MIS Quarterly

RESEARCH NOTE

PEER-TO-PEER LOAN FRAUD DETECTION: CONSTRUCTING FEATURES FROM TRANSACTION DATA[1]

Jennifer J. Xu
Department of Computer Information Systems, Bentley University
Waltham, MA, U.S.A. {jxu@bentley.edu}

Dongyu Chen
Research Center for Smarter Supply Chain, Business School, Soochow University,
Suzhou, Jiangsu, CHINA {chendongyu@suda.edu.cn}

Michael Chau
Faculty of Business and Economics, The University of Hong Kong
Pokfulam, HONG KONG {mchau@business.hku.hk}

Liting Li and Haichao Zheng
School of Management Science and Engineering
Southwestern University of Finance and Economics
Chengdu, Sichuan, CHINA {litingli-swufe@foxmail.com} {haichao@swufe.edu.cn}

Although financial fraud detection research has made impressive progress because of advanced machine learning algorithms, constructing features (or attributes) that can effectively signal fraudulent behaviors remains a challenge. In recent years, a new type of fraud has emerged on peer-to-peer (P2P) lending platforms, where individuals can borrow money from others without a financial intermediary. In these markets, the information asymmetry problem is seriously elevated. Inspired by the fraud triangle theory and its extensions, and using the design science research methodology, we construct five categories of behavioral features directly from P2P lending transaction data, in addition to the baseline features regarding borrowers and loan requests. These behavioral features are intended to capture the fraud capability, integrity, and opportunity of fraudsters based on their loan requests and payment histories, connected peers, bidding process characteristics, and activity sequences. Using datasets from real users on two large P2P lending platforms in China, our evaluation results show that combining these additional features with the baseline features significantly enhances detection performance. This design science research contributes novel knowledge to the financial fraud detection literature and practice.

Keywords: Feature construction, fraud detection, peer-to-peer lending, fraud triangle theory, machine learning

Introduction

Peer-to-peer (P2P) lending markets are usually hosted on online platforms, where individuals can acquire loans directly from strangers at arm's length without using an intermediate financial institution (e.g., a bank). In the United States, the two largest P2P platforms, Prosper and Lending Club, have facilitated a total of more than $78 billion in loans. In China, more than 6,600 P2P platforms have been launched as of 2021, originating approximately ¥9 trillion (approximately $1.4 trillion) loans (Forward View, 2021).

[1] H. Raghav Rao was the accepting senior editor for this paper. Shuk Ying Ho served as the associate editor. Dongyu Chen is the corresponding author.

P2P loan fraud occurs when a borrower deliberately creates a loan request to cheat lenders out of their money. To protect lenders from financial losses and ensure the orderly functioning of P2P lending markets, effective fraud detection measures must be developed. However, such a task is challenging. First, the information asymmetry problem (Akerlof, 1970) is seriously elevated in P2P lending markets. Unlike banks and other financial institutions, P2P platforms and individual lenders can only access very limited information about a borrower's background, financial status, and credibility. It is very difficult to find indicators for fraudulent loans with incomplete information. Second, scant research has been conducted on fraud detection in the context of P2P lending. Previous studies on the detection of traditional fraud (e.g., credit card fraud and corporate fraud) offer limited guidance for this relatively new phenomenon. Although fraud always involves some sort of information manipulation, specific fraudulent behaviors may be quite complex by nature and may vary significantly from domain to domain, and even from case to case. With the exception of domain-independent techniques such as statistical and machine learning algorithms, features[2] that can be used for fraud detection remain highly domain specific (Vlasselaer et al., 2017). For example, features that capture individuals' purchase patterns in credit card fraud detection (Bahnesen et al., 2016) are completely irrelevant to the detection of financial fraud in corporations, which needs to rely on various organizational and industry-level financial indices (Abbasi et al., 2012). Therefore, features that are effective in fraud detection in some domains are not necessarily useful for and directly applicable to P2P loan fraud detection. Unfortunately, no guidelines exist in the literature for feature construction for P2P loan fraud detection.

In this research, we propose five categories of behavioral features, in addition to baseline features (e.g., borrower demographics and loan interest rate). These behavioral features are related to some of the fraud antecedents and factors identified in the fraud triangle theory and its extensions (Albrecht et al., 2006; Cressey, 1953; Wolfe & Hermanson, 2004), and can be used to glean more information from lending transactions to mitigate information asymmetry. In particular, features that capture borrower activity sequences have never been used in previous financial fraud detection research. The key research question is: Are these proposed features effective for detecting fraudulent loan requests in P2P lending markets? We tested these features using large datasets from two leading P2P platforms in China and found the answer to our inquiry to be positive.

[2] Features are attributes (or variables) that describe the properties or characteristics of the phenomenon under study.

Literature Review

Financial Fraud Detection

Research on financial fraud detection has addressed the problem from two complementary angles: (1) method and *algorithm* development, and (2) *feature* identification and construction. Because fraud detection is primarily a classification problem, a large body of literature has focused on the development of effective machine learning algorithms (see Table 1). Although these algorithms are typically domain independent, the features used in fraud detection are usually domain specific. In credit card fraud detection, for instance, the feature set often includes a customer's demographic information, socioeconomic status, and spending patterns (Bahnesen et al., 2016). In corporate fraud detection, organizational and industry-level features (e.g., asset quality index and gross margin index) have been used to boost detection performance (Abbasi et al., 2012).

Fraud is also found in cyberspace (Liao et al., 2017). Fraudulent behaviors in online marketplaces (e.g., eBay), such as shill bidding, have long been an issue. Shill bidding usually occurs in English auctions in which fraudsters make spurious bids on an item to artificially drive its final price up (Trevathan & Read, 2012). These perpetrators may be the sellers themselves or their accomplices. In the e-commerce literature, features related to the auction process have been proposed to distinguish between shill bids and legitimate bids (Majadi et al., 2017).

A common challenge in fraud detection research is the skewed class distribution problem, which occurs when there are significantly more legitimate instances than fraudulent instances in the dataset. To address this problem, many financial fraud detection studies have used the undersampling approach to create balanced samples by randomly drawing the same number of legitimate and fraudulent instances from their datasets (see Table 1). The drawback of this approach is that classifiers trained on balanced samples often perform poorly when used to classify imbalanced samples (He & Garcia, 2009). Table 1 summarizes the feature sets, classification methods (and/or algorithms), datasets, and performance metrics in recent exemplary studies in the financial fraud detection literature.

