Improve Risk Prediction in Online Lending (P2P) Using Feature Selection and Deep Learning

Nguyen Truong Thang†, Khuat Thanh Son†, Ngo Thi Thu Trang††, Nguyen Ha Nam†††, and Tran Manh Dong†

† Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
††Posts & Telecommunications Institute of Technology, Hanoi, Vietnam
††† Institute of Information Technology, Vietnam National University, Hanoi, Vietnam

Summary:
At the beginning of the 21st century, Peer2Peer (P2P) lending was established and developed rapidly in the United Kingdom, the United States, China and some other countries. The main challenge for individual investors in the P2P lending market is to allocate their money effectively through various lendings by accurately assessing the risk level of each lending. Traditional scoring models cannot fit the needs of individual investors in P2P lending because they do not provide natural mechanisms for asset allocation since for P2P lendings there are no traditional financial institutions. In this study, the report proposes a new way to analyze data for this emerging market. We have designed a risk scoring model based on advanced machine learning methods, capable of assessing the profits and risks of personal lendings. The report also applies a feature selection method and removes extraneous features to improve the efficiency of machine learning models. We conducted experiments on real lending datasets from the P2P lending markets. Test results show that the proposed model can improve the investment efficiency compared to existing P2P lending scoring methods.

Key words:
P2P, credit scoring, AI, deep learning, Boltzmann machine, feature selection, financial risk.

1. Introduction

Traditional banks may have profound changes by the P2P lending platform in the coming decades. Peer-to-peer lending is actually lending money to individuals or businesses through online services that are appropriate for lenders and borrowers. In particular, P2P Lending companies provide an online trading platform for borrowers to connect and directly borrow with lenders. All borrowing and debt repayment activities (principals and interests) between the borrower and the lender are recorded and stored by electronic trading platforms. [1].

The main objective of P2P lending platforms is to democratize consumer financial services. For more efficient distribution and credit guarantee, technology has been heavily used by these platforms. To borrow money, individuals can send their requests to credits that can be provided by investors without the involvement of financial institutions. Some P2P lending platforms in the world are Zopa (UK), Lending Club…. Currently in Vietnam there are many models of peer-to-peer lending such as: Tima, Mosa, Mofin, …

Over the past few years, many researchers have developed and applied many models and algorithms to analyze P2P lending data. Lee-Eunkyoung et al [2] investigated the investment behavior in the P2P lending market in Korea; they found strong evidence of investment and diminished marginal efficiency as bidding increased. In the data mining (DM) approach [3], to predict the performance of a P2P lending before it was proposed by Jin and Zhu, Byanjankar Ajay et al [4] proposed a credit-scoring model using neural networks in classifying peer-to-peer lending applications into default and non-default groups. They point out that a neural network-based credit scoring model is effective at filtering out default applications. Malekipirbazari Milad and Aksakalli [5] proposed a random forest-based classification (RF) method to predict borrower status. The results from the Lending Club (LC) data show that the RF-based method outperforms the FICO credit score as well as the LC score in identifying good borrowers. H. Li [6] researched on predicting potential lenders by using the exception detection method to find out unusual lenders, and they found out outsiders with bad credit scores with the ability of high abnormality. C. Serrano et al [7] focused on predicting the expected return of investing in P2P lendings, as measured by the internal rate of return. J. Yan at el [8] provided lending project proposals by creatively exploiting large amounts of unstructured data, textual data on borrower and lender dynamics. They have solved the mismatch problem between the borrower and the lender. Recently, Y. Xia et al [9] addressed the problem of refusal to deduce in a specific P2P lending field from a semi-supervised learning perspective. They proposed a new method of reject inference (CPLE-LightGBM) by incorporating a framework to estimate contrast pessimism and classification of gradient boosted decision trees (LightGBM).

P2P lending data is collected by a P2P lending platform, often containing extraneous and redundant features. Accuracy is reduced by redundancy of data classification

Manuscript received November 5, 2019
Manuscript revised November 20, 2019

and lack of data. A feature selection method (Feature selection or Feature extraction) is applied to remove redundant features. In other words, feature selection will choose an optimal subset of the features involved. With this optimal subset, the report solves the performance enhancement problem for the credit scoring problem.

