third (i.e., $\frac{1}{3}$) share for each class, which results in 0 inequality. Moreover, a completely unbalanced distribution happens when the share of one particular class is equal to 1 and the other two classes have 0 shares. This distribution will yield the highest possible inequality value, which is equal to 1. Table 6 provides the basic statistics of the label inequality measures at a district level, while Fig. 3 illustrates the corresponding histogram.

Customer level analysis. To explore potential dependencies between a merchant's performance and the income level of its customers, we categorize the mean and median income of all merchants' customers into quartiles based on their respective distributions. This results in four categories (i.e., quartiles) for merchants based on the mean and median income of their customers. Subsequently, we conduct two chi-squared tests to examine the relationship between a merchant's performance label and the quartiles of its customers' mean and median income. The results of these tests indicate no significant dependencies between a merchant's performance class and the mean ($\chi^2(6, n = 1977) = 5.2951, p = 0.506$) or median ($\chi^2(6, n = 1977) = 2.8613, p = 0.826$) income of its customers. Additional analysis and the corresponding results regarding the customer features and their role in generating the merchant network of study are presented in the "Supplementary Information".

Based on the findings obtained from both the district-level and customer-level analyses, it can be concluded that the defined labels are independent of and unbiased towards the locations of merchants or the socio-economic status of their customers. This reaffirms the robustness and reliability of the labeling approach proposed and used in the study.

Predictive analysis. For each merchant, we gather 25 features to serve as inputs for the machine learning models. Out of these 25 features, 5 are based on the merchant's information, 3 are related to revenue, 11 are obtained from customer data, and 6 are extracted from the proposed network structure. The obtained merchant network consists of 2,011 nodes and 217,422 edges as one strongly connected component. "Supplementary Information" provides additional details concerning the network and node features. In addition to the 6 features directly obtained from the network structure, we generate a feature vector with 128 dimensions for each merchant (node) using the node2vec method.

Then we use different combinations of the feature sets as inputs for the four selected machine learning models. Table 7 summarizes the results for the classifiers we include in our analyses. Among those, Random Forest (RF) performs better than the other classifiers in most cases. This is in line with the previous findings that show tree-based models, and in particular RF, perform better in multi-class performance prediction tasks[53–55].

The computational results shown in Table 7 indicate that the performance of the network-based feature set is comparable to and almost as good as other feature sets with all classifiers. Moreover, in certain scenarios, the models utilizing feature vector representation (i.e., node2vec) of merchants within the suggested network demonstrate superior performance compared to those that solely use customer-based features. Nevertheless, the node2vec features exhibit no improvement over the network-based features, potentially due to information loss during the embedding process.

It is worth emphasizing that the inclusion of network-based features alongside the conventional feature sets enhances prediction accuracy. This suggests that certain signals, which are not captured by revenue-based and customer-based information, can be captured through features associated with a merchant's position within the proposed merchant network.

In order to demonstrate the impact of each feature on prediction accuracy, we perform a feature importance analysis using the mean decrease in accuracy. This analysis reveals the reduction in accuracy that occurs when a particular feature is absent. Figure 4 displays feature importance rankings for the RF classifier in accordance with the mean decrease in the classifier's prediction accuracy when a feature is removed after a random permutation. With the mean decrease in accuracy measures, financial features such as transaction count, revenue, and the number of unique customers, occupy the top ranks as they signify their role in accurate predictions. Among the network-based features, degree centrality, followed by betweenness and closeness centrality scores, cause the greatest deficits in accuracy. Notably, among the top ten important variables based on the mean decrease in accuracy, four belong to the network-based features.

The results of analyses shown in Table 7 and Fig. 4 demonstrate that in some prediction results as well as the feature importance ranking cases, some features from different feature sets possess close performance levels. This similarity can be the result of high correlation between the features from different feature sets, which also indicates that those features could carry similar signals and information about the merchants but obtained in different ways. For instance, features such as node degree, revenue, and the number of customers, originating from different feature sets, demonstrate significant pairwise correlations and closely align in their importance ranking. The correlation table of the features used in this study is available in "Supplementary Information".

In addition, we employ two different methods to further explore the relationship between labels and extracted features in our study. Firstly, we conducted principal component analysis (PCA)[56] on the originally extracted features, as presented in Table 2. Secondly, we utilize T-distributed stochastic neighbor embedding (t-SNE)[57] on the node2vec features. These methods were employed to visualize the placement of merchants in a two-dimensional (2D) space. The resulting 2D plots do not exhibit any discernible distribution or clustering patterns for merchants based on their performance classes. For more comprehensive information, including detailed plots, please refer to the "Supplementary Information".