Table 1. Exemplary Studies on Financial Fraud Detection

Study	Fraud type	Sample	Feature set	Classification method(s)	Results
Abbasi et al. (2012)	Financial statement fraud	9,006 firm-years, 815 frauds	420 features: 84 yearly, 336 quarterly	Meta-fraud framework with multiple classifiers	Recall: 0.83 Precision: 0.41 AUC: 0.93
Bahnesen et al. (2016)	Credit card fraud	236,735 transactions, 3,551 frauds	293 features: 12 raw, 1 periodic, 280 aggregate	Decision tree, logistic regression, random forest	>200% increase in savings compared to using only raw features
Bhattacharyya et al. (2011)	Credit card fraud	8,164–60,319 transactions, fraud rate: 2%–15%	24 features: 8 raw, 16 derived numerical features	Logistic regression, SVM, random forest	Recall: 0.25-0.81 Precision: 0.07-0.61 AUC: 0.82-0.93
Brown et al. (2020)	Financial misreporting	37,806 firm-years, fraud rate: 1.34%	40 features: 10 financial, 30 textual-style variables	LDA topic modeling, logistic regression	AUC: 0.59-0.71
Chang and Chang (2012)	Online auction fraud	312 accounts, 156 frauds	10 features: 9 seller rating related, 1 item related	Instance-based learners, decision tree	Recall: 0.87 Precision: 0.89 *F* score: 0.88
Dong et al. (2018)	Corporate fraud	128 firms, 64 frauds	105 features: 84 financial, 21 linguistic	SVM, neural network, decision tree, logistic regression	Accuracy: 0.90 Recall: 0.83 *F* score: 0.80 AUC: 0.85
Hajek and Henriques (2017)	Financial statement fraud	622 firm-years, 311 frauds	40 features: 8 linguistic, 32 financial ratios	Logistic regression, Bayesian classifiers, SVM, decision tree	Accuracy: 0.90 *F* score: 0.90 AUC: 0.98
Siering et al. (2016)	Crowdfunding project fraud	652 projects, 326 frauds	9 measures: 5 project, 4 user, 22 linguistic, 1 content-based	SVM, neural network, naive Bayes, KNN decision tree	Precision: 0.85 Recall: 0.67 *F* score: 0.75
Vlasselaer et al. (2017)	Social security fraud	230,000 companies, fraud rate: 0.09%-0.18%	21 network features: 17 direct, 4 indirect	Random logistic forest, random forest	AUC: 0.86–0.95

P2P Loan Fraud

A lending transaction typically starts with a user (*borrower*) creating a loan request (*listing*). The borrower must specify the borrowing amount and the interest rate they are willing to pay. Once approved by the platform, the listing is open for auction, during which other users (*lenders*) bid on the listing by making a pledge to contribute a certain amount. An auction typically remains active for several days. The listing is eligible to be materialized into a loan if it is successfully funded by the auction closure time. Each loan has a repayment term of up to 36 months, during which the borrower must make monthly payments according to the repayment schedule. A loan is considered to be in default if it is late for three or more months.

The risk of P2P loan fraud is rooted to a large extent in the very nature of P2P transactions. In the absence of financial intermediaries, the problem of information asymmetry (Akerlof, 1970) is substantially worsened. In traditional lending transactions (e.g., mortgage and automobile loans), the lender is typically a financial institution. With access to comprehensive records of a borrower's information (e.g., identity, income, and credit history), an institutional lender can often effectively predict the borrower's credibility and financial status. In the P2P context, however, lenders are individuals who do not have

access to information about a borrower's identity and credit history. A P2P platform, as an informational rather than financial intermediary, also has limited authority and privilege to obtain and verify a borrower's financial status. Consequently, individual lenders are highly vulnerable to loan fraud. Moreover, in countries such as China—which has no mature, nationwide credit system—measures such as credit scores may not be readily available for lenders to assess a borrower's credibility (Xu et al., 2015).

Fraud Triangle Theory and Extensions

Behavioral theories about fraud are generally lacking in the literature. The fraud triangle theory in the accounting and auditing literature identifies the conditions and antecedents for white-collar crimes. It posits that an individual may commit fraud when three conditions are present: *perceived pressure*, *opportunity*, and *rationalization* (Cressey, 1953). The perceived pressure can be financial (e.g., debts and credit crisis), personal (e.g., greed and lack of personal discipline), or social (e.g., employer expectations). Opportunity exists when there are weaknesses in the control, governance, or regulation systems. Rationalization is the perpetrator's self-justification for the offense. The fraud diamond model extends the fraud triangle by including a fourth element, that is, *capability*, to stress the importance of the traits, skills, or abilities that an offender must possess to be able to commit fraud (Wolfe & Hermanson, 2004). Other extensions of the fraud triangle theory include the following: the fraud scale model, which proposes using personal *integrity* instead of rationalization when investigating financial statement fraud (Albrecht et al., 2006); the MICE model, which decomposes pressure into *money*, *ideology*, *coercion*, and *ego* (Kranacher et al., 2011); and the SCORE model, which uses *stimulus* as an alternative factor for pressure (Vousinas, 2019). Overall, these models focus on four types of conditions and antecedents: pressure, opportunity, integrity (or self-rationalization), and capability.

Researchers have pointed out that these fraud models are useful for gathering prosecutorial evidence during post-offense investigations (Dorminey et al., 2012) or assessing fraud risks in fraud prevention, but are inadequate for fraud detection (Vousinas, 2019). They are primarily explanatory models that identify the reasons *why* individuals commit fraud, but not predictive models *per se*. As explanatory modeling and predictive modeling often have different goals and focuses (Shmueli & Koppius, 2011), the fraud triangle theory and its extensions are not necessarily directly applicable to predictive fraud detection. For example, self-rationalization is typically unobservable from the investigator's perspective and cannot be used to predict fraud in practice (Dorminey et al., 2012; Kassem & Higson, 2012; Vousinas, 2019). Despite this, these models may aid in the search for investigative leads and informational cues. For example, auditing professionals and regulators often assess the opportunities and capabilities of individuals or organizations for corporate fraud investigation and detection (Free & Murphy, 2015).

Our design and development of the behavioral features in this research are inspired by the fraud triangle theory and its extensions. Because observing and operationalizing the factors related to the pressure perceived by individual borrowers is fairly difficult in the context of our study, we propose features that relate to a borrower's *capability*, *integrity*, and *opportunity* to enhance detection performance. Note that pressure is not always unobservable in all contexts. For example, if available, the debt-to-income ratio can be a good indicator of perceived pressure.

P2P Loan Fraud Detection

We developed our P2P fraud detection artifact using the design science methodology (Hevner et al., 2004). Figure 1 presents the fraud detection process, which comprises data preparation, feature construction, training, and testing. During the first step (data preparation), empirical data are collected and preprocessed to create clean, labeled samples for supervised learning.

Feature Construction

Feature construction is the process of developing new attributes from raw data (Zheng & Casari, 2018). High-quality features can effectively capture the characteristics of the phenomenon under study. Hence, feature construction is a critical step in many machine learning tasks and often directly affects performance (Bahnesen et al., 2016; Ghaddar & Naoum-Sawaya, 2018). In P2P lending markets, the borrowers' demographic information (e.g., age and gender) and lending transaction attributes (e.g., borrowing amount and interest) are directly available, forming a set of baseline features. However, these baseline features are inadequate for capturing all the complex characteristics of loan transactions and the borrowers' behavioral characteristics. As a result, we identify and construct additional features to help detect fraudulent listings. In this research, we propose five additional categories of features: *borrowing history*, *payment history*, *connected peers*, *bidding process characteristics*, and *activity sequence*. These features are intended to reveal the *capability*, *integrity*, and *opportunity* of fraudsters based on the fraud triangle theory and its extensions (Albrecht et al., 2006; Cressey, 1953; Wolfe & Hermanson, 2004).

