UNIVERSITE BRETAGNE ECONOMIE LOIRE ET GESTION

UNIVERSITÉ DE RENNES 1

Thèse de doctorat de

L'UNIVERSITE DE RENNES 1

Commue Université Bretagne Loire

Ecole Doctorale n° 597 EDGE

Sciences Economiques et Sciences de Gestion

Spécialité : « Sciences de Gestion »

Par

Hong Hanh LE

Machine learning and Applications: New models to predict the bankruptcy of banks

Thèse présentée et soutenue à Rennes, le 21 Novembre 2018

Unité de recherche : CREM UMR CNRS 6211

Rapporteurs avant soutenance :

Laurent Weill -Professeur – Université de Strasbourg
Michael Joseph Dempsey -Professeur- RMIT Australie

Composition du Jury :

LAURENT WEILL
Professeur – Université de Strasbourg

MICHAEL JOSEPH DEMPSEY
Professeur – RMIT Australie

THI LE HOA VO
Maître de conférences HDR- Université de Rennes 1

JEAN-LAURENT VIVIANI
Professeur- Université de Rennes 1/ *directeur de thèse*

ACKNOWLEDGEMENT

It is my great pleasure to acknowledge and thank those who made this thesis possible!

At this moment of accomplishment, with a deep sense of gratitude and sincerity, I acknowledge the guidance of ***Professor Jean-Laurent VIVIANI***, who has expertly guided me since Master level. During 5 years of being his student, I am always inspired and motivated by his wise knowledge and enthusiasm. I am deeply indebted to him for his responsible guidance, support, supervision and because he shows me to be not only a researcher but also a good lecturer. Sincerely, I feel how lucky to be his PhD. Student!

My sincere thanks also go to ***Professor Franck MORAUX*** for giving me very valuable advices to improve my research. I would like to thank to the committee members: ***Professor Laurent WEILL, Professor Michael DEMPSEY, Dr. Thi Le Hoa VO*** for their time and their interest in my thesis.

I remain grateful to all members and staff of ***EDGE, IGR and CREM,*** who are always willing to help me.

I owe my heartfelt thanks to ***my grandfather, my parents, my parents-in-law, my aunt Nhi Duong, my sister and my sister in law*** for always trusting and encouraging me. I would like to deliver the special gratefulness to my parents, who always support me in both financial and mental way. I love you! Especially, this thesis is a gift for my little sister -***Hanh Phuc***, whom I care and love more than anything else in the world.

There have been many people who have walked alongside me for the last four years. Without them, this thesis would not have been possible. To: ***Fabien (and his parents)*** – who is always available as a more-than-great friend, plays chess and gives advice, ***Mansour*** - who starts and finishes PhD along with me and spends a lot of time on encouraging me, ***Wenting*** – who has finished her PhD and went back to China but always keeps in touch, ***Minh Khue, Kien Nguyen*** – who stay beside and support me by all means, ***Thu Sang*** – who has been my best friend since high-school and took flights from the UK to visit me every summer, ***Mme. Odile*** – who welcomed me to IGR and became my French mother, *and other friends who I shared great time with.*

I also want to deliver my special acknowledge to the Board of Directors of ***Ton Duc Thang University and Faculty of Finance and Banking (Vietnam),*** where I work now, for supporting and facilitating for me to spend time on this thesis. My special gratitude to the Dean of Faculty – ***Dr. Trang Hoang***, and my colleges ***Dr. Huy Pham and Ms. Van Nguyen,*** who supported me with all their best.

Last but not the least, to my husband – ***Bich Le,*** we got married since the first year of my PhD. and after 4 years long, thank you for always being patient and encouraging me to keep moving forward! Who trusted me the most in this thesis? - Of course, it is you, Honey! I do believe that one day, our kids will be proud of their parents.

LE HONG HANH

CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The thesis consists of six chapters. Each chapter can be read independently of the others, but all six chapters share the thesis's overall topic: **Using Machine learning techniques to explain and predict the bankruptcy of banks.**

Chapter 1 presents briefly the motivations and contributions of this thesis.

Chapter 2 introduces the global review of the literature: (1) The research articles and main findings in banks' failure prediction, (2) Comparison of statistical and machine learning techniques

Chapter 3 compares the accuracy of two approaches: traditional statistical techniques and machine learning techniques, which attempt to predict the failure of banks. A sample of 3000 US banks (1438 failures and 1562 active banks) is investigated by two traditional statistical approaches (Discriminant analysis and Logistic regression) and three machine learning approaches (Artificial neural network, Support Vector Machines and k-nearest neighbours). For each bank, data were collected for a 5-year period before they become inactive. 31 financial ratios extracted from bank financial reports covered 5 main aspects: Loan quality, Capital quality, Operations efficiency, Profitability and Liquidity. The empirical result reveals that the artificial neural network and k-nearest neighbour methods are the most accurate.

Chapter 4 investigates the material loss review published by the Federal Deposit Insurance Corporation (FDIC) on the U.S. failed banks from 2008 to 2015. These reports focus on explaining the causes of failure and material loss of each bank. Unlike traditional methods that provide suggestions on financial ratios, this study focuses on phrases extracted from the reports by using text mining technique. Pre-processing steps are used in this study to

'clean' the text. Bag of words technique is used for collecting the most frequent words. Topic modelling and document hierarches clustering are used for classifying these reports into groups. Our results suggest that to prevent from being the failure, banks should significantly be aware of: loan, board management, the supervisory process, the concentration of ADC (Acquisition, Development and Construction) and CRE (Commercial real estate). In addition, we find the main reasons that US banks went failure from 2008 to 2015 are covered by two main topics: Loan and Management.

Chapter 5 investigates in the efficiency of Loan loss provision of large US banks. Loan loss provision (LLP) is a significant important item in bank's financial report that can be used to (i) evaluate the level of credit risk, or (ii) smooth bank's income statement. In this research, LLP is dedicated as a cushion of credit default. Theoretically, LLP should be maintained at the appropriate amount accordingly to the potential credit loss, neither surplus nor deficit of LLP is favourable. In this study, a two-stage DEA (performance and LLP stage) is used to appraise whether LLP is effectively reserved. Then, Neural network proposes the adjustment if needed for each bank to improve LLP quality management. The dataset includes 166 large US banks in the period of 2016. The results suggest that (1) Only 12.7% of given banks are operating effectively for performance process; (2) Even fewer (only 2.4%) banks reserve LLP at the appropriate level; (3) More efficiency banks tend to have higher *(a) Number of employees, (b)Total equity , (c) Total expenses, (d) Total deposit, (e) Total loan, (f) Total investment and higher (g) Loan Loss Provision*; (4) there is significant different between efficiency and in-efficiency group for both Performance and LLP stage (5) By using neural network, we suggest 50% of banks decrease and 50% increase the amount of loan loss provision to be more efficiency.

Chapter 6 remarks the main finding of this thesis and discusses directions for future research.

PRÉFACE

La thèse se compose de six chapitres. Chaque chapitre peut être lu indépendamment des autres, mais les six chapitres partagent le thème général de la thèse : Utiliser des techniques d'apprentissage automatique pour expliquer et prédire la faillite des banques.

Le chapitre 1 présente brièvement les motivations et les contributions de cette thèse.

Le chapitre 2 présente la revue de la littérature scientifique sur : (1) les principaux résultats en matière de prévision des défaillances des banques, (2) La comparaison des techniques statistiques et d'apprentissage automatique

Le chapitre 3 compare la précision de deux approches : les techniques statistiques traditionnelles et les techniques d'apprentissage automatique, qui tentent de prédire la défaillance des banques. Un échantillon de 3000 banques américaines (1438 défaillances et 1562 banques actives) est étudié par deux approches statistiques traditionnelles (analyse discriminante et régression logistique) et trois approches d'apprentissage automatique (réseau de neurones artificiels, machines à vecteurs de support et k plus proches voisins). Pour chaque banque, les données ont été collectées sur une période de 5 ans avant de devenir inactives. 31 ratios financiers extraits des rapports financiers des banques couvraient 5 aspects principaux : qualité des crédits, qualité du capital, efficacité des opérations, rentabilité et liquidité. Les résultats empiriques révèlent que les méthodes du réseau neuronal artificiel et des k plus proches voisins sont les plus précises.

Le chapitre 4 examine le bilan des *pertes matérielles (Material Loss Review)* publié par la Federal Deposit Insurance Corporation (FDIC) sur les banques américaines en faillite de 2008 à

2015. Ces rapports visent à expliquer les causes des défaillances et des pertes matérielles de chaque banque. Contrairement aux méthodes traditionnelles qui fournissent des suggestions sur les ratios financiers, cette étude se concentre sur les expressions extraites des rapports en utilisant la technique de l'exploration de texte. Les étapes de prétraitement sont utilisées dans cette étude pour « nettoyer » le texte. La technique du sac de mots est utilisée pour recueillir les mots les plus fréquents. La modélisation de sujets et de thèmes et la classification hiérarchiques de documents sont utilisées pour classer ces rapports en groupes. Nos résultats suggèrent que pour éviter la défaillance, les banques devraient prendre conscience de l'importance : de la gestion des prêts et du conseil, de la qualité du processus de surveillance. Ils montrent les dangers de la concentration des crédits, notamment les crédits ADC (Acquisition, Développement et Construction) et CRE (Immobilier commercial). En outre, nous constatons que les principales raisons pour lesquelles les banques américaines ont échoué de 2008 à 2015 sont couvertes par deux thèmes principaux : le risque de crédit et les problèmes de gestion.

Le chapitre 5 étudie l'efficacité de la provision pour pertes sur prêts des grandes banques américaines. La provision pour pertes sur prêts (Loan Loss Provision - LLP) est un élément important du rapport financier de la banque qui peut être utilisé pour (i) évaluer le niveau de risque de crédit ou (ii) l'état des résultats de la banque. Dans cette recherche, LLP est dédiée à la protection des défauts de crédit. Théoriquement, le niveau des provisions devrait être maintenu au montant approprié en fonction de la perte de crédit potentielle. Dans cette étude, la méthode DEA en 2 étapes (performance et stade LLP) est utilisée pour évaluer si le programme LLP est effectivement réservé. Ensuite, le réseau Neural propose l'ajustement, si nécessaire, du niveau des provisions de chaque banque afin d'améliorer la qualité de la gestion des LLP. L'ensemble de données comprend 166 grandes banques américaines sur la période 2016. Les résultats suggèrent que (1) seulement 12,7% des banques opèrent efficacement pour le processus de performance ; (2) Encore moins (seulement 2,4%) de banques fixent les LLP à un niveau approprié.

(3) L'efficacité des banques augmente avec le (a) Nombre d'employés, (b) Total des fonds propres, (c) Total des dépenses, (d) Total des dépôts, (e) Total des prêts, (f) Investissement total et supérieur (Provision pour pertes sur prêts); (4) il y a des différences significatives entre le groupe efficace et le groupe inefficace, à la fois pour Performance et LLP (5). En utilisant le réseau neuronal, nous suggérons que 50% des banques diminuent et que 50% augmentent le montant des provisions pour pertes.

Le chapitre 6 commente le résultat principal de cette thèse et discute des orientations pour les recherches futures.

CHAPTER 1

CHAPTER 1: INTRODUCTION

1. BANKS AND BANKING SYSTEM

Mark Twain told a joke: "*A banker is a fellow who lends you his umbrella when the sun is shining but wants it back the minute it begins to rain*". Although many of bank's customers may get the impression that this joke is more truth than fiction, the real story is that banks today provide variety of different services to people, investors, government all over the world. Worldwide, banks contribute significantly to promote the development of consumption, construction and economy.

Banks is at the heart of financial and economic systems. Without banking infrastructure in place of reallocate capital, insurance, asset management, financial services would not be functional. In economic system, fundamentally, bank performs the following roles:

(1) **Maturity transformation**: the main task of banks is to take deposits from depositors, including both individual and organization, guarantee to return these on demand. On the other hand, banks use these deposits to make loans for longer durations. In doing so, banks have the potential to transform the different maturity from short term savings into long term investments and thus improve the productivity of the economy. This transformation is significantly important as the savers usually deposit money in short maturity, however, borrower (especially project borrowers) always want to have long maturity of loans. Savers want to be able to access their money at any time and investors want to get funds, which are committed for a long term, so they can make investment decisions, which pay off over the long term. By placing itself between the savers and investors, the bank enables productivity enhancing investments to take place.

(2) **Credit creation**: Basically, after receiving deposit from customers, banks need to split into two parts: First part stays in the bank as reserves in case the savers want to withdraw some of their money back; the second part is lent on to investors. Depending on banks' charter and centre bank requirements, however, the second

part is always much higher than the first one. The investors, after receiving the loans, in turn paying certain interest rate. This mean, certain amount of saving can be converted into much higher amount of credits, which are crucial for economy and capital market. This is a significant advantage of banks' role in economics since the individual banks cannot 'create credit' the banking system as a whole does exactly that

(3) **Credit allocation:** Demand for credit is usually higher than the amounts that the 'magic' of fractional reserve banking can create so banks typically must ration credit. Moreover, it is important to make sure that the borrowers can fully pay the original loan plus the interest amount. Hence, he decision process needs to be made careful in answering following questions.

- What is the purpose of the credit application?
- What are the risks of the financed project?
- What is the possible return?
- What is the likelihood that the client will be able to return the money in full?
- The probability of being liquid of collateral as security for repayment of a loan?

The quality of the due diligence that goes into this process is central to the dynamism of the overall economy.

(4) **Financial services provision** such as payment and clearing infrastructure: As the extremely high development of informatics and financial products, banks provide variety of financial services that support both customers and investors. These services play a significant role in social life and economics.

Nowadays, times are changing, and today's technology world is having widespread effects on both consumer behaviours as well as the services that banks provide. As remarks by Allen (2008), aside from positive aspect of the roles banks play such as delegated monitoring or sharing risk in the economy, banks are often at the centre of financial crisis that can cause fragility in the financial system.

2. WHY ARE BANKS FRAGILE?

Banks, through financial crises history, are inherently fragile institutions. Being special sensitive with the movement of market, banks may face shocks both on the asset and liability sides. Banks are more fragile than the other types of firms because of two main aspects: (1) Banks face more kinds of risks and risk affects both sides of balance sheet, and (2) Banks involve significantly to the economy and create the potential systemic consequences

Due to the special characteristics, banks take risks all the time. Aside from major types of risk that bank faces such as: credit risk, market risk, operational risk, liquidity risk, business risk, there are some out-of-reach risks, for instance: systemic risk and moral hazard. The history witnesses that each of these risks may lead bank to be fail. Kaufman (1997) suggests that the failure of an individual bank introduces the possibility of system wide failure (systemic risk) since they may spread in domino fashion throughout the banking system. In general, bank operates based on the simple rules: Taking deposits as a major funding for loans. Deposit is normally short-term, but loan is long-term. On the other hand, banks transfer short-term liabilities into long-term assets. It is the fact that, both deposits and loans are significantly sensitive to the movement of market interest rate, exchange rate, local economic and micro-economic environment.

Moreover, a shock or a bad rumour affects only one or a few institutions initially, can become systematic and affect negatively to the larger local economy. With the globalization trend in banking sector, shocks infect not only certain banks or some certain countries, but also the financial system and the whole economy in other countries. The literature proves that shocks hitting banks in one country effects on other countries. For example, Peek (2000) indicates that shocks in banks of Japan effects on the real economy in the US. Puri (2011) also suggests that the transmission of the US financial crises to the behaviour of linked German savings banks.

The history witnesses several large banks experienced sudden, massive withdrawals when any bank announced failure, which is called '**bank runs'**. Bank runs happens when depositors withdraw their deposit from banks because they are afraid of the safety of their deposit. History reported that during the Great Depression, and recent financial crisis, even with deposit insurance, there are still depositors who are most likely to run. When a bank goes failure, there will be an announcement on this failure and it could cause the uninsured depositors to run that badly affect banks system. This happens with every country all over the world. For example, in the UK, 2007, depositors lost confidence in Northern Rock banks and started a run of withdrawals that ended with the bank being taken into state ownership. This is considered as the very first sign of in Britain of the coming global financial crisis. The same phenomenon happened in the US in the year 2008, the investment bank Bear Stearns and the commercial bank Wachovia also experienced a rapid unexpected loss of funding and finally were taken over by other institutions to avoid their outright failure. These examples imply that banks, as the financial intermediaries, are inherently fragile.

In fact, bank failures are widely perceived to have greater adverse effects on the economy and are considered more important than the failure of other type of business firms. The failure of bank is also viewed to be more damaging than other failure because of a fear of the spill-over effects. The failure of an individual bank introduces the possibility of systemwide failures or systemic risk. Because the sustainability of banks depends significantly on the confidence of depositors. Unlike other fields of an economy where a failure of a competitor is usually good for business, in banking sector, a failure of one bank can cause a damage of confidence in other banks. This can happen for some reasons such as (1) Banks are involved in financial markets, the failure of one bank may drive down the markets, (2) Depositors, who are the main funding of source, might be afraid of systemic collapse, (3) The interbank relationship is large and significant, the failure of one bank may inflict large losses on others.

Moreover, if banks are just a "normal" industry with ups and downs not affecting the broader health of the entire economy, a weak banking system would hardly be a matter

of concern. However, as a sector that provides finance for entire industry, banks failure is highly relevant for all other parts of society, companies, government and investors.

In fact, in emerging countries such as Vietnam, even a bank can be failed due to insolvency, it is always acquired by centre banks. The concept 'bankruptcy' is too sensitive to this kind of market and the government is afraid of bad reputation.

3. U.S. BANK CONTEXT

Since the recent financial crisis, there has been considerable discussion about the importance of the US banking sector to the global financial system. The collapse of some banks in U.S (Lehman Brothers for example), is believed to significantly inflect to other banks around the world. Bostandzic (2013) documented that U.S banks contribute significantly more to global system risk than European banks. Dufrénot (2014) also confirmed of the spill-over effects of 2008 financial crisis in the U.S on the volatility of the Indian equity markets. Kim (2015) also found the significant correlation of the transmission of the U.S. crisis to financial markets in five emerging Asian economies: Indonesia, Korea, Philippines, Thailand, and Taiwan. Hence, this thesis ***focusses on U.S banks as the literature suggested that U.S. banks are more sensitive and contribute more to systemic risk in the global financial system than other banks*** (Bostandzic, 2018).

In fact, U.S. banking system, in some perspectives, it is similar to the banking systems in other industrialized countries, however, in other ways, it differs from them. These differences change over time, especially in recent decades:

- One feature that characterized the U.S. banking industry is banks are chartered, supervised, and regulated at both state and federal level.
- U.S. banking industry has very large number of very small banks.
- U.S. banks had more limited authority to provide securities, insurance, and real estate-related financial services.

- Banks were allowed only limited investments in industrial companies, and industrial companies were permitted only limited ownership interest in banks.

To capture the global scenario of U.S bank, it is important to introduce briefly the early history of U.S. banks and the involving of regulation:

- In 1781, Continental Congress granted the first official bank charter to the Bank of North America to provide financial support for the war of independence.
- In 1791, after the Constitution was approved, Congress moved to establish the First Bank of the USA. It acted as a central bank to promote a sound money and credit system and as the main depository for the country's gold and silver, made loans to state banks to assist with any *liquidity problems*.
- Until 1816, commercial banks did not provide a full range of services, avoided long-term securities and mortgages. To fill this gap, the first mutual savings banks – The Philadelphia Saving Fund Society was established.
- After Civil War, the number of banks had increased rapidly. However, the USA did not have a real central bank to act in the scenario of financial crisis. The history reported that in the late 19th and early of 20th century, there were numerous downturns in economic activity. In 1913, Federal Reserve System (Fed) was created as the central bank. This central bank was organized on decentralized basis with twelve regional banks and the headquarter was established in Washington DC.
- After the Great Depression of 1930, in 1933, Federal Deposit Insurance Corporation (FDIC) was established. The purpose of this corporation is to provide deposit insurance services to *protect depositors from bank failure*.
- Since then, it is noteworthy to introduce the Glass-Stegall Act or The National Banking Act of 1933. The purpose is to separate commercial banking and investment banking*, which is believed as the reason of bank failures*. Restrictions on the mixing of commercial and investment activities were mainly intended to minimize the conflicts of interest. During the same period, an interest rate ceiling on deposit was imposed to protect banks from excessive competition.

- The Federal Deposit Insurance Act of 1950 authorized the FDIC to examine national and state-member banks to determine their *insurance risk*.

-The Depository Institution Deregulation and Monetary Control Act of 1980, which established *loan-loss-reserve requirements* to create a cushion of *nonperforming loans*.

- Garn-St Germain Depository Institutions Act of 1982 enhances the powers of FDIC and Federal Savings and Loan Insurance Corporation to *provide aid to troubled institutions*.

- 1988, the central bank governors of the Group of Ten countries adopt the Basel Capital Accord, known as *Basel I,* which provides procedures for factoring on and off- balance sheet risks into the supervisory assessment of capital adequacy.

- 1991, as the consequence of hundreds of FDIC-insured banks fail, requires that *prompt corrective action* be taken against banks based on their capital levels and gives the FDIC authority to close depository institutions when capital levels fall below 2 percent.

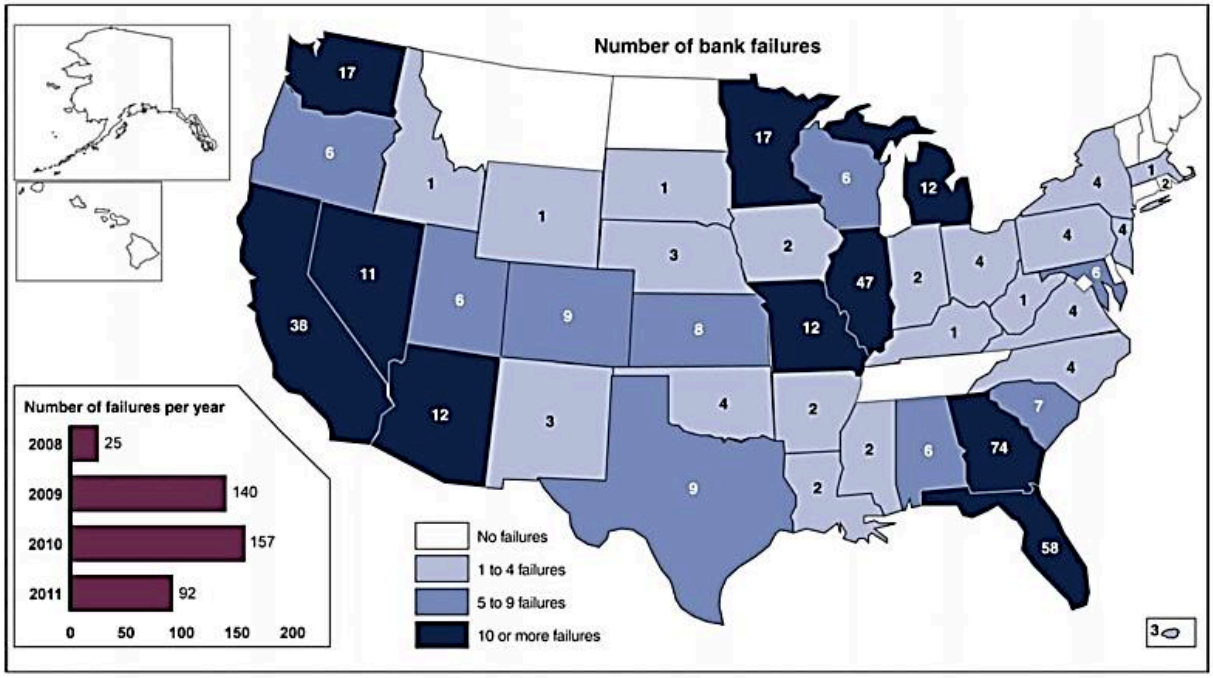
- 2003, Fair and Accurate Credit Transactions Act is released to improve the accuracy and transparency of the national credit reporting system.

- Recently, more Acts are released to prevent of being bank failure: The Dodd-Frank Wall Street Reform Act (2010) requires banks to increase their capital cushion and authorize Federal Reserve to split up large banks to avoid of "too big to fail'. In 2013, the Federal Reserve also requires big banks to add more liquid assets.

The purpose of these Acts is to protect and develop banking system as well as prevent bank of failure risk. However, it has been shown that most of the important regulations were put in place in response to various crises over time rather to prevent from occurring or mitigating their effects should they occur (Barth et al. 2009).

4. GLOBAL PICTURE OF BANK FAILURE

Figure 1: Number of U.S bank failures from 2008 to 2011



Source: GAO analysis of FDIC data: Map Resources (map).

Along the history of economics crisis, banking system accounted for its remarkable component: Panic of US in 1907 with bank failure; banking crisis of 1973-1975 in the UK, Savings and loan crisis failure from 1986 to 1995 in the US, Swedish banking crisis in 1990s, US housing bubble in US 2007-2010, or Venezuelan banking crisis of 2009–10, etc. As a crucial role in the global economics for accounting trillions of dollars in assets worldwide, banking system is particularly sensitive and important to every counterparty. The failure of bank is generally considered to be of more damaged than the failure of other types of business firms.

Historically, banks have been recognized for the variety of financial services that they provide and have tendency of expanding rapidly. However, banks are highly regulated because of several reasons; one of these reasons is a common belief that banks failures compose a significant macroeconomic cost.

Ohlson (1980) posed an interesting question: *Why forecast bankruptcy?* He agreed that this is an embarrassing question and difficult to answer. However, this question can be argued that this is *obvious* practical interest. The literature suggests that bank failures were significantly important during the financial crisis of 1930. Ashcraft (2003) wondered whether bank failure still important in the modern economics and proposed 3 aspects: (1) The deposit insurance has significantly reduced the negative effect of failure, (2) In US, establishing FDIC as receiver has minimized the illiquidity of failed banks, and (3) the US economy has likely become less bank dependent since the 1930s. The study, once again, confirm the significant and apparently permanent effects of health-bank failures on real economic activity.

Figure 1 introduces the number of bank failure in the US from 2008 to 2011. After the significant effect of financial crisis, the number of bankruptcies increases sharply. Most of states in the US experienced more than 4 bank failure events.

FDIC defined a bank failure is the closing of a bank by a federal or state banking regulatory agency. In general, when a bank is unable to meet its obligations to depositors and other, it is announced as failure. This could occur because the bank has become insolvent, or because its assets cannot be liquid to meet the payment obligations.

The literature of the bankruptcy prediction indicates the important and great concern in both academic and business community. The bankruptcy prediction can provide significant advices on making decisions and profitability to financial institutions.

5. WHY DOES BANK FAIL?

Apart from external reasons come from the adverse economic conditions, there are numerous internal reasons that lead to the failure of banks.

After the collapse of financial system in 2008, there are several papers that examines the

determinant of US bank failure before and after financial crisis. Most of the literature studies suggested that failed banks are characterized by ***significantly higher loan growth rates***, ***less capitalized*** in compared with 'healthier banks'. Some banks failed because they ***were unable to raise capital or obtain liquidity*** as the value of their portfolios declined. Another reason is believed that banks hold ***'toxic' mortgage or mortgage-related assets***, which are difficult to liquid and extremely risky, unable to value.

The report of Comptroller of the Currency also stated that management-driven weakness played a significant role in the decline of 90% of the failed and problem banks. The main difficulties that banks experienced resulted from inadequate loan policies, problem loan identification systems, and systems to ensure compliance with internal policies and banking law. The report also stated that insider abuse and fraud are significant factors that cause of more than one-third of the failed and problem banks. Unfortunately, these ***crime risk*** involved directors, managers or principal shareholders.

> *Since the global financial crisis and recession of 2007-2009, criticism of the economics profession has intensified. The failure of all but a few professional economists to forecast the episode - the aftereffects of which still linger - has led many to question whether the economics profession contributes anything significant to society.*
>
> *Robert J. Shiller- American Nobel Laureate*

GAO[1] reports on causes and consequences of recent community bank failure, that bank failure was largely correlated to ***nonperforming real estate loans*** and highlight impact of ***impairment accounting*** and ***loan loss provisioning***. Between 2008 and 2011, the failed of small and medium-size banks were largely associated with inadequate risk management and high concentration of Commercial Real Estate Loans and Acquisition,

[1] The U.S. Government Accountability Office (GAO) is an independent, nonpartisan agency that works for Congress

Development and Construction loan. Meanwhile, the failure of large banks is associated with loss from subprime and non-traditional residential mortgage loans.

In brief, the question of 'Why does bank fail", is still under debate. This thesis contributes to the failed banks' literature by answering 4 questions:

1. ***From numeric perspective (financial ratios), why does bank fail?***
2. ***What is the proper method for bank failure prediction?***
3. ***From non-numeric perspective (textual reports), what are terms that explain bank failure?***
4. ***Loan is blamed as the direct cause of bank's failure, what bank' managers should do to improve the loan management quality?***

6. CONTRIBUTION OF THIS THESIS

An early warning is necessary for stakeholders to foresee, minimize any potential damage and suggest significant adjustments. The remarkable summary of this thesis is:

Methodology: (1) Apply *multiple approaches and algorithms*: (a) Statistical techniques: Discriminant analysis, logistic regression, (b) Machine learning techniques: Support Vector Machine, Neural network, K-nearest neighbour, Data Envelopment Analysis, Bag-of-Words, Topic Modelling and Documents clustering. (2) Inspect both *numeric and non-numeric indicators* via data mining and text mining respectively.

Aims: (1) Predict the bankruptcy of banks 5 years before the failure event via financial ratios, (2) Analysis textual reports and explain the reasons that cause banks failure, (3) Suggest the adjustment on the Loan Loss Provision term.

6.1. INDUSTRIAL CONTRIBUTIONS

There is a debate on the determinant regarding causes and consequences of bank failures from the regulation and management as well as the academic literature perspective. U.S. regulators releases multifarious Acts to prevent from financial market disruption. Basel Committee on Banking Supervision (BCBS) also formulates broad supervisory standards,

guidelines and recommends statements of best practice in banking supervision in the expectation that members will take steps to implement them through their own national system.

However, these Acts or Basel standards, on the other hand, are criticized as being released *'too late'* after crisis happened. There is the fact that, even under the strict control of centre bank, the event of bankruptcy still occurred because there are many ways that banks can 'manipulate' and find their own way to blur the regulations. Hence, ***updating new approach on predicting bank's failure is always essential.***

This thesis assists the board of directors on the following aspects: (1) Foresee the probability of being failed, (2) Understand the internal problems that banks should pay significant attention to assure of bank's operation, (3) Inspect the sensitive aspects that cause banks go failure from both financial and non-financial indicators (4) Measure and adjust the Loan Loss Provision.

From regulation perspective, it is important to (1) Update and select appropriate methods for measuring the level of reserve against risk, (2) Give the up-to-date instruction on management strategy, (3) Look at both financial and non-financial indicators for auditing and supervising purposes.

