International Journal of Economics, Finance and Management Sciences
2021; 9(6): 242-249
http://www.sciencepublishinggroup.com/j/ijefm
doi: 10.11648/j.ijefm.20210906.15
ISSN: 2326-9553 (Print); ISSN: 2326-9561 (Online)

Predicting Enterprise Performance by Network-Based Risk Profiling Approach

Longda Tong[1], Song Yin[2], Dongdong Hu[2], Zhaoyuan Li[1, *]

[1]School of Data Science, the Chinese University of Hong Kong, Shenzhen, China

[2]WeBank Co., Ltd., Shenzhen, China

Email address:

118010280@link.cuhk.edu.cn (Longda Tong), sunnyin@webank.com (Song Yin), dongdonghu@webank.com (Dongdong Hu), lizhaoyuan@cuhk.edu.cn.com (Zhaoyuan Li)

[*]Corresponding author

To cite this article:

Longda Tong, Song Yin, Dongdong Hu, Zhaoyuan Li. Predicting Enterprise Performance by Network-Based Risk Profiling Approach. *International Journal of Economics, Finance and Management Sciences*. Vol. 9, No. 6, 2021, pp. 242-249. doi: 10.11648/j.ijefm.20210906.15

Received: November 4, 2021; **Accepted**: November 24, 2021; **Published**: December 2, 2021

Abstract: Deriving the firms' risk profile based on specific features has important implications in risk controlling and investment. Recently, much research demonstrates that the firms' ownership networks substantially impact the firms' risk profile. In this paper, we propose a framework of risk profiling approach built upon information retrieved from the firm's ownership networks. The method considers the non-linear relationships between firm fundamentals with network structures. To test the performance of the proposed method, we construct a new dataset of Chinese listed firms with their financials and network parameters in the period between 2005 and 2020. We show that the proposed method significantly outperforms traditional ones in predicting a firm's market value changes. Specifically, we first use the conventional linear method, like logistic regression and linear discriminant analysis, as the performance benchmark. Then, the more advanced technique based on information theory like Gradient Boosting is adopted and has shown remarkable performance with at least 85% area under the curve (AUC) compared with the 60% AUC of the traditional linear model. The proposed method has implications in risk management, portfolio management, and corporate finance. As a special implication example in risk management, we demonstrate that a network-based approach can effectively detect duplication of individual names in a unique dataset.

Keywords: Risk Profiling, Risk Management, Complex Network, Statistical Methodology

1. Introduction

1.1. Overview

How does a firm's position on the financial network matter for its risk profile? Recent studies investigate the implications of various networks, such as financial transactions [1], supply chains [2], competition relationships [3], social ties [4, 5], directorship relations [6, 7], and political connections [8], on corporate policies and stakeholder's decisions. In particular, there is emerging literature that examines the impact of ownership networks. It is well documented that ownership structure significantly shapes firm decisions [9, 10]. For example, ownership structure affects firms' loan terms and pricing [11, 12]. Recent studies examine the specific network positions that matter for firm growth and risk [13, 14]. In this study, we probe into the network parameters on firm ownership networks and combine the network parameters with fundamental firm traits to assess their implications for firm performance.

The essence of the proposed method is to exploit the valuable information hidden in the non-linear relationships between network parameters and fundamental firm characteristics. Specifically, we construct unique data that contains ownership relationships among Chinese listed firms and construct the network parameters of each firm in the networks. The impact of firm fundamentals will differ when the firm is on the different positions of the network. We construct a broad range of variables based on this approach and employ several statistical methods to show the performance. We find the proposed method outperforms other approaches.

We next explore whether this method can be used in

another tricky practice in the industry, detecting name duplications. In risk management, practitioners often face the challenge that individuals with multidimensional variables may not be from the same individual, especially in the case of China, where the duplicated individual names are prevalent. To shed light on this critical issue, we first construct a unique dataset of person-firm pairs with correct identification that whether two individuals with the same name correspond to the same person and then apply our proposed method. We find our approach is shown to outperform other existing methods. This method and the newly constructed training dataset are potentially valuable for real tasks in risk management practices.

1.2. Significance and Contribution

This paper contributes to the literature in two aspects. First, this study adds to the growing literature on the impact of network parameters. For example, recent studies document that CEO centrality affects merger outcome [15], and underwriter centrality impacts initial public offering (IPO) characteristics [16]. Our paper essentially explores the implications of the firm location on the ownership networks on risk management and firm development and relates to the studies mentioned above.