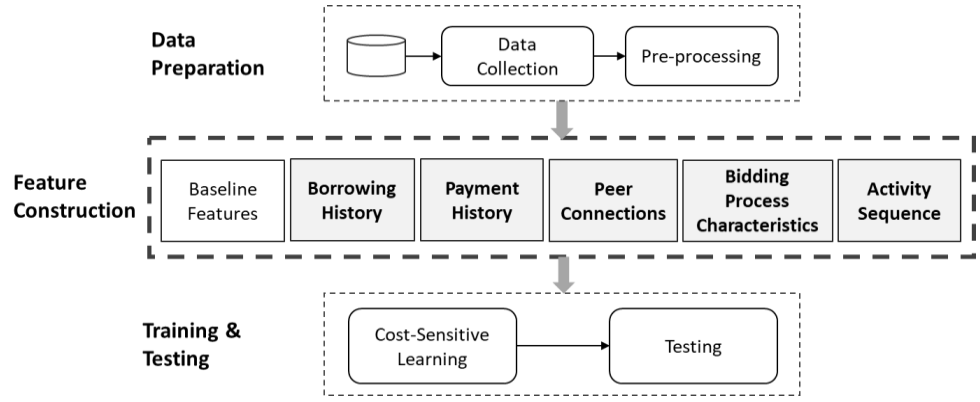

Figure 1. P2P Loan Fraud Detection Process

Borrowing History

From the perspective of a fraudster, the ultimate goal is to obtain as much money as possible with minimum costs. To commit fraud, a fraudster must have the skills and *capabilities* to get their listings funded. To learn and practice these skills and capabilities, the fraudster needs to test the waters and gain experience before committing any fraud. However, it is often difficult for a first-time user to successfully have a listing funded on a platform (Duarte et al., 2012; Lin et al., 2013; Zhang & Liu, 2012). Therefore, a fraudster may need to create a few listings and experiment with different parameter values (e.g., borrowing amount and interest rate) to learn the "best practices." Conversely, a large number of test listings can appear suspicious and reduce the chances of funding success. Moreover, if a test listing happens to be fully funded and materialized, the borrower must repay it with interest, thereby incurring extra costs. Therefore, we believe that features that capture a user's borrowing history can serve as indicators for fraud capability (not repayment capability). For each listing created by a borrower, we use four features to summarize the borrower's prior loan request activities: *n_prior_listings*, *amt_prior_listings*, *n_prior_materialized_loans*, and *amt_prior_materialized_loans*.[3] The definitions of these features can be found in Table 2.

Payment History

Although the true *integrity* traits of an individual are not directly accessible, a poor track record can be a red flag for a borrower's trustworthiness. We use four features to profile the payment history of a borrower (see Table 2): *n_prior_repaid_loans*, *amt_prior_repaid_loans*, *n_prior_delinquencies*, and *amt_prior_delinquencies*.

Connected Peers

Previous studies have reported that an individual's social network can, to a certain extent, affect their *capability* of getting a loan funded in P2P markets. For instance, endorsements made by a borrower's friends may serve as a quality signal and enhance the odds of funding success (Lin et al., 2013). Fraudsters may purposefully request or even hire their friends to bid on their listings to increase the number of endorsements and amount of support.

However, because of online user anonymity, identifying a user's friends is very difficult in practice. In our study, we constructed a network of users in which a link was created between a borrower and a lender if the lender had ever made a bid on the borrower's listings. We then counted the frequency of links between each pair of borrowers and lenders. A high-frequency value indicates that the lender has repeatedly bid on that particular borrower's multiple listings. By plotting the link frequency against the number of links with the frequency, we found that the power-law distribution curve drops dramatically when the frequency is greater than five. Specifically, 99.5% of the links have a frequency lower than five, and only 0.5% of the borrower-lender links reoccur five or more times. We thus define a borrower's (frequently) connected peer as a lender who has bid on five or more of the borrower's listings. In this study, we propose three features based on a borrower's connections (see Table 2): *borrower_degree*, *n_by_peers*, and *amt_by_peers*.

[3] We also used relative values (e.g., percentages) and the results were largely the same.

Table 2. Proposed Features and their Definitions

Category	Feature name	Definition (coding scheme)
Borrowing history	n_prior_listings	The total number of loan requests listed by the borrower previously
	amt_prior_listings	The total loan amount requested by the borrower previously
	n_prior_materialized_loans	The total number of loan requests listed by the borrower that were fully funded
	amt_prior_materialized_loans	The total loan amount requested by the borrower that was fully funded
Payment history	n_prior_repaid_loans	The number of loans fully repaid by the borrower
	amt_prior_repaid_loans	The total amount of repayments made by the borrower
	n_prior_delinquencies	The number of delinquent monthly payments by the borrower
	amt_prior_delinquencies	The total amount of delinquent monthly payments by the borrower
Connected peers	borrower_degree	The number of lenders who bid on the borrower's prior listings
	n_by_peers	The number of bids casted by the borrower's peers (frequently connected lenders) on the current listing
	amt_by_peers	The total amount of bids by the borrower's peers on the current listing
Bidding process characteristics	n_first_hour	The total number of bids during the first hour on a listing
	amt_first_hour	The total amount of bids during the first hour on a listing
	n_first_day	The total number of bids during the first day on a listing
	amt_first_day	The total amount of bids during the first day on a listing
	n_last_hour	The total number of bids during the last hour on a listing
	amt_last_hour	The total amount of bids during the last hour on a listing
	n_last_day	The total number of bids during the last day on a listing
	amt_last_day	The total amount of bids during the last day on a listing
	n_bidders	The total number of lenders who bid on a listing
	n_open_days	The number of days that a listing remains open for auction
	bidtime_std	The standard deviation of a listing's bidding timing
	amt_large_bids	The total amount of large bids on a listing (a large bid is defined as a bid with an amount two standard deviations above the mean value)
	amt_active_bidders	The total amount contributed by active bidders (an active bidder is defined as one whose number of bids is two standard deviations above the mean number of bids per bidder)
Activity sequence (coding)[4]	seq_activity_type[5]	The type of the activity: A1: borrowing; A2: lending
	seq_amt	The amount (¥) requested by the borrower (for A1), or the amount of the bid made by the lender (for A2): B1: 0–2,999; B2: 3,000; B3: 3,001–10,000; B4: 10,001–50,000; B5: >50,000
	seq_interest_rate	The interest rate (%) offered by the borrower (for A1), or the interest rate of the listing the lender bids on (for A2): C1: 0–11.9; C2: 12; C3: 12.1–17; C4: 17.1–20; C5: >20
	seq_term	The repayment period length (in months): D1: 1–5; D2: 6; D3: 7–11; D4: 12; D5: >12

[4] The interest rate, borrowing amount, and term are all distributed very unevenly. For example, the term of 48% of the listings is 12 months. Consequently, dividing a subattribute's values (e.g., 2-36 months) into five equal intervals may cause the majority of listings to be placed into a single bin (e.g., the bin containing all listings with terms of 10-16 months) and to receive the same code. Using five evenly distributed percentiles (i.e., 20%, 40%, 60%, 80%, and 100%) does not resolve the problem either because many data points take on exactly the same value. As a result, for each subattribute, we use bins of different widths to make the resulting distribution (histogram) as even as possible.