6.2. ACADEMIC CONTRIBUTIONS:

The thesis is highlighted for academic contribution of the following reasons:

- The thesis focuses on Machine learning approach and compares the efficiency of prediction accuracy between Statistical and Machine learning approach. The result suggested that Machine learning obtained higher accuracy rate. From the first article, it is noteworthy that Loan is one of the biggest problems in detecting bank failure.
- The second article analyse the Material loss reviews of 98 US banks, issued by FDIC. For this article, the financial numeric indicators are being discarded, only non-numeric (textual) information that is extracted from these reports using Text mining techniques is used. The results suggest that Loan and management are two key terms that causes bank failure.

- We hence, for the third article, measure loan quality by measuring Loan Loss Provision, which is an indicator for credit risk. This article uses Data Envelopment Analysis (DEA) to measure the efficiency of LLP and uses neural network to adjust LLP if needed.

In short, this thesis provides a global view on both numeric and non-numeric aspects. Moreover, using variety of Machine learning algorithms proves their significant advanced compare to other traditional approaches. The integration of these 3 chapters would: (1) Contribute significantly to the awareness of bank survival, (2) Predict bank failure 5 years before the bankruptcy event, (3) Provide non-numeric analysis that explain bank failure (4) Suggest the adjustment for Loan Loss Provision.

The thesis is composed by 3 articles independently which investigate on 3 aspects :

Article 1: Predicting bank failure: An improvement by implementing a machine learning approach to classical financial ratios
Article 2: Why do banks fail? – The answer from text analytics technique.
Article 3: A two-stage DEA and neural networks on measuring and estimating loan loss provision of large US banks

7. REFERENCES

1. Allen, Franklin, and Elena Carletti. "The roles of banks in financial systems." J. Wilson (2008).
2. Ashcraft, Adam B. "Are banks really special? New evidence from the FDIC-induced failure of healthy banks." American Economic Review 95, no. 5 (2005): 1712-1730.
3. Barth, James R., Tong Li, and Wenling Lu. "Bank regulation in the United States." CESifo Economic Studies 56, no. 1 (2009): 112-140.
4. Bostandzic, Denefa, and Gregor NF Weiß. "Why do some banks contribute more to global systemic risk?" Journal of Financial Intermediation (2018): 17-40
5. Dufrénot, Gilles, and Benjamin Keddad. "Spillover effects of the 2008 global financial crisis on the volatility of the Indian equity markets: Coupling or uncoupling? A study on sector-based data." International Review of Financial Analysis 33 (2014): 17-32.
6. Evans, Lawrance L. "Causes and Consequences of Recent Community Bank Failures." Testimony before the Committee on Banking, Housing, and Urban Affairs (2013).
7. Kim, Bong-Han, Hyeongwoo Kim, and Bong-Soo Lee. "Spillover effects of the US financial crisis on financial markets in emerging Asian countries." International Review of Economics & Finance 39 (2015): 192-210.
8. Kaufman, George G. "Bank failures, systemic risk, and bank regulation." Cato J. 16 (1996): 17.
9. Ohlson, James A. "Financial ratios and the probabilistic prediction of bankruptcy." Journal of accounting research (1980): 109-131.
10. Peek, Joe, and Eric S. Rosengren. "Collateral damage: Effects of the Japanese bank crisis on real activity in the United States." American Economic Review 90, no. 1 (2000): 30-45.
11. Puri, Manju, Jörg Rocholl, and Sascha Steffen. "Global retail lending in the aftermath of the US financial crisis: Distinguishing between supply and demand effects." Journal of Financial Economics 100, no. 3 (2011): 556-578.

12. Wess, Gregor NF, Denefa Bostandzic, and Felix Irresberger. "Catastrophe bonds and systemic risk." 26th Australian Finance and Banking Conference (2013).

CHAPTER 2

CHAPTER 2: LITERATURE REVIEWS

In banking sector, it looks like every topic is important, vulnerable and relates to bank's profitability or viability. Each activity may affect directly or indirectly to other activities. To improve the quality of management, avoid from being inactive, managers need to aware and consider all potential risks that banks may face. Aside from systematic risks that cannot be avoided, there are unsystematic risks that every bank needs to be aware of such as ***Credit risk (Default risk)*** - when borrowing customers fail to make some or all their promised payment, **Liquidity risk -** when a bank in the danger of running out of cash when cash is needed, ***Interest rate risk***- when the market interest rate risk will result in significant appreciation or depreciation in the market value of bank's assets, or ***Crime risk*** -when employees or directors be fraud or embezzlement. In fact, FDIC lists crime risk from employees or directors as one of the causes of recent bank closing.

However, these given risks cover only some certain aspects in banking. To avoid of being failure, banks' managers must capture the global picture and make careful decisions on each movement.

The history witnesses that when financial crisis occurred, not all banks ruined. It is convinced that, certain characteristics distinguish failed and non-failed banks. The literature introduces a vast number of research investigate and make great effort on figuring out these characteristics. The common way is to collect certain numbers of bankrupt and non-bankrupt banks, then select proper method to examine results. The review is conducted in two broad categories: (1) statistical and (2) intelligent techniques. Basically, the classical study on bank failure prediction follows the schedule:

- First, collect certain number of failed and non-failed banks (active and inactive). The data can be annually, semi-annually or quarterly in period of 1 to 5 years before the bankruptcy event. Dependent variable is dummy (active or inactive), the regressors can be either stochastic or dummy.

- Then, use appropriate methods to investigate the dataset (Traditional statistical or Intelligent Techniques)

- Finally, provide results based on: (1) compare the accuracy of different methods, (2) the accuracy accordingly to the years before bankruptcy event, (3) promote the most accuracy method.

Over the history of bank failure investigation, using financial ratios is the most common way. Moreover, case study analysis and textual reports analysis, which are also, used as additional tools.

1. FINANCIAL RATIOS – THE STORYLINE OF BANK FAILURE

Seminal works by Beaver (1966) and Altman (1968) introduced models using on-balance-sheet information to predict firm's failure. Altman, after that, continues to develop his model so-called ZETA ® (1977) credit risk model. These models are still being used by practitioners throughout the world. Literature introduced stream of research in the bankruptcy of corporate financial institutions. The researched field is then ***targeted on the area of banking sector***. Bank regulators are special interested in developing early warning system to supplement information obtained from on-site examinations and, in turn, help predict impending bank failures (Kolari, 2002).

The prediction of failure for financial firms, especially banks has been extensively researched area since late 1960s. The large number of bank failures in the US during and after financial crises has heightened the interest of researchers in explaining the reasons for these failures. Creditors, auditors, stockholders and senior managers are all interested in bankruptcy prediction because it affects all of them alike. Many reasons have been indicated for bank failures, including expanded risk-taking, interest rate volatility, insufficient management practices, inadequate accounting standards, increased competition from other financial institutions, and pervasive internal control weaknesses (Fraser, 1995).

Altman (1968), as one of the pioneers in bankruptcy prediction, introduced the discriminant function by using 5 independent variables: Working capital/Total assets, Retained Earnings/ Total assets, Earning before interest and taxes/Total assets, Market value equity/Book value of total debt, Sales/Total assets. These financial ratios represent for five aspects: Capital capacity, earning power, Liquidity position and revenue generated. These ratios have become popular and been used commonly.

In fact, these ratios are initially used for corporate firms, not targeted for bank firms. To adjust to financial institutions, aside from classical ratios, additional variables are used in most of research papers such as Current ratio, ROE or ROA, and other specifically used for banking sector such as: *non-performing loans/total assets, Net interest income/ Total Assets, Net non-interest income/Total assets, Total loans/ Total equity, Total deposit/Total Assets, etc.* This shows, from the domain point of view, research on bank failure has matured considerably as researchers started attaching more importance to other financial ratios.

Besides, some typical ratings are standardized and used worldwide. CAMELS rating system is the most popular system originally developed in the US to classify a bank's overall condition. This rating is composed of six bank's condition that are assessed through: (C) *capital adequacy, (A) asset quality, (M) management expertise, (E) earnings strength, (L) liquidity and (S) sensitivity to market risks.* In fact, CAMELS is favourable rating system. Many of the previous bank failure studies used ratios corresponding to these system criteria and found significant effect on bank failure in several markets. CAMELS is also considered as a standard of bank suspension decision by many centre bank and regulators worldwide (Persons (1999); Tam (1992); Barr (1994); Hooks (1995); Gonzalez-Hermosillo (1997)). However, Cole (2012) found that CAMELS proxies become successively less important, whereas portfolio variables become increasingly important.

Variety of ratios is used along the history of bank failure issues and convergence in the main following aspects: ***Loan quality, Liquidity position, sensitive to the market movement, the quality of management and capital adequacy***. The purpose of these

ratios is to focus on how bank manages their risks and their cushion capacity of facing these risks.

1.1. LOAN QUALITY – THE QUESTION OF ALL TIME

The foremost term i**s** *Loan quality*, which can be measured in different approaches: the diversified of loans, the concentration of typical loans, the provision for loan loss, non-performing loans, etc. As the major income, loan quality is recognized in all studies about bank failure. Loan quality is negatively correlated to bank failure probability. Literally, if credit risk does not occur, the higher of total loans, the more net income that bank can generate.

However, even with the effort of regulators, bad loans (nonperforming loans) still occur. Hence, variables to proxy for loan quality are convergent into two main aspects: Total loans to total assets and nonperforming loans to total loans. These two variables are used widely in most of the research on bank failure (Wheelock (1994); Hooks (1995); Hwang (1997); Gonzalez-Hermosillo (1997)). The empirical results indicate that large portfolio of real estate loans had a relatively large proportion of nonperforming loans and had less income diversity (Cole (2012); Lu (2013))

Recently the debate on provisioning for bad loans (or loan loss provision) has devoted specific attention. Loan loss provision is the first cushion of nonperforming loans, which helps bank smooth their cash flows and protect their liquidity position as well. Hence, some other variables related to provision is concerned, such as Allowance for loan loss to total loans, Amount taken against allowances to total loans (Timing of banks' loan loss provisioning during the crisis).

After the financial crisis in 2008, it is stated that the inappropriate allocation of loans causes bank failure. Most of failed banks face the similar trouble because of real estate loans concentration. Iturriaga (2015), Lu (2013) introduces additional variables regard to this kind of loans, such as: Real estate loans 90+ days past due, real estate loans to total

loans. Their results are homogeneous, and both confirm the significant role of real estate loans on bank's survival.

1.2. LIQUIDITY POSITION – THE TRADE-OFF

There is a trade-off relationship of bank liquidity and bank profitability, in which, the more resources are tied up in readiness to meet demand for liquidity, the mower is that bank's expected profitability. Illiquid assets and long-term investment increases profitability but exposes banks to illiquidity after intrinsic and extrinsic shocks (Cox, 2014). Market liquidity and funding liquidity are crucial in explaining bank failures.

In fact, illiquidity position has correlation with bank's assets. Cox (2014) indicated that the proportion of illiquid loans is the main predictors of bank failures. There are some ways to measure the liquidity position by taking: Total Government Securities to total investment (Dash, 2009), Liquid assets to total asset (Boyacioglu, 2009), Liquidity coverage ratio (Liquid assets to total deposit and short-term debt ratio). (Zhu, 2016)

1.3. SENSITIVE TO THE MARKET MOVEMENT – INTEREST RATE HEDGING

Several decades ago, bank regulatory agencies around the world, to include the United States, changed the singular CAMEL examination framework (or comparable structure) to the plural CAMELS (Handorf, 2016). The S represents the (S)ensitivity, to ensure examiners considered bank earning and market value of equity might be affected by a change of market movement, largely focus on market interest rate risk.

This risk can be measured by: Trading securities to total assets, Foreign exchange to Foreign exchange liabilities, net interest income to average asset or net on balance sheet position to total shareholder equity (Boyacioglu (2009); Tam (1991), Federal funds purchased-fed funds sold to total asset (Wheelock, 1995)

1.4. QUALITY OF MANAGEMENT – QUESTION OF MEASUREMENT

Apart from excessive risk-taking, or simply bad luck, banks that are poorly managed are thought to be prone to failure (Berger, 1992). However, it is challenging to measure the quality of management because it requires qualitative issues and it can take several forms (Wheelock, 1995).

Quality of management is hard to measure directly, but via the ratio of operating expense to total assets (Tam, 1992), Cost inefficiency (Wheelock, 1995), Operating expense to total expense (Canbas, 2005).

1.5. CAPITAL ADEQUACY – BASEL AND REGULATION

Banking is special sector with the lowest capital to total asset ratio. As the data announced by the World Bank, in US, on average, the bank capital to total assets is 11.6%. This figure is just a small amount in compare to manufacture firms, however, play significant role in bank failure. Capital provides the financial cushion for economic losses, protecting depositors, other creditors and official institutions, which are often forced to absorb bank losses in the interest of maintaining banking system stability (Cantor, 2001)

Aware of the crucial role of capital adequacy, the advent of Basel I in the late 1980s significantly increased the incentive for banks to find a way to reduce loans held on the balance sheet, in order to reduce the impact of capital requirement (Barth, 2009). Since the first release, until now, 4 updated versions have launched to fulfil, and record more types of risks.

Capital adequacy is measured based on total risk-based capital ratio, the higher capital implies the larger cushion against losses. Capital adequacy can be viewed by risk-weighted, leverage and gross revenue ratios. Estrella (2002) found that all three ratios are strongly informative about subsequent bank failure.

It means, banks have more equity implies that expected default costs are lower (Hellmann, 2000). Cole (2011) concludes that banks with more capital, better asset quality, higher earnings and more liquidity are less likely to fail.

To proxy capital adequacy, the regressors can be: risk-weighted capital ratio (Bouvatier, 2013), total equity (Cole, 2011), (Bell, 1997), Tier 1 Risk-based capital ratio (Bennett, 2015).

2. METHODOLOGICAL ASPECTS

To predict the failure of banks, the review is conducted in two broad categories: (1) statistical and (2) intelligent techniques. The statistical techniques covered: LDA- linear discriminant analysis, MDA- multivariate discriminant analysis, QDA- quadratic discriminant analysis, LR- Logistic regression and FA- Factor analysis. The intelligent techniques covered: NN- Neural networks, MLP- Multi-layer perception, PNN- Probabilistic neural networks, DEA- Data envelopment analysis, SVM- Support Vector Machine, k-NN- K-nearest neighbour. There are numerous papers made a comparison on the prediction accuracy among these techniques.

2.1. STATISTICAL TECHNIQUES

Collins (1982) indicated 3 most often used statistical techniques for bankruptcy prediction are: Multiple discriminant analysis, the linear probability model and logistic regression.

Multiple Discriminant analysis - MDA

Even not as popular as regression analysis, MDA has been utilized in a variety of disciplines. This is a statistical technique used to classify an observation into one or several given groups depend upon the observation's individual characteristics. Altman (1968) compliment MDA method for its advantage of considering an entire profile of characteristics for all of observations as well as its capacity in derive a linear combination of the characteristics to create the « best » discriminates between the bankruptcy and non-bankruptcy.

This is a statistical technique that reduce the differences between variables in order to classify them into a set number of broad groups, on the other hand, is to classify a data set by providing the most meaningful separation.

As one of the pioneers in bankruptcy prediction, Altman (1968) introduces the discriminant function by using 5 independent variables: Working capital/Total assets, Retained Earnings/ Total assets, Earning before interest and taxes/Total assets, Market value equity/Book value of total debt, Sales/Total assets. Altman (1973) introduces a bankruptcy classification model by investigating on 53 bankrupt and 53 non-bankrupt firms. 5 main topics of firms are considered via variables of: Profit, leverage, liquidity, capital, earning and others.

Fisher (1936) finds the linear combination of the variables that maximizes the difference between groups and minimizes the difference within groups through eigenvalues and eigenvectors analysis. Further refinement of the MDA techniques is provided by Cornfields (1967), Altman (1968), Snapinn (1985), Lam (2002)). The discriminant analysis method was utilized by Cox et al. (1988) to find the characteristics of acquisitions of the failed US banks by other banks as well as Canbas (2005) predicting the financial structure of Turkish bank failures. Canbas (2005) experiment 3 statistical models: Discriminant, Logit and Probit on the set of 40 private commercial banks in Turkish, using 49 financial ratios. Their result suggested that Discriminant model obtain the highest correct classification for each of prior year before banks go failure

Recently, Cox et al. (2014) used Discriminant analysis to trace the US bank failure. Their analysis is based on 19 financial variables including types of loan made, asset, liability and equity composition, bank size.

Linear probability model- LPM

LPM is a special case of OLS regression when the explained variable is in the binary form. The dependent variable takes value of 1 for active banks and 0 for inactive (or failed) banks. The model is of the following for:

$$y = a_0 + \sum_{m=1}^{\infty} a_m x_m + u$$

Where:

y=1 for active banks

y=0 for failed banks

a = coefficient regression

x = explanatory variable

u = residual

LPM is first introduced by Meyer (1970). Their factors explaining bank failures is divided into four groups: (1) local economic conditions, (2) general economic conditions, (3) quality of management, and (4) integrity of employees. Approximately 80% of the observations are correctly classified. This result is quite high for cross-section study.

The literature provided a special interesting on comparing LPM to other methods. The result did not remark a pre-eminence of this method. In most of study, LPM is compared with MDA.

Grammatikos (1984) experiment both MDA and LPM on predicting bankruptcy of 29 industrial firms in Greece from 1977 to 1981, using 17 variables. The result indicated that both models are very successful in predicting financial crisis for three years before the failure event. Overall, for the first, second and third year before bankruptcy announcement, MDA correctly classified the bankruptcy at the rate of 91%, 78% and 70% respectively. The corresponding rate for LPM is 91.4%, 76% and 78%.

Stone et al. (1991) compare LPM and logit regression for accounting choice studies. The result shows that logit rather than LPM may be preferable model. Moreover, LPM also may results in higher Type I error rates when it is used for prediction. Collins (1980) also made an empirical comparison of bankruptcy prediction models between multivariate statistical method and linear probability model.

Logistic regression – LR

LR is a popular method for binary classification. It is applied widely in many science field. LR predicts the probability of a binary outcome (which is: active or inactive bank in this

research). Collins (1982), appreciate LR is as good as DA.

A logistic regression provides estimates that must be in the range of 0 and 1, and be highly recommended when the explained variable do not satisfy the multivariate normality assumption. The model is as follows:

$$\log it\ (p) = a_0 + a_1 x_1\ + a_2 x_2\ + \cdots +\ a_n x_n$$

Where: a_i represent unknown parameter estimates

x_i are independent variables

Press (2012) tested two empirical studies to compare between MDA and logistic regression. The result suggested that Logistic regression outperforming classical linear discriminant analysis in both cases, but not by a large amount.

Martin (1977) used logistic regression for entire population of banks which are member of Federal Reserve System. The explanatory variable includes 25 financial ratios belong to for groups: asset risk, liquidity, capital adequacy and earnings. The result suggests that Logistic regression model predict more correctly than linear discriminant analysis.

Ohlson (1980) presents some empirical research of predicting corporate failure as evidenced by the event of bankruptcy. Ohlson also indicated the advantages of Logit regression to MDA such as (1) MDA requires certain hypothesis on the distributional properties of the predictors; (2) the output of MDA is a score and has little intuitive interpretation. In this study, he used the data from 1970-1976 on 105 bankrupt and 2058 nonbankrupt firms, using logit regression. Four conventional factors are statistically significant in affecting firm's failure: (1) company's size, (2) the measurement of financial structure, (3) the measurement of performance and (4) the measurement of current liquidity. 3 models using logit regression are used: (i) predict bankruptcy within one year, (ii) within two years, (iii) within one or two years. Zavgren (1985) used the logit technique to test the bankruptcy of American industrial firms. The result provided significant with reference to the accuracy rate in detecting bankruptcy firms up to five years. Press (1978)

experiment on breast cancer using DA and LR. Their result suggested that LR obtained superior predictive ability. Mensah (1983), Zavgren (1985) and Keasey (1987) also used logit regression for their study, and they highlighted the advantage of not requiring the independent variables to be jointly multivariate normal and do not require for prior probabilities.

However, the usage of statistical techniques in general relies on the restrictive assumption on linear separability, multivariate normality and independence of the predictive variables. Unfortunately, most of financial variable violates these assumptions. (Kumar et al. 2007). Ohlson (1980) also indicated some of the problems with this methods: (i) There are certain statistical requirement imposed on the distributional properties of the predictor (such as the variance – covariance matrices should be the same for both failed and non-failed group), (ii) The output of the application of an MDA model is a score which has little intuitive interpretation and not directly relevant, (iii) there are problems related to the "matching" procedures: Failed and non-failed firms are matched according to criteria such as size and industry, and these tend to be somewhat arbitrary.

2.2. MACHINE LEARNING TECHNIQUES FOR DATA MINING

Over the past few decades, Machine learning has become one of the main-stays of information technology that provide the vast range of applications. Mitchell (2006) explains the field of Machine Learning seeks to answer the question:

> *"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"*

To make machine learning valuable, it is necessary to split the whole dataset into training and testing set. The training sets are used to set up initial functions, which are, in turn, evaluated by the corresponding testing set. The training set is considered as 'learning process' and the testing set is considered as 'generalization process'.

There are frequently-used algorithms of Machine learning techniques for bank analysis: Neural networks, Support Vector Machines, K-nearest neighbour and Data Envelopment Analysis. The first three algorithms are favourite tools for pattern

recognition or pattern classification. The literature review introduces vast numbers of studies that compare the accuracy of each method. Besides, Data Envelopment Analysis is a useful tool for evaluate the efficiency.

Support Vector Machines, in very simple terms, corresponds to a linear method in a very high dimensional feature space that is nonlinearly related to the input space (Amendolia, 2003). *Neural networks*, as the "neural" part of their name suggests, they are brain-inspired systems which are intended to replicate the way that humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. *K-nearest-neighbor* classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data (Peterson, 2009). K-NN is described as: Given a collection of data points and a query point in an m-dimensional metric space, find the data point that is closest to the query point (Beyer, 1999). On the other hand, *Data Envelopment Analysis* is a method for measuring efficiency of DMUs using linear programming techniques to envelop observed input–output vectors as tightly as possible. Because it requires very few assumptions, DEA has also opened possibilities for use in cases which have been resistant to other approaches because of the complex (often unknown) nature of the relations between the multiple inputs and multiple outputs involved in DMUs (Cooper, 2004).

Neural networks

Neural networks is used widely in finance sector. As being one of the most popular tool of Machine learning, this technique usually received the compliment for giving better average correct classification rates than others.

Neural networks is used since 1990s for bankruptcy prediction and become one of the most popular tool for bankruptcy prediction. Empirical results show that neural network is a promising method of evaluating in terms of predictive accuracy, adaptability and robustness (Tam, 1992). Empirical results also show that neural network is a competitive

method among existing ones in assessing the likelihood of bank failures, especially in reducing type I misclassification rate.

Wilson (1994) predicts the bankruptcies of firm by comparing neural network and classical multivariate discriminant analysis. The same ratios as Altman (1968) are used. Their study suggested that neural networks perform significantly better than the other.

In compared with Logistic Regression, Salchenberger (1992) indicated that NN performed significantly more accuracy. LR achieves 83.3-85.4% accuracy, meanwhile NN achieves 91.7%. This result is homogenous with Zhang (1999). Many studies compared between NN and MDA. Most of these studies suggested that NN perform better than MDA (Coasts et al. (1993), However, there are some studies indicated that MDA was slightly better than NN. The history of literature showed the huge interest of researcher for NN method, it can be seen that NN are generally superior than other methods.

K-Nearest Neighbors – K-NNs

Compared to other machine learning technique, k-NNs is simple, easy to interpret and can achieve adequate rate of accurate. k-NNs is one the simplest non-parametric pattern classification methods. In the k-NNs algorithm a class is assigned according to the most common class amongst its k-nearest neighbours (Chen et al. (2013)). This method is developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine (Leif, 2009).

K-NNs classifier calculates the distance between points, then assigns each point to the class among its k nearest neighbours (k is an integer). The literature shows that using different value of k may produce different results. The question of "What is the standard value of k, is still an open-ended question".

Support vector machine

Among techniques of Machine learning that are used for bankruptcy prediction, Support vector machine is highly appreciated and popular. Literature shows the numerous of study that compare Neural network and support vector machine. Shin et al (2005)

indicated the limitations of using Neural network to support vector machine: (1) It is an art to find an appropriate NN model, (2) the empirical risks minimization principle that seeks to minimize the training error does not guarantee good generalization performance.

SVM is used commonly in bankruptcy prediction. Recently, the SVM approach has been introduced to several financial applications such as credit rating, time series prediction, and insurance claim fraud detection (Fan (2000); Van (2001); Huang (2004); Kim (2003); Tay & Cao (2001); Viaene (2002)). These studies reported that SVM was comparable to and even outperformed other classifiers including NN, MDA, and Logit in terms of generalization performance.
Shin (2005) employed the data of Korea Credit Guarantee Fund that included 2320 medium-size manufacturing firms, in which 1160 bankruptcy and 1160 non-bankruptcy randomly selected firms from 1966 to 1999. The authors split data into two subsets: 80% of training set and 20% of validation set. The result reported that SVM model present the best prediction performance and has higher accuracy than NN.

Min (2005) also used SVM to predict the bankruptcy. The database is also collected in Korea, from 2000 to 2002 with 944 bankrupt and 944 non-bankrupt cases. The result is not consistent with Shin (2005): SVM showed to be an attractive prediction power compared to the classical existing methods (MDA, LR). Meanwhile, SVM does not statistical significantly outperform NN.

Bose (2006) experiment on 240 click-and-mortar traded corporation from 1993-2003, using 24 financial ratios. MDA, NN and SVM are compared on the accuracy. They concluded that NN and SVM are always better than MDA in predicting. However, their result supports NN and reported that this method outperform SVM.

Ding (2008) research on Chinese market to predict on the "*Special treated"* companies by Chinese Stock Exchange. SVM is also compared with other method BPNN, MDA and Logit. The result suggested that SVM outperforms NN, MDA and Logit for about 6%.

Depending on the study, NN and SVM are always reported as one of the most accuracy method and superior than classical techniques. However, in comparison with neural networks, SVM has some major advantages: SVM has only two free parameters (upper bound and kernel parameter). SVM also guarantee the existence of unique, optimal and global solution.

2.3. TEXTUAL ANALYSIS

The advancement of technology provides state-of-the-art tool to analysis data. One of the current streams is to use textual analysis. Textual analysis furnishes new point of view by investigating in non-numeric data.

Textual analysis was introduced in the 1960s by document classification and became popular in 1990s. This method has found a variety of applications in diverse domains (Kumar, 2016). Especially, in the decades of social media, big data nowadays, text mining has become a leading trend to analyse text content not only on Facebook, Twitter, blog or other social networks but also through news and reports (Wu He, 2013). It is admitted as an interesting and potential approach and becoming popular in finance field.

There are significant characteristics that distinguish between numeric and non-numeric data. In consequence, the steps in dealing with data set are different. The following states a classical step for textual analysis:

(1) Collect non-numeric data: it can be news, annual reports. This data is in the format of unstructured data and unstandardized
(2) Transfer unstructured data into structured data, which so-called pre-process
(3) Apply analysing process and generate information.

To achieve information from textual data, there are 3 frequently-used methods: Bag-of-words, Topic modelling and Document clustering.

Bag-of-words

Since first introduced by Sivic and Zisserma (2007), this model has been widely used in many studies. The bag-of-words model is one of the most popular representation methods for object and text categorization. The approach is relatively simple: organize words into 'bags'. In this model, a sentence or document is considered as a 'Bag' containing words without any order. Basically, it will take into account the words and their frequency of occurrence in the text documents disregarding semantic relationship in the sentences.

Using bag-of-words transfer from unstructured data into structured data by creating Document Term Matrix.

Topic modelling via Latent Dirichlet Allocation

Topic modelling is built based on the assumption that banks go failure for certain groups of reasons. These reasons can be shared among banks. Hence, topic modelling classifies given banks into certain groups of reasons (topics).

Topic modelling indicates the latent semantics in the document corpus and identify document groups, which is more useful than raw term features. Latent Dirichlet Allocation (Blei et al., 2003) provides an approach to modelling text corpora. This is one of the most popular algorithms for topic modelling. Without diving into math equations behind the model, we can understand it as being conceptualized by two principles: LDA estimates the ***mixture of words*** that compose a ***topic*** and determine the topics that describe each ***document***.

However, the most challenge question is to define the optimal numbers of topics before classification process. The literature has not provided a homogeneous suggestion on the optimal number of topics and remained an open-end question (Arun 2010).

Document clustering

Document clustering is a powerful technique for large-scale topic discovery from text (Larsen 1999). This technique has not been used widely in finance sector, however, is

used in law, web page analysis. (Ramage et al. 2009, Wong et al. 2002). Similarity to topic model, the goal of document clustering is to assign documents to different clusters (Aggarwal and Zhai, 2012; Lu et al., 2011; Xu and Gong, 2004). By grouping, document clustering crucial for document organization, browsing, summarization, classification and retrieval. There are two common algorithms: using the ***hierarchical*** based algorithm and using the ***K-means*** algorithm and its variants. The main different between these two algorithms is at the number of clustering defined step. K-means, like topic modelling, define the number of clusters before running the algorithms. Meanwhile, hierarchical designates number of clusters accordingly to the research purpose after applying algorithms.

Generally, hierarchical algorithms produce more in-depth information for detailed analyses, while algorithms based on variants of the K-means algorithm are more efficient and provide sufficient information for most purposes (Qin et al. 2017).

The brief approach of the hierarchical clustering is: (1) Place each data point (document) into its own cluster, then (2) Identify the closet 2 clusters and combine them into one cluster, and (3) Repeat step (2) till all documents are in merged into a single cluster. HC is typically visualized as a dendrogram.

Meanwhile, the goal of K-means clustering is to find groups in the corpus with the number of groups defined by the given variable K. For this approach, it is important to define the number of topics and iteratively redistribute the documents into topics until some termination condition is set. The disadvantage in k-means is the accuracy and efficiency depend on the choice of initial clustering centre.

2.4. COMPARISON OF STATISTICAL AND MACHINE LEARNING

In general, there are some different approaches in the comparison of Machine learning and statistical methods (Table 1). In fact, the two are highly related and share some underlying mechanism; however, these methods have different purposes. As Brian D. Ripley, the British Statistician, who won several prizes and awards used to say that: "To

paraphrase provocatively, ***machine learning*** is statistics minus any checking of models and assumptions".

The review of (Kumar, 2007) Concludes that almost machine learning techniques are used to solve bank failure problem, statistical techniques in stand-alone mode are no longer employed.

Basically, Machine learning is an algorithm that can learn from data without relying on rules-based programming. Meanwhile, Statistical modelling is the formalization of relationships between variables in the form of mathematical equations.

As mentioned above, the major distinguish between machine learning and statistics is their purpose. Machine learning focuses on prediction; meanwhile statistics are designed for explanation about the relationship between independent and dependent variables. In fact, researchers made great efforts in using statistics models (such as linear regression and logistics regression) for predict purpose; however, the predictive accuracy is not their strength.