Second, this paper joins the immense literature on statistical methodologies for firm risk profiling. The recent studies in this field include methods based on natural language processing (NLP) [17, 18], high dimensional statistics [19], etc.

This study also has essential industry implications. First, it provides a framework that can apply to multiple occasions. The framework proposed can also be extended to supply chain networks, social networks, and other types of networks. The outcome of interests in the framework can also be generalized to portfolio management and individual risk management situations. For example, although we illustrate our method using a dataset of firms, in practice, the process can also be applied to asset the risks of individuals.

Second, we construct a unique training database for the detection of duplicated individual names. Although the primary purpose of the dataset is to illustrate the performance of the proposed method, the dataset itself contains merits for practitioners. On the on hand, practitioners could use this dataset to train their models. For example, a firm identifies several individuals with the same name and needs to identify duplicated names. On the other hand, practitioners can extend the variables in the dataset by constructing the firm's fundamentals in their situation for their purposes.

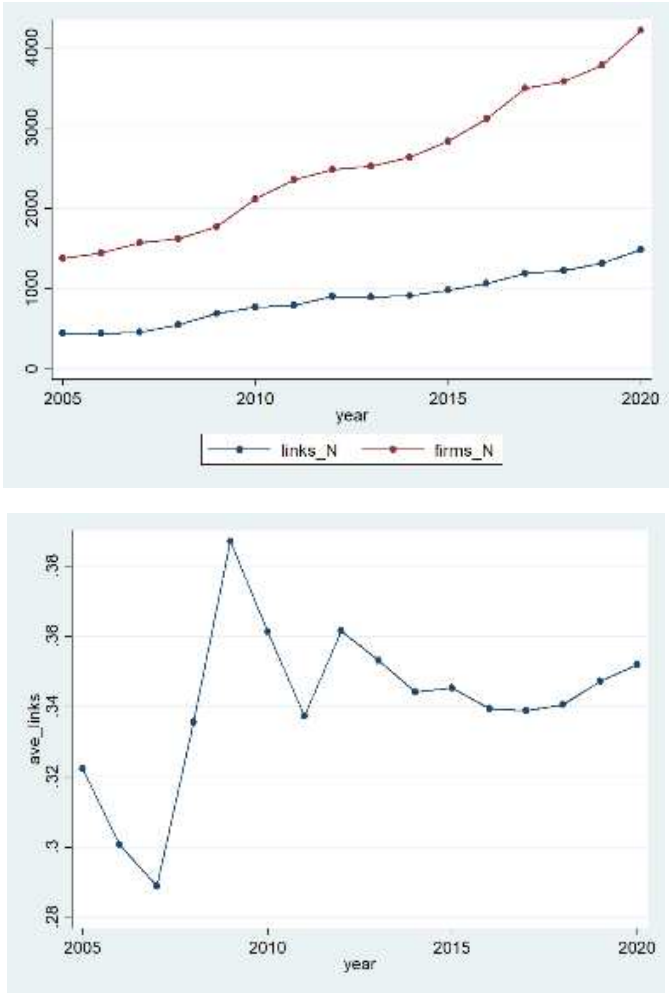

Figure 1. *Sample Firms by Time.*

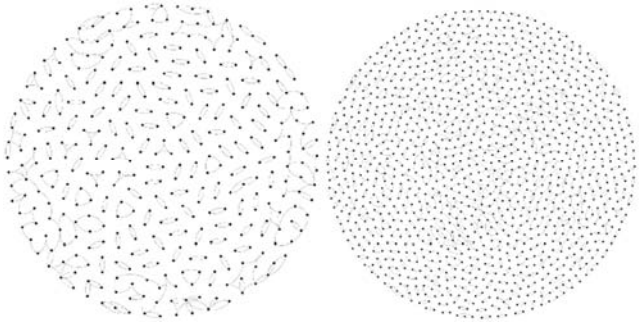

Figure 2. *Evolution of Ownership Networks.*

2. Data and Methods

2.1. Evolution of Ownership Networks

We construct the data from annual reports to gauge the firm's ownership relationships. The original data is from 205 to 2020. Later we will restrict the sample to observations with essential financial data. We report the trajectory of the network evolution in Figure 1. The left panel suggests both the number of firms and the number of links grow over the years. The right panel shows that the network intensity peaks in 2009 after the financial crisis and remained stable after 2012.

We visualize the network structures of two years in Figure 2 for illustration. The left panel plots the ownership network in 2005, and the right panel plots the ownership network in 2018, showing that the network intensity on average increases over the years.