[5] Note that a borrower is also allowed to lend money to others on P2P platforms.

Bidding Process Characteristics

Fraudsters may not only take advantage of existing *opportunities* in the marketplace (e.g., information asymmetry and weaknesses in the governance systems) but also deliberately create opportunities for fraud by manipulating the bidding process. On P2P platforms, lenders are more likely to invest in listings that have already received bids from others, which is a behavior called herding (Berkovich, 2011; Burtch, 2011; Lee & Lee, 2012; Zhang and Liu, 2012). The first and last periods of loan auctions often see large crowds of lenders (Shen et al., 2010). From a fraudster's perspective, if the fraudster can manage to initiate herding momentum, the odds of funding success may be substantially boosted. To achieve this goal, a fraudster may register as a lender using a fake identity or hire other people to register as lenders. The fraudster and the confederates can then bid on the fraudster's own listings within the first day or hour to stimulate bids from other real lenders, thereby creating herding momentum. Similarly, to prevent a funding failure at the time of closure, the fraudster and the confederates may contribute to the listing in the last period (day/hour) to successfully close and materialize it into a loan. However, because hiring confederates incurs costs (e.g., communication and coordination cost, labor cost, and opportunity cost) and increases the risks of being caught, a fraudster may not hire an exceptionally large number of people to make fake bids. Therefore, the scale of shill bidding on P2P platforms may be limited. In this study, we propose eight features to measure the herding momentum (number of bids and amount of bids) at the beginning (in the first hour and on the first day) of a bidding process and in periods toward the end of it (in the last hour and on the last day). In this category, we also include five additional measures that describe the bidding process characteristics (e.g., number of bidders and number of days an auction remains open). The names and definitions of these 13 features are provided in Table 2.

Activity Sequence

In addition to the 24 numerical features, we also constructed a sequence feature to record how a borrower learned to increase their fraud *capability* through a series of borrowing and lending practices and activities over time. When encoding the past activities of a borrower, we first encode each activity as a *word*, following the method proposed by Grbovic and Cheng (2018), and then adopt the word2vec approach (Mikolov et al., 2013) to compute the word embeddings. Specifically, each *word* comprises four subattributes that describe a past activity of a borrower: activity type, amount, interest rate, and repayment term. Each subattribute of the event is represented by a letter (e.g., A and B). Continuous attribute values are discretized into several bins (e.g., 1, 2, and 3). For instance, a loan of ¥4,000 (approximately $616) with an interest rate of 12% and a 6-month payment term can be encoded as the word "A1B3C2D2." Table 2 provides the coding scheme.

For each listing, we retrieved all the borrowing and lending activities that the listing borrower had been involved in prior to the present listing. Each activity is coded as a word, and the sequence of these words forms a *sentence*. The sentence is then used as the sequence feature for the listing and fed into a long short-term memory (LSTM) neural network. We posit that because the sequence feature encodes more information about the activity history of a borrower, it can help further improve the detection performance.

Learning and Testing

In fraud detection applications, the costs of false positive (FP) errors and false negative (FN) errors are different for different types of stakeholders (Abbasi et al., 2012). Although mistakenly classifying a legitimate instance as fraudulent (an FP error) may incur unnecessary investigation and audit expenses, failing to signal a fraudulent instance (an FN error) results in significantly larger losses to victims. Therefore, in fraud detection, FN errors are considered costlier than FP errors. However, many classification algorithms assume equally important classes and treat the two types of errors equally. In our research, we used a cost-sensitive learning strategy to impose a higher penalty on FN errors than on FP errors (Elkan, 2001). The higher the FN-to-FP ratio, the more instances the cost-sensitive learning algorithm classifies as fraudulent.

The classification algorithms we use in this research are four state-of-the-art machine learning methods: random forest (RF), XGBoost (XGB), deep neural network (DNN), and LSTM. RF and XGB are decision tree-based algorithms, and both have achieved outstanding performance in classification applications (Bhattacharyya et al., 2011; Chen & Guestrin, 2016). DNN and LSTM are neural network algorithms with deep learning architectures, and LSTM is particularly effective for capturing relationships and patterns embedded in sequences (Hochreiter & Schmidhuber, 1997). Appendix A[6] provides descriptions of the architectures of the DNN and LSTM models used in our study.

[6] Appendices are located at OSF.io. DOI: 10.17605/OSF.IO/PKA3C.

Evaluation

Data

We evaluated the design artifact using empirical data from EasyLoans,[7] which is one of the largest P2P lending platforms in China. Launched in 2007, this platform has attracted over 103 million users, and more than ¥178 billion (approximately $25 billion) loans have been funded. Users borrow money for various reasons, ranging from covering travel expenses and paying medical bills to purchasing automobiles and funding business startups.

EasyLoans provided us with a proprietary dataset that comprised all 1,406,404 listings made by 922,453 borrowers on the platform from the platform's inception in June 2007 to December 2014. The sample also contained 11,254,118 records about all the bids made on listings during this period. The largest amount requested was ¥708,686 ($109,028). The number of loan requests skyrocketed in 2014 and accounted for 81.2% of all listings in the platform's history (as of the end of 2014). We selected the records in 2014 as our evaluation sample, which contained 1,142,739 loan requests made by 828,086 borrowers in this single year. Approximately 20% (227,479 of 1,142,739) of the listings were successfully funded (by 52,452 unique lenders) and materialized into loans, and 11.2% (25,519 of 227,479) of those loans were in default by the time we acquired the dataset.

Class Labeling

Although we could directly identify defaulted loans in our dataset, we could not simply label all the defaults as fraud. A borrower may fail to make a monthly payment on time for various reasons (e.g., unemployment and medical emergency). If the borrower eventually repays the loan in full, it is very unlikely that the borrower intended to deceive or cheat lenders in the first place. To identify true fraud, we consulted the blacklist published by EasyLoans on its website. This blacklist exposes information about malevolent borrowers, who sternly refused to repay their loans even after the platform exhausted all payment collection measures at its disposal (e.g., multiple email reminders, phone calls, and third-party debt collectors). Each entry on this list contained a borrower's user ID, real name, and default amount, in addition to the year in which the borrower defaulted. Note that if a borrower defaulted but paid off the loan later, the borrower's entry would have been removed from this list. As the time when we retrieved the blacklist (June 2018) was far beyond the repayment deadlines of the latest loans in our dataset (December 31, 2014), we considered this blacklist to be a fairly accurate list of deceitful borrowers on this platform. This assumption was also confirmed by the platform.

Note that a borrower's presence on the blacklist does not necessarily mean that all loan requests made by the borrower were fraudulent. We found 3,202 blacklisted borrowers in 2014 whose total loan amount was more than ¥21 million (approximately $3.2 million). For each loan by a blacklisted borrower, we carefully examined the loan information (i.e., default vs. repaid status, total overdue amount in the data, the default amount posted on the blacklist, and the default percentage of the borrowing amount). Eventually, we identified 3,002 fraudulent loans. This number accounted for 1.3% (3,002 of 227,479) of all materialized loans and 11.8% (3,002 of 25,519) of all defaulted loans in 2014. The class distribution was extremely skewed, with a legitimate-to-fraud ratio as high as 75:1 (224,477:3,002). In this case, detecting a small number of fraudulent instances in such a large volume of data was like finding a needle in a haystack.