Another important point is about the data. As rule of thumps, for statistics methods the model can require very small dataset (30 observations for example). However, for machine learning models, it requires a large database. Moreover, it is noteworthy that, for statistics, when we increase the number of observation, it could decrease the R^2 but for Machine learning models, the more data, the better predictive accuracy.

The dependence on hypothesis is also noteworthy. Statistics require strict hypothesises, meanwhile Machine learning does not.

Table 1: Comparison of machine learning and statistical methods

	Machine Learning methods	**Statistical methods**
	- Create model from data - Work with data to solve problems	
	- is a subfield of computer science and artificial intelligence - Emphasize on prediction than analysis	- is a subfield of mathematics and statistical. - Focus on estimation and inference
Purpose	- Analyses the data examples and generate a procedure that, given a new unseen example, can accurately predict its class.	- Provide the mathematical framework needed to make estimations and predictions.
Data	- Easily process with a large data sets	- Also run on large amount of data, but originally not as large as Machine learning.
Process	- Learn from data without relying on rules-based programming. - Necessary to divide data into Training and Test set	- Formalize the relationships between variables in the form of mathematical equations
Assumption	– Freed from model assumption or diagnostics – Freed from justify model choice or test assumption	– Large number of strict assumptions regarding, for example, multicollinearity, linear relation, homoscedasticity, etc.
Birthday	Machine learning was defined more recently by computer scientists like Arthur Samuel and	Statistical modelling has been around for centuries.

	Machine Learning methods	**Statistical methods**
	Tom Mitchell in the mid-to-late 1950's.	
Output	"The model is 85% accurate in predicting Y, given a, b and c."	"The model is 85% accurate in predicting Y, given a, b and c; and I am 90% certain that you will obtain the same result."

In conclusion, the literature provides a comprehensive view on bank failure by analysing variety types of financial ratios and using mixture of methods (table 2). Bank failed for several reasons, however, all these reasons relate directly or indirectly to Loan quality and bank management: (1) If loan quality is improved, it can reduce the probability of being failure, (2) If provision for loan loss is adequately estimated, it can be a firm cushion of liquidity risk, (3) If bank managers manage bank effectively, it is certain that the bank will not be failed, (4) If equity capital is good enough, it can cover loan loss and protect liquidity position. And the question of "Bad luck and Bad Management" is still on debate.

Table 2: Brief reviews of papers

Authors	Year	Sources of data	Samples	Techniques	Period
Martin	1977	US	5700	Logit regression	1970-1976
West	1985	US	1900	FA, logit	1980-1982
Tam et al.	1990	Texas	202	LDA, logit, k-NN, BPNN	1985-1987
Tam et al.	1991	Texas	188	BPNN, Factor logistic, DA, k-NN, ID3	1985-1987
Haslem et al.	1992	US	176	Canonical correlation	1987
Barr et al.	1994	US	930	DEA	1984-1987
Bell et al.	1997	Texas, FDIC	2067	NN, logit	1985-1986
Olmeda et al.	1997	Spanish	66	BPNN, Logit, MARS, DA	1997-1985
Alam et al.	2000	US, FDIC annual reports	100	CNN, SONN, Fuzzy clustering	1991
Swicegood et al.	2001	US	1741	MDA, BPNN, Regulators	1993
Kolari et al.	2002	US	8977	Logit, trait recognition	1989-1992
Cielen et al.	2004	Belgium	366	MSD, DEA, C5.0	1994-1996
Kao et al.	2004	Taiwan	24	DA, BPNN	2000

Authors	Year	Sources of data	Samples	Techniques	Period
Canbas et al.	2005	Turkish	40	PCA, MDA, logit, probit	1994-2001
Boyacioglu et al.	2009	Turkey	65	NNs, SVM, LDA,	1997 -2003
Zhao et al.	2009	US	240	LR, decision tree, NNs, and *k*-nearest neighbo	1991-1992
Jin et al.	2011	US	25,428 bank-quarter	simple univariate analysis, multivariate analysis	2006-2007
Amadasu	2012	Nigerian	6	MDA, LR, OLS, Z-score	2003-2007
Cox et al.	2014	US	322	MDA	2007 -2010
Maghyereh	2014	Gulf Cooperation Council countries	70	Hazard	2000-2009
Behbood et al.	2015	20news group dataset		Multistep Fuzzy, ANN, SVM,	
Iturriaga et al	2015	US	52	ANNs	2002-2012
Mare et al.	2015	Italia	434	Hazard	1993-2011
Geng et al.	2015	China	109	ANN, SVM	2001-2008
Cleary et al	2016	US	132	MDA	2002- 2009

3. REFERENCES

1. Amendolia, Salvator Roberto, Gianfranco Cossu, M. L. Ganadu, Bruno Golosio, G. L. Masala, and Giovanni Maria Mura. "A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening." Chemometrics and Intelligent Laboratory Systems 69, no. 1-2 (2003): 13-20.
2. Alam, P., Booth, D., Lee, K., & Thordarson, T. (2000). The use of fuzzy clustering algorithm and self-organizing neural networks for identifying potentially failing banks: an experimental study. Expert Systems with Applications, 18(3), 185-199.
3. Altman, Edward I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." The journal of finance 23, no. 4 (1968): 589-609.
4. Altman, Edward I., Robert G. Haldeman, and Paul Narayanan. "ZETATM analysis A new model to identify bankruptcy risk of corporations." Journal of banking & finance 1, no. 1 (1977): 29-54.
5. Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." In Mining text data, pp. 77-128. Springer, Boston, MA, 2012.
6. Amadasu, David E. "Bank Failure Prediction." AFRREV IJAH: An International Journal of Arts and Humanities 1, no. 4 (2012): 250-265.
7. Arun, Rajkumar, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. "On finding the natural number of topics with latent dirichlet allocation: Some observations." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 391-402. Springer, Berlin, Heidelberg, 2010.
8. Barr, Richard S., Lawrence M. Seiford, and Thomas F. Siems. "Forecasting bank failure: A non-parametric frontier estimation approach." Recherches Économiques de Louvain/Louvain Economic Review 60, no. 4 (1994): 417-429.
9. Barth, James R., Tong Li, and Wenling Lu. "Bank regulation in the United States." CESifo Economic Studies 56, no. 1 (2009): 112-140.

10. Beaver, William H. "Financial ratios as predictors of failure." Journal of accounting research (1966): 71-111.
11. Behbood, Vahid, Jie Lu, Guangquan Zhang, and Witold Pedrycz. "Multistep Fuzzy Bridged Refinement Domain Adaptation Algorithm and Its Application to Bank Failure Prediction." IEEE Trans. Fuzzy Systems 23, no. 6 (2015): 1917-1935.
12. Bell, Timothy B. "Neural nets or the logit model? A comparison of each model's ability to predict commercial bank failures." Intelligent Systems in Accounting, Finance & Management 6, no. 3 (1997): 249-264.
13. Bennett, Rosalind L., and Haluk Unal. "Understanding the components of bank failure resolution costs." Financial Markets, Institutions & Instruments 24, no. 5 (2015): 349-389.
14. Berger, Allen N., and David B. Humphrey. "Megamergers in banking and the use of cost efficiency as an antitrust defense." The Antitrust Bulletin 37, no. 3 (1992): 541-600.
15. Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. "When is "nearest neighbor" meaningful?" In International conference on database theory, pp. 217-235. Springer, Berlin, Heidelberg, 1999.
16. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.
17. Bose, Indranil, and Raktim Pal. "Predicting the survival or failure of click-and-mortar corporations: A knowledge discovery approach." European Journal of Operational Research 174, no. 2 (2006): 959-982.
18. Bouvatier, Vincent, Michael Brei, and Xi Yang. "The determinants of bank failures in the United States: Revisited." (2013).
19. Boyacioglu, Melek Acar, Yakup Kara, and Ömer Kaan Baykan. "Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey." Expert Systems with Applications 36, no. 2 (2009): 3355-3366.

20. Canbas, Serpil, Altan Cabuk, and Suleyman Bilgin Kilic. "Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case." European Journal of Operational Research 166, no. 2 (2005): 528-546.
21. Cantor, Richard. "Moody's investors service response to the consultative paper issued by the Basel Committee on Bank Supervision "A new capital adequacy framework"." Journal of Banking & Finance 25, no. 1 (2001): 171-185.
22. Cao, Juan, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. "A density-based method for adaptive LDA model selection." Neurocomputing 72, no. 7-9 (2009): 1775-1781.
23. Chen, Hui-Ling, Chang-Cheng Huang, Xin-Gang Yu, Xin Xu, Xin Sun, Gang Wang, and Su-Jing Wang. "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach." Expert systems with applications 40, no. 1 (2013): 263-271.
24. Cielen, Anja, Ludo Peeters, and Koen Vanhoof. "Bankruptcy prediction using a data envelopment analysis." European Journal of Operational Research 154, no. 2 (2004): 526-532.
25. Cleary, Sean, and Greg Hebb. "An efficient and functional model for predicting bank distress: In and out of sample evidence." Journal of Banking & Finance 64 (2016): 101-111.
26. Coats, Pamela K., and L. Franklin Fant. "Recognizing financial distress patterns using a neural network tool." Financial management (1993): 142-155.
27. Cole, Rebel A., and Lawrence J. White. "Déjà vu all over again: The causes of US commercial bank failures this time around." Journal of Financial Services Research 42, no. 1-2 (2012): 5-29.
28. Collins, Robert A., and Richard D. Green. "Statistical methods for bankruptcy forecasting." Journal OF Economics and business 34, no. 4 (1982): 349-354.
29. Cooper, William W., Lawrence M. Seiford, and Joe Zhu. "Data envelopment analysis." In Handbook on data envelopment analysis, pp. 1-39. Springer, Boston, MA, 2004.

30. Cornfield, Jerome. "Discriminant functions." Revue de l'Institut International de Statistique (1967): 142-153.
31. Cox, Raymond AK, and Grace W-Y. Wang. "Predicting the US bank failure: A discriminant analysis." Economic Analysis and Policy 44, no. 2 (2014): 202-211.
32. Cox, Raymond AR, and Rose M. Prasad. "Characteristics of acquisitions of failed banks: methodological considerations." The Review of Research in Banking and Finance 4 (1988).
33. Dash, Mihir, and Annyesha Das. "A CAMELS analysis of the Indian banking industry." (2009).
34. Deveaud, Romain, Eric SanJuan, and Patrice Bellot. "Accurate and effective latent concept modeling for ad hoc information retrieval." Document numérique 17, no. 1 (2014): 61-84.
35. Ding, Yongsheng, Xinping Song, and Yueming Zen. "Forecasting financial condition of Chinese listed companies based on support vector machine." Expert Systems with Applications 34, no. 4 (2008): 3081-3089.
36. Dörre, Jochen, Peter Gerstl, and Roland Seiffert. "Text mining: finding nuggets in mountains of textual data." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 398-401. ACM, 1999.
37. Estrella, Arturo, Sangkyun Park, and Stavros Peristiani. "Capital ratios and credit ratings as predictors of bank failures." Federal Reserve Bank of New York: Economic Policy Review (2002): 33-52.
38. Fan, Alan, and Marimuthu Palaniswami. "Selecting bankruptcy predictors using a support vector machine approach." In Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on, vol. 6, pp. 354-359. IEEE, 2000.
39. Fraser, D. R., Gup, B. E., & Kolari, J. W. (1995). Commercial bank: the management of risk. New York: West Publishing.
40. Geng, Ruibin, Indranil Bose, and Xi Chen. "Prediction of financial distress: An empirical study of listed Chinese companies using data mining." European Journal of Operational Research 241, no. 1 (2015): 236-247.

41. Gonzalez-Hermosillo, Brenda, Ceyla Pazarbaşioğlu, and Robert Billings. "Determinants of banking system fragility: A case study of Mexico." Staff Papers44, no. 3 (1997): 295-314.
42. Grammatikos, Theoharry. "Intervalling effects and the hedging performance of foreign currency futures." Financial Review 21, no. 1 (1986): 21-36.
43. Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.
44. Handorf, William C. "CAMEL to CAMELS: The risk of sensitivity." Journal of Banking Regulation 17, no. 4 (2016): 273-287.
45. Haslem, John A., Carl A. Scheraga, and James P. Bedingfield. "An analysis of the foreign and domestic balance sheet strategies of the US banks and their association to profitability performance." MIR: Management International Review (1992): 55-75.
46. Hellmann, Thomas F., Kevin C. Murdock, and Joseph E. Stiglitz. "Liberalization, moral hazard in banking, and prudential regulation: Are capital requirements enough?" American economic review 90, no. 1 (2000): 147-165.
47. Hooks, Linda M. "Bank asset risk: Evidence from early-warning models." Contemporary Economic Policy 13, no. 4 (1995): 36-50.
48. Huang, Zan, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. "Credit rating analysis with support vector machines and neural networks: a market comparative study." Decision support systems 37, no. 4 (2004): 543-558.
49. Hwang, Dar-Yeh, Cheng F. Lee, and K. Thomas Liaw. "Forecasting bank failures and deposit insurance premium." International Review of Economics & Finance 6, no. 3 (1997): 317-334.
50. Iturriaga, Félix J. López, and Iván Pastor Sanz. "Bankruptcy visualization and prediction using neural networks: A study of US commercial banks." Expert Systems with applications 42, no. 6 (2015): 2857-2869.
51. Jin, Justin Yiqiang, Kiridaran Kanagaretnam, and Gerald J. Lobo. "Ability of accounting and audit quality variables to predict bank failure during the financial crisis." Journal of Banking & Finance 35, no. 11 (2011): 2811-2819.

52. Kao, Chiang, and Shiang-Tai Liu. "Predicting bank performance with financial forecasts: A case of Taiwan commercial banks." Journal of Banking & Finance 28, no. 10 (2004): 2353-2368.
53. Keasey, Kevin, and Robert Watson. "Non-financial symptoms and the prediction of small company failure: A test of Argenti's hypotheses." Journal of Business Finance & Accounting 14, no. 3 (1987): 335-354.
54. Kolari, James, Dennis Glennon, Hwan Shin, and Michele Caputo. "Predicting large US commercial bank failures." Journal of Economics and Business 54, no. 4 (2002): 361-387.
55. Kumar, P. Ravi, and Vadlamani Ravi. "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review." European journal of operational research 180, no. 1 (2007): 1-28.
56. Kumar, P. Ravi, and Vadlamani Ravi. "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review." European journal of operational research 180, no. 1 (2007): 1-28.
57. Lam, Kim Fung, and Jane W. Moy. "Combining discriminant methods in solving classification problems in two-group discriminant analysis." European Journal of Operational Research 138, no. 2 (2002): 294-301.
58. Larsen, Bjornar, and Chinatsu Aone. "Fast and effective text mining using linear-time document clustering." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 16-22. ACM, 1999
59. Leif, E. P., and K-Nearest Neighbor. "Scholarpedia 4 (2): 1883." (2009).
60. Lu, Caimei, Xiaohua Hu, and Jung-ran Park. "Exploiting the social tagging network for web clustering." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41, no. 5 (2011): 840-852.
61. Lu, Wenling, and David A. Whidbee. "Bank structure and failure during the financial crisis." Journal of Financial Economic Policy 5, no. 3 (2013): 281-299.

62. Maghyereh, Aktham I., and Basel Awartani. "Bank distress prediction: Empirical evidence from the Gulf Cooperation Council countries." Research in International Business and Finance 30 (2014): 126-147.
63. Mare, Davide Salvatore. "Contribution of macroeconomic factors to the prediction of small bank failures." Journal of International Financial Markets, Institutions and Money 39 (2015): 25-39.
64. Martin, Daniel. "Early warning of bank failure: A logit regression approach." Journal of banking & finance 1, no. 3 (1977): 249-276.
65. Mensah, Yaw M. "The differential bankruptcy predictive ability of specific price level adjustments: some empirical evidence." Accounting Review (1983): 228-246.
66. Meyer, Paul A., and Howard W. Pifer. "Prediction of bank failures." The Journal of Finance 25, no. 4 (1970): 853-868.
67. Min, Jae H., and Young-Chan Lee. "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters." Expert systems with applications 28, no. 4 (2005): 603-614.
68. Mitchell, Tom Michael. The discipline of machine learning. Vol. 9. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.
69. Ohlson, James A. "Financial ratios and the probabilistic prediction of bankruptcy." Journal of accounting research(1980): 109-131.
70. Olmeda, Ignacio, and Eugenio Fernández. "Hybrid classifiers for financial multicriteria decision making: The case of bankruptcy prediction." Computational Economics 10, no. 4 (1997): 317-335.
71. Pang, Shaoning, Daijin Kim, and Sung Yang Bang. "Membership authentication in the dynamic group by face classification using SVM ensemble." Pattern Recognition Letters 24, no. 1-3 (2003): 215-225.
72. Persons, Obeua. "Using financial information to differentiate failed vs. surviving finance companies in Thailand: an implication for emerging economies." (1999).
73. Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4, no. 2 (2009): 1883.
74. Philbin, James, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. "Object retrieval with large vocabularies and fast spatial matching." In Computer

Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pp. 1-8. IEEE, 2007.

75. Press, S. James, and Sandra Wilson. "Choosing between logistic regression and discriminant analysis." Journal of the American Statistical Association 73, no. 364 (1978): 699-705.
76. Qin, Zemin, Hao Lian, Tieke He, and Bin Luo. "Cluster Correction on Polysemy and Synonymy." In Web Information Systems and Applications Conference (WISA), 2017 14th, pp. 136-138. IEEE, 2017.
77. Ramage, Daniel, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. "Clustering the tagged web." In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 54-63. ACM, 2009.
78. Salchenberger, Linda M., E. Mine Cinar, and Nicholas A. Lash. "Neural networks: A new tool for predicting thrift failures." Decision Sciences 23, no. 4 (1992): 899-916.
79. Salton, Gerard, and Michael J. McGill. "Introduction to modern information retrieval." (1986).
80. Shin, Kyung-Shik, Taik Soo Lee, and Hyun-jung Kim. "An application of support vector machines in bankruptcy prediction model." Expert Systems with Applications 28, no. 1 (2005): 127-135.
81. Snapinn, Steven M., and James D. Knoke. "An Evaluation of Smoothed Classification Error–Rate Estimators." Technometrics 27, no. 2 (1985): 199-206.
82. Stone, Mary, and John Rasp. "Tradeoffs in the choice between logit and OLS for accounting choice studies." Accounting review (1991): 170-187.
83. Swicegood, Philip, and Jeffrey A. Clark. "Off-site monitoring systems for predicting bank underperformance: a comparison of neural networks, discriminant analysis, and professional human judgment." Intelligent Systems in Accounting, Finance & Management 10, no. 3 (2001): 169-186.

84. Tam, Kar Yan, and Melody Kiang. "Predicting bank failures: A neural network approach." Applied Artificial Intelligence an International Journal 4, no. 4 (1990): 265-282.
85. Tam, Kar Yan, and Melody Y. Kiang. "Managerial applications of neural networks: the case of bank failure predictions." Management science 38, no. 7 (1992): 926-947.
86. Tam, Kar Yan. "Neural network models and the prediction of bank bankruptcy." Omega 19, no. 5 (1991): 429-445.
87. Tay, Francis EH, and Lijuan Cao. "Application of support vector machines in financial time series forecasting." omega29, no. 4 (2001): 309-317.
88. Thomson, James B. "Predicting bank failures in the 1980s." Federal Reserve Bank of Cleveland Economic Review 27, no. 1 (1991): 9-20.
89. Van Gestel, Tony, Johan AK Suykens, D-E. Baestaens, Annemie Lambrechts, Gert Lanckriet, Bruno Vandaele, Bart De Moor, and Joos Vandewalle. "Financial time series prediction using least squares support vector machines within the evidence framework." IEEE Transactions on neural networks 12, no. 4 (2001): 809-821.
90. Viaene, Stijn, Richard A. Derrig, Bart Baesens, and Guido Dedene. "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection." Journal of Risk and Insurance 69, no. 3 (2002): 373-421.
91. West, Robert Craig. "A factor-analytic approach to bank condition." Journal of Banking & Finance 9, no. 2 (1985): 253-266.
92. Wilson, Rick L., and Ramesh Sharda. "Bankruptcy prediction using neural networks." Decision support systems 11, no. 5 (1994): 545-557.
93. Wong, Wai-Chiu, and Ada Wai-Chee Fu. "Incremental document clustering for web page classification." In Enabling Society with Information Technology, pp. 101-110. Springer, Tokyo, 2002.
94. Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." In Proceedings of the 26th annual international

ACM SIGIR conference on Research and development in information retrieval, pp. 267-273. ACM, 2003.

95. Zavgren, Christine V. "Assessing the vulnerability to failure of American industrial firms: a logistic analysis." Journal of Business Finance & Accounting 12, no. 1 (1985): 19-45.
96. Zedeh, L_A. "Fuzzy sets." Information and control 8, no. 3 (1965): 338-353.
97. Zhang, Guoqiang, Michael Y. Hu, B. Eddy Patuwo, and Daniel C. Indro. "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis." European journal of operational research 116, no. 1 (1999): 16-32.
98. Zhao, Huimin, Atish P. Sinha, and Wei Ge. "Effects of feature construction on classification performance: An empirical study in bank failure prediction." Expert Systems with Applications36, no. 2 (2009): 2633-2644.
99. Zhu, Wenyu, and Jiawen Yang. "State ownership, cross-border acquisition, and risk-taking: Evidence from China's banking industry." Journal of Banking & Finance 71 (2016): 133-153.

CHAPTER 3

CHAPTER 3:

PREDICTING BANK FAILURE: AN IMPROVEMENT BY IMPLEMENTING A MACHINE-LEARNING APPROACH TO CLASSICAL FINANCIAL RATIOS

(This article is published in:

Research in International Business and Finance 44 (2018) 16–25)

ABSTRACT

This research compares the accuracy of two approaches: traditional statistical techniques and machine learning techniques, which attempt to predict the failure of banks. A sample of 3000 US banks (1438 failures and 1562 active banks) is investigated by two traditional statistical approaches (Discriminant analysis and Logistic regression) and three machine learning approaches (Artificial neural network, Support Vector Machines and k-nearest neighbours). For each bank, data were collected for a 5-year period before they become inactive. 31 financial ratios extracted from bank financial reports covered 5 main aspects: Loan quality, Capital quality, Operations efficiency, Profitability and Liquidity. The empirical result reveals that the artificial neural network and k-nearest neighbour methods are the most accurate.

Keywords: Failure, prediction, intelligent techniques, Artificial neural network, Support vector machines, K-nearest neighbours, banks

JEL: G21, G33

1. INTRODUCTION

According to the Federal Deposit Insurance Corporation (FDIC), during 2008–2014 more than 500 banks declared as failures in the United States of America. The cost of failure per dollar of failed-bank assets is already high and may continue to rise. Consequently, the more banks go bankrupt, the higher the cost of resolving after-failure events. Year-end 2013, FDIC estimated that the total cost to the deposit insurance funds of resolving these failed banks is as high as 30 billion US dollars.

Banks are considered as failures if the state or bank regulator forces them to close because of insolvency problems. Because of the strong interconnection between banks and their essential role in financing the economy, the failure of banks is more threatening for the economy than the failure of other business firms. In some cases, the bankruptcy of one bank can cause a knock-on effect, which can spread quickly and have a negative impact on other banks (systemic risk). Hence, detection of bank failure before it occurs and try to avoid them is mandatory. In this research, we execute machine learning techniques which have been claimed to improve the prediction of bank failure.

Starting with seminal research studies, Beaver (1966) and Altman (1968) built statistical models to predict firm failure based on accounting ratios. Since then, numerous studies have been advanced using different financial ratios, samples and periods. In parallel with the development of computational sciences, many different interesting approaches were explored to promote the power of technology. In order to help researchers better understand this complex field, Ravi Kumar and Ravi (2007) present a comprehensive review of the applications of prediction techniques to solve bankruptcy prediction problems of banks and firms. One of the intelligent technique families known as 'Machine learning' becomes more popular among researchers and practitioners. A commonly cited formal definition of machine learning, proposed by a computer scientist (Lantz, 2013, p. 10) explained that a machine is said to learn if it is able to take experience and utilize it such that its performance improves up on a similar experience in the future.

More formally, according to Mitchell (1997), a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. In this study, we want to examine the effectiveness of these methods compared with more traditional statistical techniques.

The main contributions of the paper are the following. First it proposes a comparison of traditional statistical techniques: Linear Discriminant analysis (LDA) and Logistic regressions (Logistic) to machine learning techniques on predicting the failure of US banks. Machine learning techniques (k- Nearest neighbours (k-NNs), artificial neural networks (ANN) and Support Vector Machines (SVMs)) had not been systematically compared to predict bank failure. Moreover, the empirical results of previous studies are unclear (see the recent paper of López and Pastor Sanz, (2015)). Secondly, we use a large number of financial ratios (31) for 5 years before banks become inactive (inactivity can be due to bankruptcy, illiquidity, merging, or insolvency). The large number of ratios provides a means of covering all bank financial characteristics: loan quality, capital quality, operations, profitability and liquidity and of determining the ratios with the best failure prediction power. This diversity is justified by our analysis of the "Material loss review" of 102 banks from FDIC reports since 2009–2015, which shows that banks fail for various reasons (loan problems, profit reduction, credit risk, ineffective board of directors and management). Thirdly, we test these various techniques on a large sample of 3000 US banks (1438 failed, 1532 active) during the crisis and post-crisis period (2008–2014). This period deserves an in-depth study due to the change in financial environment and banking techniques: fall of real estate prices, biased pricing methods, new financial products and risks (Demyanyk and Hasan, 2010; López and Pastor Sanz, 2015), the reasons for bank failures could be different (or not) from those previously observed. Better knowledge of bank failure determinants is also very important for regulators (in the Basel 3, 4 reforms perspective).

This paper is constructed as follows: part 2 introduces a literature review. The methodologies are presented more detail in part 3. Part 4 is about the data and variables. The final result is mentioned in part 5. Finally, conclusion and discussion are included in part 6.

2. LITERATURE REVIEW ON BANK FAILURE PREDICTION

Taking into account the fact that bankruptcy prediction is an important and widely studied topic, we will concentrate our literature review on the prediction of bank failure using financial ratios. Moreover, we will focus on the studies that implement at least one of the five techniques compared in this paper. The history of bankruptcy prediction originated from predicting the failure of businesses. The important contribution of Altman (1968) motivated researchers to use multivariate analysis to predict the bankruptcy of firms. He provided an original Z-score formula (1968) and showed its advantage by analysing five main financial and economic aspects of a firm: the liquidity, size dimensions; operating efficiency and profitability of the assets, financial leverage as well as considering the capability of management in dealing with competitive conditions (total asset turnover). Sinkey (1975) employed discriminant analysis to predict bank failures.

In comparison, Martin (1977) and Ohlson (1980) employed logistic regression to predict failures of firms and bank. Martin (1977) attempted to predict the US commercial bank failure within 2 years during 1970 and 1976 by using 25 financial ratios of asset risk, liquidity, capital adequacy and earning. He suggested that logistic regression has a higher percentage of correctly classified than linear discriminant. Since these initial studies, empirical studies have been conducted to compare the prediction accuracy of these two approaches (Boyacioglu et al., 2009).

Nevertheless, empirical studies do not demonstrate a clear advantage for one of the two main traditional techniques: discriminant analysis versus Logit and Probit models. But Canbas et al. (2005) on a sample of 40 privately owned Turkish commercial banks showed, using 49 ratios, that discriminant analysis obtains slightly better results than

Probit and Tobit models. On the same vein, a recent study Chiaramonte et al. (2015) revealed, on a big sample of 3242 banks across 12 European countries, that Z-score is a good predictive model to identify banks in distress (better than the Probit model) and also has the great advantage of simple calculation. According to the empirical study by Lo (1986), the equivalence between LDA and LR may not be rejected.

However, in some standpoints, statistical techniques are no longer preferred in view of their relatively low accuracy (Ravi Kumar and Ravi, 2007). The attention to and confidence in machine learning has increased enormously during the past 5–10 years. Numerous studies suggest that intelligent techniques perform more effectively than traditional statistical techniques. The main difference between intelligent and statistical techniques is that statistical techniques usually require researchers to define the structures of the model a priori, and then to estimate parameters of the model to fit the data with observations, while with intelligent techniques the particular structure of the model is learned directly from the data (Wang et al., 2015). Moreover, the statistical analysis depends on strict assumptions (normal distribution, no correlations between independent variables), that can result in poor prediction accuracy.

Among several machine-learning methods, the artificial neural network seems to be the most favored tool in prediction issues. Ky (1991) was among the first to implement a neural network on 118 banks (59 failed and 59 non-failed banks) in Texas during 1985–1987 and indicated that the neural network performed more effectively than other methods (Discriminant Analysis, factor-logistic, k-NNs and Decision tree). Several studies (Miguel et al., 1993; Bell, 1997; Olmeda and Fernandez, 1997; Swicegood and Clark, 2001; Aktas et al., 2003; Wu and Wang, 2000) compare ANN and the classical statistical techniques (Discriminant Analysis and Logistic Model) to predict bank failure. They generally conclude in the superiority of the neural network approach. In their survey Vellido et al. (1999), also suggest that ANN is better than the logit model for predicting commercial bank failures. More recently Lee and Choi (2013) compared the prediction accuracy of neural networks and linear discriminant analysis on a sample of Korean companies. Their results indicated that the bankruptcy prediction accuracy using neural networks is greater than that of LDA. Finally, a meta-analysis performed by Adya

and Collopy (1998) reveals that neural networks outperformed alternative approaches in 19 out of the 22 analysed studies.

Unlike Neural networks and Support Vector Machines, the k-nearest neighbour algorithm is not implemented widely in finance. This technique is implemented widely in biological and transportation fields. This method, however, can function appreciatively and obtain high prediction accuracy. Min and Lee (2005)proposed support Vector Machines for bankruptcy prediction. Boyacioglu et al. (2009) examined ANNs, SVMs and multivariate statistical methods to predict the failure of 65 Turkish financial banks. 20 financial ratios belonging to 6 main groups were chosen: Capital adequacy, Asset quality, Management, Earning, Liquidity and the sensitivity to the market risk. Overall, the result proved that SVMs achieved the highest accuracy. They concluded that this method outperforms neural network, discriminant analysis and logit methods. SVM was also proved to work better than neural networks through the research of Chiaramonte et al. (2015) for a sample of 3242 EU banks. Park and Han (2002)used k-nearest neighbour for company bankruptcy prediction but we do not find empirical studies specifically dedicated to the use of k-nearest neighbour to predict bank failure.