In this article, we have proposed a method based on feature selection that is integrated with Restricted Boltzmann Machines (RBMs) in credit scoring tasks. The proposed method of the report is to select the top features according to RMSE ranking criteria, then apply that method to the credit scoring problem. Integrating with GPU Tensorflow, the method shows better results and is faster than the original RBM.

The rest of the paper is presented as follows: Section II, overview of feature selection method, deep learning and RBM; Section III, description of proposed model; Section IV, analysis of obtained results. Section V concludes and proposes development directions.

2. Technology Overview

A. Deep learning

In recent years, AI - Artificial Intelligence, and more specifically Machine Learning, has emerged as evidence of the fourth industrial revolution. Machine Learning is a small area of Computer Science. It is capable of self-learning based on input without having to be specifically programmed. In recent years, as the computing power of computers increased to a new height and the enormous amount of data collected by technology companies has grown, Machine Learning has taken a long step and a new field was born called Deep Learning [10 - 11]. Deep learning is the technique of using neural networks similar to those of the human brain to build a machine learning system. This is a great combination between math and neuroscience. Its results are tremendous and can be considered as the beginning of a new industry. Deep learning relies on a set of algorithms to try to model high-level abstraction data using multiple processing layers with complex structures.

One of the famous and popular applications of Deep Learning is virtual voice-based assistant. Not only applications in virtual assistants, many other applications of deep learning have become common in life such as detecting human face in photos (Face Detection), recoginizing images and human faces (Image/Face Recognition) on social networks Facebook and Google, self-driving cars of Google, etc.

The Botlzmann Machine is a regression neural network, where the output of a neuron can be an input of a neuron on the same or previous layers. It has a closed cycle of Figure 1 consisting of random binary units and symmetric connections discovered by David Ackley, Geofrey Hinton, and Terrence Sejnowski [13].

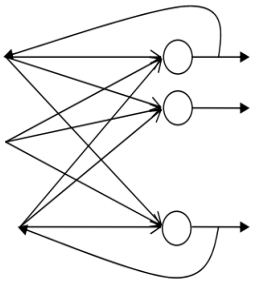

Fig. 1 Single-layer regression neural network

The units of a BM are usually divided into two layers, one is a set of visible units v which may have data held on it, and a set of hidden units h that act as potential variables. The units are connected by symmetrical connections in any arbitrary way, except for themselves (Figure 2). Each unit i has a binary state si and two states "off" or "on" (si=0 or si=1) with a probability that is a logistic function of the inputs it receives from other units j it is connected to.

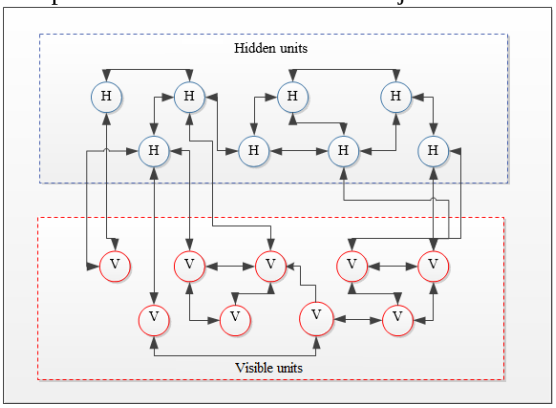

Fig. 2 A Boltzmann Machine, the units can be connected in any way.

$$p(s_i = 1) = \frac{1}{1 + \exp\left(-b_i + \sum s_i w_{ij}\right)_i} \quad (1)$$

In which, bi is the threshold (also known as bias deviation: This threshold is often included as a component of the transfer function of unit i and the weight wij is the associated weight of symmetry between units i and j

$$E(v,h) = -\sum_i b_i s_i^{(v,h)} - \sum_{i<j} s_i^{(v,h)} s_j^{(v,h)} w_{ij} (2)$$

If units are randomly selected and constantly updated using formula (1) it can be shown that the network will eventually achieve a stable (or equilibrium) probability distribution, in which the probability of finding the network in any global configuration (v, h) is determined by the energy of the relative configurations compared to the power sources of all possible configurations [13]:

$$p(v,h) = \frac{\exp(-\mathrm{E}(\mathrm{v},\mathrm{h}))}{\sum_{u,g}\exp(-E(u,g))} \quad (3)$$