Feature set	NB	SVM	LR	RF
(A) Revenue	0.565	0.557	0.587	0.586
(B) Customer	0.558	0.561	0.579	0.571
(C) Network	0.561	0.556	0.575	0.579
(D) node2vec	0.537	0.556	0.567	0.572
(A) + (B)	0.566	0.566	0.588	0.591
(A) + (B) + (C)	0.569	0.567	0.598	0.609

Table 7. AU-ROC produced by four classifiers using different combinations of input feature sets.

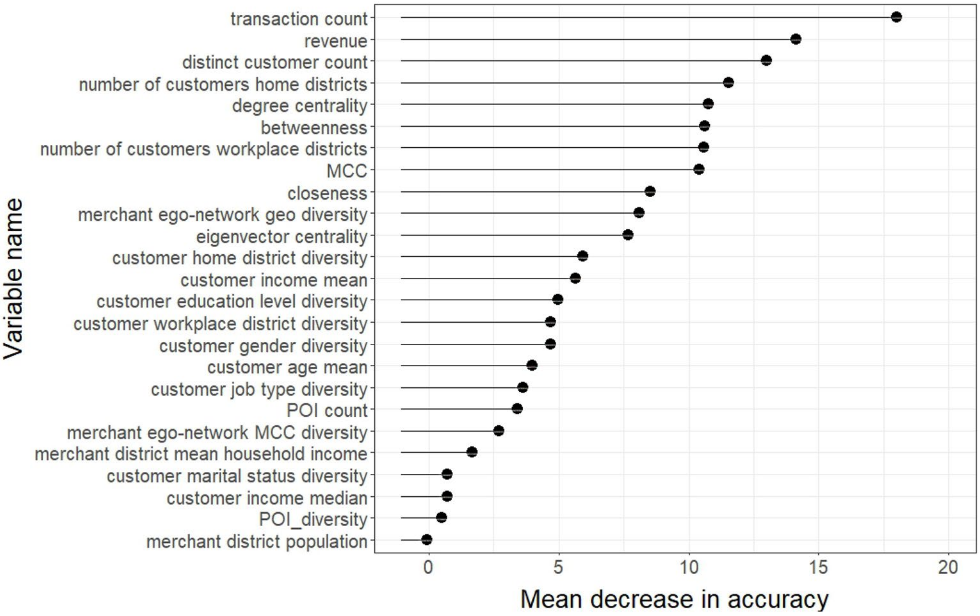

Figure 4. Feature importance ranking for Random Forest classifier based on the mean decreasing values of accuracy after randomly permuting the relevant feature. A higher decreasing value of a feature indicates more contribution.

Privacy implications. Within a highly competitive environment, the proprietary financial information of merchants assumes a heightened level of sensitivity. Consequently, the custodians of such data (e.g., banks), exercise caution in divulging such confidential information to external entities. However, financial institutions and investors find themselves compelled to evaluate the stability of merchants and predict their future performance in order to make informed determinations regarding business loans and investments. As a result, accessing this confidential information becomes imperative for these stakeholders. However, given the context of data sharing, safeguarding merchants' revenue- and customer-related information from unauthorized disclosure assumes paramount importance and is of utmost priority.

Our methodology offers a twofold advantage. Firstly, the network-based features generated through our approach exhibit prediction accuracy on par with conventional feature sets. Secondly, these features are presented in a tabular format, enhancing privacy beyond that offered by merely anonymizing raw financial records. This unique property enables the establishment of a more secure and expeditious data-sharing process with third-party organizations. Such a mechanism holds immense potential for enhancing data privacy while facilitating efficient collaboration and knowledge exchange.

The nature of network-based features, which primarily reveal the interconnectedness within the merchant network, presents a challenge in directly inferring and estimating customer- and revenue-related information in the absence of raw data or relevant statistical measures (e.g., average spending per customer transaction) which are never released by a financial institution that owns the raw data. Nevertheless, due to the significant correlation between certain financial or customer-related variables (e.g., revenue and the number of customers) and network-based features (e.g., node degree), there is a possibility of inferring sensitive information, such as revenue ranking and comparison.