Finally, some empirical studies compare the various predictions methods. Tam and Kiang (1992) compare discriminant analysis, Logit, k-nearest neighbour and artificial neural networks on bank failure prediction and find that the latter outperforms the other techniques. Martínez (1996) compares neural network back propagation methods with discriminant analysis, logit analysis and the k-nearest neighbour for a sample of Texan banks and concludes that the first set of methods outperforms. Zhao et al. (2009) compare Logit, ANN and k-NN. They find that ANN > Logit > k-NN when financial ratios rather than row data are used. These studies support neural networks as being the best methods of predicting bank failure. Serrano-Cinca and Gutierrez-Nieto (2011) compared 9 different methods to predict the bankruptcy of USA banks during the financial crisis, including Logistic Regression, Linear Discriminant Analysis, Support vectors Machines, k-nearest neighbour and Neural Networks. It can be concluded that no technique is clearly better than the others. Performance depends on the performance

measure chosen; some techniques have more accuracy but less recall (1 minus Type II error rate).

Among numerous studies on predicting the bankruptcy of banks, history has shown that intelligent techniques (and specifically artificial neural networks) seem to work more effectively than statistical techniques. This study will execute both families of techniques in different methods and in a new attempt to make a comparison on two aspects: the accuracy and the importance of each ratio.

3. DATA AND METHODOLOGY

3.1. STATISTICAL TECHNIQUES

Linear discriminant analysis and Logistic regression are popular methods for classifying objects based on their characteristics. These methods have been applied widely to predict the failure of firms and banks.

3.1.1. Linear discriminant analysis (LDA) and logistic regression (LR)

Logistic regression (LR) is a regression model where the outcome is categorical (in our case the bank is active or inactive). More technically, the model assumes a linear relationship between the logarithm of the odds ratio (ratio of probabilities, see equation below) and one or more independent variables (bank characteristics, x_j).

$$g(x) = \ln \frac{P(y=1)}{P(y=0)} = \sum_{j=1}^{m} \beta_j x_j + \beta_0$$

Linear Discriminant Analysis (LDA) derives a linear combination of ratios which best discriminate between failed and non-failed firms. Observations are assigned to one of the two groups in some 'optimal' way, for example, so as to minimize the probability or cost of misclassification. Logistic is often preferred to LDA as it is more flexible in assumptions and types of data that can be analysed.

Canbas et al. (2005) propose an integrated model that combines LDA and LR in order to help predict bank failure. They demonstrate that this combination improves the prediction accuracy. Serrano-Cinca and Gutierrez-Nieto (2011)combine LDA with Partial Least Square analysis in order to predict the failure of US banks during the 2008 financial crisis.

Although the LDA and LR have become the most commonly used in bankruptcy prediction, their inherent drawbacks of statistical assumptions such as linearity, normality and independence among variables have constrained their practical applications (Lee and Choi, 2013). To solve the limitation of a linear approach, intelligent techniques (in this paper considered as machine learning approaches) achieve a forward movement by introducing nonlinear separation between groups.

Several methods have been implemented to classify companies or financial institutions and predict bankruptcy or failure. In this paper, three machine learning algorithms are applied: k-Nearest Neighbours, Artificial Neural Network and Support Vector Machines. Neural network is a well-known model and is considered to be one of the most powerful tools in prediction even when their conceptions are not easy to be translated. These models are referred to as black box processed because the mechanism that transforms the input into the output is obfuscated by a figurative box. On the contrary, k-nearest neighbours is regarded as a lazy learning technique (meaning that generalization beyond the training data is delayed until a query is made to the system). The idea is to classify unlabelled examples by assigning to them the class of the majority of its neighbours. Support Vector Machines are in between since, being not overly-complex, it is possible to enter the black box.

3.1.2. K-Nearest Neighbours (k-NN)

The k-Nearest Neighbour (k-NN) is an instance-based method, meaning that it assigns a new case to the majority class among the k-closest cases in the training set (Hand et al., 2001). In a brief description, nearest neighbour classifies by mapping the different characteristics of the dataset closely to different label groups, the given data with

common features will then be placed in the same group. Each new case is classified based on the outcome of the majority of its neighbours.

Each bank is represented by a vector of its characteristics. Banks with similar characteristics tend to be placed closely together. The distance between each point to each group must be calculated in order to find out which group (active or inactive) the banks belong to. If the majority of neighbours of a given bank are classified as failed (active), this bank will be classified as failed (active).

There are three major decisions in the k-NN method: the set of stored cases, the distance metric used to compute the distance between cases, and the value of k (Weiss and Indurkhya, 1998).

There are several ways of calculating this distance. Traditionally, the k-NN algorithm deploys Euclidean distance. If *p* and *q* are two vectors of characteristics (two banks), each of them has *n* features. The Euclidian distance between p and q is calculated as:

$$\text{Dist (p,q)}=\sqrt{(p_1 - q_1)^2 + \ \ldots + (p_n - q_n)^2} \quad (1)$$

To classify the bank as active or inactive, we should begin by assigning the number of neighbours, k. We can select any value of *k* to find the best grouping method. There are divergent hypotheses on selecting the 'best' *k*. Some researchers suppose that *k* should be the square root of number of features. However, others assume that *k* performs the best if it is between (2, 10). In this research, we experiment with various value of *'k'* to find the optimal value.

3.1.3. Artificial neural networks (ANNs)

Taking advantages of computer potential, Artificial Neural Networks (ANNs) are inspired by biological neural networks. ANN is applied widely on a variety of tasks such as: computer vision, speech recognition, etc. ANN is a machine learning technique, which

can simulate any relationship. Although ANN is not the only technique that can do this it is often preferred for the ability to obtain a solution in a reasonable time.

The idea is to learn from examples using several algorithms just as a human being learns new things. The advantages of ANNs are their flexible nonlinear modelling capability, strong adaptability, as well as their learning and massive parallel computing abilities (Ticknor, 2013). However, they cannot explain the causal relationship among variables, which restricts its application to managerial problems (Lee and Choi, 2013).

A fully connected network includes series of neuron layers. While each unit in the same layer cannot interconnect, each layer can. The connection between one unit in a given layer and another in the following layer is represented by a number call a weight, which can be positive or negative. There are two ways to transfer information: feed-forwarding and back-propagation. Feed-forwarding will forward information from input layer to output layer and this processing can lead to a wrong result. However, back-propagation can fix these errors by sending back the information to optimize the outcome.

When designing a multilayer network, the decision on choosing the number of hidden layers is very important. Lee et al. (2005) and Zhang et al. (1999) show that one hidden layer is sufficient for most classification problems. Meanwhile, Vasu and Ravi (2011) suggested choosing 2 hidden layers in order to be sure that the network architecture will be sufficiently complex to cope with the complexity of bank failure prediction. In our study, we applied for both 1 hidden layer and 2 hidden layers to examine which one performs better.

To eliminate the possibility of being linear, we use an activation function which creates a non-linear decision boundary. Various types of activation function exist, for example: sigmoid, tanh, rectified linear unit, leaky rectified linear unit or max out. We decided to use a sigmoid function since its characteristics are suitable for our output. It is the most widely used function.

3.1.4. Support Vector Machines (SVM)

The great advantage of Support Vector Machines (SVM) is that they combine the strengths of theory-driven conventional statistical methods and data-driven machine learning methods (Min and Lee, 2005). The method is based on the Vapnik's (1995) structural risk minimization principle. SVM is highly appreciated for successful applications in many fields such as bioinformatics, text, image recognition, etc. SVMs are supervised learning models that analyse data used for classification and regression analysis. This method is developed from Statistical Learning Theory (Boser et al., 1992). The basic idea is that input vectors (a vector represents the financial characteristics of a given bank) are non-linearly mapped to a very high dimension feature space. A linear decision surface is constructed in this feature space thus SVMs transform complex problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions.

Unlike numerous other methods which focus on whole training data, SVM pays attention to the most difficult to recognize data point based on the idea that if SVMs can figure out the toughest points, the others will be seen easily. The vectors most difficult to recognize are located close to the hyperplane separating active and failed banks, they are called support vectors. These points can be easily misclassified. The distance from the closest data points in each respective class to the hyperplane is called the margin. SVM will attempt to maximize these margins, so that the hyperplane is at the same distance from the 2 groups (failed and active banks). Intuitively, the more distant vectors are from the hyperplane, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

3.2. IMPLEMENTATION OF STATISTICAL TECHNIQUES

We use WEKA software to apply the listed statistical methods. WEKA is a collection of algorithms for data mining tasks. It can be run quickly for a big database. For more

detailed information about the Weka package, the reader is referred to Witten and Frank (2005).

Firstly, 70% of the data from the 6-year period (5 year before failure plus the failure year) will be used on WEKA for training. The remaining 30% (5400 observations) will be used in order to test the prediction accuracy of each model. For each bank, we then checked whether it is correctly classified in the right group (active or inactive) at the right time (how many years before being inactive).

For the k-NN method, we tested several values of k to determine the appropriate value. As mentioned in part 3, there are several hypotheses on selecting k, for example: k should be the square root of total number of observation (in this paper are 43) or k should be between 1 and 10. We then tested both values of k from 1 to 10 and 43 For ANNs method, we set the hidden layer is equal to 1 or 2. The default training time is 500 on WEKA.

3.3. DATA AND VARIABLES

Initially, we collected over 5000 banks from Bankscope database. However, we set the condition that the number of inactive and active banks should be equivalent to test the performance of the machine learning approach. Finally, therefore, we select randomly a sample of 3000 banks including 1438 inactive and 1562 active banks. 6 year-periods include: year when banks go bankrupt and 5 years before being inactive was selected. Active banks were selected randomly with the criterion of being a US bank and still active until the first quarter of 2016.

After collecting and importing data in panel, we shuffled the order. Shuffling data is important to prevent bias learning process and predict more intelligently and in an integrated way. We then, divide the dataset into 2 subsets: a Training set (70% of data) and a Test set (30% of data). Theoretically, each method will learn 70% first training to create significant models. The remaining 30% will examine accuracy.

From bank financial statements we extract or construct 31 ratios. Zhao et al. (2009) demonstrate that the use of financial ratios, instead of raw accounting variables, significantly improves the performance of prediction techniques. Detailed accounting information is taken to forecast the status of banks and provide more adequate points of view. Ratios were selected by comparison with the lists of ratios used in previous empirical studies. Before 2007, these lists were presented in detail in the review by Ravi Kumar and Ravi (2007), after 2007 we take into account the lists presented in the papers referred in the literature review section.

Finally, the selected ratios cover: (i) loan quality, (ii) capital quality, (iii) operation efficiency, (iv) profitability and (v) liquidity. To shorten the name, we label each ratio by Z from Z1 to Z31. Each of these financial ratios is expected to have a strong influence on bank performances as well as possibly helping to predict the failure. For 31 ratios, we have dissimilar expectation signs on the bank's survival. Positive signs suggest that the higher the ratio the better the influence on the bank's survival. Negative signs indicate the contrary (Table 3).

Table 3. Expected sign of ratios on bank's survival.

Variables	**Variables description**	**Expected sign**
	Loan quality	
Z1	Loan Loss reserve/Gross Loans	Negative
Z2	Loan Loss provision/Net interest revenue	Negative
Z3	Impaired Loans/Gross Loans	Negative
Z4	Net charge off/Average Gross Loans	Negative
Z5	Impaired Loans/Equity	Negative

Variables	Variables description	Expected sign
	Capital quality	
Z6	Tier 1 capital ratio	Positive
Z7	Total capital ratio	Positive
Z8	Equity/Total assets	Positive
Z9	Equity/Net Loans	Positive
Z10	Equity/Customer & short-term funding	Positive
Z11	Equity/Liabilities	Positive
Z12	Capital funds/Total assets	Positive
Z13	Capital funds/Net loans	Positive
Z14	Capital funds/Deposit & Short-term funding	Positive
Z15	Capital funds/Liabilities	Positive
	Operations	
Z16	Net interest margin	Positive
Z17	Net interest revenue/Average Assets	Positive
Z18	Other Operation income/Average Assets	Positive
Z19	Non-Interest expense/Average Assets	Negative

Variables	Variables description	Expected sign
Z20	Pre-tax Operating Income/Average Assets	Positive
Z21	Non-Operating Items & taxes/Average Assets	Negative
	Profitability	
Z22	Return on Average Assets	Positive
Z23	Return on Average Equity	Positive
Z24	Inc. Net of Dist/Average Equity	Positive
Z25	Cost to Income Ratio	Negative
Z26	Recurring Earning Power	Positive
	Liquidity	
Z27	Net Loans/Total Asset	Negative
Z28	Net loans/Deposit & Short term funding	Negative
Z29	Net Loans/Total Deposit and Borrowing	Negative
Z30	Liquid Assets/Deposit & Short term Funding	Positive
Z31	Liquid Assets/Total Deposit & Borrowing	Positive

Note: A positive sign indicates that when the ratio increases, the probability to fail decreases.

4. EMPIRICAL RESULTS

4.1. DESCRIPTIVE STATISTICS

Table 4 presents the means and standard deviations of the 31 selected financial ratios for active and failed banks groups for one year before becoming inactive. As in Canbas et al. (2005) the last two columns present the F-test for the equality of means among the two groups and the significance levels. We find that **25 of the 31 ratios** have a significant different mean (for failed and non-failed banks) at a level than 5%. Hence, the null hypothesis that the two-group means are equal is rejected at the 5% significance level for these ratios. We find that one year before failure, the loan quality is significantly lower for inactive banks (especially Z2 and Z4). Equity can be seen as a general buffer against risk and we observe that these banks have less equity whatever the measure of equity and the comparison point (assets, loans, liability) (see Z6, Z7, Z8, Z12, Z15). Operational efficiency is also lower for inactive banks compared with active banks (Z19, Z20). However, contrary to expectations, liquidity is higher for inactive banks (Z29, Z30, Z31), possibly because the banks in the sample became inactive for solvability problems rather than for liquidity problems. Note that our results are quite similar to those of Canbas et al. (2005). On a sample of 40 Turkish banks during the period 1997–2003, they find that the ratios that are the most different between failed and active banks are interest expenses on assets and interest income on interest expenses, equity/TA, liquid assets total assets, standard capital ratio. López and Pastor Sanz (2015) employed data from the FDIC between 2002 and 2012, their results state that failed banks are more concentrated in real estate loans and have more provisions.

Table 4 : Descriptive statics for the 31 financial ratios for active and failed banks – one year before being inactive.

Ratio	Inactive banks		Active banks		F	Sig.
	Mean	SD	Mean	SD		
Z1	1.60	0.944	1.45	0.913	20.058	0.000
Z2	13.45	24.780	4.09	8.065	200.135	0.000
Z3	2.03	3.015	2.00	2.115	0.115	0.735
Z4	0.53	1.077	0.19	0.654	110.766	0.000
Z5	14.88	24.632	12.55	14.892	10.005	0.002
Z6	13.17	4.952	15.14	6.743	81.813	0.000
Z7	14.53	4.827	16.40	6.713	75.745	0.000
Z8	10.09	3.213	11.10	3.398	70.150	0.000
Z9	16.61	8.880	18.29	13.153	16.502	0.000
Z10	12.26	4.964	13.41	7.119	25.630	0.000
Z11	11.40	4.295	12.71	5.188	56.090	0.000
Z12	10.37	3.221	11.48	3.313	85.988	0.000
Z13	17.08	9.010	18.90	13.129	19.327	0.000
Z14	12.63	5.148	13.87	7.186	29.358	0.000
Z15	11.73	4.326	13.14	5.132	66.062	0.000
Z16	3.95	1.088	3.76	1.436	16.040	0.000

Ratio	Inactive banks		Active banks		F	Sig.
	Mean	SD	Mean	SD		
Z17	3.52	0.967	3.35	1.213	17.886	0.000
Z18	1.00	1.190	1.12	1.474	6.184	0.013
Z19	3.52	1.810	3.18	1.605	30.231	0.000
Z20	1.00	1.663	1.30	0.998	38.089	0.000
Z21	−0.29	0.622	−0.24	0.572	4.716	0.030
Z22	0.67	1.238	0.98	0.780	67.328	0.000
Z23	7.28	11.764	8.83	6.525	20.391	0.000
Z24	1.88	11.626	5.27	5.961	103.154	0.000
Z25	69.85	32.613	67.83	14.711	4.909	0.027
Z26	1.46	1.416	1.48	1.260	0.081	0.777
Z27	65.16	13.438	65.97	13.338	2.763	0.097
Z28	78.41	20.623	78.06	17.342	0.263	0.608
Z29	73.60	15.244	75.59	15.583	12.394	0.000
Z30	9.72	10.426	7.71	8.077	35.414	0.000
Z31	9.24	9.080	7.50	7.843	31.809	0.000

Note: Table presents the means and standard deviations (SD) of the 31 ratios (Z1 to Z31) used to compare active and inactive banks. The F-test (F) is used for comparison of means. The p-value for the F-test (Sig.) is given in the last column.

4.2. COMPARISON OF ACCURACY

To analyse in detail the predictive performance of each method, we use several indicators (see Powers (2011) for more details on these measures). Precision is the fraction of those predicted positive by the model that are actually positive. Recall, also referred to as sensitivity, is the fraction of those that are actually positive which were predicted positive. F-measure is the harmonic mean of precision and sensitivity. The value of F-measure ranges from 0 to 1. A value of 1 indicates perfect prediction. MCC (Matthews Correlation Coefficient) is the measurement of the quality of binary classification. This indicator was first introduced by the biochemist Brian Matthews in 1975. The value of MCC is between [−1, 1]. MMC = 1 indicates a prefect prediction; MMC = 0 indicates that the prediction is not better than random prediction; MMC = −1 indicates disagreement between prediction and observation. ROC Area (Receiver Operation Characteristic curve) is usually used for a binary classifier with the value between [0, 1]. This curve is created with y-axis is true positive rate, and x-axis is false positive rate. The closer to 1 the values of ROC are, the better the prediction. PRC Area (Precision/Recall plots): this indicator is used less frequently than others. However, Saito and Rehmsmeier (2015) suggested that PRC is more informative than a ROC plot when evaluating binary classifiers. PRC plots evaluate the fraction of true positives among positive predictions and hence can provide an accurate prediction of future classification performance.

4.2.1. Choice of parameters

Firstly, we made a decision on choosing the number of hidden layer for ANNs and the number of neighbours for the k-NN method. Regarding ANN methods, we test whether the number of hidden layers is 1 or 2. The result in Table 5 shows that for 1 or 2 hidden layers, the difference is small. Overall, with 1 and 2 hidden layers, ANNs can recall 74.4% and 75% respectively and the precision ratio is 75.7% and 75.7% respectively. Consequently, we may conclude as in Lee et al. (2005) and Zhang et al. (1999) that using

1 hidden layer is sufficient. Finally, we will use the result from ANNs with 2 hidden layers to compare with other methods.

Table 5: The comparison of ANNs 1 hidden layer and 2 hidden layers.

Method	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ANNs_1	0.758	0.741	0.734	0.493	0.771	0.739
ANNs_2	0.757	0.753	0.75	0.506	0.819	0.803

Note: Table gives the accuracy measures for ANN with 1 hidden layer (ANNs_1) and two hidden layers (ANN_2). ***Precision*** *is the fraction of those predicted positive that are actually positive.* ***Recall*** *is the fraction of those that are actually positive which were predicted positive.* ***F-measure*** *is the harmonic mean of precision and sensitivity.* ***MCC****: Matthews correlation coefficient.* ***ROC area****: Receiver Operation Characteristic curve.* ***PRC Area****: Precision/Recall plot.*

For the k-NN method, we also implemented various values of *k* in order to try to find the 'best k'. The first assumption is that k should equal the square root of the total number of observations, which is 43. The second assumption is that k should be between 1 and 10. The first one brings only 72.9% precision, while the others obtained around 74% (see Table 6). We therefore state that the number of k-nearest neighbours should be between 1 and 10. In this case, we choose the k with the greatest precision which is k = 8 and denote it 8_NN.

Table 6: Comparison of k-NNs with different number of neighbours.

Method	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
1_NN	0.731	0.731	0.73	0.459	0.728	0.668
2_NN	0.74	0.712	0.698	0.442	0.774	0.719
3_NN	0.741	0.74	0.739	0.477	0.791	0.743

Method	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
4_NN	0.736	0.722	0.714	0.451	0.8	0.759
5_NN	0.736	0.736	0.734	0.469	0.804	0.768
6_NN	0.738	0.727	0.721	0.459	0.808	0.775
7_NN	0.741	0.739	0.737	0.476	0.81	0.78
8_NN	0.741	0.731	0.725	0.467	0.811	0.783
9_NN	0.741	0.739	0.737	0.476	0.81	0.783
10_NN	0.739	0.73	0.724	0.464	0.812	0.787
43_NN	0.729	0.724	0.72	0.448	0.806	0.798

Note: Table gives the accuracy measures of k-NNs with a number of neighbours from 1 (1_NN) to 43 (43_NN). ***Precision*** *is the fraction of those predicted positive that are actually positive.* ***Recall*** *is the fraction of those that are actually positive which were predicted positive.* ***F-measure*** *is the harmonic mean of precision and sensitivity.* ***MCC****: Matthews correlation coefficient.* ***ROC area****: Receiver Operation Characteristic curve.* ***PRC Area****: Precision/Recall plots.*

4.2.2. Comparison of the five bank failure prediction methods

Following Vasu and Ravi (2011) we decompose the accuracy ratio into two dimensions: false positive (Type I error, the classifier misclassifies an actual active bank as a failed bank, FP in Table 7) and false negative (Type II error, the classifier misclassifies a failed bank as an active bank, 1-TP rate in Table 7). Note that for banks false negative is considered by banking regulators to be far costlier than false positive. As foreseeable from the literature review and from the previous results, Logistic and ANN obtain the best performance with the lowest values of type 1 and type 2 errors for both ratios. Table 7 highlighted that ANNs performed better than all the other methods whatever the performance measure (75.3% of TP rate and 25.9% of FP rate, which lead to 75.7% precision and 75.3% recall). As can be seen, k-NN and LR achieved similar results: around 74% precision. SVM and LDA obtain the lowest performance with only 71.6% and 72% precision respectively (TP). The distance among these results is not too

significant, however we notice that the traditional logistic approach can predict more accurately than some of the machine learning approaches (as already observed by Zhao et al., 2009).

We then extract the result into years as in Table 8 to summarize the total errors of each method by year. The error here is defined when the active is classified as inactive and vice versa. ANNs make fewer errors than the others in the year that banks go inactive and make the most errors 3 years before. Meanwhile, other methods can make errors evenly over the years. As noticed from the previous comment, SVMs make the most mistakes for most of every year. Surprisingly, the maximum number of incorrect classification occurred at the year or one year before failure.

Table 7: Performance of bank failure prediction methods.

Method	**Confusion matrix**		**Precision**	**Recall**	**ROC Area**	**PRC Area**
ANNs_2	2415	444	75.7%	75.3%	81.9%	80.3%
	892	1649				
8_NN	2455	404	74.1%	73.1%	81.1%	78.3%
	1048	1493				
LDA	2185	674	72.0%	72.0%	77.6%	75.8%
	836	1705				
LR	2235	624	73.9%	73.9%	79.6%	77.3%
	785	1756				
SVM	2121	738	71.6%	71.6%	71.5%	65.5%
	794	1747				

Note: Table gives the accuracy measures for the five bank failure prediction techniques: ANN with two hidden layers (ANNs_2), k-NN with 8 neighbours (8_NN), Linear discriminant analysis (LDA), Logistic Regression (LR), Support Vector Machine, (SVM). ***Precision*** *is the fraction of those predicted positive that are actually positive.* ***Recall*** *is the fraction of those that are actually positive which were predicted positive.* ***ROC area****: Receiver Operation Characteristic curve.* ***PRC Area****: Precision/Recall plots.*

Table 8: Number of banks misclassified by year for the 5 predictions techniques.

Year	ANNs_2	KNN_8	LDA	Logistic	SVM	Total
0	199	252	285	262	295	1293
1	231	252	272	255	279	1290
2	220	258	248	212	225	1165
3	240	261	258	238	279	1279
4	234	213	218	221	219	1109
5	212	216	229	221	235	1118
Total	**1336**	**1452**	**1510**	**1409**	**1532**	

We investigate to find out which method can recognize the failure when the other methods predict wrongly (Table 9). This criterion is important and not used widely to date. Our purpose is to observe how dominant the method is. Surprisingly, ANNs can recognize 469 instances while the other methods cannot. After ANNs, k-NNs can also predict 92 observations while the other methods cannot. However, LDA can recognize only 3 banks, which is very poor.

Table 9: Right when other methods are wrong.

Methods	ANNs_2	8_NN	LDA	Logistic	SVM
Total	469	92	3	14	12
Year 0	73	24	1	3	2
Year 1	85	24	0	2	1
Year 2	94	10	0	3	1
Year 3	82	15	0	0	1
Year 4	67	9	1	4	5
Year 5	68	10	1	2	2

Note: Number of failed banks that one method can detect when all the others cannot.

5. CONCLUSION

This paper proposes an empirical study on the prediction of bank failure through 2 approaches: machine learning and two traditional statistical approaches. We observed firstly that machine learning, ANNs and k-NN methods perform more effectively than traditional methods. However, the difference in prediction accuracy between ANNs and k-NN methods and the traditional logistic regression method is not very big. In addition, we observed that SVM does not perform better than traditional methods. Nevertheless, ANN and k-nearest neighbour demonstrate their remarkable ability when they can detect the failure correctly, but the other methods cannot.

All 31 ratios are important to predict bank failures. Each group has at least one significant ratio that affects the survival of the banks. Among them, three groups play a more important role, namely operation efficiency, profitability and liquidity. Notably, the ratios Z3 (Impaired Loans/Gross Loans), Z6 (Tier 1 capital ratio), Z12 (Capital funds/Total assets), Z18 (Other Operation Income/Average Assets), Z17 (Net interest

revenue/Average Assets), Z21 (Non Operation Items&taxes/Average Assets), Z22 (Return on Average Assets), Z25 (Cost to income ratio) Z27 (Net Loans/Total Asset), Z28 (Net loans/Deposit &Short Term funding) and Z29 (Net Loans/Total Deposit& Borrowing) are more relevant than the others.

Our results have important institutional and policy implications. In effect, banks and bank supervisors developed early warning systems to prevent individual bank failure and banking crisis. Sahajwala and Van den Bergh (2000) provide an overview of the different approaches that are being used or developed in this field. Our study can help banks and bank supervisors to design such early warning systems because it shows that the traditional logistic regression models perform quite well, and they can be complemented by machine learning techniques (ANNs and k-NN) to detect the most difficult cases. Moreover, these methods are based on ratios analysis and our study provides some information on the financial ratios that could help to better predict bank failures.

The limitation of this study is that we emphasize accounting information and ignore bank market data. Moreover, we could not determine the role of each ratio in machine learning techniques.

6. REFERENCES

1. Adya, M., Collopy, F., 1998. How effective are neural networks at forecasting and prediction? A review and evaluation. Int. J. Forecast. 17 (5–6), 488–495.
2. Aktas, R., Doganay, M., Yildiz, B., 2003. Predicting the financial failure: a comparison of statistical methods and neural networks. Ankara Univ. J. SBF 58, 1–24.
3. Altman, E.I., 1968. Financial ratios: discriminant analysis and the prediction of corporate bankruptcy. J. Finance 23, 589–609.
4. Beaver, W.H., 1966. Financial ratios as predictors of failures. Emp. Res. Account. 4, 71–111.
5. Bell, T.B., 1997. Neural nets or logit model?: A comparison of each model's ability to predict commercial bank failures. Int. J. Intell. Syst. Account. Finance Manag. 6, 249–264.
6. Boser, B.E., Guyon, I., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Pittsburgh, ACM. Proceedings of the Fifth Annual Workshop of Computational Learning Theory 5. pp. 144–152.
7. Boyacioglu, M.A., Kara, Y., Baykan, O.K., 2009. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: a comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. Expert Syst. Appl. 36, 3355–3366.
8. Canbas, S., Cabuk, A., Kilic, S.B., 2005. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: the Turkish case. Eur. J. Oper. Res. 166, 528–546.
9. Chiaramonte, L., Croci, E., Poli, F., 2015. Should we trust the Z-score? Evidence from the european banking industry. Global Finance J. 28, 111–131.
10. Demyanyk, Y., Hasan, I., 2010. Financial crises and bank failures: a review of prediction methods. Omega 38, 315–324.

11. Fethi, M.D., Pasiouras, F., 2009. Assessing Bank Performance with Operational Research and Artificial Intelligence Techniques: A Survey. University of Bath School of Management (Working Paper Series).
12. Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT Press, Cambridge, MA.
13. Ky, T., 1991. Neural network models and the prediction of bank bankruptcy. Omega 19 (5), 429–445.
14. López, F.J., Pastor Sanz, I.I., 2015. Bankruptcy visualization and prediction using neural networks: a study of U.S. commercial banks. Expert Syst. Appl. 42, 2857–2869.
15. Lantz, B., 2013. Machine Learning with R. Packt Publishing Ltd.
16. Lee, S., Choi, W.S., 2013. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Syst. Appl. 40, 2941–2946.
17. Lee, K., Booth, D., Alam, P., 2005. A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. Expert Syst. Appl. 29, 1–16.
18. Lo, A.W., 1986. Logit versus discriminant analysis. a specification test and application to corporate bankruptcies. J. Econometr. 31 (2), 151–178.
19. Martínez, I., 1996. Forecasting company failure: neural approach versus discriminant analysis: an application to Spanish insurance companies. In: Sierra Molina, G.,
20. Bonsón Ponte, E. (Eds.), Intelligent Systems in Accounting and Finance, pp. 169–185 Huelva.
21. Martin, D., 1977. Early warning of bank failure: a logit regression approach. J. Bank. Finance 1 (3), 249–276.
22. Min, J.H., Lee, Y.C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst. Appl. 28, 603–614.
23. Mitchell, T.M., 1997. Machine Learning. McGraw-Hill.

24. Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. 18, 109–131.
25. Olmeda, I., Fernandez, E., 1997. Hybrid classifiers for financial multicriteria decision making: the case of bankruptcy prediction. Computational Econ. 10, 317–335.
26. Park, C.-S., Han, I., 2002. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. Expert Syst. Appl. 23 (3), 255–264.
27. Powers, D., 2011. Evaluation: from precision, recall and F-Measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2 (1), 37–63.
28. Ravi Kumar, P., Ravi, V., 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. Eur. J. Oper. Res. 180, 1–28.
29. Sahajwala, R., Van den Bergh, P., 2000. Supervisory Risk Assessment and Early Warning Systems. Basel Committee on Banking Supervision Working Papers. 53p.
30. Saito, T., Rehmsmeier, M., 2015. The precision – recall plots is more informative than the ROC Plot when evaluating binary classifiers on Imbalanced datasets. Plos
31. One J. 11/024, 1–22. http://dx.doi.org/10.1371/journal.pone.0118432.
32. Serrano-Cinca, C., Gutierrez-Nieto, B., 2011. Partial Least Square Discriminant Analysis (PLS-DA) for Bankruptcy Prediction. CEB Working Paper. (n° 11/024).
33. Sinkey, J.F., 1975. A multivariate analysis of the characteristics of problem banks. J. Finance 30, 21–36.
34. Swicegood, P., Clark, J.A., 2001. Off-site monitoring for predicting bank under performance: a comparison of neural networks, discriminant analysis and professional human judgement. Int. J. Intell. Syst. Account. Finance Manag. 10, 169–186.
35. Tam, K.Y., Kiang, M., 1992. Predicting bank failures: a neural network approach. Decis. Sci. 23, 926–947.