More specifically, the probability of finding the network at stationary level with a configuration v with visible units given by:

$$p(v) = \frac{\sum_{h}\exp(-E(v,h))}{\sum_{u,g}\exp(-E(u,g))} \quad (4)$$

The Boltzmann machine is a special type of Markov random field log-linear. The energy function of a Boltzmann machine is linear in free parameters. The limit parameter is set to no parameter making the function strong enough to represent complex distributions. Some features are hidden because they are never observed. Boltzmann Machine's modeling capabilities can be increased thanks to more hidden variables with hidden unit names. The RBM model is a Botlzmann Machine with a limited architecture. With BM, visible units and hidden units can be connected. In RBM, it also consists of two layers, one is visible units and one is hidden units but there is no connection between visible units and visible units (visible – visible) or between hidden units and hidden units (hidden – hidden), they are completely independent [5]. The limited functionality between those units makes it easy to teach RMB. An example of an RBM model is described in Figure 3.

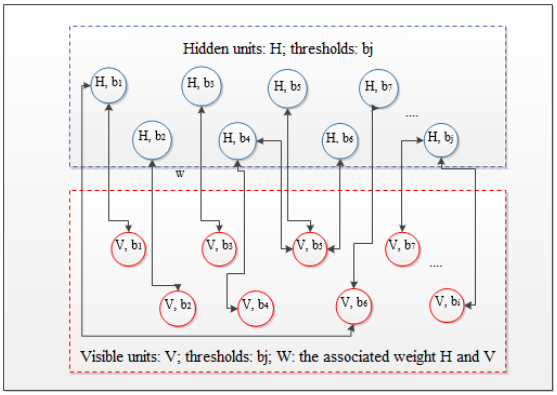

Fig. 3 A Restricted Boltzmann Machine only has a link between hidden units and visible units.

B. Feature Selection

One of the most important topics in machine learning is determining which data features will be used for forecasting in your model. Today, storing large amounts of data is easy. However, too many features can confuse the model and make the analysis meaningless. Benefits of reduced anticipation include reduced computational complexity, lower storage costs, and easier interpretation of results.

Selecting fearures (Feature Selection or Feature Extraction) [14] is a very important task in the preprocessing stage of data when deploying data mining models. One problem is that the datasets used to build Datamining Models often contain unnecessary (or even confusing) information for building models. For example, a dataset of hundreds of features used to describe the customers of a business is collected. However, when building a data mining model (Data mining model) only need about 50 features from hundreds of those features. If we use all (hundreds of) the features of our customers to build models, we need more CPU, more memory in the Training model process, even those unnecessary features that reduce the accuracy of the model and make it difficult to detect knowledge.

Feature selection methods usually calculate the score of features and then select only features with the best score to use for the model. These methods allow to adjust the threshold to get features with Score above the allowed threshold. The feature selection process is always performed before the Training Model process.

C. Some Feature Selection Methods

There are many methods to select features depending on the structure of the data used for the model and the algorithm used to build the model. The following are some common methods used in feature selection.

1. Interestingness score: Is used to rank features for features with continuous data types. An feature is considered to be interesting if it has some useful information (what is useful information depends on the problem being analyzed). To measure the level of interestingness, people often rely on entropy (a feature with a random distribution has a higher entropy and has a lower information gain) so those features are called less interesting).

Entropy of an feature will be compared with the entropy of all the remaining features by the following formula:

Interestingness (Attribute) = - (m - Entropy (Attribute)) * (m - Entropy (Attribute))

In which m is called the Central entropy, which means the entropy of the entire set of features)

2. Shannon's Entropy: Used for discretized data.

Shannon's entropy measures the uncertainty of a random variable for a particular outcome. For example, the entropy of tossing a coin can be expressed as a function of the probability of occurence of heads or tails.

Shannon's entropy is calculated using the following formula:

$$\mathrm{H}(\mathrm{x}) = \sum \mathrm{p}_{(\mathrm{x_i})} \log(\mathrm{p}(\mathrm{x_i}))$$

3. Other methods: In addition to the interestingness score and Shannon's entropy, several other methods are also commonly used in feature selection such as Bayesian with K2 Prior, Bayesian Dirichlet Equivalent with Uniform Prio.