FN-to-FP Cost Ratios

To apply the cost-sensitive learning strategy, we estimated the costs of classification errors for lenders. If a classification algorithm fails to signal a fraudulent loan, an FN error occurs. The lenders' costs included the lost principal (average = ¥7,108; approximately $1,093; see the statistics in Table C1 in Appendix C on OSF.io) and the opportunity cost of not investing in a legitimate loan from which they could earn interest (average = 14%). Because other costs (e.g., lost time for searching and examining the list, and negative emotions caused by the financial losses) were intangible and difficult to measure, we did not include them in the cost estimation. Therefore, for lenders, the costs of an FN error were ¥7,108 + ¥7,108 × 0.14 = ¥8,103 (approximately $1,246). In the case of an FP error, in which a legitimate request (average = ¥5,334) was mistakenly marked as fraudulent, the lenders' opportunity cost was the lost interest that could have been earned: ¥5,334 × 0.14 = ¥747 (approximately $115). As a result, the FN-to-FP cost ratio was roughly 11:1 for lenders on this platform. Because we did not have data about the various costs for the platform (e.g., investigation costs) and for legitimate borrowers, we could not estimate their cost ratios.

[7] To protect the platform's confidentiality, we use a fictitious company name.

Feature Sets

Seven baseline features about a borrower's demographic and listing information were directly available in the dataset: *age*, *gender*, *education level*, *occupation*, *borrowing amount*, *interest rate*, and *repayment term*. To test the performance of the proposed features, we constructed three feature sets: Set A (7 baseline features), Set B (7 baseline + 24 numerical features), and Set C (7 baseline + 24 numerical + 1 sequence features). Although EasyLoans assigned a letter credit grade from AAA (high quality) to F (high risk) to each user based on the user's background information (e.g., education level and degrees) and previous repayment records on this platform, if applicable, the credit grade was an ex post measure in this dataset: at the time of our data collection, the platform had already lowered the credit grades of the default borrowers (including fraudulent borrowers). Therefore, we did not include the credit grade in the baseline set.

Performance Metrics

As in most supervised machine learning studies, we selected *accuracy* as the first performance metric. However, because the sample was extremely imbalanced, with only 1.3% of the loans (3,002 of 227,479) labeled as fraudulent, accuracy was as high as 98.7%, even without using any features and any supervised classifiers (i.e., all the loans are blindly classified as legitimate). As a result, we also selected the *recall*, *precision*, and *F score* for the fraud class, in addition to accuracy, to measure how well a classifier identified fraudulent loan requests.

$$Recall = \frac{\text{Number of Correctly Identified Fraudulent Listings}}{\text{Number of True Fraudulent Listings}} = \frac{TP}{TP + FN}$$

$$Precision = \frac{\text{Number of Correctly Identified Fraudulent Listings}}{\text{Number of Listings Classified as Fraudulent}} = \frac{TP}{TP + FP}$$

$$F\ score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In these definitions, *TP* and *TN* represent the numbers of true positives and true negatives, respectively. As these three metrics only correspond to the fraud class, recall is the same as sensitivity; that is, the higher the recall, the more sensitive the classifier is to detecting fraud within a set of listings. The higher the precision, the less likely it is for a classifier to mislabel a legitimate listing as fraudulent. The *F* (or *F*1) score is the harmonic mean between precision and recall and provides a balanced view between the two metrics. We also report the AUC, which is the area under the receiver operating characteristics curve. Generally, the higher the AUC, the more capable the classifier is at outperforming random guessing. The closer the AUC is to 1.0, the better the classifier.

Analyses and Results

Blacklisted versus Legitimate Borrowers

Our underlying assumption for fraud detection is that fraudsters' characteristics and behaviors deviate from those of legitimate borrowers. To confirm this assumption, we first compared the means of features between the two groups of borrowers by running independent-sample *t*-tests. Table C1 in Appendix C reports the means, standard deviations, and statistical significance of the baseline and our proposed numerical features. To keep the description and discussion straightforward, we express all the monetary amounts in the following text in Chinese currency (¥) without converting them into U.S. dollars ($).

With the exception of the interest rate and borrower age, blacklisted and legitimate borrowers had significantly different characteristics measured by these features ($p < 0.001$). For example, on average, the requested loan amount per listing by blacklisted borrowers (¥7,108) was significantly greater than the amount requested by legitimate borrowers (¥5,334). Blacklisted borrowers also had significantly more delinquencies (0.21) than legitimate borrowers (0.02).

The bidding process features reveal drastic differences between blacklisted and legitimate borrowers. For instance, the number and amount of first-day and first-hour bids, as well as those for the last day and last hours, were much higher for blacklisted borrowers than for legitimate borrowers.

Effects of the Skewed Class Distribution

To investigate how the skewed class distribution affected detection performance, we used the undersampling approach, which is widely adopted in fraud detection studies (see Table 1). By randomly removing a number of legitimate instances from the original dataset, we constructed a series of samples, labeled Samples 1 to 7, with a legitimate-to-fraud ratio of 1:1, 10:1, 20:1, 30:1, 40:1, 50:1, and 60:1, respectively. Sample 1 was a completely balanced sample in which only 3,002 legitimate loans were extracted to match the fraudulent loans, and the majority (97.4%) of records were discarded.

We found that the performance (recall, precision, *F* score, and AUC) of the classifiers continued to deteriorate as the sample size increased and the skew of the class distribution worsened. Figure B1 in Appendix B (on OSF.io) plots the *F* scores for Samples 1 to 7 and the original dataset using RF with feature Sets A (baseline features) and B (baseline + numerical features). It shows that when the entire dataset (with a legitimate-to-fraud ratio of approximately 75:1) was used, the

F score was very low (0.22 for Set A and 0.33 for Set B). This means that as a sample becomes more imbalanced, the classifier becomes less capable of correctly distinguishing between fraudulent and legitimate loans. The AUC also decreased as the legitimate-to-fraud ratio increased. However, its accuracy remained high (approximately 0.96 for both feature sets), even with the 75:1 ratio. We obtained similar results using the XGB and DNN algorithms. As the class distribution became more skewed, the performance gap between Sets A and B became wider, thereby demonstrating the potential of our proposed features.

Performance of Features

Following the common practice of financial fraud detection research (see Table 1), we first selected Samples 1 and 2—with legitimate-to-fraud ratios of 1:1 and 10:1, respectively—when testing classification performance. For each sample, we compared the performance between the baseline features (Set A) and combinations of the baseline and proposed features (Sets B and C). We performed a series of *t*-tests using the 10-fold cross-validation approach to establish the statistical significance of the performance difference.

Numerical Features

We first sought to determine whether the proposed numerical features could provide more information about borrower behavior in addition to the baseline features. Table 3 reports the comparison results between Set A (baseline features) and Set B (baseline + numerical features).