36. Ticknor, J., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Syst. Appl. 40 (14), 5501–5506.
37. Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.
38. Vasu, M., Ravi, V., 2011. Bankruptcy prediction in banks by principal component analysis threshold accepting trained wavelet neural network hybrid. In: Proceedings of the 7th International Conference on Data Mining. Las Vegas, July, 18–21.
39. Vellido, A., Lisboa, P., Vaughan, J., 1999. Neural networks in business: a survey of applications (1992–1998). Expert Syst. Appl. 17, 51–70.
40. Wang, G., Ma, J., Yang, S., 2015. Improved boosting based on feature selection for corporate bankruptcy prediction. Expert Syst. Appl. 41, 2353–2361.
41. Weiss, S.M., Indurkhya, N., 1998. Predictive Data Mining: A Pratical Guide. Morgan Kaufmann, San Francisco, CA.
42. Witten, I.H., Frank, E., 2005. Data Mining: Pratical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufamn, San Fransisco, CA.
43. Wu, C., Wang, X.M., 2000. A neural network approach for analyzing small business lending decisions. Rev. Quant. Finance Account. 15, 259–276.
44. Zhang, H., Hu, M.Y., Patuwo, B.E., Indro, D.C., 1999. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. Eur. J. Operational Res. 116, 16–32.
45. Zhao, H., Sinha, A.P., Ge, W., 2009. Effects of feature construction on classification performance: an empirical study in bank failure prediction. Expert Syst. Appl. 36 (2), 2633–2644.

CHAPTER 4

"In financial services, if you want to be the best in the industry, you first have to be the best in risk management and credit quality. It's the foundation for every other measure of success. There's almost no room for error"

- John Stumpf, chairman and CEO of Wells Fargo.

CHAPTER 4

WHY DO BANKS FAIL? - THE EXPLANATION FROM TEXT ANALYTICS TECHNIQUE

Jean-Laurent Viviani and Hanh-Hong LE

(This article is accepted for the 31st Australasian Finance & Banking Conference 2018)

ABSTRACT:

This study investigates the material loss review published by the Federal Deposit Insurance Corporation (FDIC) on the U.S. failed banks from 2008 to 2015. These reports focus on explaining the causes of failure and material loss of each bank. Unlike traditional methods that provide suggestions on financial ratios, our study focuses on phrases extracted from the reports by using text mining technique. Pre-processing steps are used in this study to 'clean' the text. Bag of words technique is used for collecting the most frequent words. Topic modelling and document hierarches clustering are used for classifying these reports into groups. Our results suggest that to prevent from being the failure, banks should significantly be aware of: loan, board management, the supervisory process, the concentration of ADC (Acquisition, Development and Construction) and CRE (Commercial real estate). In addition, we find the main reasons that US banks went failure from 2008 to 2015 are covered by two main topics: Loan and Management.

Keywords: text mining, US failed bank, BoW, k-means, topic modelling, hierarches clustering

JEL code: G01, G21, G28

1 INTRODUCTION

1.1 THE MAIN QUESTION AND CONTEXT

Ashcraft (2005) posed the question of whether a bank failure is important, and whether they have significant effects on real economic activities. The reasons behind these questions regarding the existence of deposit insurance, FDIC and the low dependent of US economy on banking system. This study has provided pieces of evidence that failed banks have significant and apparently permanent effects on real activity.

It is reported that the number of active commercial banks in the United States had declined roughly and continues to fall. As the report of FDIC, 2427 depository institutions have failed in the United States from 1986-2007. A bank is closed when it is unable to meet its obligations to depositors. In this case, as the definition of FDIC, the bank is announced as failure by a federal or state banking regulatory agency.

Numerous studies have been seeking for the reasons that cause banks to fail. Many Banks can fail for a numerous of reasons including undercapitalization, liquidity, safety and soundness and fraud. Causes of failure are mainly due to the deterioration of internal conditions which are the result of a bank's misguided policy for a number of previous years (Wheelock et al., 2000). They also suggested that less well capitalized banks are at greater risk of failure, as these banks are more likely to possess with higher ratios of loans to assets and evidence of poor-quality loan portfolios and banks with low earnings. Tussing (1967) presented that the main hazards faced by banks are typically illiquidity, bad assets, overbanking, and mismanagement.

Financial ratios analysis can provide meaningful quantitative information about the changes of internal conditions of the banks. Numerous studies in the literature used financial ratios as a tool to assess the bank's performance as well as to predict the failure of banks. Using this approach is able to bring some significant advantage aspects by calculating the probability of being failure and give some suggestions on investigating the financial ratios. However, it is important to note that these articles also face significant

limits on (i) determining reasonable ratios, especially on assessing the quality of management, (ii) pointing out the concrete problems in bank failure, (iii) as a 'special sector', the reasons that banks go bankruptcy cannot be fully described only through numbers or ratios. In past, there were some articles that analysing on bank failure without showing evidence of financial ratios (See Tussing (1967); Caprio et al. (1996), for example). This sort of analyse capture the scenarios for specific period.

Along with financial ratios analysis, ***text analysis*** should be considered in an earnest. In fact, the textual information is as important as numeric information. Textual information is easier to understand and help readers generate information.

For example, in the research of Wheelock (2000) on 'Why do banks disappear? The determinant of US bank failure and acquisition', they concluded that '*banks with little equity relative to assets are at significantly greater risk of failure than other banks'*. As a researcher in banking sectors, we may raise the in-depth questions, such as: *What are the reasons that cause ratio of equity to asset is little? Why don't banks' managers control this ratio?* etc. The answer for these questions cannot be fully explained by using couples of ratios, it is necessary to assess in the context and to look at on several aspects by extracting reports, news or reviews. Hence, text analysis is significant important and supplement financial indicators analysis.

The literature also presented remarkably the majority of research papers that analysed financial ratios. However, it is the fact that text information is also as important as qualitative information and people every day are working hard for understanding and extracting text information via journals, news, Twitter, Facebook, etc.

However, text data has some characteristics that are different in compared with numeric data: (1) the document can be very large (store larger in Megabyte in compare with file that contains numeric data), but it is composed by only few hundred words. For the corpus that contains numerous documents, it also may contain that certain amount of words. (2) Even the large dimensionality of a given corpus, the number of concepts is small as the words are typically correlated with one another, (3) The number of words (or non-zero entries) in the different documents may vary widely, hence, it is important

to normalize the document representations appropriately during the processing tasks, (4) The joining structure of words express different meaning.

Text mining, in short, is an artificial intelligence technique for solving problems can generalize the main idea in short time and calculate the correlation between words. Das (2014) defined *text mining as the large-scale, automated processing of plain text language in digital form to extract data that is converted into useful quantitative or qualitative information.* Text analytics are derived from automated text analysis and applied to digital texts using elements from natural language processing and machine learning such as latent semantic analysis, bags of words, or support vector machine.

1.2 SAMPLE CHARACTERISTICS

According to *speed-reading test* sponsored by Staples, an adult reads 300 words per minute on average. In general, there are 400-500words per page. Hence, to read a document of 20 pages, people need at least 30 minutes, and it takes even more time for understanding and analysing. As the volume of information available on the Internet continues to increase, there is growing need for tools helping people better find, filter, and manage these resources (Aas, 1999). Text mining has been successfully used in information exploring and retrieval. We use text mining for extract information from 98 US banks failure reports.

US Banking system plays a significant role in the global banking system, as the statistic of Global banking sector ranking based on the domestic, bank assets of US ranked 2 worldwide. As a leading economy, US financial system has many types of banks and there are several US banks goes failure each year, especially during and after the financial crisis. At the year of 1933, as seriously being affected of financial crisis, FDIC was created by the 1933 Banking Act. FDIC provided deposit insurance for more than 5500 financial institutions (2017), examines and supervises for safety and soundness.

Most of US banks are 'insured bank'. The term "Insured bank" indicates that bank is insured by Federal Deposit Insurance Corporation (FDIC). As a bank goes failure, the

FDIC acts in two capacities: Pays insurance to the depositors based on the insurance limit and, assumes the task of selling/collecting the assets of the failed bank and settling its debts.

When a bank goes failure, FDIC invests and publishes a material loss review to determine the causalities of failure of each bank. There are more than 30 pages of each report. Aside from determining the causes of loss, FDIC evaluates the supervision of failure bank that covered 10 years before failure announcement.

Information obtained from these reports is valuable and meaningful because of the detail assessment from FDIC. The investigation is frank and indicates the criteria which are considered as hard-to-measure such as board oversight, examination qualities, etc. For example, in the report AUD-15-007 that determines the causes of failure of Doral bank, FDIC states that the quality of an institution's management is not adequate by providing several reasons such as '*The board failed to ensure that policies were being properly implemented. For instance, the 2013 examination stated that Doral lacked proper Board oversight to ensure management fully resolved the repetitive weakness identified by regulators and auditors'*. This type of text that contains opinions is not easy to measure by financial ratios but can presented easily via news or reports. The statement from the report also helps the followers gain experience and more understanding the real other reason that leads to failure. Moreover, the reasons that banks go failure are varying; cover different sources of causes regarding governance, portfolio management, board oversight, earnings and so on. Hence, to generate the key reason that causes most banks go failure is challenging for researchers regarding summarizing and analysing terms. Using text mining technology can shorten the time and find out the most important popular key terms.

To the best of our knowledge, there are no similar recent studies in the field of bank failure analysis. Furthermore, we believe that this study will contribute to the literature of significantly enhances the knowledge regarding **(*1) explain the reasons that banks fail in the aspect of text analysis, (2) the supplementation of text analysis to financial ratios analysis.***

This article is organized as follows: Section 2 provides an overview of literature concerning failure recognized and textual representation techniques. Section 3 introduces our data corpus and the methodology that we utilized to extract the key terms. The main results are represented in section 4. Section 5 suggests some main remarks. The brief conclusion is in section 6.

2. LITERATURE REVIEWS

Text mining is applied popularly in the field of business management such as opinion mining and sentiment analysis (See more at Pang et al., 2008). This technique, however, has not been used widely in finance and banks' failure field.

Numerous finance research papers addresses on the ***financial indicators***. The financial indicators are arranged in a matrix of numbers. The researchers made a great effort to quantify non-numeric information into numeric format. For example, to measure the effect of governance, researchers need to consider one or some variables as the representative of 'governance' such as the gender of CEO, number of children that CEO has, the composition of the board of directors, etc. This approach brings the result with the equation measuring of parameters. Most of these kinds of algorithm requires statistical test, hypothesis or robustness check to assure that the method performs well. Some financial ratios are used widely such as CAMELS rating, coverage ratios, management quality (via ratios such as CEO duality, the percentage of independent directors, current ratio, ROE/ROA, etc.). Applying text mining to extract the most popular ratios, Ravi Kumar (2007) reported that among 128 given papers, most of the paper mentioned current ratios, quick ratio, income ratio, EBIT/total assets, ROA or ROE. These ratios are also considered as the **"core ratios"** that affect bank's performance.

In recent decades, the question on the value of non-numeric data is addressed. Text mining was introduced in the 1960s by document classification and became popular in 1990s. This method has found a variety of applications in diverse domains (Kumar, 2016).

Especially, in the decades of social media, big data nowadays, text mining has become a leading trend to analyse text content not only on Facebook, Twitter, blog or other social networks but also through news and reports (He, 2013). This information is potentially very valuable to decision makers, their partners, competitors and shareholders. It is believed that text in a context bears more diverse information than numbers do (Kloptchenko et al., 2004). However, in most of the previous researches using text mining, which is an artificial intelligence technique for solving problems, researchers mostly analysed text data based on word frequency calculated by morphologically analysed text. The extracted word might be lack of important information that was included in the original text, such as word-to-word dependencies and the contexts around high-frequency words.

In the field of finance, prolific work is reported in using text mining to solve problems such as predict FOREX rate, stock market or customer relationship management (Kumar, 2016). Data is collected from headline news, financial reports. However, in comparison with the number of finance research that based on financial ratios, the number of research based on text mining is the minority. Regarding FOREX rate prediction, the research suggested that based on the historical trend (Goodhart, 1989), news report (Fung et al., 2002), macro news (Evans et al., 2008) or even Twitter messages (Vu et al., 2012) might effect on Forex rate and help investor predict the movement of foreign currency. Besides, there are more papers on stock market prediction that used news headlines, annual report, or financial news from Bloomberg, Yahoo to foresee the trend of stock price (Back et al.,2001; Wang et al., 2011; Koppel et al., 2006; Mellouli, 2010; Nassirtoussi et al, 2015, Wang, 2008; Wang,2011).

One common process of basic text mining is: Firstly, collect the data by acquiring articles, news or reports from the internet. Secondly, extract and retrieve the given data by reporting the frequency of most common vocabularies as a baseline of the framework

(Mirończuk et al., 2018). Finally, calculate the correlation among words. *However, this approach is more about statistical than giving the true meaning of the text*.

To maximize the efficiency of the use of text mining, document analysis is introduced. The most two popular tools are Document clustering and topic modelling. These are two closely related tasks which can mutually benefit each other (Xie et al., 2013).

Topic modelling is one of the most powerful techniques in text mining and gaining attention from researchers. Blei (2009) proposed ***topic modelling***, by discovering patterns of word use and connecting documents that exhibit similar patterns, topic models have emerged as a powerful new technique for finding useful structure in an otherwise unstructured collection. This technique is applied in various fields such as customer analysis, political science, etc. A topic contains a cluster of words that frequently occurs ensemble and can connect words with similar meanings and distinguish between uses of words with multiple meanings (Alghamdi et al., 2015). There is various type of topic model's algorithms such as Latent Sematic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Correlated topic model (CTM), Latent Dirichlet Allocation (LDA). Among them, Latent Dirichlet allocation (LDA), an algorithm based on statistical (Bayesian) topic models, is one of the most popular tools in this topic modelling and be considered as a standard tool. Various studies in the field of social network (McCallum et al., 2005; Wang et al.2013; Yu et al., 2015; Kim, 2014, Cohen, 2014), Political science (Cohen, 2013), Linguistic science (Bauer et al.,2015) used LDA. In the field of finance, the dataset is mainly used for financial news and financial report (Kumar et al., 2016).

Document clustering is a method for automatic cluster textual documents. There are many algorithms for document clustering like K-means algorithm (Hartigan et al., 1979) and hierarchical clustering (Jain et al., 1988). This algorithm is widely applied in many field such as document organization, browsing, summarization or classification (Lu, 2011).

It is obvious that text mining is becoming more popular and draw special attention from researchers. Using text mining is not only for saving time from reading the thousands of documents but also help researchers have a general idea effectively. The text and data mining are now considered complementary techniques required for efficient business management, text mining tools are becoming even more important.

3. DATA AND METHODOLOGY

3.1. DATA

The corpus contains 98 official reports of bank failure collect from the website of FDIC (https://www.fdicoig.gov/reports-bank-failures) from 2009 to 2015. 69 over 98 banks go failure as the result of Global Financial crisis 2008. The detail description of the corpus is in table 10. These reports are announced by the Federal Deposit Insurance Corporation- Office of Inspector General (FDIC OIG), an independent office that conducts audits, evaluations, investigations and other reviews of FDIC to prevent, deter and detect waste, fraud, abuse and misconduct in FDIC programs and operation, and to promote efficiency and effectiveness at the agency. These audits are subject to: Determine the causes of the financial institution's failure and resulting material loss to the DIF and evaluate the FDIC's supervision of the institution, including implementation of the PCA provisions. The report provides both numeric and text information on each bank. At some other aspects, textual part of material loss review contains richer information than the financial ratios.

The structure of each report is Causes of failure and Material loss, and the FDIC's supervision. Each reason is analysed in detail paragraph. For example, the report of failure of The Bank of Union, El Reno, Oklahoma 2014: *the CEO occasionally presented information to the Board about certain borrowing relationships and the bank's overall lending strategy, but in some cases, subsequent management actions would deviate from the materials presented.*

Table 10: Descriptive on bank's failure reports

Year	Number of banks
2009	22
2010	47
2011	19
2012	5
2014	1
2015	4
Total	**98**

3.2. METHODOLOGY

Unlike the traditional financial ratio analysis that numbers are organized as a structured matrix, the primary challenge in applying Text Mining is to investigate the unstructured format of data. **Text mining** applies mostly similar techniques than data mining; the different is to deal with the corpus of textual data (Dörre, et al., 1999). Basically, ***Pre-processing step*** to structure the data is important. The corpus is converted into the **Document-term matrix** after removing stop words, stemming, punctuation, number and strip whitespace as proposed by Dillon (1983).

3.2.1. Pre-process and Bag-of-words (BoW)

BoW model is purely based on raw documents. In this model, features are extracted from text as a bag of words, discarding order, grammar or the important role of each word. The BoW is the frequency used when the occurrence of the word is considered as a feature. However, it is insufficient to capture all semantics.

Meyer (2013) introduced the text mining infrastructure in R with **tm** package which provides a framework for text mining applications. This package presents methods for data import, corpus handling, pre-processing, data management, and creation of term-document matrices.

For example; below is a paragraph in the song "If I Had a Million Dollars" by Barenaked Ladies:

"*If I had a million dollars, I'd build a tree fort in our yard*

If I had a million dollars, you could help it would not be that hard

If I had a million dollars, maybe we could put a refrigerator in there"

We can make a list of all the words in our vocabulary as following:

- If
- I
- Had
- A
- Million
- Dollars
- Build
- Tree
- fort
- In
- Yard
- Could
- Help
- Would
- That
- Hard
- Maybe
- Put
- Refrigerator
- There
- Not
- Our

We can easily recognize that there are some words that do not bring valuable meaning in the context, like 'A, in, that, there, our'. These words should be removed from the feature to reduce noises. The algorithm considers the occurrence of appearance of each word and put all in a 'bag' discarding the order, grammar: "If, I, Had, Million, Dollars, etc.".

3.2.2. Topic modelling via Latent Dirichlet Allocation

We hypothesise that a bank went failure for given reasons, and the other banks may face similar reasons. Hence, we apply document classification tools in the effort of categorizing reports into groups.

Topic modelling indicates the latent semantics in the document corpus and identifies document groups, which is more useful than raw term features. Latent Dirichlet Allocation (Blei, 2003) provides an approach to modelling text corpora. This is one of the most popular algorithms for topic modelling. Without diving into math equations behind the model, we can understand it as being conceptualized by two principles: LDA estimates the ***mixture of words*** that compose a ***topic*** and determine the topics that describe each ***document***.

Mathematically, LDA calculates the based on a conditional distribution. A corpus **D** contains **d documents** distributed into **T topics including z_t** single latent topic. Each of **z_t topics** composed by each word **w_t.** LDA assumes the following generative process for each document:

1. Choose N ~ Poisson(ξ).

2. Choose θ ~ Dir(α).

3. For each of the **w** words **w_t**:

(a) Choose a topic **z_t** ~ Multinomial(θ)

(b) Choose a word **w_t** from p(**w_t** | **z_t**,β), a multinomial probability conditioned on the topic **z_t**.

In which, θ is a multinomial distribution over topics for the document, N is the number of words in the document

For brief explanation, we pick 3 sentences from 3 reports (documents):

Sentence 1 (Report of Doral Bank, San Juan, Puerto Rico): *Poor asset quality was the underlying cause of Doral's failure*

Sentence 2 (Report of Vantage Point Bank, Horsham, Pennsylvania): *Board and management did not effectively manage the risks associated with the bank's rapid expansion of its mortgage*

banking operation.

Sentence 3 (Valley Bank, Moline, Illinois): *VBI failed primarily because of lax oversight by its Board and a dominant CEO that implemented a risky business strategy.*

As the goal of LDA is to automatically discovers topics that these document (in this example is sentence) contain. Given these sentences, LDA might classify into 2 topics: G (for Governance) and R (for Risk).

Topic G contains: Board, management, strategy, business, oversight, CEO

Topic R contains: Poor, asset, risk, mortgage, operation, lax, expansion

LDA will then spread of each sentence by a word count:

Sentence 1: 100% Topic R

Sentence 2: 42% Topic G and 58% Topic R

Sentence 3: 32% Topic G and 68% Topic R

A similar process is used for entire documents. LDA will automatically classify each document into given topics. The crucial point is to assign the number of topics.

However, selecting the appropriate value of number of topics is essential. The literature has not find out a consistent answer on the standard number of topics and has remained an open-end question (Arun. 2010).

In the experiment of Arun et al (2010), for example, the number of optimal topics is varying as presented in table 11. In this study, the authors test with different corpus, however the suggestion on the number of topics is not consistent.

Table 11:The experiments of Arun et al. (2010)

Dataset	**Number of Documents**	**Number of topics**
Toy data set	12	3
Authorship data set	834	15 to 25
NIPS Dataset	1500	100-120
Associated Press	2246	140

The literature suggested 4 algorithms that estimate the optimized number of topics: Griffiths (2004), Cao et al. (2009), Kumar (2010) and Deveaud (2014).

Griffiths et al. (2004) suggested that the number of topics that maximizes the harmonic mean of the sampled log-likelihoods. Deveaud et al. (2014) maximize the average Jensen Shannon distance between all pairs of topic distributions. Cao et al. (2009) estimate the average cosine similarity between topic distributions and chooses the value of that minimizes this quantity. Kumar et al. (2010) minimize the symmetric Kullback Liebler divergence between the singular values of the matrix representing word probabilities for each topic and the topic distribution within the corpus. While there are several methods of determining the optimal number of topics empirically in the literature, a rigorous treatment of their effectiveness is lacking.

3.2.3. Document clustering

Document clustering is a powerful technique for large-scale topic discovery from text (Larsen et al.,1999). This technique has not been used widely in finance sector, however, is used in law, web page analysis. (Ramage et al. 2009, Wong et al. 2002)

The goal of document clustering is to assign documents to different topics. (Aggarwal et al., 2012; Lu et al., 2011; Xu et al., 2004) categorize documents with common features into groups. By grouping, document clustering crucial for document organization, browsing, summarization, classification and retrieval. There are two common algorithms: using the ***hierarchical*** based algorithm and using the ***K-means*** algorithm and its variants.

The first algorithm, the hierarchical clustering, includes single link, complete linkage, group average and Ward's method. By aggregating or dividing, documents can be clustered into the hierarchical structure, which is suitable for browsing. However, this algorithm may suffer from efficiency problems.

The other algorithm is developed using the K-means algorithm and its variants. Generally, hierarchical algorithms produce more in-depth information for detailed analyses, while algorithms based on variants of the K-means algorithm are more efficient and provide sufficient information for most purposes (Qin et al. 2017).

We experiment both algorithms for document clustering. Noted that, K-mean requires us to specify the number of groups before classification, meanwhile hierarchical does not- which means any number of clusters may be picked at any level of the tree.

The brief approach of the hierarchical clustering is: (1) Place each data point (document) into its own cluster, then (2) Identify the closet 2 clusters and combine them into one cluster, and (3) Repeat step (2) till all documents are in merged into a single cluster. HC is typically visualized as a dendrogram.

Meanwhile, the goal of K-means clustering is to find groups in the corpus with the number of groups defined by the given variable K. For this approach, it is important to define the number of topics and iteratively redistribute the documents into topics until

some termination condition is set. The disadvantage in k-means is the accuracy and efficiency depend on the choice of initial clustering centre.

4. DESIGN OF THE EMPIRICAL MODEL

4.1. FEATURES SELECTION

Figure 2 presents the essential steps in mining documents.
The first step after collecting documents is to transform documents into statements appropriate for text algorithms and the mining tasks.

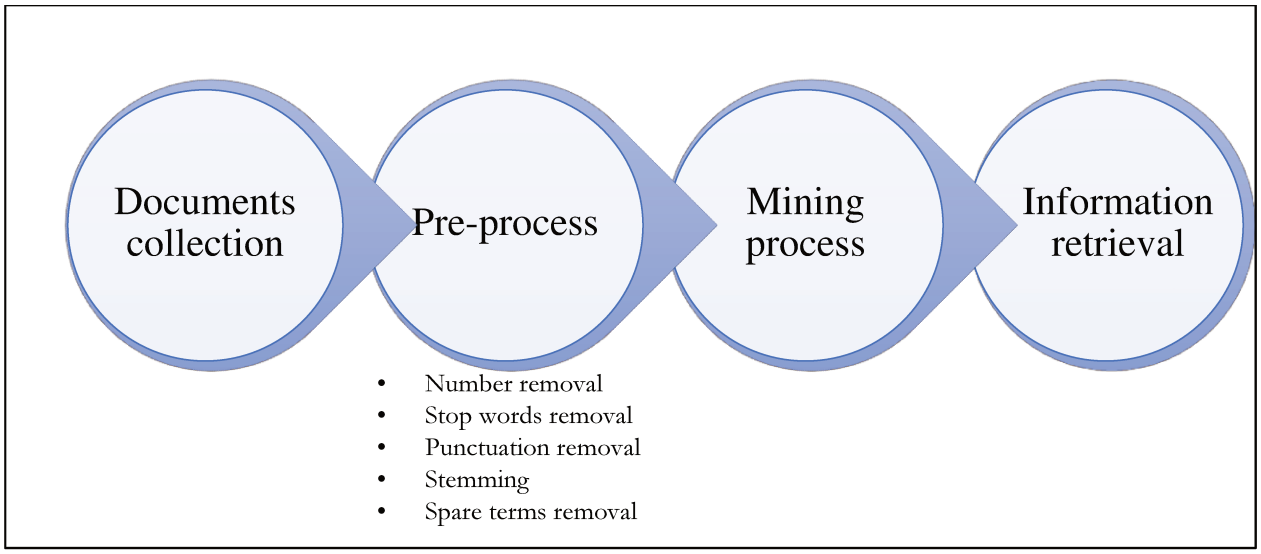
The reports are in the format of the .pdf file, we converted into text form and 'cleaned' before processing. The quality of text mining method is highly dependent on the noisiness of the features. For instance, commonly used words such as 'the', 'for', 'of', may not be very useful in improving the algorithm; hence, it is critical to select the feature effectively to remove the noisy words in the corpus. These are steps that we used for feature selection:

- *Remove number*: In this research, we focus on investigating the text information; the number in each report will be removed.
- *Remove stop words*: A list of stop word is provided in the package of 'stop words' in R-software. The list included 175 words that are frequently occurring but transmit no significant meaning, such as *I, our, his, was, is, are, will, etc.* The recurrent appearance of these words may interfere with the analysis process, we hence, remove words belongs to this list. Moreover, we also create and remove our own stop-words list, such as *bank, fdic, also, because, the, etc.*
- *Stem words*: there are words that have the similar meaning but have different word form, such as *banks and banking, institution and institutions, managing and manager or management, etc.* We convert different word form into similar canonical form. For example, *failure and fail or failing in to fail, examination and exams or examine into exam*,

etc. This process reduces the data redundancy and simplifies the later computation.

- *Remove punctuation:* All punctuations are removed from the text. The purpose of this step is to make the statements appropriate for text algorithms
- *Remove spare terms:* We remove the spare terms that appear in only one report.

Figure 2: Text-mining steps



4.2. MODEL DESIGNED

- Bag-of-words: This step is simply to count the number of times that words are mentioned in each document of the corpus. We then sum to obtain the total. We hypothesise that the more repeated occurring of the words, the more important of the words to the reports.

- Correlation analysis: Correlation among words is measured in a binary form - either the words show up together or they do not. A common measure for binary correlation is the phi coefficient. Table 12 presents the matrix of words X and Y combination.

Table 12.Words correlation

	Word Y	**No word Y**	**Total**
Word X	N_{11}	N_{10}	$N_{1.}$
No word X	N_{01}	N_{00}	$N_{0.}$
TOTAL	$N_{.1}$	$N_{.0}$	N

In which:

N_{11}: the number of documents where both word X and word Y appear

N_{10} and N_{01}: where one appears without the other

N_{00}: the number where neither appears

In terms of this table, the phi coefficient is:

$$\emptyset = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{1.}N_{0.}N_{.0}N_{.1}}}$$

The high value of $\emptyset$ suggests the high correlation between words X and Y.
The literature suggested that counting the number of appearance times does not bring high value for analysing. Finding phrases via words correlation is a progression for text mining technique.

- Topic modelling: We classify reports into topics. We make a hypothesis that among 98 failed banks, there are main topics which can be considered as main reasons. Each topic is composed of weighed words. Grouping helps information retrieval process bring the higher value. LDA and document clustering techniques are applied to classify the reports into sub-groups.

- Document clustering with K-means and hierarches: partition reports into groups.

5. RESULTS

5.1. DESCRIPTIVE STATISTIC

Table 13 introduces **30 words that appear most frequently** in the corpus. These words are set in stem forms. Majority of these words describe important bank's activities such as Loan, Deposit, Credit, Insurance. These words focus on **Loan issues**: Loan, Loss, review, ADC, CRE, ALLL[2], Lend, Estate. Another important issue is about **governance**: management, report, supervisory, board, exam. Figure 2 describes in chart top 15 frequent words. The repetition of words "Exam" and "Loan" (nearly 15000 times in 98 reports) is significantly higher than others. From financial analysis perspective, most of these words are considered as sensitive and representative for bank's failure analysis.

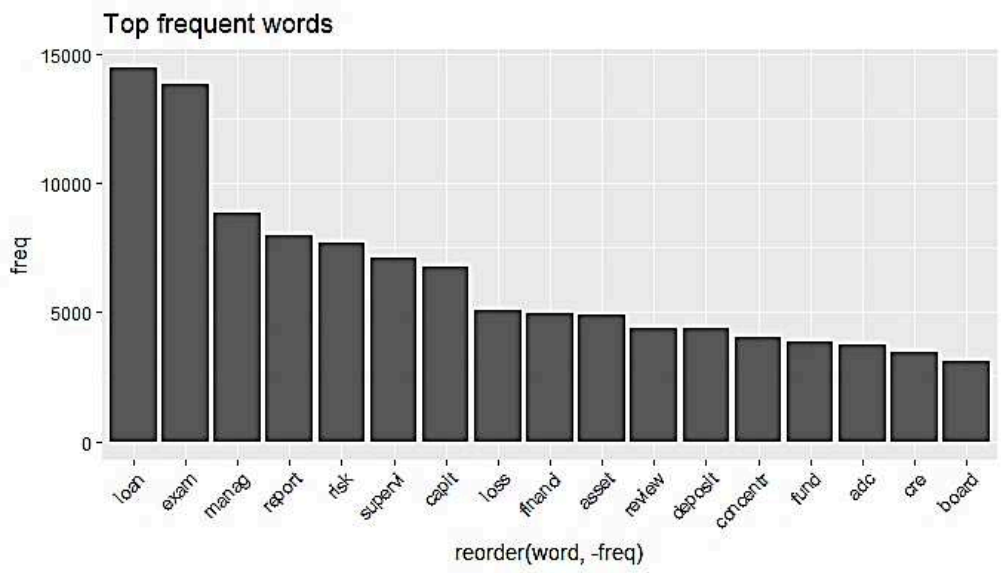
Table 13: The 30 most frequent words

Word	Count	Word	Count	Word	Count
loan	14492	asset	4908	portfolio	2482
exam	13847	review	4390	growth	2452
manag	8885	deposit	4378	audit	2406
report	7961	fund	3885	perform	2400
risk	7706	adc	3767	increas	2399
supervi	7135	cre	3481	level	2052
capit	6788	board	3117	lend	2050

[2] See Appendix 1 for full written forms

Word	Count	Word	Count	Word	Count
concentr	5765	plan	2992	market	1978
loss	5061	credit	2948	alll	1958
financi	4994	signific	2804	portfolio	2482

Figure 3: Top 15 frequent words.