2.2. Firm Basic Characteristics

Before we dive into the discussions on network parameters, we select several firm characteristics, which have an important impact on firm performance and growth. We first construct time-invariant variables such as industry dummies and location dummies. Then we consider ownership types, such as variables about the private, state, and foreign ownership. Last, we consider a broad range of financial variables, such as firm size, leverage, profitability, related party transactions, productivities, government subsidies, etc. We also construct the quadratic terms of those financial variables to capture the non-linear effects. This group of variables is labeled as variable list A.

We mainly consider the change in market value for the primary outcome variables, as the change in market value is forward-looking has implications for investment and risk management [26]. To facilitate model testing, we define a dummy variable set to one when the market value change is in the bottom decile and zero otherwise.

2.3. Network Parameters

We consider several classic network parameters to pinpoint a firm's location in the network. The local parameters we consider include degree centrality, eigenvector centrality, and closeness centrality.

The method we propose has several procedures as follows. First, we construct a list of variables for the firm's direct neighbors on the network. This group of variables is labeled as variable list B. Second, we calculate the additional list of variables that interact with those above-constructed variables and local network parameters. This group of variables is labeled as variable list C.

2.4. Statistical Models

We employ several classic statistical methods and high dimensional methods in the data and compare the performance. The baseline model candidates include Logistic Regression (LR), K-Nearest Neighborhood with K=3 (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Random Forest (RF) [22]; 6) AdaBoost [24]; 7) Light GBM [21]; 8) XGBoost [20]. We also consider several high dimensional models, including high-Dimensional LDA (RLDA & OIILDA) [25], SROAD1, and SROAD2 [23]. In what follows, we will test the performance of the model based on different combinations of the datasets.

Table 1. *Performance of the Methods in Predicting Market Value.*

Model	Data Groups	AUC
LR	A	0.668
LR	A+B	0.583
LR	A+B+C	0.558
KNN	A	0.557
KNN	A+B	0.601
KNN	A+B+C	0.577
NB	A	0.534
NB	A+B	0.527
NB	A+B+C	0.520
LDA	A	0.681
LDA	A+B	0.651
LDA	A+B+C	0.664
RF	A	0.719
RF	A+B	0.714
RF	A+B+C	0.768
AdaBoost	A	0.620
AdaBoost	A+B	0.605
AdaBoost	A+B+C	0.734
LGBM	A	0.702
LGBM	A+B	0.710
LGBM	A+B+C	0.844
XGBoost	A	0.707
XGBoost	A+B	0.722
XGBoost	A+B+C	0.848
RLDA	A	0.687
RLDA	A+B	0.679
RLDA	A+B+C	0.682
OIILDA	A	0.565
OIILDA	A+B	0.613
OIILDA	A+B+C	0.561
SROAD1	A	0.642
SROAD1	A+B	0.545
SROAD1	A+B+C	0.642
SROAD2	A	0.642
SROAD2	A+B	0.545
SROAD2	A+B+C	0.545

3. Results on Risk Profiling

3.1. Data Preprocessing

We prepare the data in the following steps:

1. We remove all the missing values in the original datasets;
2. To address multicollinearity issues, we remove one feature in a pair of features with the Pearson correlation greater than 0.9;
3. We construct the training and testing set by splitting the dataset.

After preprocessing, the remaining dataset consists of 28790 instances with 970 features. The dataset includes both continuous and discrete features. We categorize the variables according to the rules above and construct variable lists Group A (152 features), Group B (148 features), and Group C (287 features).

3.2. Performance

We use three combinations of the datasets to show the performance of each of the classic methods. First, we only consider variable list A, which only contains basic firm fundamentals, such as industry, location, ownership type, basic financial traits, etc. Second, we add the same set of variables based on the feature of the neighboring firms of a given firm according to the network structure. The variable list considered here is the variables in group A plus group B. Last but not the list, we use the full set of variables (A+B+C). We add all interactions between variables in B and the firm's network parameters discussed in section 2.3.

The key element of our proposed method is that we construct additional variables based on network features. We conjecture that the inclusion of the full set of variables constructed would improve the predictive power in risk profiling.