The results show that the overall accuracy using Set B was significantly higher than that using Set A across all algorithms. In terms of recall (sensitivity to fraud), the performance difference between the two sets was not significant; that is, the algorithms were very sensitive to fraudulent instances given even a small number of baseline features. This is because of the use of a cost-sensitive learning strategy that imposed a penalty 11 times greater on FN errors than on FP errors. Indeed, an extremely high FN-to-FP cost ratio could cause an algorithm to classify nearly all instances as fraudulent, thereby resulting in a very high recall value close to 1.0. However, such a classifier is useless because it fails to rule out legitimate cases; that is, high recall is at the cost of precision: although most fraudulent listings are correctly detected, many red-flagged requests are actually legitimate.

Precision for the fraud class using Set B was significantly better than that using Set A. With the imbalanced Sample 2, precision dropped below 50%. Although the algorithms erred on the side of caution by identifying as many fraudulent listings as possible, the FP error rate elevated quickly. Given the trade-off between precision and recall, the *F* score provided a more balanced picture. Specifically, Set B's *F* scores were significantly better than those of Set A, except for DNN in Sample 1. The AUC values were also significantly higher using Set B than using Set A.

To summarize, the performance of Set B in terms of accuracy, precision, *F* score, and AUC was significantly better than that of Set A. When the sample was perfectly balanced, the performance of the two sets was nearly the same. However, when the sample was more skewed, the role of our proposed numerical features became more important.

To evaluate the performance of the feature sets in real-world applications, we also used the original dataset with a legitimate-to-fraud ratio of 75:1. The first two groups of rows in Table 4 report the 10-fold cross-validation comparison results between Sets A and B using the models learned from under-samples on the original dataset. It shows that performance deteriorated significantly, particularly for the precision and *F* score. On average, about 94% of the identified fraudulent instances were actually legitimate. This is consistent with the findings in previous research that models learned from samples using the undersampling approach perform poorly when applied to real datasets with very skewed class distributions (He & Garcia, 2009).

We then tested the performance of the two feature sets on the original sample without undersampling. The last group of rows in Table 4 reports the results, which show that Set B consistently outperformed Set A across all metrics and algorithms.

To assess the quality of the features, we performed chi-squared tests (see Table C2 in Appendix C) and found that all the features were significant ($p < 0.001$).

Activity Sequence Feature

We only used Set C (baseline + numerical + sequence features) with the LSTM neural network algorithm, which is capable of processing sequence input. To save space, in Table 5, we only report the comparison results for Set C using LSTM against Set B (baseline + numerical features) using DNN, which is a neural network without sequence handling.[8]

[8] We did not test the features using LSTM on the original dataset because this would have required constructing the sequences from the transaction histories for all 224,477 instances, which would have required too much time.

Table 3. Performance Comparison between Feature Sets A and B

Sample (L-to-F Ratio)		Accuracy		Recall		Precision		*F* score		AUC	
		Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
Sample 1 (1:1)	RF	0.887	0.892*	0.997	0.997	0.818	0.824**	0.899	0.902*	0.929	0.944**
	XGB	0.888	0.894*	0.996	0.997	0.821	0.827**	0.899	0.904*	0.930	0.955***
	DNN	0.888	0.883	0.998	0.997	0.818	0.812**	0.899	0.895	0.930	0.937***
Sample 2 (10:1)	RF	0.824	0.849***	0.967	0.948	0.337	0.371***	0.510	0.533***	0.937	0.949***
	XGB	0.823	0.845***	0.964	0.960	0.336	0.365***	0.503	0.529***	0.937	0.954***
	DNN	0.821	0.848***	0.953	0.919	0.332	0.367***	0.496	0.524***	0.937	0.941**

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 4. Performance Comparison between Feature Sets A and B on the Original Dataset

Sample (L-to-F Ratio)		Accuracy		Recall		Precision		*F* score		AUC	
		Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
(1:1)	RF	0.778	0.786***	0.992	1.0***	0.056	0.058***	0.105	0.11***	0.929	0.943***
	XGB	0.779	0.790***	0.997	1.0***	0.056	0.059***	0.107	0.112***	0.938	0.954***
	DNN	0.777	0.771	0.996	0.992	0.056	0.054	0.106	0.103	0.924	0.934***
(10:1)	RF	0.818	0.851***	0.962	0.958	0.066	0.078***	0.123	0.145***	0.937	0.956***
	XGB	0.818	0.840***	0.954	0.957	0.065	0.073***	0.121	0.136***	0.938	0.955***
	DNN	0.818	0.849***	0.942	0.896	0.064	0.074***	0.12	0.137***	0.93	0.934
Original sample (75:1)	RF	0.962	0.973***	0.398	0.487***	0.150	0.242***	0.218	0.323***	0.939	0.961***
	XGB	0.952	0.965***	0.489	0.550***	0.136	0.199***	0.213	0.292***	0.938	0.955***
	DNN	0.967	0.971*	0.303	0.414*	0.151	0.211**	0.192	0.275***	0.932	0.94*

Note: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 5. Performance Comparison between Feature Sets B and C

Sample (L-to-F Ratio)	Accuracy		Recall		Precision		F Score		AUC	
	Set B (DNN)	Set C (LSTM)	Set B (DNN)	Set C (LSTM)	Set B (DNN)	Set C (LSTM)	Set B (DNN)	Set C (LSTM)	Set B (DNN)	Set C (LSTM)
Sample 1 (1:1)	0.895	0.891	0.998	0.995	0.829	0.825	0.905	0.902	0.958	0.952
Sample 2 (10:1)	0.832	0.858***	0.977	0.963	0.350	0.390***	0.515	0.554***	0.951	0.957**

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

In Sample 1, Set C did not perform significantly better than Set B. However, in Sample 2, Set C significantly outperformed Set B in terms of accuracy, precision, *F* score, and AUC. This implies that in the sample with a balanced class distribution (Sample 1), the sequence feature provided little additional help, whereas in the sample with a skewed distribution (Sample 2), Set C helped improve all the metrics, except for recall. Therefore, we believe that the combination of all features, including the sequence feature, generally outperforms the baseline features (except for recall), particularly regarding imbalanced datasets with skewed class distributions.

Generalizability Check

To ascertain whether the proposed features were generalizable to other platforms, we obtained proprietary data from another P2P lending platform. Founded in 2010, this platform has attracted over 18 million users, with more than ¥63 billion (approximately $9.84 billion) of funded loans. The dataset provided by this second platform comprised all 795,758 listings made by 594,481 borrowers from 2010 to 2015. The largest amount requested during this period was ¥3,000,000 (approximately $469,000). In 2015, there was a drastic increase in the number of loan requests, accounting for 50.3% of all listings in the platform's history. We selected the records

for 2015 as our experimental dataset, which contained 400,745 loan requests made by 329,479 borrowers in this single year. Among these listings, 114,742 requests were funded by 146,308 unique lenders. We identified 1,454 fraudulent loans, which was 1.26% (1,454 of 114,742) of all materialized loans and 77.8% (1,454 of 1,868) of all defaulted loans in 2015. The legitimate-to-fraud ratio was approximately 79:1, and the FN-to-FP cost ratio was approximately 5:1 on this platform. Table 6 reports the comparison of the all-feature Set C against Set A on a sample with a legitimate-to-fraudulent ratio of 10:1. Similar to the results from EasyLoans, except for recall, the performance using all features was significantly better than that using the baseline features alone.