There are three types of feature selection methods [15-17]: filter, wrapper and filter-wrapper. In preprocessing steps,

the filter methods will filter out irrelevant, redundant or jamming features based on some univariate or multivariate measures. In wrapper method, the predictor is a black box and the predictor performance is the objective function to evaluate the features of a subset. Wrapper methods perform better than filter methods because the process of selecting a feature is optimized for classification. However, the wrapper method requires more computational cost when used for large feature spaces because of the high computational cost and each feature set must be evaluated by a trained classifier. The filter method has lower and faster computational costs, but with lower reliability in classification efficiency than wrapper methods and it will be more suitable for high sample space datasets. Filter-wrapper methods have recently been developed in which they utilize both filter and wrapper methods. A filter-wrapper approach that uses both performance evaluation and independent testing of feature sets.

3. Proposed Method

In this section, we have presented the main RBM workflow proposed by the report with the feature selection method. The diagram frame is displayed the sequence of tasks to perform the classification as shown in Figure 4.

1. Data Preprocessing: In this frame, the credit data is cleaned through a preprocessing step by the scaling, standardization, and conversion functions in the preprocessing modules.
2. Features Selection with RBMs: The proposed method of the report is to use TensorFlow with GPU support to improve its runtime and accuracy. The proposed method is divided into two phases.
   - Phrase 1: To select the datasets that are ranked by the features, the dataset has been trained by RBMs. After model training, we calculated the rating value for each feature by using RMSE as the most important information capture in the RBMs according to data levels of measurement. Based on the rating values of the features, we select only the important features and remove the unrelated feature(s). The output of this phrase is provided the n-th features as the top-ranked ones.
   - Phrase 2: the top-ranked features in the Phase 1 are used as the filters for the test set.
3. Classification: In this study, we used several classification techniques that can be divided into three categories based on the types of algorithms used. The classifications using linear, nonlinear, or rule-based algorithms. Logistic regression (LR) and Linear Discriminant Analysis (LDA) are the classification techniques based on linear algorithms. The above two types of linear classifications are presented at first. A further classification category is the classifiers using non-linear algorithms. Those are Support Vector Machine (SVM), Artificial Neural Network (ANN) and the k-nearest neighbors (k-NN). The final classification category using rule-based algorithms is Random Forest (RF).

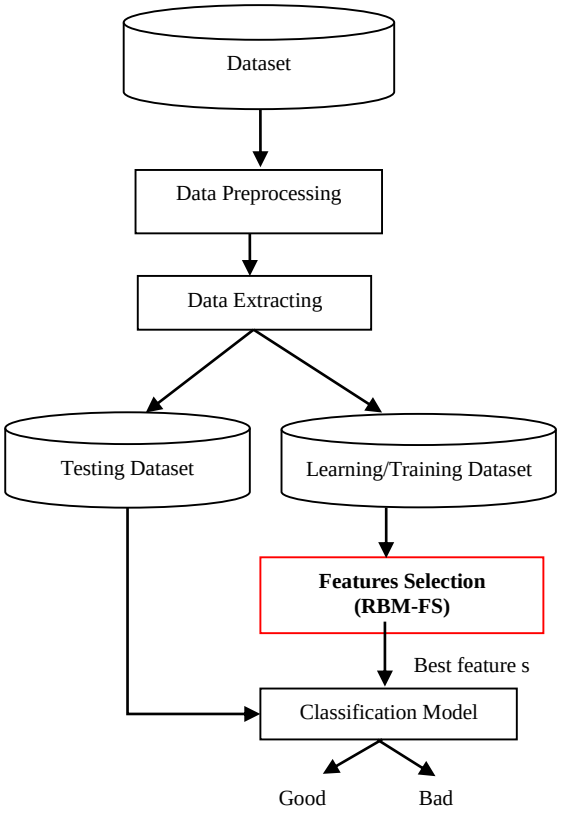

Fig. 4 Proposed method

4. Experiments and Results

Our proposed method has been implemented in Python language with Tensorflow GPU. We used datasets from the datasets for credit from Germany, Australia and the lending club (LC) (USA) to test the method. LDA classification was used with cross-tests in 5 times to evaluate selected features.

A.The test datasets

1. Australian credit dataset: Australia's credit database consists of 690 candidates of which 468 are good credit and 222 are bad credit. Each version contains 16 features, including loan_number, amount_borrowed, term, borrower_rate, last_payment_date, next_payment_due_date, days_past_due, etc.