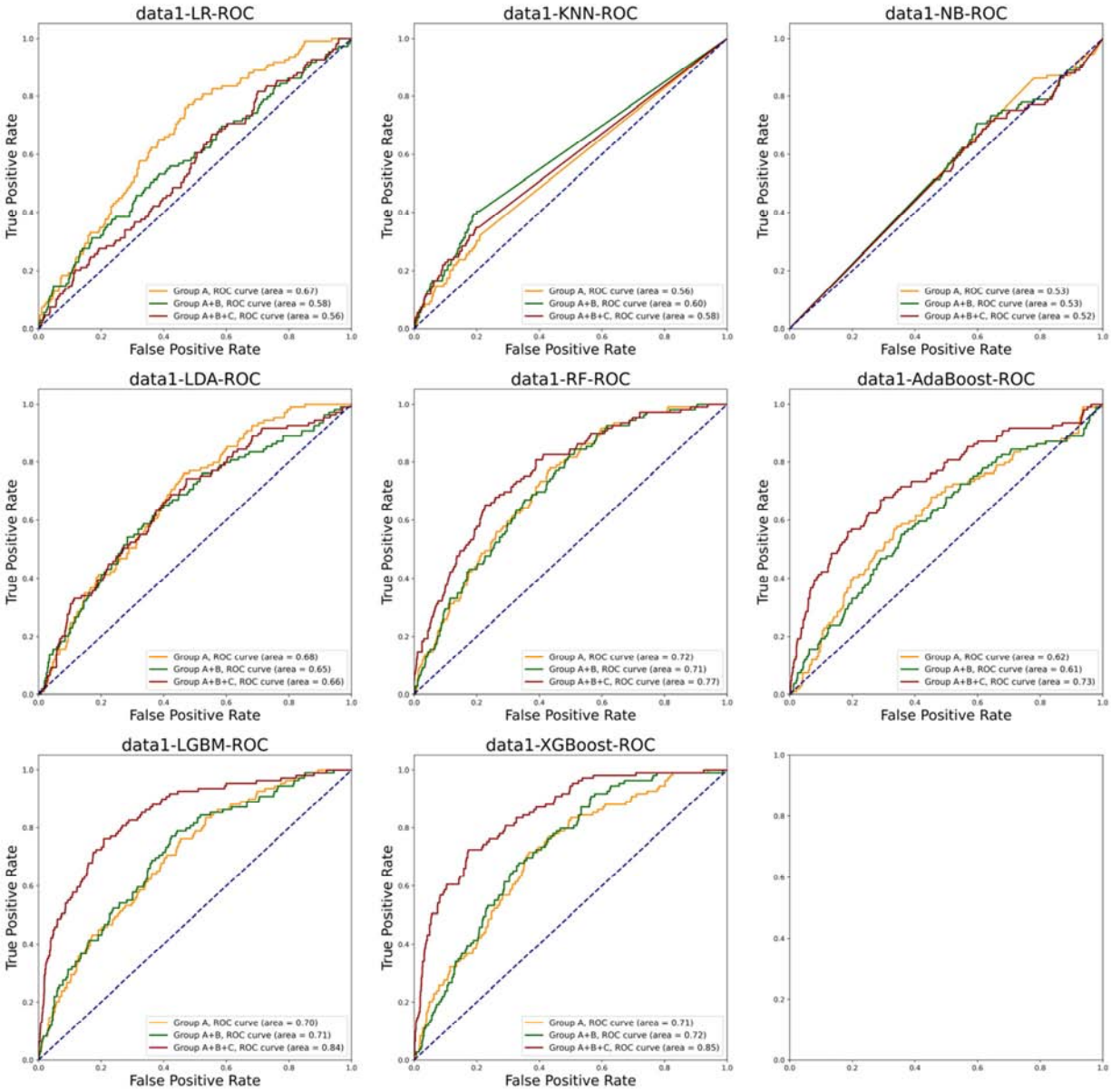

Figure 3. *Performance of Classic Models.*

We first consider traditional statistical methods and then discuss the potential benefits of high-dimensional models. We use the receiver operating characteristic (ROC) curve as the evaluation metrics. A greater area under the ROC curve. (AUC) indicates the superior performance of a model. The results are summarized in Table 1. The ROC plots are presented in Figure 3 and Figure 4.

We consider several high-dimensional models and present the ROCs in Figure 3.