Discussion

It has been a challenge in fraud detection research to construct effective features for signaling fraudulent behaviors (Pourhabibi et al., 2020; Vlasselaer et al., 2017). In this research, we proposed a set of features constructed from P2P lending transactions to detect loan fraud. Based on the fraud triangle theory and its extensions, we identified five categories of features that capture the behavioral patterns of loan fraudsters, in addition to baseline features. To a certain extent, these features reflect a fraudster's capability of committing fraud, personal integrity, and opportunity. Testing these features using four state-of-the-art classification algorithms (RF, XGB, DNN, and LSTM), we found that these additional features enhanced detection performance. Our research has important implications for both fraud detection research and practice.

Implications for Research and Practice

Our research contributes to the literature on fraud theories and models. Despite their role in fraud prevention and investigation, the fraud triangle theory and its extensions have been found to be inadequate for fraud detection (Dorminey et al., 2012; Kassem & Higson, 2012; Vousinas, 2019). The five categories of behavioral features proposed in our research bridge this gap by providing operationalization of the three antecedents of fraud (capability, integrity, and opportunity) in the P2P lending context, thereby demonstrating these models' potential in assisting fraud detection. Our research also broadens and deepens our understanding of fraud conditions and manipulation tactics. It shows that fraudsters may exploit opportunities in P2P lending markets. Given the information asymmetry and identity anonymity on these platforms, offenders can engage lenders' trust by building a track record deliberately before committing fraud. They may manipulate the bidding process by recruiting shill bidders (Wang et al., 2002) to create herding momentum during the first or last period of a loan auction.

Furthermore, our research adds to the discussion regarding predictive analytics in information systems (IS) research. Although theories are usually expected to achieve both explanation and prediction, there have been calls to consider these two goals separately (Forster, 2002) because explanatory modeling seeks to establish causal relationships and predictive modeling aims at "predicting new/future observations or scenarios" (Shmueli & Koppius, 2011, p. 555). This raises a question about the role of theory in predictive modeling; and, indeed, many predictive analytics studies in IS do not connect to any theory (e.g., Ben-Assuli & Padman, 2020; Geva et al., 2017). However, we believe that although the explanatory power of a theory does not always imply predictive power, explanatory theories may still offer useful directions and hints for predictive modeling. In our research, the fraud models are clearly explanatory, yet they helped us identify additional behavioral features to enhance predictive performance in P2P fraud detection.

As a proof of concept, this study provides useful guidance for constructing effective features for the practice of P2P lending fraud detection. Features used in the detection of traditional fraud (e.g., credit card fraud, see Bahnesen et al., 2016; corporate fraud, see Abbasi et al., 2012) have limited applicability to P2P loan fraud detection. The feature set that we propose in this research includes not only "raw" attributes about P2P lending transactions, which are similar to those used in traditional fraud detection research (e.g., features for borrowing and payment history), but also novel features for capturing bidding process characteristics and activity sequences, which have been seldom or never used in financial fraud detection. More importantly, all these features can be constructed directly from a P2P lending platform's historical and current data, which makes them readily accessible for a platform's risk control and management. Using appropriate machine learning methods with these features, a platform can monitor the marketplace and spot suspicious loan requests in real time. With these countermeasures, lenders are better protected from and less vulnerable to loan fraud; and the financial risks of investment are reduced.

Theoretically, after training and testing, a learned classification model can be used in real time to detect fraudulent transactions. In practice, however, some implementation issues must be addressed. For instance, some of the bidding process-related features (e.g., number of bids in the last hour) are not available before the auction closes. Thus, a platform must determine *when* to apply the detection methods and *how* to operationalize the countermeasures.

Table 6. Performance Comparison Between Feature Sets on Platform 2

	Accuracy		Recall		Precision		*F* score		AUC	
	Set A	Set C (LSTM)	Set A	Set C (LSTM)	Set A	Set C (LSTM)	Set A	Set C (LSTM)	Set A	Set C (LSTM)
RF	0.935	0.941**	0.988	0.984	0.583	0.611**	0.733	0.753**	0.975	0.983**
XGB	0.936	0.941**	0.987	0.984	0.590	0.611**	0.738	0.753**	0.978	0.983*
DNN	0.934	0.941**	0.994	0.984	0.581	0.611**	0.733	0.753**	0.977	0.983*

Note: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Although it is most cost-effective for a platform to wait until all the features become available, some platforms may wish to monitor listings before, during, and after auctions. If a platform chooses to act sooner, we recommend a three-stage strategy. The first stage occurs after a borrower submits a loan request for approval and before the auction starts. At this stage, the platform can use features constructed from historical records (i.e., borrowing and payment histories) to determine if it is legitimate. The second stage occurs after a listing is approved and the auction starts. During the auction period, the platform can use some of the bidding process features (e.g., number of first-day bids) to monitor the auction. The last stage starts immediately after the auction is closed and all the bidding process-related features become available. The platform can suspend the transaction right away if a loan request is red-flagged at any stage. To use this three-stage strategy, the platform can implement a standard policy for a "holding period" before, during, and after the auction. This policy allows the platform to investigate any suspicious listings and/or request additional supporting documents or proof during the holding period. In fact, on the two platforms in our study, even if a listing is fully funded after the auction, it must be approved before it materializes into a loan. If a listing fails to be approved by the platform (because of reasons such as shill bidding), the listing is rejected, and the lenders' funds, frozen during the bidding, are released back to the lenders. In the case in which a loan is found to be fraudulent after it is already materialized and the lenders' funds have already been transferred to the borrower, the platform can still protect the lenders to some degree. It has been found that the implementations of certain low-cost behavioral mechanisms (e.g., sending the borrowers text message reminders conveying the lenders' positive expectations) can significantly increase loan repayment rates on P2P lending platforms (Du et al., 2020).

While we are developing advanced methods and approaches to fighting fraud, fraudsters are adaptive human beings—they also update their tactics, hone their skills, and adjust their strategies. Therefore, fraud detection practitioners must keep watching the marketplace and continuing to identify emerging patterns, changes, and trends to ensure a healthy, orderly investment environment.

Limitations

This research has several limitations. First, our datasets were from single sources, and most of the features were constructed based on transaction records. No data from other sources (e.g., social media) or of a different nature (e.g., geographical locations and IP addresses) were included in our dataset. Consequently, the information captured by our proposed features was limited, which resulted in moderate recall levels (around 50%) in the original datasets. Second, both platforms under study are in China, in which the financial environment, laws, regulations, and culture are quite different from those in other developed and developing countries. Whether our proposed features can also be applied to P2P platforms in other countries remains to be seen. For example, with the availability of credit scores, platforms in other countries may be able to better detect fraudulent loans. Third, because we use supervised learning models, our approach is only applicable to platforms that have already been suffering from fraud; that is, the data used for training must have confirmed fraudulent instances. New platforms, which do not have rich sets of historical records with both legitimate and fraudulent transactions, cannot use these supervised learning methods. Another potential limitation is that our proposed features rely heavily on the past transaction records of borrowers. If a borrower is new to a platform and has no previous transactions, many of the features (e.g., number of previous listings and delinquencies) are zero. This is similar to the "cold-start" problem in recommender systems research (Basu et al., 1998). In this case, the classification must rely only on the baseline and bidding process-related features.