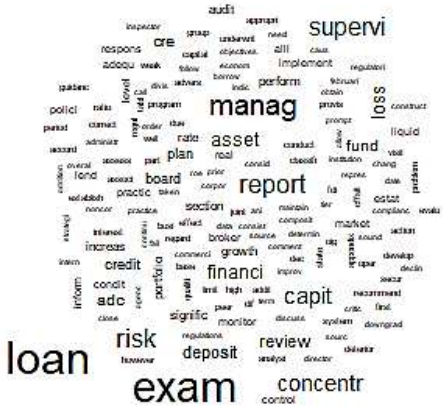


Figure 4: Word cloud

At first glance, the statistic presents 'not-surprising' words such as Loan, exam (or examination), management, risk, report. These words are always considered as **'core reasons'** of bank's failure. In the history of researching about the bank, these reasons can be found regularly (Alam et al.,2000; Bell, 1997; Haslem et al., 1992; Kolari et al 2002; Martin, 1977)

However, when going further, there are words that are significantly important and remarkable: **Concentr (or concentration), Adc (or Acquisition, Development and Construction), Cre (Commercial real estate), board**.

By looking at this list, the readers can have a global scenario of these banks during this period. **Figure 2** shows the order of frequent word. **Figure 3** presents in word-cloud all the words that appear more than 800 times in the corpus. **Figure 4** also presents in the word-cloud of the top frequent words. For figure 3 and figure 4, the bigger size of the word shows the more frequency.

This step suggested general ideas about bank's failure. The second step analyses the correlation of words bring more profound result.

5.2. THE WORDS CORRELATION MATRIX

Counting number of words, however, cannot reflect fully the picture of the context. We apply in R software to find out the correlation matrix among words. 20 most correlated words among most 50 frequent words are visualized as figure 5. The correlation matrix suggested the connection between words.

Table 14 presents the correlation coefficient according to Cohen (1988). The correlation is between 0.10 and 0.29 are "small", those greater than 0.30 and smaller than 0.49 are "medium" and those greater than 0.50 are "large" in terms of the magnitude of effect sizes. We hence, follow Cohen (1988) find the words that their correlation at the minimum as *medium (Correlation must be greater than 0.3)*

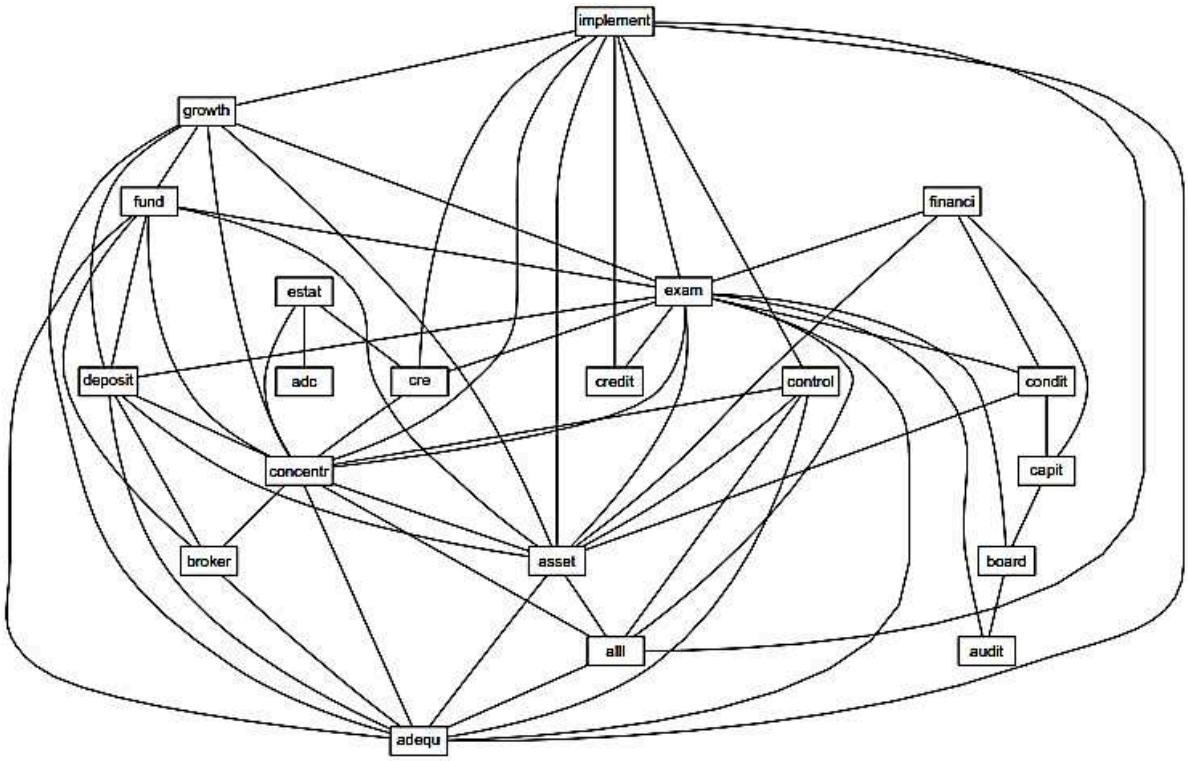
Table 14: Cohen (1988) correlation coefficient

Interpretation	**Correlation**
Small	0.10 -0.29
Medium	0.30-0.39
Large	0.50-1.00

Figure 5 shows the correlation matrix of words. The linking is intersection and complicated. The matrix is created by the important words, which is considered as "core nodes" that most of the other words must 'cross' them. These "core nodes" are significantly important as they are (i) in the most frequent words list and (ii) are considered as dominant factors that connect and control others. The **'core nodes'** are**: Exam, concentr, implement, asset, adque**. Via 'Core nodes", we can generate meaningful phrases, such as: "increase loan loss", "credit loss insurance", "implement credit exam", "growth concentr estate adc", "implement control asset concentr growth", 'implement control Alll", etc.

Compare with the simple descriptive statistic, this step brings the more extensive picture of what has happened for bank failure during 2008 to 2015.

Figure 5: Correlation matrix created by words



5.3. TOPIC MODELLING

Latent Dirichlet Allocation (LDA), a generative model for documents in which each document is viewed as a mixture of topics and each topic is a composition of words. The number of topics is crucial to the performance; however, finding the appropriate value for it is a challenge (Cao 2009). To find the suitable number of latent topics in a given corpus has remained an open-ended question We assume that there will be no less than 2 reports per topic. For 98 given report, the range of number of topics is from 1 to 50 topics.

Figure 6: Optimal topics suggested by Griffths 2004, Cao 2009, Arun 2010 and Deveaud 2014.

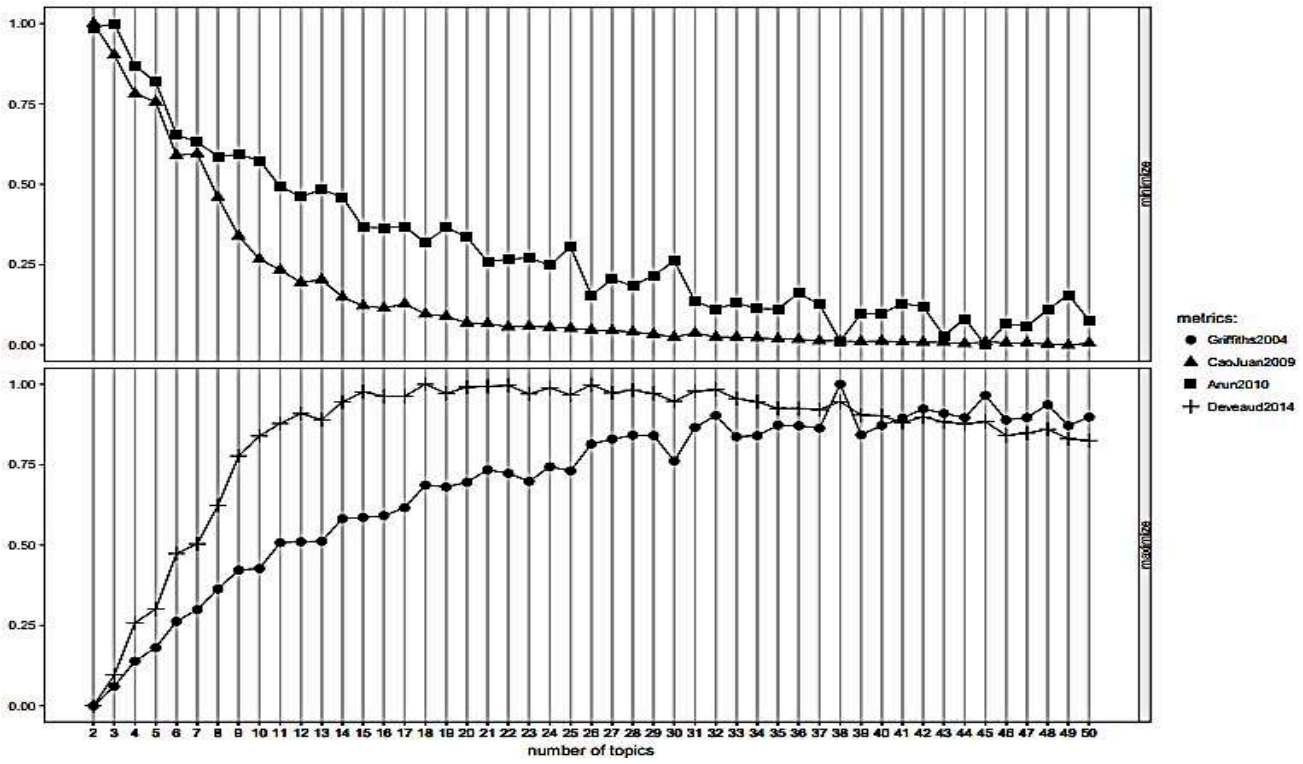


Figure 6 suggests the number of optimal topics based on the algorithm of Griffths 2004, Cao 2009, Arun 2010 and Deveaud 2014. As the Deveaud's algorithm, we should categorize into 18 topics. Griffths and Juan proposed 38 topics. However, Juan's algorithms indicate the optimal number of topics is 50. The question of 'How many topics for text classification' is still an ongoing question and the answer typically depends on the characteristics of each corpus. Hence, we then, experiment from 1 to 50 topics. Our result indicates that the optimal number of topics for **this corpus is 2**. As the numbers of topic increases, the distinction among topics becomes unclear. Figure 8 is an example of the classification of 3 topics: The words are similar in all 3 topics, the only one different is the weight of each word.

Figure 7 shows 2 topics of the given corpus. These 2 topics included some common words: **loan, exam, concentr, risk**. These words are also included in **"core nodes"** of the correlation matrix. Topic 1 focus on loan related issues and the other focus on management related issues. There are 65 banks belonging to Topic 1 and 33 banks belong to topic 2.

Figure 7: 2 topics by words distribution

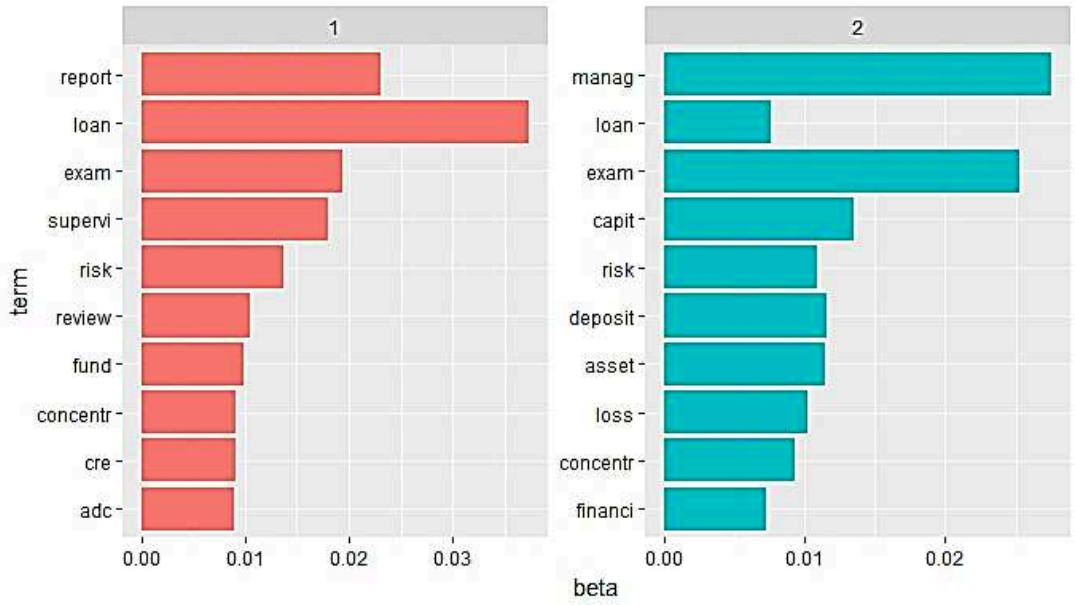
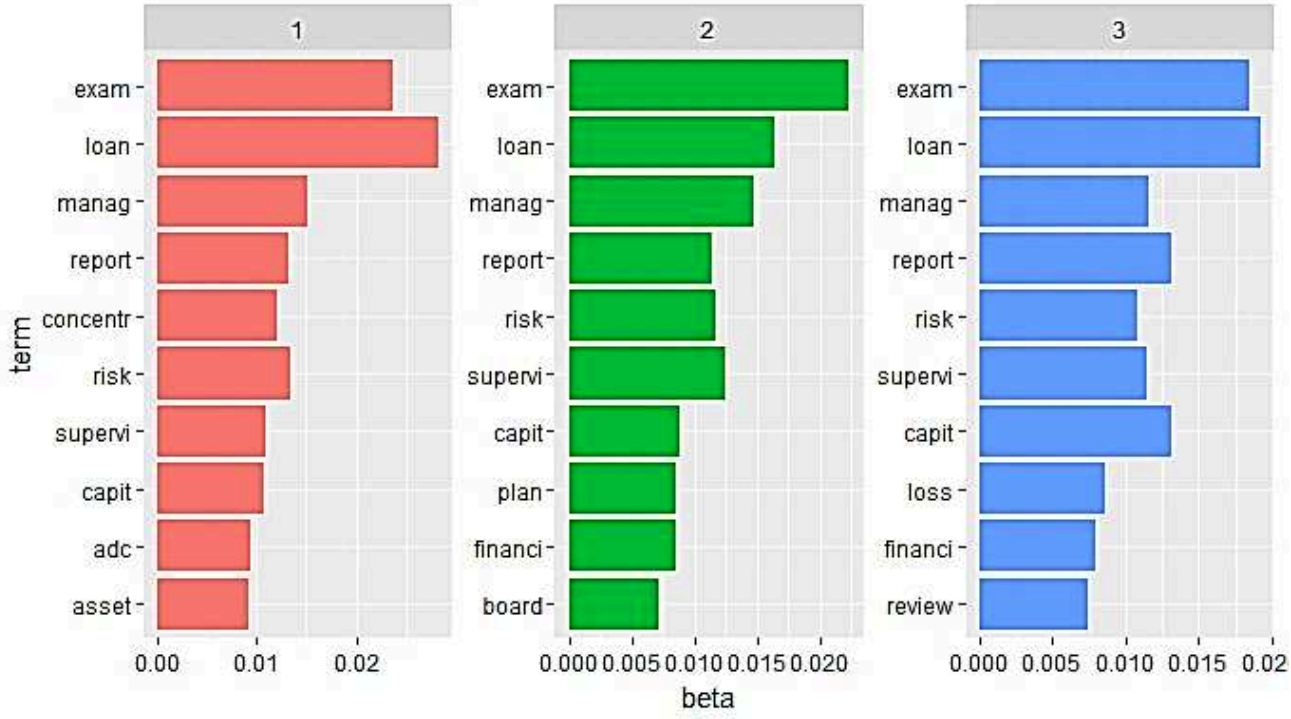


Figure 8: 3 topics by words distribution



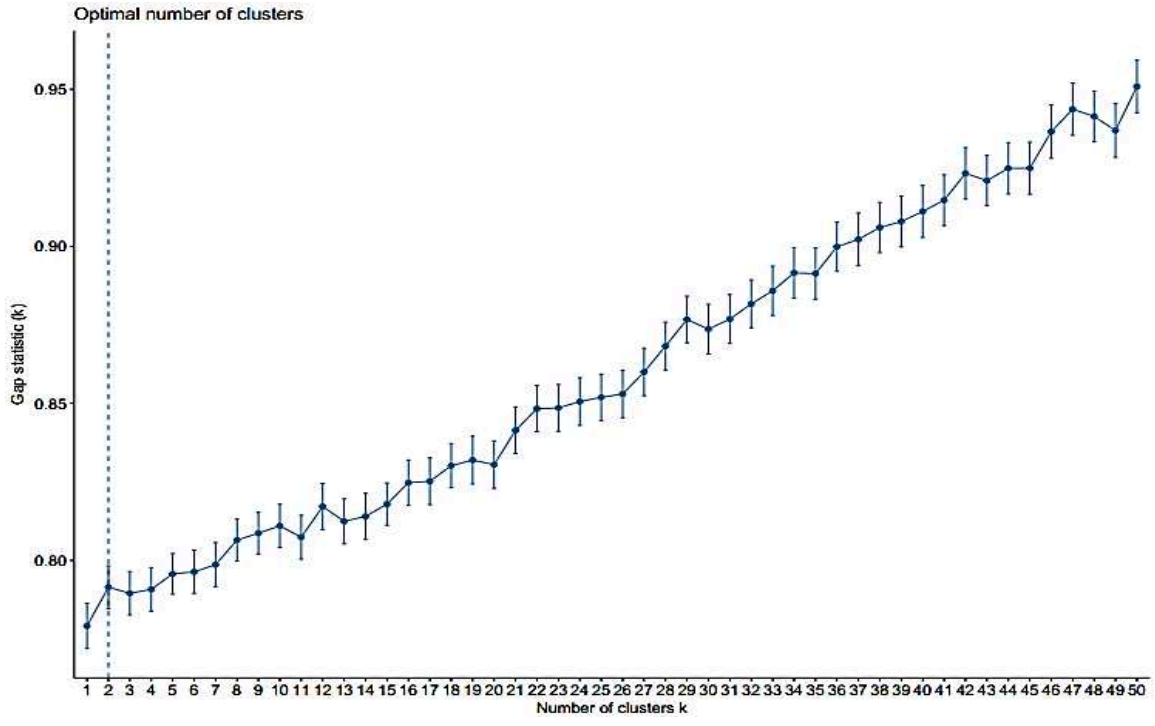
5.4. DOCUMENT CLUSTERING

- **The k-means algorithm** is applied to find out the optimal number of topics by documents clustering. The calculation is based on Euclidean methods. With *p* and *q* are two random points, each of them has *n* features. The distance between p and q can be calculated as:

$$\text{Dist (p,q)}=\sqrt{(p_1 - q_1)^2 + \ ... + (p_n - q_n)^2}$$

By applying package 'factoextra' (Alboukadel Kassambara and Fabian Mundt (2017)) in R, the result suggests that for 98 documents, it should be divided into **2 groups** to optimize the clustering. (Figure 9)

Figure 9: K-means algorithm for optimal topics



- Hierarches clustering: One of the advantages of Hierarches clustering is to specify the number of topics at any level. The result (Figure 10) suggested that we can cluster into **2 groups** at the highest level of tree. This classification is consistent with K-means

algorithm and topic modelling with LDA. The dendrogram reports that: 18 documents are placed in the first group, 80 for the second group.

Figure 10: Hierarches clustering result

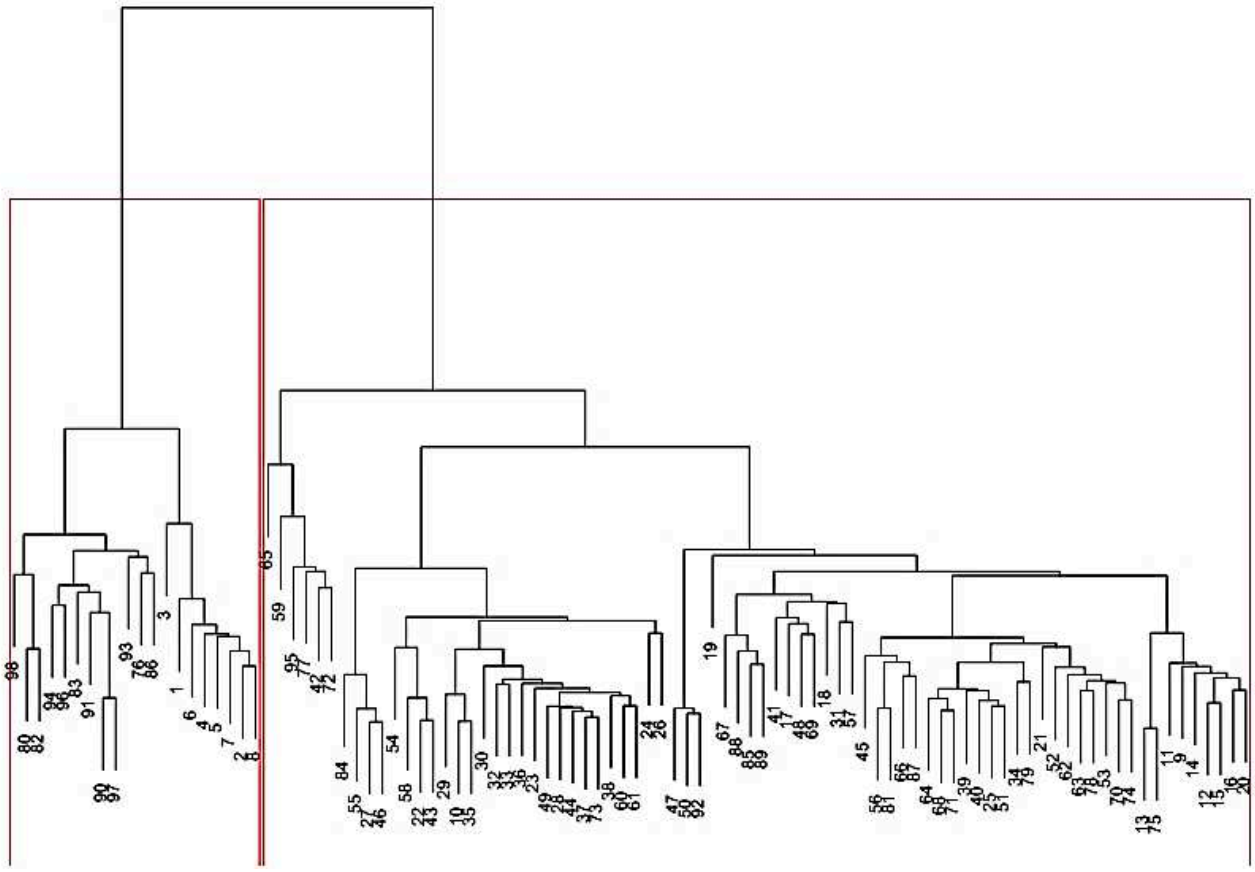


Table 15 presents the most common frequent vocabularies of 2 sub-groups. This list is corresponding to the list in table 13.

Table 15: The common frequent words of sub-groups

The common frequent vocabularies of 2 sub-group

loan	loss
exam	deposit
manag	review
report	concentr
risk	fund
supervi	cre
capit	plan
financi	board
asset	credit

There are words in top 20 common words that appear in the 'bigger' group but does not appear in the second group is *adc (Acquisition, Development and Construction), alll (allowance for loan and lease Losses), liquid, policy, broker*. These 'sensitive terms' imply the factors that distinguish 2 groups.

6. CONCLUSION

As the important role of banking system in economics, studying on bank's failure has become a topic of interest. By suggesting issues that banks must beware of, text analytics can be a complementary action for profound bank's financial analysis. It makes possible that the text analytics has captured a global tendency to foresee the features before they injure bank's financial condition.

It is noteworthy that ADC and CRE are mentioned significantly. Under Basel III, CRE (Commercial real estate loan) is a mortgage loan secured by a lien on commercial, rather

than residential, property. This type of loan is typically made to business entities formed for the specific purpose of owning commercial real estates. ADC (Acquisition, Development and Construction) loan, considered as the riskiest type of commercial real estate (CRE) lending, is a loan which allows the borrower to purchase real property (such as land), put in the necessary infrastructure and then build stores or other buildings. This type of loans is often used by developers of large properties such as strip malls or shopping centre. As our best of knowledge, rarely ADC and CRE are criticized as the reasons of bank's failure. One of the reason is the difficulty on obtaining the numeric data of ADC and CRE due to complication in their calculation.

Aside from loan problems, it is important to note that board and management is extremely important. As mentioned in *The Directors' book* released by the OCC, "A bank's board of directors is ultimately responsible for the conduct of the banks' affairs. The board controls the bank's direction and determines how the bank will go about its business. A board must be strong, independent, and actively involved in the bank's affairs". However, as the report of OCC, nearly 60% of failed banks had directorates that either lacked necessary banking knowledge or were uninformed or passive in the supervision of the bank's affairs. This report was released in the year of 1988, however, as can be seen from this study, for the database from 2007 to 2016, the similar problem is still addressed.

We have demonstrated a Bag-of-words technique, a statistical inference algorithm for LDA, topic modelling and document clustering for analysing 98 banks' material loss reviews. Our research contributed by using text analytics on 4 major aspects:

- **Core words:** The result suggested that there are some core words, which are considered as some main reasons causes the bank's failure, appear in most of the reports. We classify them into 4 groups:

Loan: Loan, ADC, CRE, credit, rate, ALLL

Management: Exam, management, report, supervise, review, board, audit

Capital: Capital, deposit, asset, fund, portfolio

Magnitude: Increase, Significant, growth, concentration

The given words are significantly sensitive to the banking system. Our result is comparable to financial ratios aspects. Moreover, it is noteworthy that there are some terms that are hard-to-measure and therefore have not been mentioned as a reason in the literature on bankruptcy, but that have a significant influence on banks' survival: Management, supervision, concentration on ADC or CRE.

- Core nodes: *Exam, Concentration, asset, implement, adequate*. We do suggest that the bank must increase the supervisory process, seriously pay attention to the allocation of loans, especially on the ADC and CRE loan.

- Number of optimal topics in text mining: Even clustering has been assessed in many ways, there is a little agreement on the optimal number of topics. Our experiment, once again, raises a question on this issue. In fact, the number of topics should depend on the features and component of each given corpus, there **should not be a standard** for every experiment. Our research suggests dividing the reasons that banks go failure into 2 main sub-groups: **Loan** and **Governance** related issues.

- **The consistent of Topic modelling with LDA, k-means and hierarches clustering:** By experimenting these approaches, we obtained the consistent suggestion on the number of clusters. 3 algorithms suggested that for this corpus, the number of topics should be divided in 2.

There is scope for further research as our study has some limitations: We focus only on the loss material reviews; the numeric information is discarded and by looking at this text analysis, the movement of financial condition is not mentioned.

In brief, this research has shown that utilizing text analytics bring some advantages than financial ratios analysis approach. Text analytics is relevant to data analytics for the main reasons that bank goes failure via **core words** such as loan, capital, deposit. Moreover, text analytics contributes to the literature of bank failure that **the concentration on ADC and CRE loan,** which is rarely considered in previous research.

7. APPENDIX

Word	Full written form
manag	Manage / manager / management
supervi	Supervisory / Supervise
capit	Capital
concentr	Concentration / Concentrated
adc	Acquisition, Development and Construction
cre	Commercial real estate
signific	Significant /significantly
financi	Financing
increas	increase
alll	Allowance for loan loss and lease

8. REFERENCES

1. Aas, Kjersti, and Line Eikvil. "Text categorisation: A survey." (1999).
2. Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." In Mining text data, pp. 77-128. Springer, Boston, MA, 2012.
3. Alghamdi, Rubayyi, and Khalid Alfalqi. "A survey of topic modeling in text mining." Int. J. Adv. Comput. Sci. Appl.(IJACSA) 6, no. 1 (2015).
4. Andrews, Nicholas O., and Edward A. Fox. "Recent developments in document clustering." (2007).
5. Arun, Rajkumar, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. "On finding the natural number of topics with latent dirichlet allocation: Some observations." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 391-402. Springer, Berlin, Heidelberg, 2010.
6. Ashcraft, Adam B. "Are banks really special? New evidence from the FDIC-induced failure of healthy banks." American Economic Review 95, no. 5 (2005): 1712-1730.
7. Back, Barbro, Jarmo Toivonen, Hannu Vanharanta, and Ari Visa. "Comparing numerical data and text information from annual reports using self-organizing maps." International Journal of Accounting Information Systems 2, no. 4 (2001): 249-269.
8. Bauer, Sandro, Anastasios Noulas, Diarmuid O. Séaghdha, Stephen Clark, and Cecilia Mascolo. "Talking places: Modelling and analysing linguistic content in foursquare." In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 348-357. IEEE, 2012.
9. Blei, David M., and John D. Lafferty. "Topic models." In Text Mining, pp. 101-124. Chapman and Hall/CRC, 2009.
10. Caprio, Gerard, and Daniela Klingebiel. "Bank insolvency: bad luck, bad policy, or bad banking?" In Annual World Bank conference on development economics, vol. 79. 1996.