To address this issue, the network data owner has the option to generate node2vec features locally and share them with third parties for the downstream task, specifically merchant performance classification. This approach is justifiable, as demonstrated by the outcomes presented in Table 7, wherein node2vec features exhibit comparable results to network-based features while ensuring a higher level of privacy.

While there exist adversarial attack methods aimed at reconstructing the graph from the node embeddings of the original graph, it is important to note two significant limitations. Firstly, these methods are unable to fully recover the original graph, thereby compromising the reliability of the reconstructed network in providing accurate insights about merchants. Secondly, data holders can effectively mitigate privacy risks associated with integrating node embeddings for downstream analysis by employing suitable defense mechanisms.

In the realm of defense mechanisms, two widely employed approaches are perturbation of the node2vec matrix, and generating embeddings in lower dimensions[58,59]. While these methods effectively mitigate inference attacks' accuracy, they come at the expense of degrading the accuracy of merchant performance prediction tasks. To tackle this issue, one can employ a tentative defense mechanism that involves an iterative process of removing the least significant feature vector (i.e., column) from the node2vec matrix until no substantial changes are observed in the classification accuracy. This technique is both straightforward and effective, while ensuring that our proposed approach does not compromise the accuracy of the downstream classification task.

Discussion

Although SME merchants contribute significantly to employment and economic activity, historical data indicates that only approximately 50% of them manage to sustain their businesses beyond the initial five years. Economic downturns and financial crises can further compound the challenges faced by SMEs, leading to a significant negative impact on the business continuity of the majority of SME merchants.

Furthermore, SME merchants heavily rely on business financing and external investments. Given the high failure rate among SMEs, banks and other financial institutions face a dire need for effective models and tools to assess the current state and predict the future performance of their clients. Complicating matters, the privacy concerns surrounding SMEs often result in the unavailability of their financial data and reports for sharing with external entities, and therefore, the data-driven methods and tools become ineffective and useless when the necessary data is not accessible. This fact underscores the significance and urgency of developing methods for sharing information swiftly and securely, without compromising privacy.

To tackle these challenges, one potential solution is to utilize credit card transaction information, which is accessible to banks and select financial institutions. In this study, we offer insights to banks and financial institutions regarding the utilization of credit card transaction data in order to enhance the accuracy of predicting a merchant's future performance. We present a new approach that involves constructing a merchant network, based on customer co-purchase patterns, derived from credit card transaction data. This network structure possesses distinct characteristics that enable the extraction of various signals, including four different centrality measures. These measures indicate the position of a specific merchant within the broader network of merchants and its level of connection to other merchants, whether from the same or different business categories within its ego network. Notably, we discover that these features provide sufficient signals to predict a merchant's future business performance. This approach fundamentally differs from the commonly employed methods found in current state-of-the-art techniques.

Our computational results show that the network-based features are capable of revealing new information about a merchant's performance level that is on par with more straightforward measures such as demographic characteristics and composition of customers visiting the merchant, and financial indicators of the merchant such as total revenue or the number of transactions generated. We find that, with the use of effective machine learning modeling techniques demonstrated in this study, analysts can more accurately predict the future performance level of merchants using the features extracted from the proposed network. Such predictions are important for financial institutions because a major portion of their clients are SMEs, which may exhibit erratic financial behavior and performance, and may not always disclose the details of their financial records. Considering the substantial number of SME merchants and their elevated default rate, even marginal improvements in prediction accuracy can significantly aid financial institutions in mitigating risks, particularly those with limited budget allocations.

The network-based and node2vec features not only exhibit comparable performance to conventional feature sets (i.e., revenue- and customer-based) in predicting merchants' future performance, but they also offer a higher level of privacy when presented in a tabular format, as opposed to raw financial records. This privacy-enhancing characteristic enables the establishment of a secure and efficient information-sharing process with third-party organizations. Although inferring comprehensive financial and customer-related information solely from network-based features presents challenges, certain financial and customer indicators exhibit correlations with these features, making estimations like revenue ranking feasible. As a result, generating node2vec embeddings can facilitate safer sharing. It is worth noting that while inference attacks targeting node2vec embeddings can reconstruct parts of the original network, their success is limited. Data holders can further mitigate privacy risks associated with these attacks by employing the defense mechanisms such as those discussed in this paper.