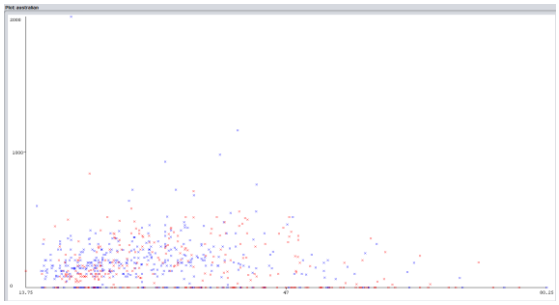

Fig. 5 Allocation of data according to the amount_borrowed feature of the Australian dataset

2. German credit dataset: The German credit database consists of 1,000 candidates, with 700 accepted persons and 300 rejected persons. Each candidate is described by 24 features that were retrieved for the classification model. Features include: name, ca_status, duration, credit_history, purpose, savings, job, etc.

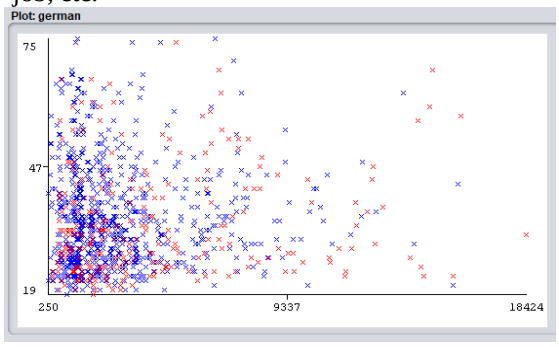

Fig. 6 Allocation of data by credit_amount / age feature of the German dataset

3. Lending Club Dataset (LC): The data used in the experimental section is available at: https://www.lendingclub.com/info/doad-data.action . After downloading, syncing, and undergoing data pre-processing, the final dataset selected contains more than 60,000 records and 123 features from 2007 to 2015. It all comes from the personal information provided on their loan application.

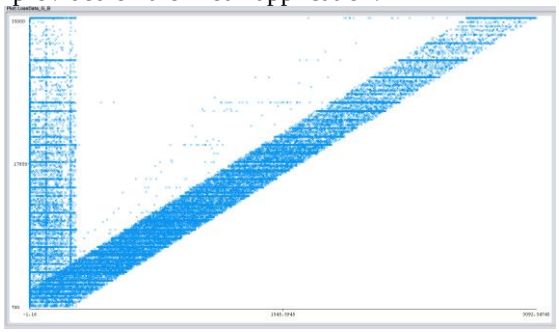

Fig. 7 Allocation of data according to loan_amount / installment feature of LC dataset

Among these features, not all of them are valid for the method proposed in this report, we eliminate the fields with low scoring weight through step 1 of preprocessing stage and select important features that affect the credit scoring method of the reporting method. Example: personal loan history, personal assets, personal payment period, etc

B. Test results

Firstly, we train algorithms based on the full subset of our predictive factors. For all models, we use cross-test in 5 times to create test samples or training samples that ensure our accuracy.

Table 1. Performance on the full model

Method	Predicted performance (%)		
	Germany	Australia	LC
IDA.	76.50	85.80	81.20
Logistic regression	76.40	85.51	81.05
ANN	75.80	71.45	66.08
k-NN	67.10	65.94	72.55
SVM Linear	74.80	84.35	76.60
Random forest	67.72	67.72	67.72

Obviously, the LDA model is superior to other models. In the LC dataset, the logistic regression model with an accuracy score of 81.20%. Next is the SVM linear model with an accuracy of 73.55%. k-NN, ANN and RF models have lower scores of 73.55% and 66% respectively. Therefore, LDA is chosen for our model in the feature selection model.

C. Model with feature selection

With such a large number of variables, we can intuitively feel that not all of them will be useful in predicting risks for a given loan. One solution to this problem would be to reduce the number of features we use in the model. An ideal reduction model will easily explain our predictions while still ensuring high accuracy.

Firstly, we choose from 2 to n features that we believe are the most informative related to lending.