11. Cheng, Victor C., Clement HC Leung, Jiming Liu, and Alfredo Milani. "Probabilistic aspect mining model for drug reviews." IEEE transactions on knowledge and data engineering 26, no. 8 (2014): 2002-2013.
12. Chi, Ed H., Lichan Hong, and Stuart K. Card. "Method for automatically performing conceptual highlighting in electronic text." U.S. Patent 7,702,611, issued April 20, 2010.
13. Cohen, Raviv, and Derek Ruths. "Classifying political orientation on Twitter: It's not easy!" In ICWSM. 2013.
14. Das, Sanjiv Ranjan. "Text and context: Language analytics in finance." Foundations and Trends® in Finance 8, no. 3 (2014): 145-261.
15. Dillon, Martin. "Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., $32.95 ISBN 0-07-054484-0." (1983): 402-403.
16. Dörre, Jochen, Peter Gerstl, and Roland Seiffert. "Text mining: finding nuggets in mountains of textual data." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 398-401. ACM, 1999.
17. Evans, Martin DD, and Richard K. Lyons. "How is macro news transmitted to exchange rates?" *Journal of Financial Economics* 88, no. 1 (2008): 26-50.
18. Frydman, Halina, Edward I. Altman, and Duen-Li Kao. "Introducing recursive partitioning for financial classification: the case of financial distress." The Journal of Finance 40, no. 1 (1985): 269-291.
19. Fung, Gabriel Pui Cheong, Jeffrey Xu Yu, and Wai Lam. "News sensitive stock trend prediction." In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 481-493. Springer, Berlin, Heidelberg, 2002.
20. Goodhart, Charles. News and the foreign exchange market. No. dp71. Financial Markets Group, 1990.
21. Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." Proceedings of the National academy of Sciences 101, no. suppl 1 (2004): 5228-5235.

22. Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, no. 1 (1979): 100-108.
23. Haslem, John A., Carl A. Scheraga, and James P. Bedingfield. "An analysis of the foreign and domestic balance sheet strategies of the US banks and their association to profitability performance." MIR: Management International Review (1992): 55-75.
24. He, Wu. "A survey of security risks of mobile social media through blog mining and an extensive literature search." Information Management & Computer Security 21, no. 5 (2013): 381-400.
25. Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." Machine learning 42, no. 1-2 (2001): 177-196.
26. Kolari, James, Dennis Glennon, Hwan Shin, and Michele Caputo. "Predicting large US commercial bank failures." Journal of Economics and Business 54, no. 4 (2002): 361-387.
27. Jain, Anil K., and Richard C. Dubes. "Algorithms for clustering data." (1988).
28. Kao, Chiang, and Shiang-Tai Liu. "Predicting bank performance with financial forecasts: A case of Taiwan commercial banks." Journal of Banking & Finance 28, no. 10 (2004): 2353-2368.
29. Kim, Younghoon, and Kyuseok Shim. "TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation." Information Systems 42 (2014): 59-77.
30. Kloptchenko, Antonina, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. "Combining data and text mining techniques for analysing financial reports." Intelligent systems in accounting, finance and management 12, no. 1 (2004): 29-41.
31. Koppel, Moshe, and Itai Shtrimberg. "Good news or bad news? let the market decide." In Computing attitude and affect in text: Theory and applications, pp. 297-301. Springer, Dordrecht, 2006.
32. Kumar, B. Shravan, and Vadlamani Ravi. "A survey of the applications of text mining in financial domain." Knowledge-Based Systems 114 (2016): 128-147.

33. Larsen, Bjornar, and Chinatsu Aone. "Fast and effective text mining using linear-time document clustering." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 16-22. ACM, 1999.
34. Lu, Caimei, Xiaohua Hu, and Jung-ran Park. "Exploiting the social tagging network for web clustering." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41, no. 5 (2011): 840-852.
35. M. Kwast, J. Rose, Pricing, Operating Efficiency and Profitability Among Large Commercial Banks, Journal of Banking and Finance 6 (1982) 233–254.
36. Martin, Daniel. "Early warning of bank failure: A logit regression approach." Journal of banking & finance 1, no. 3 (1977): 249-276.
37. McCallum, Andrew, Andres Corrada-Emmanuel, and Xuerui Wang. "Topic and role discovery in social networks." (2005).
38. Mellouli, Sehl, Faouzi Bouslama, and Aichath Akande. "An ontology for representing financial headline news." Web Semantics: Science, Services and Agents on the World Wide Web 8, no. 2-3 (2010): 203-208.
39. Meyer, David, Kurt Hornik, and Ingo Feinerer. "Text mining infrastructure in R." Journal of statistical software 25, no. 5 (2008): 1-54.
40. Mirończuk, Marcin Michał, and Jarosław Protasiewicz. "A recent overview of the state-of-the-art elements of text classification." Expert Systems with Applications (2018): 36-45
41. Nassirtoussi, Arman Khadjeh, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. "Text mining for market prediction: A systematic review." Expert Systems with Applications 41, no. 16 (2014): 7653-7670
42. Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." In LREc, vol. 10, no. 2010, pp. 1320-1326. 2010.
43. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2, no. 1–2 (2008): 1-135.
44. Qin, Zemin, Hao Lian, Tieke He, and Bin Luo. "Cluster Correction on Polysemy and Synonymy." In *Web Information Systems and Applications Conference (WISA), 2017 14th*, pp. 136-138. IEEE, 2017.

45. Ramage, Daniel, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. "Clustering the tagged web." In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 54-63. ACM, 2009.
46. Ravi Kumar, P. and Ravi, V., (2007), Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review, *European Journal of Operational Research*, 180, issue 1, 1-28.
47. Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24, no. 5 (1988): 513-523.
48. T.B. Bell, Neural nets or the logit model? A comparison of each model's ability to predict commercial bank failures, International Journal of Intelligent Systems in Accounting, Finance and Management 6 (1997) 249–264.
49. Tussing, A. Dale. "The Case for Bank Failure." *The Journal of Law and Economics* 10 (1967): 129-147.
50. Vu, Tien-Thanh, Shu Chang, Quang Thuy Ha, and Nigel Collier. "An experiment in integrating sentiment features for tech stock prediction in twitter." (2012): 23-38.
51. Wang, Shanshan, Kaiquan Xu, Long Liu, Bing Fang, Shaoyi Liao, and Huaiqing Wang. "An ontology-based framework for mining dependence relationships between news and financial instruments." Expert Systems with Applications 38, no. 10 (2011): 12044-12050.
52. Wang, Shanshan, Zhang Zhe, Ye Kang, Huaiqing Wang, and Xiaojian Chen. "An ontology for causal relationships between news and financial instruments." Expert Systems with Applications 35, no. 3 (2008): 569-580.
53. Wang, Y.-C., Burke, M. & Kraut, R. E. Gender, topic, and audience response: an analysis of user-generated content on facebook. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013. ACM, 31-34.
54. Wheelock, David C., and Paul W. Wilson. "Why do banks disappear? The determinants of US bank failures and acquisitions." Review of Economics and Statistics 82, no. 1 (2000): 127-138.

55. Wong, Wai-Chiu, and Ada Wai-Chee Fu. "Incremental document clustering for web page classification." In Enabling Society with Information Technology, pp. 101-110. Springer, Tokyo, 2002.
56. Xie, Pengtao, and Eric P. Xing. "Integrating document clustering and topic modeling." arXiv preprint arXiv:1309.6874 (2013).
57. Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 267-273. ACM, 2003.
58. Yu, Rose, Xinran He, and Yan Liu. "Glad: group anomaly detection in social media analysis." ACM Transactions on Knowledge Discovery from Data (TKDD) 10, no. 2 (2015): 18.

CHAPTER 5

"Adventure is the life of commerce, but caution is the life of banking."

-- Walter Bagehot- founder of The Economist

CHAPTER 5

A TWO-STAGE DEA AND NEURAL NETWORK ON MEASURING AND ESTIMATING LOAN LOSS PROVISION OF LARGE US BANKS

Jean-Laurent Viviani, Hanh- Hong LE

(This article is presented at the 4th International Conference of Economics and Finance 2017)

ABSTRACT: Loan loss provision (LLP) is a significant important item in bank's financial report that can be used to (i) evaluate the magnitude of credit risk, or (ii) smooth bank's income statement. In this study, LLP is dedicated as a measurement of credit default. Theoretically, LLP should be maintained at a proper amount accordingly to the estimated credit loss, neither surplus nor deficit of LLP is favourable. In this research, a 2-stage DEA method (performance and LLP stage) is used to measure the efficiency of LLP. Then, Neural network suggests an adjustment if needed for each bank to improve LLP quality management. The dataset includes 166 large US banks in the period of 2016. The results suggest that (1) 12.7% banks are operating effectively for the performance process; (2) Only 2.4% banks reserve LLP at the proper level; (3) More efficiency banks tend to have higher *(a) Number of employees, (b)Total equity , (c) Total expenses, (d) Total deposit, (e) Total loan, (f) Total investment and higher (g) Loan Loss Provision*; (4) There is significant different between efficiency and in-efficiency group for both Performance and LLP stage; (5) By using neural network, we suggest 50% of banks should decrease the amount of LLP .

Keyword: LLP, Bank performance, DEA, Neural networks

JEL: G21, H81

1 INTRODUCTION

1.1. BANK, LOANS AND THE POTENTIAL RISKS

Traditionally, banks conduct their business by taking deposit from depositors and making loans using these funds. By gaining the disparity of deposit interest (expense) and loan interest (revenue), net interest income is generated. Hence, creating loans is a significant important mission and contribute to the majority of banks' profit. Every bank makes great effort in diversifying types of loans as well as attracting investors. However, in both non-crisis and crisis scenarios, loans contain a lot of potential risks, especially default risk, is when borrowers do not pay a part or full of the loan. If this risk occurs, it affects not only bank's liquidity position but also bank's profit.

Based on the customer's performance, loans are classified by increasing magnitude of risk, mostly depending on the number of late payment days: Criticized, scheduled, adversely, substandard, doubtful and loss loan. For each type, certain portion is taken to cover for potential risk. For instance, 5% of criticized, 10% of scheduled 20% of adversely, 100% of loss loan. This cover for uncollectible loans is called ***allowance for loan loss and lease*** (ALLL). This charge shows up on bank's income statement as a non-cash expense category named ***Loan Loss Provision*** (hereinafter LLP) to cover all, or a portion of the loss. Theoretically, if the loan loss is recognized properly by Loan loss provision, LLP will be the first cushion to cover this loss.

LLP is an engrossing item and can be viewed from some perspectives:

(i) US accounting standards for loan losses, FAS 114, indicate that, LLP should be recognized only when probable that loans have been impaired. Probable, means that a creditor will be unable to collect all amounts due according to the contractual terms of the loan agreement. The LLP is an expense that is

adjusted on screening of bank loan portfolios. As the non-cash expense in bank's income statement, the increase (decrease) in LLP would decrease (increase) bank earning.

(ii) From balance sheet perspective, a loss on a loan is a loss of an asset. The loan loss provision ensures that banks will have sufficient cushion to provide services to its depositors. However, on an operating basis, because loan loss provision is a non-cash expense, cash flow remains available.

(iii) From management perspective, LLP is a part of non-cash cost which is extracted from total revenue. The literature addresses the hypothesis that LLP is discretionary, to fulfil managerial objectives such as tax evasion, income smoothing or capital management (Beaver and Engel (1995)).

As an index to measure the potential loss, or default risk, LLP should not be less or more compared to expect future losses. If the reserve is surplus, it leads to potential trade-off between reserve and profit. In the contrary, if the deficit occurs, bank may face unexpected liquidity risk. Hence, LLP should be maintained properly and effectively at a level that is sufficient to absorb the potential losses.

However, the argument regarding how much should bank set aside, even has already guided by accounting standard became the main subject of debate, especially after recent financial crisis. Although banks provision for bad loan losses at all times, the allowances for bad loan losses were not sufficient to absorb all losses since the crisis began (De Haan, 2018).

1.2. INTEGRATION OF DEA & NEURAL NETWORKS

It is obvious that appropriately measure and provision for credit losses will firmly establish a cushion and bank can enjoy improved performance***. Data envelopment analysis*** (hereinafter DEA) is a popular tool for analysing and measuring the efficiency. DEA is an

approach to estimate the production function of organizations and organizational units and enables the assessment of their efficiency (Mostafa, 2009). Moreover, to measure 2 stages via an intermediate layer, a two-stage DEA is suggested by several prior researches (Kwon, 2015; Wanke, 2014; Zha, 2010). Although DEA is highly appreciated for measuring and optimizing goal, it is lack of predictive capacity. To compliment this weakness, a conjunction of DEA and Neural network is suggested (Kwon, 2015; Yang, 2012; Zha, 2010).

In this study, our aim is to (i) measure the production process of bank, then (ii) measure the efficiency of loan loss provisioning. In parallel, due to the lacks predictive capacity of DEA, we will (iii) apply neural networks (hereinafter NNs) to predict the adjusted value of loan loss provision.

More specifically, the purpose of this research is twofold:

(1) How efficiency of bank's operation via 2 stages: Production process and Loan Loss provisioning process via a two-stage DEA,
(2) Propose an adjustment for loan loss provision for each bank based on their efficiency using NNs.

The remainder of this paper is organized as follows: part 2 reviews briefly about the prior literature on DEA, NNs and features selection. The data and methodology will be introduced in part 3. Part 4 summaries the empirical result followed by concluding remarks in part 5.

2. LITERATURE REVIEWS

The literature presents the vast number of research in the effort of analysing the efficiency of banks' operation and loan loss provision. As loan loss provision can be created based on both quantitative and qualitative estimation, this item becomes more important and attracts researchers.

2.1. THE DETERMINANT OF LOAN LOSS PROVISION

Loan Loss Provision (LLP) is a reserve created to provide for losses that a bank expects to take because of uncollectable or troubled loans. It results in a noncash charge to earnings and includes transfers to bad debt reserves due to write-offs (in Japan), impairments charges, and impairments reversals.

The centre bank designs the legal requirement for the minimum reservation to protect banks from loss. However, the ratio is nothing than the minimum amount which is determined by regulators. In fact, Loan loss provision is significantly decided by bank managers. Most of research use LLP as an independent variable to show its affection on three main arguments: Management hypothesis, Signalling hypothesis and Income smoothing (Ahmed,1999; Anandarajan,2003; Beaver, 1997; Berger, 1997)

The criteria of loan-loss provisions is a topic of interest in the literature (Greenawalt, 1988; Keeton,1987; Ma, 1988; Wetmore, 1994). Ma (1988) and Greenawalt (1988) figure out an income smoothing effect in the determination of the loan-loss provision when bank managers can adjust loan-loss provision according to the financial conditions, for example: take a large loan-loss provision in a good year so that extra reserves are available for bad years. However, in the contrary, bank managers also may accurately disclose loan losses if income levels are low, resulting in misleading information about the bank's condition.

Ma (1999) also concluded that the poor performance of African-American banks may be attributable to inadequate assessment of risk as measured by adjustments to the provision for loan losses. From this point of view, LLP is a direct charge against earnings; thus, inappropriate assessments of risk on the part of bank managers are directly reflected in earnings.

Kim (1998) also found an evidence of LLP manipulated. The result shows that banks with low capital ratios reduced their loan loss provisions, meanwhile, banks with high capital ratios exhibited no difference in loss provision.

Dahl (2013) considered loan loss depends on: Performing loans and non-performing loans, Logarithm of asset, CAMELS rating, Equity, Allowance and Monitoring. The

result suggested that examinations had a significant and positive effect upon commercial and industrial loan-loss recognition and auditors tended to have a significantly positive effect on provisions for loan losses.

Laeven (2003) argues that loan loss provision needs to be an integral component of capital regulation, moreover, empirical evidence is found that many banks delay provisioning for nonperforming loans until too late, when cyclical downturns have already set in. As a result, loan loss reserves would increase in good times and decrease in bad times (for an example, see Kim (1993)).

Research on the determinants of loan-loss provisioning proved that the decision to set aside provisions depend on not only loan portfolio but also banks' capacity such as total asset, total equity or monitoring (Dahl, 2013).

However, Jill (1994) estimated the tie between relevant criteria to the actual provision. Contrary to the results of previous studies, Jill (1994) found no evidence of income smoothing and loan loss provision is determined by bank managers based on past loan risk, loan quality deterioration, and foreign risk. However, bank managers do not take into account off-balance-sheet exposure.

Beatty et al and Collins et al. indicated that there is positive relationship between LLP and earning before LLP, especially when the earnings are lower. Hence, LLP, which is shown in income statement, can be discretionary or non-discretionary accounting item. The question how much the bank should set aside for LLP is still the subject to be investigated

2.2. BANK'S EFFICIENCY AND THE INTEGRATION OF DEA AND NEURAL NETWORK

Basically, there are two approaches to assess bank efficiency: The stochastics efficiency frontier analysis and the deterministic frontier analysis. So far, DEA, a part of deterministic frontier analysis is the most used technique (Staub, 2010). The initial idea

of DEA namely the CCR model (Charnes et al.,1978) that evaluated DMU[3] from given inputs and one /multiple outputs. This is a management tool in identifying inefficiencies (Yang, 2012). Expanded DEA models have been subsequently presented by A two-stage DEA model.

The pioneering study of Charnes (1978) has motivated hundreds of papers using DEA for measuring the efficiency. Later, Charnes (1986) published an application using a two-stage DEA. Since then, two-stage DEA is applied widely (see Kao, 2011).

Seiford (1999) utilized two-stage DEA model to examine the performance of the top 55 US commercial banks via 2 production process that separate profitability and marketability. Their result suggests that smaller banks perform better to marketability, whereas larger banks tend to perform better to profitability. Kwon(2015) measure the production process by using set of inputs includes: Employee, Equity, Expenses and set of outputs includes: Loans, Deposit, Investment. Nath (2001) measures buy DEA with inputs: Total deposit, Labor, Other non-interest expenses, equity and outputs: Total loans, investment and total non-interest income. Several researches used A two-stage DEA model by following Seiford (1999); (Berger, 1997; Fethi,2010; Liang, 2011; Wang, 2014, Yang 2011).

In a brief explanation, two-stage DEA model is the way to combine 2 separate DEA model: The output variables of the first DEA are the inputs variables of the second DEA. This overlap, as an unavoidable effect, leads to the potential conflicts when the 1st stage increases the outputs, while the 2nd stage decreases the inputs during the optimization process. However, in the context of this research paper, this incompatibility will not be considered; we focus on related reviews in selecting input and outputs variables.

Literature shows the growing trend in utilizing the integration of DEA and BPNN in many sectors. Kwon (2015) utilized DEA neural network approach for measuring and predicting the profit (net income) of 181 US large banks which have more than 3 billion

[3] Decision Making Unit - all individuals and groups that take part in the decision-making process relating to the negotiation of products /services (Philip Kotler)

USD for assets. Stewart (2016) applied a DEA double bootstrap approach to examine efficiency in the Vietnamese banking system from 1999 to 2009. Wu (2006) use DEA-neural network approach to evaluate branch efficiency of large Canadian bank. The results are comparable to the normal DEA results overall. However, DEA-NN approach produces a more robust frontier and identifies more efficient units since more good performance patterns are explored. Furthermore, DEA-NN approach provides worse performers the guidance on how to improve their performance to different efficiency ratings. Mostafa (2009) also investigates the efficiency of top Arab banks. Results indicate that the predictive accuracy of NN models is quite like that of traditional statistical methods. The study also shows that the NN models have a great potential for the classification of banks' relative efficiency due to their robustness and flexibility of modelling algorithms. Ohsato (2015) measure efficiency in Japanese Regional banks by using a network DEA.

As DEA is a proper method for measuring the efficiency but lack of predict capacity, and Neural network is proved to be one of the most accuracy prediction method, the integration of DEA and Neural network would achieve suitable advices for bank managers.

3. DESCRIPTION OF THE METHODOLOGY

3.1. DEA

DEA involves in several applications thanks to its advantages since it was first introduced by Charnes, Cooper and Rhodes (1978). One consideration is to include as many DMUs as possible because with a larger population there is a greater probability of capturing high performance units that would determine the efficient frontier and improve discriminatory power. The other conflicting consideration with a large data set is that the homogeneity of the data set may decrease, meaning that some exogenous impacts of no interest to the analyst or beyond control of the manager may affect the results (Golany 1989). Within the scope of this paper, we will apply DEA as a nonparametric linear

programming method for measuring the efficiency of 166 large banks- which is 166 decision making units (hereinafter DMUs).

Assume that there are 166 banks; the stakeholders need to identify which of these are efficient and more efficient than the others. DEA creates an efficient frontier that envelope the best banks (which are considered as efficiency). Eventually, the banks which are not located at the border are inefficiency, or less efficiency than the others. By determining the inefficiency DMUs, DEA proposes the potential adjustment of inputs and outputs to make that inefficiency becomes efficiency. It is noteworthy that the DMU which is efficiency is in comparison with other units. The best practice units obtain the efficiency score (hereinafter ES) ES=1. The inefficiency units obtain the ES less than 1.

The proposed calculation by Cooper-Charnes- Rhodes (CCR model, 1978) as followed:

With y_{rj} and x_{ij} are called as inputs and outputs of j^{th} bank and u_r, $v_i>0$ are variable weights to be determined by the solution of this problem:

$$\max h_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}}$$

Subject to: $\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, j = 1, \dots, n$

$$u_r, v_i \geq 0, r = 1, \dots . s, i = 1, \dots, m$$

It is noteworthy that using DEA, the researchers should have answered 2 important questions on: Orientation and Return to scale questions.

Under DEA model, there are 2 types of Orientation: Input and Output orientation. Input orientation is to minimize the uses input at given output. Output orientation is to maximize the output at the given input

There is an argument related to the returns to scale. There are 2 popular methods in DEA is CRS (Constant returns to scale) and VRS (Variable returns to scale). CRS reflects that outputs will change by the same proportion as inputs are changed. Meanwhile, VRS reflects that the output may increase/decrease/constant when the inputs are changed.

The choice of parameters for this study is explained in part 4.

3.2. NEURAL NETWORKS

Artificial Neural network model is popular for its flexibility in creating prediction, classification, pattern recognition process, etc. This model is also well-known for the ability to transform inputs into outputs to the best of its capacity. The output of a neuron is a function of the weighted sum of the inputs plus a bias. Denote the set of input is: $I=\{i_1, i_2, \ldots, i_n\}$, suppose that the corresponding weight of each input is $W=\{w_1,w_2,\ldots,w_n\}$. The output will be transformed from the input set via activation function:

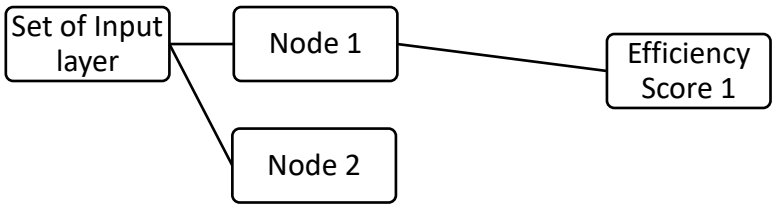
$$Output = f(i_1.w_1 + i_2.w_2 + \cdots . + i_n.w_n + bias\ )$$

Where n is the number of input variable.

The Back propagation neural network is a systematic training method for a Multi-Layer perceptron (MLP). This method is popularly applied in prediction and classification tasks.

In using neural networks, it is important to split the dataset into 2 independent sets: Training set and testing set. The main role of training set is to 'train' the neural network. Testing set is a group of samples used to verify the performance of the neural network.

Figure 11: The process of calculating Efficiency Score 1



3.3. DESIGN THE MEASUREMENT AND PREDICTION WITH DEA AND BPNN

There are two separate processes in this research paper: DEA measurement and NNs prediction process. (Figure 12)

DEA measurement process includes 2 sub-processes with structure of 3-3-1 is built by forming 3 layers: Input layer (Number of Employee, Total equity and Total expense), intermediate layer (Total deposit, Total loan and Total investment) and output layer (Loan Loss provision). The first stage represents the bank's performance evaluation; the second stage measures the efficiency of Loan loss provisions (LLP)

Figure 12: The two-stage DEA steps

In parallel with 2 -stage DEA, BPNN also predicts with 2 steps: First step, BPNN is trained to predict the Efficiency Score with data set include: Input layer and Intermediate layer. The second stage BPNN model is trained by the data set composes of the Efficiency Score and the Intermediate layer to predict the amount of Loan Loss Provision to total loan. The 2nd stage, moreover, is designed to estimate the incremental / decremental loan loss provision.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. DATA

The dataset is collect from Bloomberg database for the period of 2016. There are 166 banks with total asset higher than 3 billion US dollar. The descriptive statistic of each variable is mentioned in table 16.

The descriptive statistic reported that there is the big gap between the minimum and maximum in most of variables.

Table 16: Descriptive statistic

Variable	Mean	S.E	S.D	Minimum	Maximum
I1	3397	623	8029	166	71191
I2	2822	501	6452	157	47933
I3	701	126	1627	48	13315
M1	18117	3102	39972	1050	334590
M2	15994	2660	34276	1129	278033
M3	21327	3665	47224	1683	387308
O1	46	10	131	-16	1324

Table 17: Variable explanation

Variable	Variable	Explanation
I1	Number of Employee	Total number of employee
I2	Total equity	Total equity capital
I3	Total expense	Total expense
M1	Total deposit	Total deposit
M2	Total Loan	Total Loan
M3	Total Investment	Total Investment
O1	Provision for Loan loss	Provision for loan losses over the year

4.2. DEA EFFICIENCY MEASUREMENT

Two separate DEA analyses have been applied to measure the efficiency of 2 separate sets of outputs. The first stage measures the efficiency of the banks' performance based on the input include:

- Governance variable: Number of employee

- Bank's financial capacity variable: Total equity

- Management's efficiency variable: Total expenses

The output of the first stage based on 3 main activities of bank:

- Deposit: Total customer deposit

- Loan: Total loan

- Investment: Total investment

The second stage measures the Loan loss provision based on the outputs of the first stage.

4.3. BACK PROPAGATION NEURAL NETWORKS

Stage 1: The outputs are expected to be the 'Prediction Efficiency scores'. To predict the Efficiency score, the required inputs include: I1, I2, I3 and sets of M1, M2, and M3. Prediction the first stage efficiency:

$$ES1 = f(I1, I2, I3, M1, M2, M3)$$

Stage 2: The loan loss provision (O1) is predicted based on the input: M1, M2, M3 and ES1 from DEA first stage. Prediction the second stage efficiency:

$$O1 = f(M1, M2, M3, ES1)$$

Where O1 denotes the Loan loss provision, and ES1 represents the efficiency score from the first prediction stage.

4.4. DEA EFFICIENCY ASSESSMENT

One main advantage of DEA is to allow several inputs and outputs to be considered at the same time. Charnes–Cooper–Rhodes (CCR) (1978) expanded the model Farrell's efficiency measurement to the concept of multiple inputs and multiple outputs. By applying CCR, an efficiency frontier is created from the ratios to measure the relative efficiency of each DMU (Lin. 2009). In CCR model, DEA can be conducted under the

assumption of constant or variable returns to scale. CRS, proposed by CCR (1978) or constant return to scale is appropriate when all banks (or DMU) are operating at an optimal scale. Otherwise, VRS or variable returns to scale is proposed by Banker, Charnes and Cooper (1984) is the other option.

The literature provides an argument regarding the efficiency of CRS and VRS. Reddy (2015) made a comparison between CRS and VRS models of OC mines. His study strongly recommends of using VRS model. This result is consistent with Seiford (1999) and Chen (2005). As mentioned in Chen (2000), most of the applications on DEA are based upon the CRS assumption. There are a few using VRS model in literature, however, Chen (2005) recommend using the VRS technology. Demsetz (1997) show that the large bank holding companies have better diversified than small ones. Which means, the assumption of CRS that assume there is no significant between the size of bank and their efficiency is not accepted here. Hence, VRS is used

As mentioned in part 2, it is important to point out which of orientation is selected for the study. Output orientation is to maximize the output given the input. Input orientation is to minimize the input at the given outputs. Input orientation is more appropriate selection because we consider the Loan Loss provision as the credit risk measurement. Hence, LLP with the certain amount of reserve, the input should be as minimum as possible.

Table 18: The result from two-stage DEA

	1st stage	**Proportion**	**2nd stage**	**Proportion**
Mean Efficiency	**0.775**		**0.373**	
Efficiency DMUs	**21**	**12.7%**	**4**	**2.4%**
0.9 – 1	16	9.6%	0	0%
0.8 – 0.9	31	18.7%	3	1.8%
0.7-0.8	45	27.1%	4	2.4%
0.6-0.7	35	21.7%	17	10.2%
0.5-0.6	14	8.4%	16	9.6%

	1st stage	**Proportion**	**2nd stage**	**Proportion**
0.4-0.5	3	1.8%	21	12.7%
0.3-0.4	1	28	16	16.9%
0.2-0.3	0	0%	33	19.9%
0.1-0.2	0	0%	32	19.3%
0.0-0.1	0	0%	8	4.8%
TOTAL	**166**	**100%**	**166**	**100%**

Table 18 shows the result by using VRS technology and input orientated efficiency. For the first stage, 21 of 166 banks are consider as "efficiency" with ES=1. Most of the remain have ES from 0.6 to 0.9 and only 4 banks have ES smaller than 0.5. This result is opposite to the second stage. 110 banks (73.6%) have ES smaller than 0.5. Only 4 banks are considered as efficiency.

The mean of efficiency score for the first stage is 0.755, which is 40% higher than the second stage. 20 out of 21 efficiency DMUs of the first step is no longer efficiency in the 2nd step. 157 over 166 banks have efficiency score of the first stage higher than in the second stage. This result reveals that the Loan Loss Provision process of these banks is less effective than the first performance step. Table 19 demonstrates the correlation between the efficiency score of stage 1 and 2. The correlation of the efficiency score of both steps is low degree of correlation. (Table 19)

Table 19: The correlation of Efficiency score of step 1 and step 2

	ES1	***ES2***
ES1	1	
ES2	0.124	1

Table 20: DEA- The first stage: Comparison in the mean of sub-sample

	ES	TOTAL	In-ES
I1	8606.762	3396.711	2642.152
I2	7637.915	2822.346	2124.918
I3	1858.821	701.3439	533.7093
M1	50736.52	18116.9	13392.68
M2	43812.58	15993.92	11965.01
M3	59968.7	21327.18	15730.82
O1	144.6719	46.36157	32.12353
ES1	1	0.77479	0.742173

Note: ES=1: Efficiency sample: the group with ES=1; ES< 1: Inefficiency sample: the group with ES<1, TOTAL= Total data sample.

t-Test: Paired Two Sample for Means	
Pearson Correlation	0.99
P(T<=t) two-tail	0.047***
t Critical two-tail	2.36

Table 20 and Table 21 show the mean comparison among 3 groups: Efficiency group, Inefficiency group and Total. Efficiency group includes bank that obtains ES=1 for equivalent step. Inefficiency group includes bank that has ES <1 and total is included both Efficient and Inefficient group. Table 20 shows that on average, efficiency group tends to have higher I1, I2, I3, M1, M2, M3 and O1 in comparison with in-efficiency group. The comparison in table 21 shows that on average, efficiency group tends to have higher M1, M2, and M3 than other groups. However, the differences among groups in step 2 are not significant as in the step 1. **It is noteworthy that for each group, the Total mean is always in the middle of Efficiency and Inefficiency groups.**

Table 21: DEA- The second stage: Comparison in the mean of each sample

	ES	**TOTAL**	**In ES**
M1	86424.19	18116.9	16430.3
M2	71988.34	15993.92	14611.34
M3	100008.9	21327.18	19384.42
O1	378.8643	46.36157	38.15163
ES2	1	0.373423	0.357951

Note: ES=1: Efficiency sample: the group with ES=1; ES< 1: Inefficiency sample: the group with ES<1, TOTAL= Total data sample.

t-Test: Paired Two Sample for	
Pearson Correlation	0.99
P(T<=t) two-tail	0.04***
t Critical two-tail	2.36

4.5. BPNN prediction experiments

When designing a multilayer network, the decision on choosing the number of hidden layers is very important. Lee et al. (2005) and Zhang et al. (1999) suggested that one hidden layer is sufficient for most classification problems. However, Vasu (2011) show that utilizes 2 hidden layers in order to be sure that the network architecture will be sufficiently complex to cope with the complexity of prediction. In our study, we choose 2 hidden layers for our experiment.