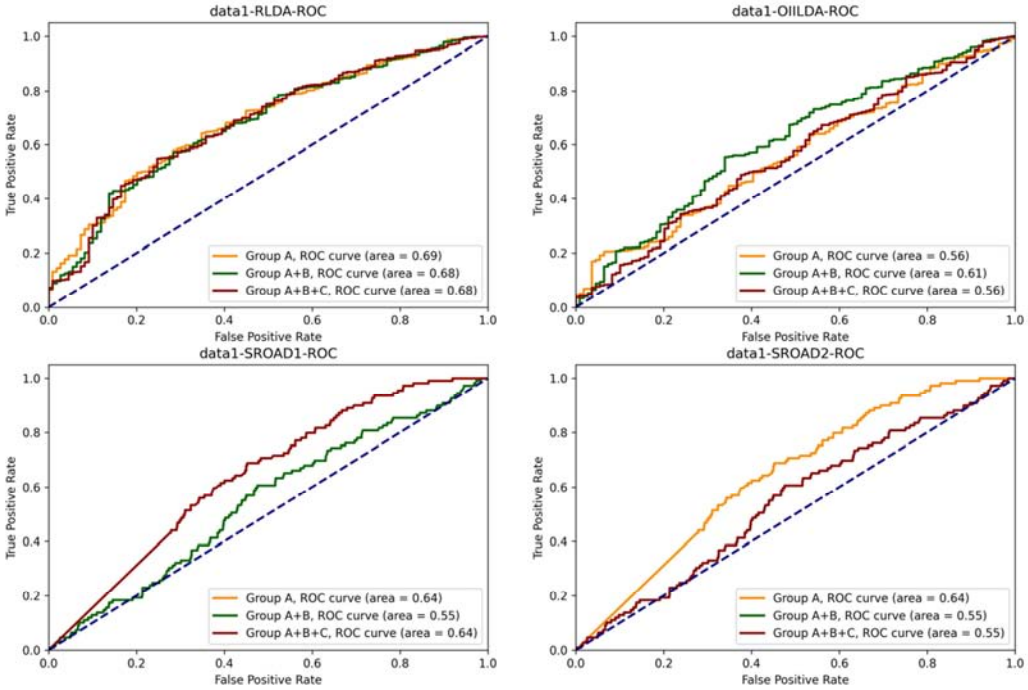

Figure 4. *Performance of High-dimensional Models.*

We report our findings in the following. First, recent machine learning approaches outperform traditional models. For example, in general, the performance of Light GBM [21] and XGBoost [20] outperforms traditional methods, such as Logistic Regression (LR) and K-Nearest Neighborhood methods (KNN). After comparing the performance across different models, we find XGBoost is the best performing method in predicting changes in market value. Third, we find the newly added variables increase the AUC by a lot, leading to a superior performance of the method. For instance, as reported in Table 1, XGBoost has an AUC of 0.848 when the full set of variables is used, compared with an AUC of 0.717 when basic features are considered. The proposed method based on network parameters increases the performance by 18.3%.

There are several reasons for our findings. For example, the dataset contains both the continuous and discrete features, and the tree-based model alleviates some problems of mixed-up feature spaces. The best performing method XGBoost is based on machine learning algorithms under the Gradient Boosting framework and provides a parallel tree boosting that solves many data science problems quickly and accurately [27]. Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

We also consider several high-dimensional models including, high-Dimensional LDA (RLDA & OIILDA) [25], SROAD1, and SROAD2 [23]. As shown in Table 1 and Figure 3, the high-dimensional method does not perform well in the dataset and thus is not competitive to the tree-based boosting algorithm, XGBoost. One potential reason is that the number of observations in this task is large, and the concern arising from high-dimensional issues is mitigated.

4. Detection of Duplications in Individual Names

One practical issue in risk management is that it is challenging to identify whether individuals sharing the same name correspond to the same person. The failure to detect the correct individuals leads to measurement errors in assessing individual risk profiles and generates losses in business operations. This issue becomes severe as individual Chinese names have more common first names as well as family names, leading to the issue prevalent in the real world.

As there are few previous studies in this field, we thus are among the first studies that tackle this vital issue. In particular, we employ the framework discussed above and demonstrate the performance in this specific case by constructing a unique training dataset.

4.1. A Unique Dataset of Individual-Firm Pairs

We first construct a unique dataset that consists of individual-firm pairs. In reality, a typical issue is that

individual names are associated with two firms, but it is not clear whether they are referring to the same person. We collect all personnel characteristics for Chinese listed companies and identify the correct identities. To fit our research purpose, we construct individual-firm pairs. In the data, each name is associated with more than one firm. The outcome variable is a dummy set to one of those records point to the same individual and zero otherwise.

We then collect data to construct variables at the firm level. The idea is that two individual names are more likely to correspond to the same person when the associated firms are similar across some characteristics or firms are interconnected in some aspects.

4.2. Variables and Design

Similar to the previous task in predicting changes in market value, we also construct three sets of variables. Group A contains name traits, such as the length of the names, the popularity of the first name, and the family name's popularity. Then we construct Group B, which contains firm characteristics such as industry, location, firm size categories, etc. Last but not least, we construct nuanced measures that capture firm interdependence. Specifically, we consider ownership relationships and supply chain relationships. The underlying mechanism is that firms along the same supply chains or ownership chains are more likely to have the same individual involved.