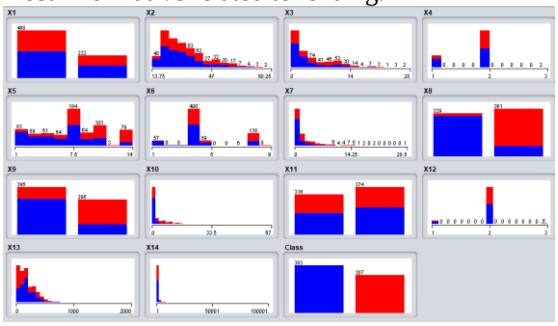

Fig. 8 Features of Australian dataset before being included in the model

It can be seen immediately that, based on the data description charts prior to the introduction of the Australian dataset, there are features that do not have much value for credit scoring. Therefore, the use of our method will help minimize costs when calculating because the method will select the most weighted and informative features related to lending.

After running the feature selection, in the case of the Australian credit dataset, our best subset included the 12 selected features and the method achieved the accuracy of 86.09%. The best feature set in the case of the German credit dataset consists of 22 features and the accuracy of 76.7%.

In the case of using LC data, the best subset consists of 80 features and an accuracy of 81.53%. Our models outperform modern and more powerful machine learning algorithms such as SVM linear, ANN, K-NN and random forests.

The highest weighted features in all three datasets besides the borrower's information related features are as follows:

- Purpose: What is the borrower's loan purpose?
- Housing: The borrower's house is a rented or owned house
- Savings: the savings the borrower has.
- Credit amount: credit limit of the borrower.
- Credit_history: payment history of the borrower.

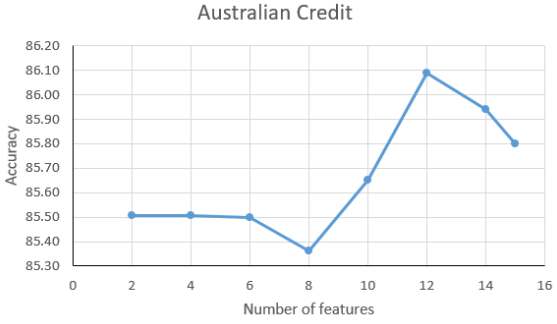

Fig. 9 Accuracy with Australian credit dataset

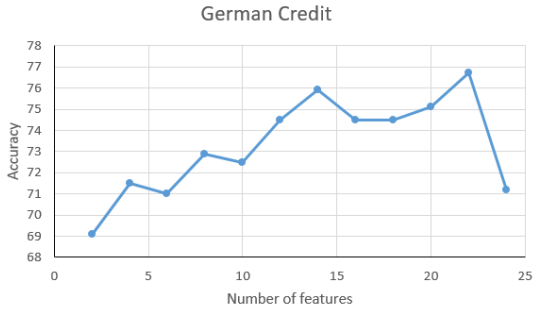

Fig. 10 Accuracy with German credit dataset

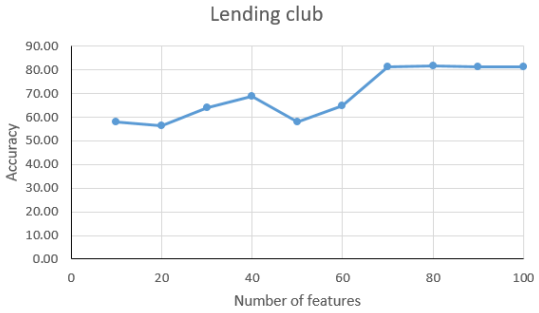

Fig. 11 Accuracy with LC credit dataset

5. Conclusion

In this report, we have introduced a relatively effective method based on feature selection and deep learning. Our method identifies n important features that affect risk prediction with the highest degree of accuracy. The accuracy of using LDA classification to select features is better than other methods. The performance of P2P lending platforms can be improved to get less risk. Process using GPU leads to reduced run time. Therefore, the workload of a credit officer in the risk assessment task may be reduced, because they do not have to include in the database a large number of unnecessary features during the assessment process. Test results show that our method is effective in credit risk analysis.

Acknowledgement

We would like to thank the research PTNTD19.06 “Market research in the era of 4.0 and prediction of consumer tastes based on network data collection”, the Institute of Information Technology and National Key Laboratory of Networking and Multimedia, Vietnam Academy of Science and Technology has supported the funding and the input data for this study. The article is also partially supported by FIRST Central Project Management Unit, Ministry of Science and Technology, Vietnam in Digital Grant Agreement No.12/FIRST/2a/IoIT.

Reference

[1] P2P Lending: What is an Expected Return? A Survey of Industry Voices”. LendingMemo. September 27th 2013. Publish March,28th, 2017.