Our study does have certain limitations that should be acknowledged. One limitation stems from our reliance on on-premise credit card transactions for constructing and validating our models. It is important to note that a portion of consumers opt for cash or online transactions, which are not included in our datasets. The number of such transactions may be non-negligible, especially in the country under study. This issue becomes even more pronounced in economies with less transparency and a higher prevalence of informal businesses that primarily operate on a cash basis. In such economies, an alternative approach could involve utilizing mobility data and visit patterns collected passively from users' cellphones to create the merchant network[60]. However, despite these limitations, we believe that our dataset, which comprises over 2 million transactions, captures a substantial portion of the economic activity during the specified time period. Moreover, it is worth noting that there is a significant correlation between cash and card spending in larger economies[61].

Our study possesses another limitation concerning the absence of a clear indicator or identifier within our dataset to differentiate SMEs from non-SME merchants. While certain variables such as transaction volume or revenue may suggest the classification of certain merchants as SMEs, we intentionally refrained from using such designations. As a result, our prediction models are aimed at all merchants collectively. Undoubtedly, tailoring our models specifically for predicting SME performance could have potentially enhanced their prediction accuracy. However, it is crucial to acknowledge that the network structure in which SMEs operate also involves interactions with non-SME merchants. Nonetheless, future research endeavors could focus on acquiring a dataset where SMEs can be accurately identified, allowing for the development of prediction models specifically tailored to their unique characteristics.

To summarize, our study has the following contributions. Our first contribution is anchored on the current literature about merchant risk and bankruptcy prediction using machine learning models. We build on these theories to introduce a new method in labeling merchant performance levels, which takes into account three financial objective measures considering the dynamic nature of business performance in a highly competitive environment over time.

The main contribution of this study is inspired by computational social science literature, and is built on the premise that just as we humans live our lives in networks [23], merchants also run their businesses in networks. Accordingly, we propose a novel approach to build a network of merchants based on their customer credit card transactions, and use the extracted features from the merchant network structure for predicting their future performance. While we show that the network-based features improve the prediction accuracy when added to the conventional revenue-based and customer-based features, they also possess a higher privacy level in comparison to the conventional feature sets, facilitating more secure and efficient data sharing among financial entities.

Our approach and methodology can further be incorporated into a decision support system or integrated with a legacy system along with exploratory visual analytics tools[62] to be safely shared and used by various organizations including banks and financial institutions. Risk analysts and decision makers can utilize our models to rank their clients in order of decreasing estimated future risk and flag those that rank lower to take further action. The system can be configured to be a "learning" system that is continuously fed with freshly incoming transaction data and the models can periodically be re-trained to account for changes in trends and customer behavior.

Our proposed approach offers banks and financial institutions valuable new insights into predicting the future performance of their merchant clients. By leveraging network-based analysis, these institutions can enhance their ability to assess loan risks and identify investment opportunities. This study serves as an initial step toward exploring the potential of network-based methodologies for assessing merchant performance. It also highlights the possibility of developing methods that enable businesses to share information derived from data without the need to share the actual data itself.

Data availability

All pre-processed data and code necessary to replicate this study and reproduce this work are available at: https://github.com/alppboz/credit-card-research Additional analyses details and results are available in Supplementary Information of this article.

Received: 29 January 2023; Accepted: 7 June 2023
Published online: 21 June 2023

References

1. Commission, E. et al. Annual report on European SMEs 2021/2022 : SMEs and environmental sustainability: Background document (Publications Office of the European Union, 2022).