4.3. Performance

We apply the same preprocessing process. The raw data consists of 119479 instances and 510 features. Dataset 1: The feature space contains both continuous and discrete features. After preprocessing, *the remaining dataset consists of* 49286 instances and 509 features. The outcome variable identifies 36505 cases corresponding to the same person and 12781 records not. The ROC and results are presented in Table 2, Figure 5, and Figure 6.

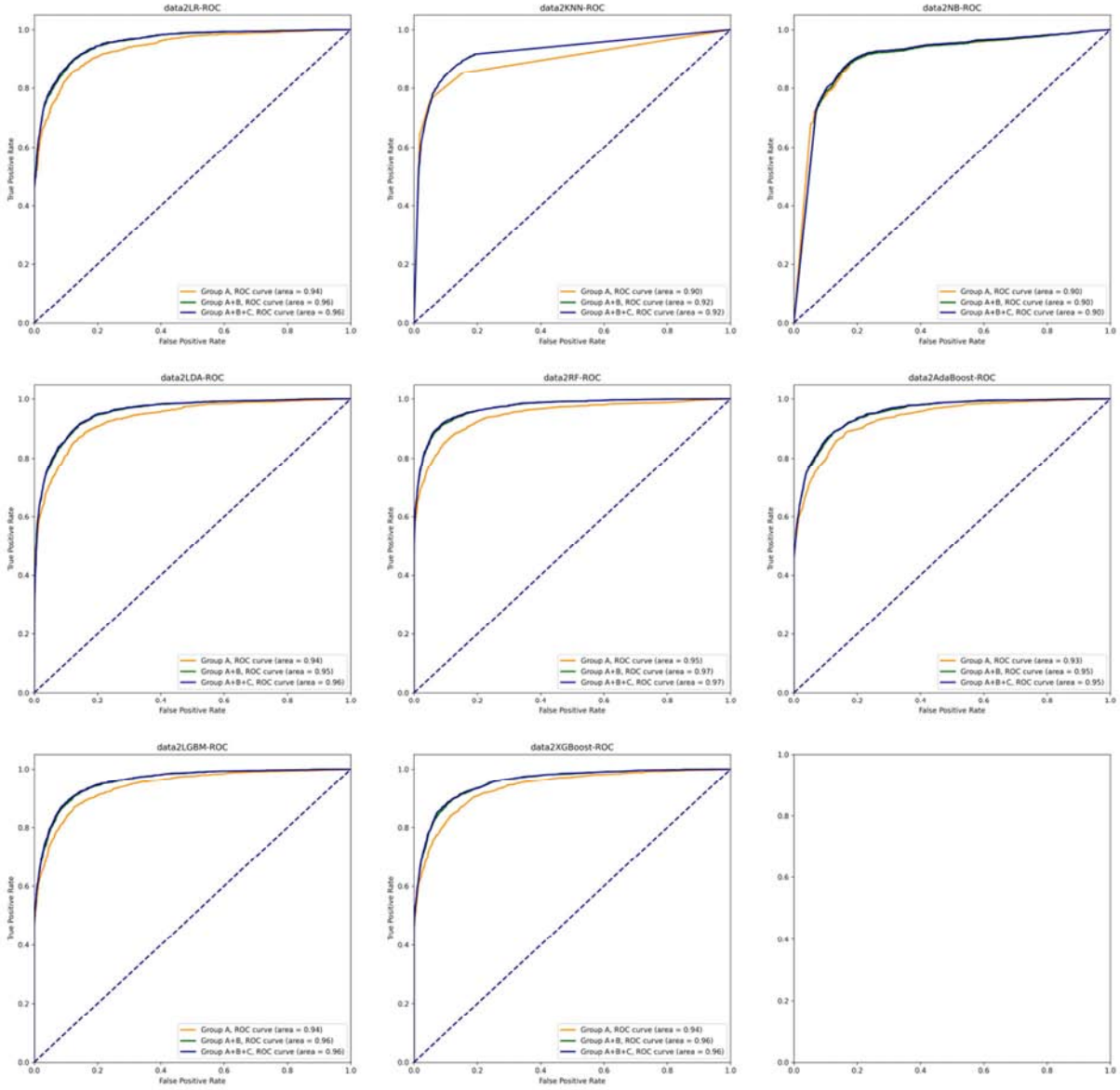

Figure 5. *Performance of Classic Models.*

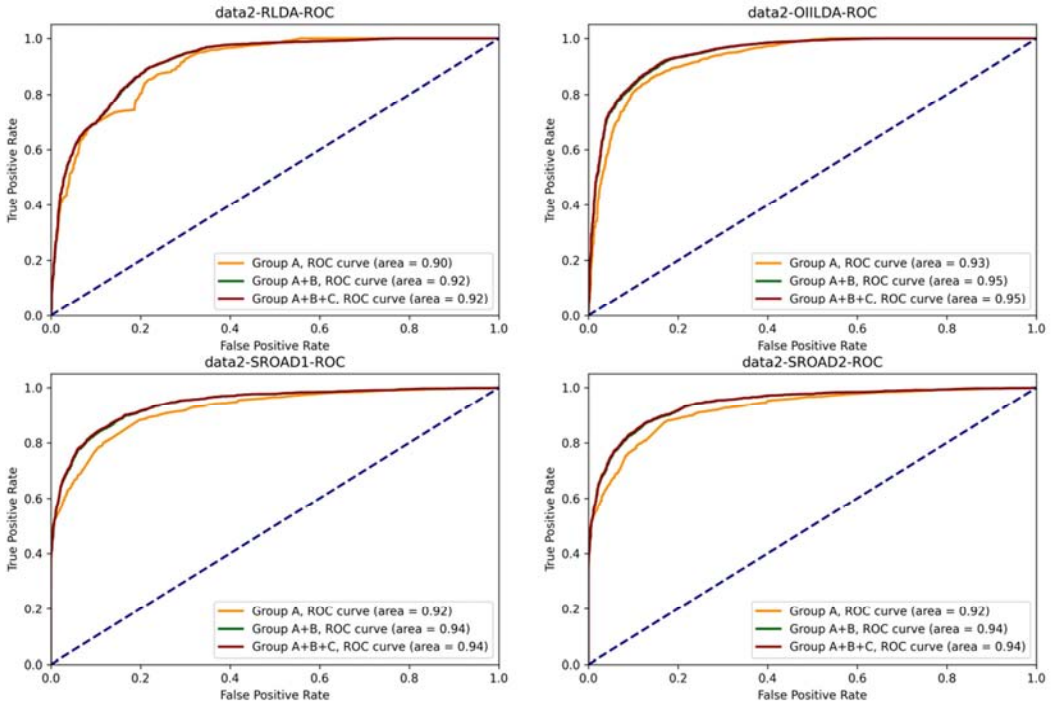

Figure 6. *Performance of High-dimensional Models.*

Table 2. *Performance of the Methods in Predicting Duplicated Names.*

Model	Data Groups	AUC
LR	A	0.937
LR	A+B	0.955
LR	A+B+C	0.957
KNN	A	0.895
KNN	A+B	0.923
KNN	A+B+C	0.923
NB	A	0.905
NB	A+B	0.900
NB	A+B+C	0.904
LDA	A	0.935
LDA	A+B	0.954
LDA	A+B+C	0.955
RF	A	0.946
RF	A+B	0.970
RF	A+B+C	0.971
AdaBoost	A	0.934
AdaBoost	A+B	0.952
AdaBoost	A+B+C	0.954
LGBM	A	0.941
LGBM	A+B	0.958
LGBM	A+B+C	0.959
XGBoost	A	0.937
XGBoost	A+B	0.955
XGBoost	A+B+C	0.956
RLDA	A	0.903
RLDA	A+B	0.916
RLDA	A+B+C	0.917
OIILDA	A	0.930
OIILDA	A+B	0.945
OIILDA	A+B+C	0.947
SROAD1	A	0.921
SROAD1	A+B	0.942
SROAD1	A+B+C	0.942
SROAD2	A	0.923
SROAD2	A+B	0.943
SROAD2	A+B+C	0.943

We discuss the performance briefly as follows. First, in general, the training dataset could generate very good results. The AUC is within the range of 0.9 to 0.97. Second, the best performing method is Random Forest (RF). Last, the addition of the network variables increases the performance by roughly 2.6%, which is sizable given the predictive performance is already above 90%.

5. Conclusion

This paper proposes a framework for risk profiling based on network parameters. First, we detail the mechanism of the method. Second, we demonstrate how traditional machining learning combined with the network parameters can increase the predictive power in assessing a firm's risk profile. Then, we apply our proposed method in another important task in risk management, the detection of duplicated individual names. By constructing a unique dataset, we show that machining learning methods can generate decent performance, and the method based on additional network features improves performance. Specifically, we find that the model based on information theory, like XGBoost with Decision Trees, shows exceptional performance on both datasets. This finding demonstrates risk profiling based on ownership network indeed provide important information in determining the firm’s risk profile.