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

[3] Jin, Y., & Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. Proceedings - 2015 5th International Conference on Communication Systems and Network Technologies, CSNT 2015, 609–613.

[4] Byanjankar, A., Heikkila, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015, 719–725.
[5] Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621–4631.
[6] Li, H., Zhang, Y., Zhang, N., & Jia, H. (2016). Detecting the Abnormal Lenders from P2P Lending Data. Procedia Computer Science, 91(Itqm), 357–361.
[7] Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decision Support Systems, 89, 113–122.
[8] Yan, J., Wang, K., Liu, Y., Xu, K., Kang, L., Chen, X., & Zhu, H. (2017). Mining social lending motivations for loan project recommendations. Expert Systems with Applications
[9] Xia, Y., Yang, X., & Zhang, Y. (2018). A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. Electronic Commerce Research and Applications, 30, 111–124.
[10] Deep Learning 101 - Part 1: History and Background. (2017, February 23). Retrieved from https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html
[11] Li Deng and Dong Yu (2016). Deep Learning Methods and Applications (2/2016) - published as Foundations and Trends® in Signal Processing Volume 7 Issues 3-4, ISSN: 1932-8346.
[12] Deep Learning in a Nutshell: History and Training. (2018, September 04). Retrieved from https://devblogs.nvidia.com/deep-learning-nutshell-history-training/
[13] Geoffrey Hinton. A Practical Guide to Training Restricted Boltzmann Machines, Aug 2, 2010, UTML TR 2010–003. http://www.cs.toronto.edu/ ~hinton/absps/guideTR.pdf
[14] Bermingham, Mairead L.; Pong-Wong, Ricardo; Spiliopoulou, Athina; Hayward, Caroline; Rudan, Igor; Campbell, Harry; Wright, Alan F.; Wilson, James F.; Agakov, Felix; Navarro, Pau; Haley, Chris S. (2015). "Application of high-dimensional feature selection: evaluation for genomic prediction in man". Sci. Rep. 5: 10312. Bibcode:2015NatSR...510312B.
[15] Hammon. Optimisation combinatoire pour la sélection de variables en régression en grande dimension : Application en génétique animale. November 2013 (in French)
[16] https://www.aaai.org/Papers/ICML/2003/ICML03 -111.pdf (PDF). External link in |title= (help)

Truong-Thang Nguyen received a Ph.D. in 2005 at the Japan Advanced Institute of Science and Technology (JAIST), Japan. Currently working at the Institute of Information Technology, Vietnam Academy of Science and Technology. Research fields: software quality assurance, software verification, program analysis.

Thanh-Son Khuat received the B.S. degrees in Information Technology from University of Engeneering and Technology, Vietnam National University in 2016, respectively. During 2016-2017, he stayed in Samsung Vietnam Mobile R&D Centre, Samsung Electronic Vietnam, to study mobile and application for samsung mobile. Currently working at the Institute of Information and Technology, Vietnam Academy of Science and Technology. Research fields: software quality assurance, software verification, program analysis.

Thi-Thu-Trang Ngo received B.E degree of Telecommunications and Electronics Engineering from Vietnam National University, Hanoi (VNUH) in 2002, and M.E degree of Computer and Communication Engineering from Chungbuk National University, Korea in 2005. Now, she is a lecturer and PhD student in Telecommunications Faculty of Posts and Telecommunications Institute of Technology (PTIT). Her research interests include optical communication, digital signal processing, and broadband networks.

Ha-Nam Nguyen received his BSc in Information Technology from Hanoi University of Science and Technology (HUST) in 1998, MSc in Computer Science from Chungwoon University (CWU), Korea in 2003, and PhD in Software Applications from Korea Aerospace University (KAU), Korea in 2007, respectively. At present, he is a vice-president of Information Technology Institute, Vietnam National University in Hanoi. He has more than 10 years of teaching and research experience in the areas of financial risk analysis, bioinformatics, behavior analysis, maritime logistics/transport and developing information systems using techniques from data analysis, modelling, and software engineering.

Manh-Dong Tran received a M.S. degree in 2013 at the University of Engineering and Technology, Vietnam National University, Hanoi. Currently working at the Institute of Information Technology, Vietnam Academy of Science and Technology. Research fields: software quality assurance, software verification, program analysis