2. U.S. Small Business Administration Office of Advocacy. 2018 Small Business Profile. https://www.sba.gov/sites/default/files/advocacy/2018-Small-Business-Profiles-US.pdf (2018). Accessed: 2023-01-14.
3. U.S. Small Business Administration Office of Advocacy. Frequently Asked Questions About Small Business, 2021. https://advocacy.sba.gov/2021/11/03/frequently-asked-questions-about-small-business-2021/ (2021). Accessed: 2023-01-14.
4. Plattner, D. Why firms go bankrupt. The influence of key financial figures and other factors on the insolvency probability of small and medium sized enterprises. *KfWResearch* **28**, 37–51 (2002).
5. Berger, A. N. & Frame, W. S. Small business credit scoring and credit availability*. *J. Small Bus. Manage.* **45**, 5–22 (2007).
6. Yoon, J. S. & Kwon, Y. S. A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information. *Expert Syst. Appl.* **37**, 3624–3629 (2010).
7. Ciampi, F. & Gordini, N. Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises. *J. Small Bus. Manage.* **51**, 23–45 (2013).
8. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 855–864 (2016).
9. Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H. & Wu, S. Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decis. Support Syst.* **37**, 543–558 (2004).
10. Fantazzini, D. & Figini, S. Random survival forests models for SME credit risk measurement. *Methodol. Comput. Appl. Probab.* **11**, 29–45 (2009).
11. Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y. & Ryu, K. H. An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability* **11**, 699 (2019).
12. Kim, S. Y. & Upneja, A. Majority voting ensemble with a decision trees for business failure prediction during economic downturns. *J. Innov. Knowl.* **6**, 112–123 (2021).
13. Berger, A. N. & Frame, W. S. Small business credit scoring and credit availability. *J. Small Bus. Manage.* **45**, 5–22 (2007).
14. Chi, G. & Meng, B. Debt rating model based on default identification: Empirical evidence from Chinese small industrial enterprises. *Manage. Decis.* **57**, 2239–2260 (2018).
15. Christopoulos, A. G., Dokas, I. G., Kalantonis, P. & Koukkou, T. Investigation of financial distress with a dynamic logit based on the linkage between liquidity and profitability status of listed firms. *J. Oper. Res. Soc.* **70**, 1817–1829 (2019).
16. Gallucci, C., Santullli, R., Modina, M. & Formisano, V. Financial ratios, corporate governance and bank-firm information: A Bayesian approach to predict SMEs' default. J. Manage. Govern. 1–20 (2022).
17. Son, H., Hyun, C., Phan, D. & Hwang, H. J. Data analytic approach for bankruptcy prediction. *Expert Syst. Appl.* **138**, 112816 (2019).
18. Tang, T. T. Information asymmetry and firms' credit market access: Evidence from Moody's credit rating format refinement. *J. Financ. Econ.* **93**, 325–351 (2009).
19. Te, Y.-F. Predicting the Financial Growth of Small and Medium-Sized Enterprises using Web Mining. Doctoral Thesis, ETH Zurich (2018).
20. Fernandes, G. B. & Artes, R. Spatial dependence in credit risk and its improvement in credit scoring. *Eur. J. Oper. Res.* **249**, 517–524 (2016).
21. Pentland, A. *Social Physics: How Social Networks can make us Smarter* (Penguin, 2015).
22. Ball, P. *Why Society is a Complex Matter: Meeting Twenty-First Century Challenges with a New Kind of Science* (Springer Science & Business Media, 2012).
23. Lazer, D., Brewer, D., Christakis, N., Fowler, J. & King, G. Life in the network: The coming age of computational social. *Science* **323**, 721–723 (2009).
24. Dong, X. *et al.* Social bridges in urban purchase behavior. *ACM Trans. Intell. Syst. Technol.* **9**, 1–29. https://doi.org/10.1145/3149409 (2017).
25. Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E. & Pentland, A. Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task. Available at SSRN 1130251 (2008).
26. Granovetter, M. The impact of social structure on economic outcomes. *J. Economic Perspect.* **19**, 33–50 (2005).
27. Perc, M. Diffusion dynamics and information spreading in multilayer networks: An overview. *Eur. Phys. J. Spl. Topics* **228**, 2351–2355 (2019).
28. Reagans, R. & Zuckerman, E. W. Networks, diversity, and productivity: The social capital of corporate R &D teams. *Organ. Sci.* **12**, 502–517 (2001).
29. Chong, S. K. *et al.* Economic outcomes predicted by diversity in cities. *EPJ Data Sci.* **9**, 17 (2020).
30. Alvarez-Rodriguez, U. *et al.* Evolutionary dynamics of higher-order interactions in social networks. *Nat. Hum. Behav.* **5**, 586–595 (2021).
31. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
32. Organisation for economic co-operation and development. https://www.oecd.org/. Accessed: 2023-01-14.
33. ISO 18245 Merchant Codes. https://www.iso.org/standard/33365.html. Accessed: 2023-01-14.
34. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
35. Bianconi, G., Pin, P. & Marsili, M. Assessing the relevance of node features for network structure. *Proc. Natl. Acad. Sci.* **106**, 11433–11438. https://doi.org/10.1073/pnas.0811511106 (2009).
36. Perc, M. The social physics collective. *Sci. Rep.* **9**, 16549 (2019).
37. Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How correlated are network centrality measures ?. *Connections* **28**, 16 (2008).
38. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1**, 215–239 (1978).
39. Wasserman, S. *et al. Social Network Analysis: Methods and Applications* Vol. 8 (Cambridge University Press, 1994).
40. Bonacich, P. Technique for analyzing overlapping memberships. *Sociol. Methodol.* **4**, 176–185 (1972).
41. Golbeck, J. *Analyzing the Social Web* (Newnes, 2013).
42. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
43. Singh, V. K., Bozkaya, B. & Pentland, A. Money walks: Implicit mobility behavior and financial well-being. *PLOS ONE* **10**, 1–17. https://doi.org/10.1371/journal.pone.0136628 (2015).
44. Leković, B. & Marić, S. M. Measures of small business success/performance—Importance, reliability and usability. *Industrija* **43** (2015).
45. Mariooryad, S. & Busso, C. The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier. *IEEE Trans. Affect. Comput.* **8**, 119–130 (2015).
46. Anderson, E., Lin, S., Simester, D. & Tucker, C. Harbingers of failure. *J. Market. Res.* **52**, 580–592 (2015).
47. Simester, D. I., Tucker, C. E. & Yang, C. The surprising breadth of harbingers of failure. *J. Market. Res.* **56**, 1034–1049 (2019).
48. Kaya, E., Alpan, E., Balcisoy, S. & Bozkaya, B. Quantifying insurance agency channel dynamics using premium sales big data and external factors. *Big Data* **9**, 116–131 (2021).
49. Netto, C. F. S. *et al.* Disaggregating sales prediction: A gravitational approach. *Expert Syst. Appl.* **217**, 119565. https://doi.org/10.1016/j.eswa.2023.119565 (2023).