Acknowledgements

This project is partially supported by CCF-Tencent Rhino Bird Grant (CCF-Webank RAGR20200105) and the University Development Fund of CUHK-SZ (UDF01001160).

References

[1] Acemoglu, D., Ozdaglar, A. and Tahbaz-Salehi, A., 2015. Systemic risk and stability in financial networks. American Economic Review, 105 (2), pp. 564-608.

[2] Ostrovsky, M., 2008. Stability in supply chain networks. American Economic Review, 98 (3), pp. 897-923.

[3] Hoberg, G. and Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124 (5), pp. 1423-1465.

[4] Ishii, J. and Xuan, Y., 2014. Acquirer-target social ties and merger outcomes. Journal of Financial Economics, 112 (3), pp. 344-363.

[5] Gompers, P. A., Mukharlyamov, V. and Xuan, Y., 2016. The cost of friendship. Journal of Financial Economics, 119 (3), pp. 626-644.

[6] Pugachev, L. and Schertler, A., 2021. Neglecting Peter to Fix Paul: How Shared Directors Transmit Bank Shocks to Nonfinancial Firms. Journal of Financial and Quantitative Analysis, pp. 1-38.

[7] Bouwman, C. H., 2011. Corporate governance propagation through overlapping directors. The Review of Financial Studies, 24 (7), pp. 2358-2394.

[8] Acemoglu, D., Hassan, T. A. and Tahoun, A., 2018. The power of the street: Evidence from Egypt's Arab Spring. The Review of Financial Studies, 31 (1), pp. 1-42.

[9] Demsetz, H. and Lehn, K., 1985. The structure of corporate ownership: Causes and consequences. Journal of political economy, 93 (6), pp. 1155-1177.

[10] La Porta, R., Lopez-de-Silanes, F. and Shleifer, A., 1999. Corporate ownership around the world. The journal of finance, 54 (2), pp. 471-517.

[11] Lin, C., Ma, Y., Malatesta, P. and Xuan, Y., 2012. Corporate ownership structure and bank loan syndicate structure. Journal of Financial Economics, 104 (1), pp. 1-22.

[12] Lin, C., Ma, Y., Malatesta, P. and Xuan, Y., 2013. Corporate ownership structure and the choice between bank debt and public debt. Journal of Financial Economics, 109 (2), pp. 517-534.

[13] Mizuno, T., Doi, S. and Kurizaki, S., 2020. The power of corporate control in the global ownership network. Plos one, 15 (8), p. e0237862.

[14] Garcia-Bernardo, J., Fichtner, J., Takes, F. W. and Heemskerk, E. M., 2017. Uncovering offshore financial centers: Conduits and sinks in the global corporate ownership network. Scientific Reports, 7 (1), pp. 1-10.

[15] El-Khatib, R., Fogel, K. and Jandik, T., 2015. CEO network centrality and merger performance. Journal of Financial Economics, 116 (2), pp. 349-382.

[16] Bajo, E., Chemmanur, T. J., Simonyan, K. and Tehranian, H., 2016. Underwriter networks, investor attention, and initial public offerings. Journal of Financial Economics, 122 (2), pp. 376-408.

[17] Hoberg, G. and Maksimovic, V., 2015. Redefining financial constraints: A text-based analysis. The Review of Financial Studies, 28 (5), pp. 1312-1352.

[18] Bodnaruk, A., Loughran, T. and McDonald, B., 2015. Using 10-K text to gauge financial constraints. Journal of Financial and Quantitative Analysis, 50 (4), pp. 623-646.

[19] Fan, J., Lv, J. and Qi, L., 2011. Sparse high-dimensional models in economics. Annu. Rev. Econ., 3 (1), pp. 291-317.

[20] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, pp. 3146-3154.

[22] Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2 (3), pp. 18-22.

[23] Li, Z. and Yao, J., 2016. On two simple and effective procedures for high dimensional classification of general populations. Statistical Papers, 57 (2), pp. 381-405.

[24] Schapire, R. E., 2013. Explaining adaboost. In Empirical inference (pp. 37-52). Springer, Berlin, Heidelberg.

[25] Sifaou, H., Kammoun, A. and Alouini, M. S., 2020. High-dimensional linear discriminant analysis classifier for spiked covariance model. Journal of Machine Learning Research, 21 (112), pp. 1-24.

[26] Marcus, A. J., Bodie, Z. and Kane, A., 2004. Essentials of investments. McGraw Hill.

[27] XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/.