Conclusion

To the best of our knowledge, we are among the first few groups to study P2P loan fraud empirically. In future work, we will construct additional features when data from diverse sources and of different natures (e.g., income-to-debt ratio, employment history, online purchase records, and social media posts) are available and will operationalize other factors identified in the fraud theories (e.g., pressure).

Acknowledgments

The authors thank the senior editor, associate editor, and three anonymous reviewers for their constructive comments and advice. This research was partly funded by the Natural Science Foundation of China under grants #71771159 and #72071160, and by the Chinese Ministry of Education under grant #18YJAZH142.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, *36*(4), 1293-1327.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488-500.

Albrecht, W. S., Albrecht, C. C., & Albrecht, C. O. (2006). *Fraud examination*, Thomson South-Western, New York, NY.

Bahnesen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, *51*(1), 134-142.

Basu, C., Hirsh, H., & Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. *Proceedings of the 5th National Conference on Artificial Intelligence* (pp. 714-720).

Ben-Assuli, O., & Padman, R. (2020). Trajectories of repeated readmissions of chronic disease patients: Risk stratification, profiling, and prediction. *MIS Quarterly*, *44*(1), 201-226.

Berkovich, E. (2011). Search and herding effects in peer-to-peer lending: Evidence from prosper.com. *Annals of Finance*, *7*, 1-17.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602-613.

Brown, N.C., Crowley, R. M., & Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*(1), 237-291.

Burtch, G. (2011). Herding behavior as a network externality. In *Proceedings of the 32nd International Conference on Information Systems*.

Chang, W.-H., & Chang, J.-S. (2012). An effective early fraud detection method for online auctions. *Electronic Commerce Research and Applications*, *11*(4), 346-360.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Cressey, D. R. (1953). *Other people's money: A study of the social psychology of embezzlement*. Free Press.

Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems*, *35*(2), 461-487.

Dorminey, J., Fleming, A.S., Kranacher, M. J., & Riley Jr., R. A. (2012). The evolution of fraud theory. *Issues in Accounting Education*, *27*(2), 555-579.

Du, N., Li, L., Lu, T., & Lu, X. (2020). Prosocial compliance in P2P lending: A natural field experiment. *Management Science*, *66*(1), 315-333.

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), 2455-2484.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 973-978).

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, *69*(3), S124-S134.

Forward View. (2021). *Annual Report on P2P Markets in China, 2020*. Available at https://bg.qianzhan.com/trends/detail/506/200408-9067ee5d.html.

Free, C., & Murphy, P. R. (2015). The ties that bind: The decision to co-offend in fraud. *Contemporary Accounting Research*, *32*(1), 18-54.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high-involvement projects. *MIS Quarterly*, *41*(1), 65-82.

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, *265*(3), 993-1004.

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud: A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139-152.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284.

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Kassem, R., & Higson, A. (2012). The new fraud triangle model. *Journal of Emerging Trends in Economics and Management Sciences*, *3*(3), 191-195.

Kranacher, M. J., Riley Jr., R. A., & Wells, J. T. (2011). *Forensic accounting and fraud examination*. Wiley.

Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications*, *11*(5), 495-503.

Liao, R., Balasinorwala, S., & Rao, H. R. (2017). Computer assisted frauds: An examination of offender and offense characteristics in relation to arrests. *Information Systems Frontiers*, *19*, 443-455.

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, *59*(1), 17-35.

Majadi, N., Trevathan, J., Gray, H., & Estivill-Castro, V. (2017). Real-time detection of shill bidding in online auctions: A literature review. *Computer Science Review*, *25*, 1-18.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111-3119.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support System*, *133*, 1-15.

Shen, D., Krumme, C., & Lippman, A. (2010). Follow the profit or the herd? Exploring social effects in peer-to-peer lending. In *Proceedings of IEEE International Conference on Social Computing* (pp. 137-144).

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553-572.

Siering, M., Koch, Jascha-Alexander, and Deokar, A. V. (2016). Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. *Journal of Management Information Systems*, *33*(2), 421-455.

Trevathan, J., & Read, W. (2012). Detecting shilling bidding in online English auctions. In *Cyber crime: Concepts, methodologies, tools and applications* (pp. 618-640). IGI Global.

Vlasselaer, V. V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2017). GOTCHA! Network-based fraud detection for social security fraud. *Management Science*, *63*(9), 3090-3110.

Vousinas, G. L. (2019). Advancing theory of fraud: The S.C.O.R.E. model. *Journal of Financial Crime*, *26*(1), 372-381.

Wang, W., Hidvegi, Z., & Whinston, A. B. (2002). Shill bidding in multi-round online auctions. In *Proceedings of the 35th Hawaii International Conference on System Sciences*.

Wolfe, D. T., & Hermanson, D. R. (2004). The fraud diamond: Considering the four elements of fraud. *CPA Journal*, *74*(12), 38-42.

Xu, J., Lu, Y., & Chau, M. (2015). P2P lending fraud detection: A big data approach. in *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics* (pp. 71-81).

Zhang, J., & Liu, P. (2012). Rational herding in microloan markets. *Management Science, 58*(5), 892-912.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning*, O'Reilly Media.

About the Authors

Jennifer J. Xu is a professor in the Computer Information Systems Department at Bentley University. Her research interests include artificial intelligence and business analytics, data science, fintech, e-commerce, social network analysis, and human-computer interaction. She has published more than 80 articles in information systems journals, books, and conference proceedings. She is currently serving on the editorial boards of *Journal of the Association for Information Systems* and *Communications of the Association for Information Systems*. She received her Ph.D. degree in management information systems from the University of Arizona.

Dongyu Chen is a professor of management information systems at the Soochow University. His research interests include online lending, crowdfunding, and information management. His work on these topics has been published in *MIS Quarterly*, *Communications of the Association for Information Systems*, *International Journal of Electronic Commerce*, *Journal of Global Information Management*, *Information Technology and Management*, *Journal of Global Information Technology Management*, *Electronic Commerce Research*, among other outlets.

Michael Chau is a professor in the Faculty of Business and Economics (HKU Business School) at the University of Hong Kong. His research focuses on the cross-disciplinary intersection of information systems, computer science, business analytics, and information science, with an emphasis on the applications of data, text, and web mining in various business, education, and social domains. He has received multiple awards for his research and is a member of the AIS College of Senior Scholars. His research has resulted in more than 160 publications in top journals and conference proceedings and he is highly ranked in many research productivity studies. He received his Ph.D. degree in management information systems from the University of Arizona and his B.Sc. degree in computer science and information systems from the University of Hong Kong.

Liting Li is a Ph.D. candidate in technology economics and management at the School of Management Science and Engineering, Southwestern University of Finance and Economics. Her research interests are in the areas of crowdsourcing, P2P lending, and human-machine hybrid systems. Her work on these topics has been published in *Electronic Commerce Research* and other outlets.

Haichao Zheng is a professor of management information systems at the School of Management Science and Engineering, Southwestern University of Finance and Economics. His research interests include crowdsourcing, open innovation, and human-machine hybrid system. He has published in *Decision Support Systems*, *Information Systems Journal*, *European Journal of Information Systems*, *Information and Management*, *International Journal of Electronic Commerce*, and *Electronic Commerce Research*.