50. Gu, W., Tandon, A., Ahn, Y.-Y. & Radicchi, F. Defining and identifying the optimal embedding dimension of networks. Preprint at arXiv:2004.09928 (2020).
51. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*, vol. 1. Springer Series in Statistics. (Springer, 2001).
52. Moro, E., Calacci, D., Dong, X. & Pentland, A. Mobility patterns are associated with experienced income segregation in large us cities. *Nat. Commun.* **12**, 1–10 (2021).
53. Finlay, S. Multiple classifier architectures and their application to credit risk assessment. *Eur. J. Oper. Res.* **210**, 368–378 (2011).
54. Jones, S., Johnstone, D. & Wilson, R. Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *J. Bus. Fin. Account.* **44**, 3–34 (2017).
55. Son, H., Hyun, C., Phan, D. & Hwang, H. Data analytic approach for bankruptcy prediction. *Expert Syst. Appl.* **138**, 112816 (2019).
56. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics Intell. Lab. Syst.* **2**, 37–52 (1987).
57. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 1–27 (2008).
58. Shen, Y. et al. Finding mnemon: Reviving memories of node embeddings. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp 2643–2657 (2022).
59. Zhang, Z., Chen, M., Backes, M., Shen, Y. & Zhang, Y. Inference attacks against graph neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, pp 4543–4560 (2022).
60. Solmaz, G. & Turgut, D. A survey of human mobility models. *IEEE Access* **7**, 125711–125731 (2019).
61. Chetty, R., Friedman, J. N., Hendren, N., Stepner, M. et al. The economic impacts of COVID-19: Evidence from a new public database built using private sector data. *National Bureau of Economic Research* (2020).
62. Boz, H. A., Bahrami, M., Suhara, Y., Bozkaya, B. & Balcisoy, S. An Exploratory Visual Analytics Tool for Multivariate Dynamic Networks. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pp 19–23, https://doi.org/10.2312/eurova.20201081 (2020).

Acknowledgements

The authors are grateful to the financial institution that provided the credit card transaction data for this research. The authors are also indebted to Professor Vivek Singh and Julius Adebayo for their constructive comments and great contribution to idea generation and primary analysis. The authors are indebted to the anonymous reviewers for their remarkable contribution to the content and quality of this paper. Hasan Alp Boz would like to thank TUBITAK (Scientific and Technological Research Council of Turkey) for their support.

Author contributions

M.B., H.B., and Y.S. were involved in idea generation, design and implementation of analyses, drafting the article, and writing. B.B. and A.P. were involved in idea generation, results, discussion, feedback, and final revision of the article. Whereas S.B., B.B., and A.P. were involved in discussion, feedback, and drafting and final revision of the article. All authors reviewed the manuscript. M.B., H.B., made equal contributions to this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-36624-0.

Correspondence and requests for materials should be addressed to M.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.