There is argument in determining the proportion of training and test set. We follow Kwon (2015) to divide data into 60% training set and 40% is testing set. The quantity of data sample is also very important. A large dataset is usually preferred for BPNN, however, for modelling, as Kwon (2014) suggested, small scale of data is also satisfactory. Moreover, to avoid from overfitting issues, as rule of thumb, researchers suggested that the size of training set must be at least 10 times the number of variables.

The model includes total 7 variables for the first and second prediction process r; hence, our splitting data meet the hypothesis. 166 banks (DMUs) are split randomly into 2 subsets: 99 banks for the training subset and 67 banks for the testing subset. Figure 13 and figure 14 present graphically the neural network for stage 1 and stage 2.

Figure 13: Plot of neural network for Bank's performance stage

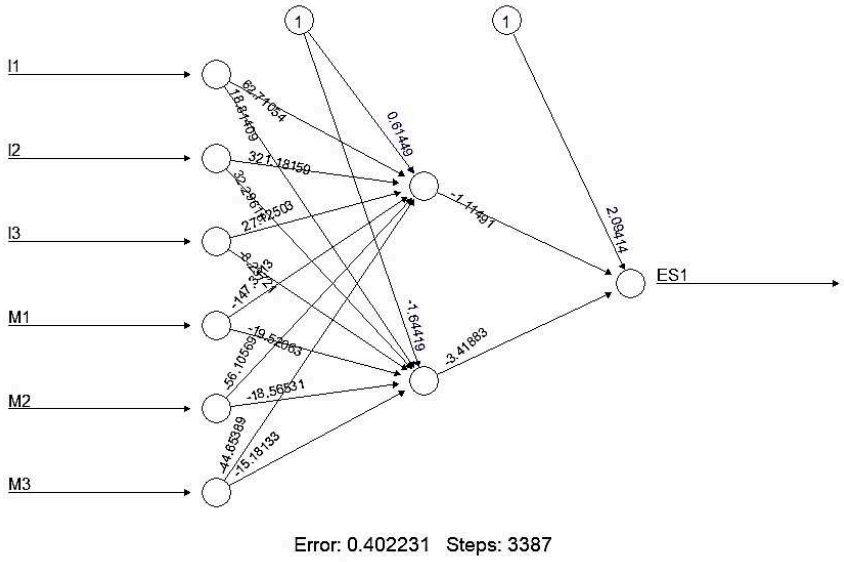


Figure 14: Comparison of real and predicted Efficiency score 1 and LLP

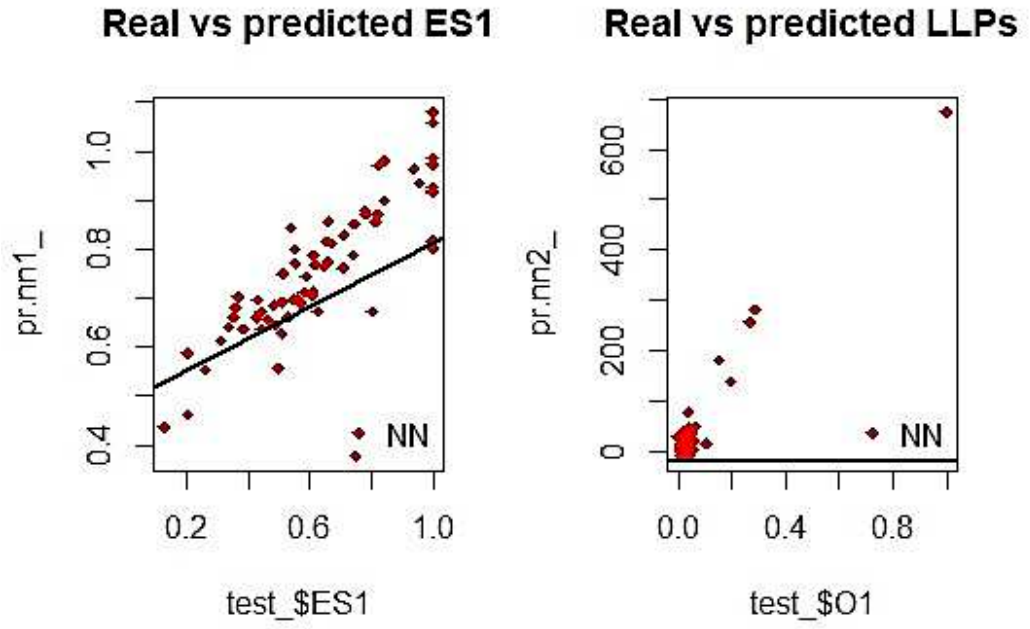
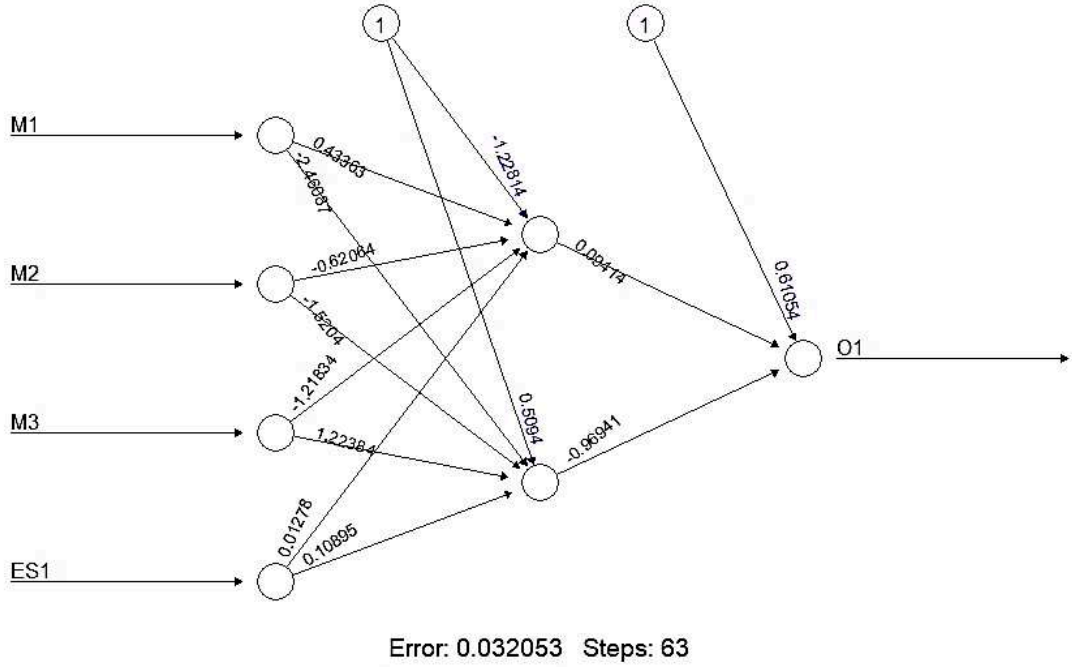


Figure 15: Plot of neural network for Bank's LLP stage



The first prediction process is to estimate Efficiency Score of each bank based on the same sets of input and outputs. The second prediction process is built to predict the level of Loan Loss Provision ratio.

By applying BPNN to predict the efficiency score, we obtained the comparison as in table 22 and table 23. The Pearson correlation in table 22 suggests that there is low correlation between Predicted and real value for efficiency score of the first stage (performance process). However, as the result from table 23, the correlation is very high correlation.

Table 22: The comparison of real and the predicted value for the ES1

	ES1-predicted	***ES1- by DEA***
Mean	0.753	0.730
Variance	0.020	0.034
Pearson Correlation	0.250	

Table 23: The comparison of real and the predicted value for LLP

	O1- predicted	***O1-real***
Mean	38.93	53.56
Variance	7431.34	30233.82
Pearson Correlation	0.93	

Table 24: BPNN learning result

Stage	**Mode**	**DMUs**	**MAE**	**MAPE**
NN-1st stage	Test	67	14.30	25.70
NN-2nd stage	Test	67	3052.63	323.41

Note: MAE: Mean absolute error, MAPE: Mean absolute percentage error

5. CONCLUSION

Basel I, II, III are released with one of those scopes is to advice banks well-managed their capital, reduce credit risks. In any circumstance, if bank reserve more or less than the amount-should-be, may cause harmful for bank profit discarding the fact that there is always trade-off between reserved amount and profitability.

Bank manager always makes effort to minimize the possible Loss Loan, however, default risk still occurs and become one of the biggest problems. Hence, loan policy should be reviewed and updated frequently to avoid from this risk. Bank managers should also consider carefully each loan contract before deciding to reserve for future loss.

The priority target of this study is to measure the efficiency of the reserve for loan loss in large US banks. By taking advantages from the complemented features of both DEA and BPNN, this research has conducted a measurement and prediction process to measure and predict the efficiency of 166 top big banks in US.

This study suggest that (1) Only 12.7% of banks are operating efficiency for performance process, (2) Even fewer (only 2.4%) banks reserve money for loss provision at efficiency level (3) efficiency group of bank tends to have higher I1 (Number of employee), I2 (Total equity), I3 (Total expense), M1 (Total deposit), M2 (Total loan) and M3 (Total investment) and higher O1 (Loan Loss Provision) (4) there is significant different between efficiency group and in-efficiency group for both Performance and LLP stage (5) By using neural network, we suggest 50% of banks to decrease and 50% to increase the amount of loan loss provision to be more efficiency. However, like other research papers, our major findings are not warranted without limitations that may be investigated for future studies. In this pilot paper, the objective is to measure and estimate the proper amount of loan loss provision discarding the fact that LLP can be used for earning management.

6. REFERENCES:

1. Aebi, Vincent, Gabriele Sabato, and Markus Schmid. "Risk management, corporate governance, and bank performance in the financial crisis." *Journal of Banking & Finance* 36, no. 12 (2012): 3213-3226.
2. Ahmed, Anwer S., Carolyn Takeda, and Shawn Thomas. "Bank loan loss provisions: a reexamination of capital management, earnings management and signaling effects." *Journal of accounting and economics* 28, no. 1 (1999): 1-25.
3. Anandarajan, Asokan, Iftekhar Hasan, and Ana Lozano-Vivas. "The role of loan loss provisions in earnings management, capital management, and signaling: The Spanish experience." *Advances in International Accounting*16 (2003): 45-65.
4. Beaver, William H., Stephen G. Ryan, and James M. Wahlen. "When is "bad news" viewed as "good news"?." *Financial Analysts Journal* 53, no. 1 (1997): 45-54.
5. Beatty, Anne, Sandra L. Chamberlain, and Joseph Magliolo. "Managing financial reports of commercial banks: The influence of taxes, regulatory capital, and earnings." *Journal of accounting research* (1995): 231-261.
6. Berger, Allen N., and David B. Humphrey. "Efficiency of financial institutions: International survey and directions for future research." *European journal of operational research* 98, no. 2 (1997): 175-212.
7. Berger, Allen N., and Robert DeYoung. "Problem loans and cost efficiency in commercial banks." *Journal of Banking & Finance* 21, no. 6 (1997): 849-870.
8. Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. "Measuring the efficiency of decision making units." *European journal of operational research* 2, no. 6 (1978): 429-444.
9. Cook, Wade D., Liang Liang, and Joe Zhu. "Measuring performance of two-stage network structures by DEA: a review and future perspective." *Omega* 38, no. 6 (2010): 423-430.

10. Collins, Julie H., Douglas A. Shackelford, and James M. Wahlen. "Bank differences in the coordination of regulatory capital, earnings, and taxes." *Journal of accounting research* (1995): 263-291.
11. Dahl, Drew. "Bank audit practices and loan loss provisioning." Journal of Banking & Finance 37, no. 9 (2013): 3577-3584.
12. De Haan, Leo, and Maarten RC Van Oordt. "Timing of banks' loan loss provisioning during the crisis." *Journal of Banking & Finance* 87 (2018): 293-303.
13. Demsetz, Rebecca S., and Philip E. Strahan. "Diversification, size, and risk at bank holding companies." Journal of money, credit, and banking (1997): 300-313.
14. Fethi, Meryem Duygun, and Fotios Pasiouras. "Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey." *European journal of operational research* 204, no. 2 (2010): 189-198.
15. Flannery, M., J. Houston, Ed Kane Hadlock, Andy Naranjo, and Mike Ryngaert. *The Value of a Government Monitor for Firms with Hard-to-Value Assets*. Working Paper, University of Florida, 1995.
16. Golany, Boaz, and Yaakov Roll. "An application procedure for DEA." *Omega* 17, no. 3 (1989): 237-250.
17. Greenawalt, Mary Brady, and Joseph F. Sinkey. "Bank loan-loss provisions and the income-smoothing hypothesis: an empirical analysis, 1976–1984." *Journal of financial services research* 1, no. 4 (1988): 301-318.
18. Kim, Myung-Sun, and William Kross. "The impact of the 1989 change in bank capital standards on loan loss provisions and loan write-offs." *Journal of Accounting and Economics* 25, no. 1 (1998): 69-99.
19. Keeton, William R., and Charles S. Morris. "Why do banks' loan losses differ?." *Economic Review* 72, no. 5 (1987): 3-21.
20. Kwon, He-Boong, and Jooh Lee. "Two-stage production modeling of large US banks: A DEA-neural network approach." *Expert Systems with Applications* 42, no. 19 (2015): 6758-6766.

21. Kwon, He-Boong. "Performance modeling of mobile phone providers: A DEA-ANN combined approach." *Benchmarking: An International Journal* 21, no. 6 (2014): 1120-1144.
22. Kyereboah-Coleman, Anthony. "Corporate governance and shareholder value maximization: An African perspective." *African Development Review* 19, no. 2 (2007): 350-367.
23. Laeven, Luc, and Giovanni Majnoni. "Loan loss provisioning and economic slowdowns: too much, too late?." *Journal of financial intermediation 12*, no. 2 (2003): 178-197.
24. Liang, Liang, Zhao-Qiong Li, Wade D. Cook, and Joe Zhu. "Data envelopment analysis efficiency in two-stage networks with feedback." *IIE Transactions* 43, no. 5 (2011): 309-322.
25. Lin, Tyrone T., Chia-Chi Lee, and Tsui-Fen Chiu. "Application of DEA in analyzing a bank's operating performance." *Expert systems with applications* 36, no. 5 (2009): 8883-8891.
26. Ma, Christopher K. "Loan loss reserves and income smoothing: The experience in the US banking industry." *Journal of Business Finance & Accounting* 15, no. 4 (1988): 487-497.
27. Mostafa, Mohamed M. "Modeling the efficiency of top Arab banks: A DEA–neural network approach." *Expert Systems with Applications* 36, no. 1 (2009): 309-320.
28. Mostafa, Mohamed. "Benchmarking top Arab banks' efficiency through efficient frontier analysis." *Industrial Management & Data Systems* 107, no. 6 (2007): 802-823.
29. Nath, Prithwiraj, Avinandan Mukherjee, and Manabendra Nath Pal. "Identification of linkage between strategic group and performance of Indian commercial banks: a combined approach using DEA and Co-Plot." (2001).
30. Ohsato, Satoshi, and Masako Takahashi. "Management efficiency in Japanese regional banks: A network DEA." *Procedia-Social and Behavioral Sciences* 172 (2015): 511-518.

31. Reddy, G. Thirupati. "Comparison and Correlation C CRS and VRS models o." Management (2015).
32. Salim, Ruhul, Amir Arjomandi, and Juergen Heinz Seufert. "Does corporate governance affect Australian banks' performance?" *Journal of International Financial Markets, Institutions and Money* 43 (2016): 113-125.
33. Seiford, Lawrence M., and Joe Zhu. "Profitability and marketability of the top 55 US commercial banks." *Management science* 45, no. 9 (1999): 1270-1288.
34. Shin, Kyung-Shik, Taik Soo Lee, and Hyun-jung Kim. "An application of support vector machines in bankruptcy prediction model." *Expert Systems with Applications* 28, no. 1 (2005): 127-135.
35. Staub, Roberta B., Geraldo da Silva e Souza, and Benjamin M. Tabak. "Evolution of bank efficiency in Brazil: A DEA approach." *European journal of operational research* 202, no. 1 (2010): 204-213.
36. Stewart, Chris, Roman Matousek, and Thao Ngoc Nguyen. "Efficiency in the Vietnamese banking system: A DEA double bootstrap approach." *Research in International Business and Finance* 36 (2016): 96-111.
37. Vasu, Madireddi, and Vadlamani Ravi. "Bankruptcy prediction in banks by principal component analysis threshold accepting trained wavelet neural network hybrid." In *International Conference on Data Mining, USA*. 2011.
38. Wang, Ke, Wei Huang, Jie Wu, and Ying-Nan Liu. "Efficiency measures of the Chinese commercial banking system using an additive two-stage DEA." *Omega* 44 (2014): 5-20.
39. Wanke, Peter, and Carlos Barros. "Two-stage DEA: An application to major Brazilian banks." *Expert Systems with Applications* 41, no. 5 (2014): 2337-2344.
40. Wetmore, Jill L., and John R. Brick. "Loan-loss provisions of commercial banks and adequate disclosure: A note." *Journal of Economics and Business* 46, no. 4 (1994): 299-305.
41. Wu, Cheng-Ru, Chin-Tsai Lin, and Pei-Hsuan Tsai. "Evaluating business performance of wealth management banks." *European Journal of Operational Research* 207, no. 2 (2010): 971-979.

42. Wu, Desheng Dash, Zijiang Yang, and Liang Liang. "Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank." *Expert systems with applications*31, no. 1 (2006): 108-115.
43. Yang, Chyan, and Hsian-Ming Liu. "Managerial efficiency in Taiwan bank branches: A network DEA." *Economic Modelling* 29, no. 2 (2012): 450-461.
44. Yang, Feng, Dexiang Wu, Liang Liang, Gongbing Bi, and Desheng Dash Wu. "Supply chain DEA: production possibility set and performance evaluation model." *Annals of Operations Research* 185, no. 1 (2011): 195-211.
45. Zha, Yong, and Liang Liang. "Two-stage cooperation model with input freely distributed among the stages." *European Journal of Operational Research* 205, no. 2 (2010): 332-338.
46. Zhang, Guoqiang, Michael Y. Hu, B. Eddy Patuwo, and Daniel C. Indro. "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis." *European journal of operational research* 116, no. 1 (1999): 16-32.

CHAPTER 6

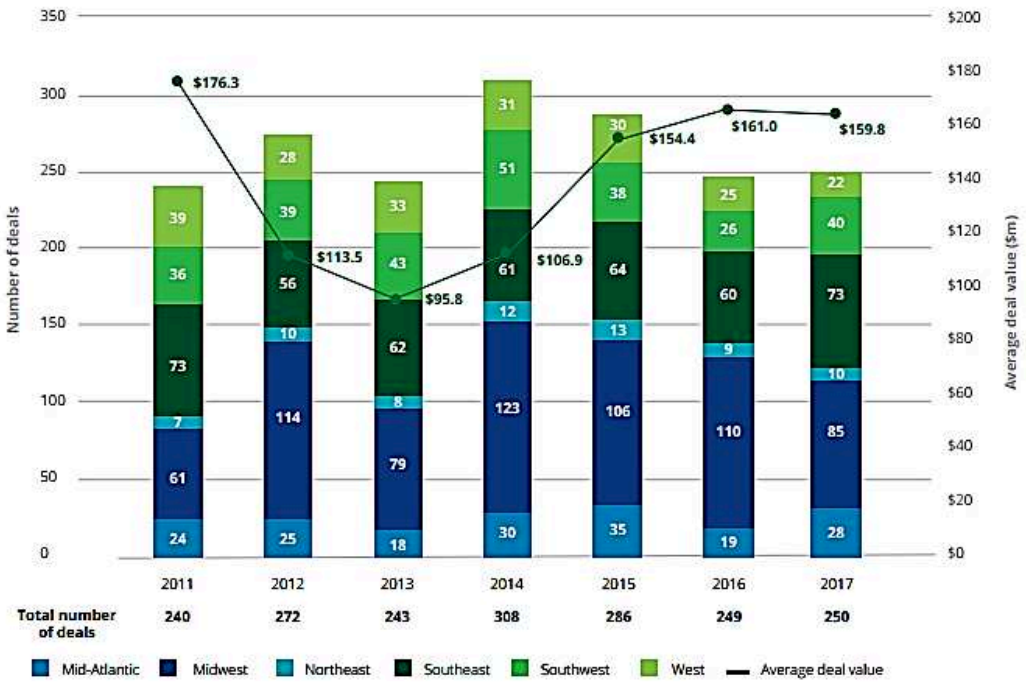
CHAPTER 6: CONCLUSION

1. CONCLUSION

FDIC has summarised the bank closing from 2001 to 2018. There is the sharp decline trend after financial crisis of 2008. However, as the cycle of economic, we will never know when the next crisis would come back, and each bank should be cautious and well protected from all kinds of risk.

Moreover, the US banking system is recognized as the system of too many small banks. As the resulted of consolidation, there has been a decline in the number of U.S. banks from more than 18.000 to less than 5.800 banks and continue to fall. Banks disappear primarily for one of two reasons: either they fail, or they are acquired by another bank. Figure 16 presents the number of banking deals (by region) and average deal value of US bank in the period of 2011-2017. Even when the bank is acquired by another bank, it includes both pros and cons for local as well as global economy.

Figure 16:Banking deals (by region) and average deal value



This thesis focused on predicting and explaining the reasons why U.S banks go failure by using variety of methods and from two points of views: Numeric and non-numeric data.

The thesis predicts and compares the accuracy of traditional statistical techniques and machine learning techniques. A sample of 3000 US banks includes 1438 failures and 1562 active banks is investigated by two traditional statistical approaches (Discriminant analysis and Logistic regression) and three machine learning approaches (Artificial neural network, Support Vector Machines and k-nearest neighbours). For each bank, data were collected for a 5-year period before they become inactive. 31 financial ratios extracted from bank financial reports covered 5 main aspects: Loan quality, Capital quality, Operations efficiency, Profitability and Liquidity. It is observed that machine learning technique (i) ANNs and (ii) k-NN predict more accuracy than traditional methods. However, the difference in prediction accuracy between ANNs and k-NN methods and the traditional logistic regression method is not too significant. In addition, SVM does not perform better than traditional methods. Nevertheless, ANN and k-nearest neighbour demonstrate their remarkable ability when they can detect the failure correctly, but the other methods cannot. Among 31 ratios, notably, the ratios (1) Impaired Loans/Gross Loans, (2) Tier 1 capital ratio, (3) Capital funds/Total assets, (4) Other Operation Income/Average Assets, (5); Net interest revenue/Average Assets, (6) Non Operation Items &taxes/Average Assets, (7) Return on Average Assets, (8) Cost to income ratio, (9) Net Loans/Total Asset, (10) Net loans/Deposit &Short Term funding, (11) Net Loans/Total Deposit& Borrowing are more relevant than the others.

By analysing the textual information that collect from the Material Loss Review published by FDIC, the thesis suggests that ADC (Acquisition, Development and Construction loan) and CRE (Commercial Real Estate Loan) are of paramount importance to the bank's survival. Aside from loan problems, it is important to note that board and management is extremely important. To prevent from being failure, bank must pay attention on the following issues:

Loan: Loan, ADC, CRE, credit, rate, ALLL

Management: Exam, management, report, supervise, review, board, audit

Capital: Capital, deposit, asset, fund, portfolio

Magnitude: Increase, Significant, growth, concentration

Moreover, ***Exam, Concentration, asset, implement, adequate*** on ADC and CRE loan also effect on bank's management. The thesis also suggests that the reasons that US banks went failure are classified into 2 main groups: **Loan and Governance.** This result is consistent among Topic Modelling and Document clustering method.

Finally, this thesis proposes of using DEA and Neural networks to manage the amount of Loan Loss Provision of 166 top big banks in US. This study suggest that (1) Only 12.7% of banks are operating efficiency for performance process, (2) Even fewer (only 2.4%) banks reserve money for loss provision at efficiency level (3) efficiency group of bank tends to have higher Number of employee, Total equity, Total expense, Total deposit, Total loan, Total investment and Loan Loss Provision (4) there is significant different between efficiency group and in-efficiency group for both Performance and LLP stage (5) By using neural network, we suggest 50% of banks to decrease and 50% to increase the amount of loan loss provision to be more efficiency.

2. DISCUSSION AND FUTURE RESEARCH

Many of the difficulties the banks experiences resulted from inadequate loan policies, loan identification system and quality management. Moreover, economic decline contributed significantly to the difficulties of many failed banks. A bank failed for couples of certain reasons. There is an opinion that a bank may fail because of 'bad luck', however, if bank manages their risks well enough, 'bad luck' cannot knock it down.

Although the research has reached its aims, there were some unavoidable limitations.

One of its limitation in chapter 4 is the collected reports. As these reports are released after the event of bankruptcy by FDIC, the achieved information is for the explanation

purpose more than predicted value. To develop and use the advantage of textual analysis, we will collect news relating to bank's condition, since then predict the bank failure probability via non-numeric analysis.

Chapter 5 is about Loan Loss Provision, however, we assumed that LLP is a proxy of credit risk absorb. In fact, there is smoothing theory that consider LLP as a tool of bank managers. For the future research, the role of LLP should be clarified and examined before giving adjustment advices.

However, research on the bank's failure using Machine learning techniques should continue as Ohlson used to say: "*Forecast bankruptcy is obvious practical interest*".

APPENDIX:

PUBLICATION IN: RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE

Research in International Business and Finance 44 (2018) 16–25

Contents lists available at ScienceDirect

Research in International Business and Finance

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

Full length Article

Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios

Hong Hanh Le, Jean-Laurent Viviani*

IGR-IAE-Rennes, CREM, 11 rue J. Macé, 35700 Rennes, France

ARTICLE INFO

JEL classification:
G33

Keywords:
Failure prediction
Intelligent techniques
Artificial neural network
Support vector machines
K-nearest neighbors
US banks

ABSTRACT

This research compares the accuracy of two approaches: traditional statistical techniques and machine learning techniques, which attempt to predict the failure of banks. A sample of 3000 US banks (1438 failures and 1562 active banks) is investigated by two traditional statistical approaches (Discriminant analysis and Logistic regression) and three machine learning approaches (Artificial neural network, Support Vector Machines and k-nearest neighbors). For each bank, data were collected for a 5-year period before they become inactive. 31 financial ratios extracted from bank financial reports covered 5 main aspects: Loan quality, Capital quality, Operations efficiency, Profitability and Liquidity. The empirical result reveals that the artificial neural network and k-nearest neighbor methods are the most accurate.

1. Introduction

According to the Federal Deposit Insurance Corporation (FDIC), during 2008–2014 more than 500 banks declared as failures in the United States of America. The cost of failure per dollar of failed-bank assets is already high and may continue to rise. Consequently, the more banks go bankrupt, the higher the cost of resolving after-failure events. Year-end 2013, FDIC estimated that the total cost to the deposit insurance funds of resolving these failed banks is as high as 30 billion US dollars.

Banks are considered as failures if the state or bank regulator forces them to close because of insolvency problems. Because of the strong interconnection between banks and their essential role in financing the economy, the failure of banks is more threatening for the economy than the failure of other business firms. In some cases, the bankruptcy of one bank can cause a knock-on effect, which can spread quickly and have a negative impact on other banks (systemic risk). Hence, detection of bank failure before it occurs and try to avoid them is mandatory. In this research, we execute machine learning techniques which have been claimed to improve the prediction of sbank failure.

Starting with seminal research studies, Beaver (1966) and Altman (1968) built statistical models to predict firm failure based on accounting ratios. Since then, numerous studies have been advanced using different financial ratios, samples and periods. In parallel with the development of computational sciences, many different interesting approaches were explored to promote the power of technology. In order to help researchers better understand this complex field, Ravi Kumar and Ravi (2007) present a comprehensive review of the applications of prediction techniques to solve bankruptcy prediction problems of banks and firms. One of the intelligent technique families known as 'Machine learning' becomes more popular among researchers and practitioners. A commonly cited formal definition of machine learning, proposed by a computer scientist (Lantz, 2013, p. 10) explained that a machine is said to learn if it is able to take experience and utilize it such that its performance improves up on a similar experience in the future. More formally, according to Mitchell (1997), a computer program is said to learn from experience E with respect to some class of tasks T

* Corresponding author.
E-mail addresses: hanh_vn90@yahoo.com (H.H. Le), jean-laurent.viviani@univ-rennes1.fr (J.-L. Viviani).

http://dx.doi.org/10.1016/j.ribaf.2017.07.104
Received 17 February 2017; Received in revised form 3 June 2017; Accepted 4 July 2017
Available online 13 July 2017

and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. In this study, we want to examine the effectiveness of these methods compared with more traditional statistical techniques.

The main contributions of the paper are the following. First it proposes a comparison of traditional statistical techniques: Linear Discriminant analysis (LDA) and Logistic regressions (Logistic) to machine learning techniques on predicting the failure of US banks. Machine learning techniques (k- Nearest neighbors (k-NNs), artificial neural networks (ANN) and Support Vector Machines (SVMs)) had not been systematically compared to predict bank failure. Moreover, the empirical results of previous studies are unclear (see the recent paper of López and Pastor Sanz, (2015)). Secondly, we use a large number of financial ratios (31) for 5 years before banks become inactive (inactivity can be due to bankruptcy, illiquidity, merging, or insolvency). The large number of ratios provides a means of covering all bank financial characteristics: loan quality, capital quality, operations, profitability and liquidity and of determining the ratios with the best failure prediction power. This diversity is justified by our analysis of the "Material loss review" of 102 banks from FDIC reports since 2009–2015, which shows that banks fail for various reasons (loan problems, profit reduction, credit risk, ineffective board of directors and management). Thirdly, we test these various techniques on a large sample of 3000 US banks (1438 failed, 1532 active) during the crisis and post-crisis period (2008–2014). This period deserves an in-depth study due to the change in financial environment and banking techniques: fall of real estate prices, biased pricing methods, new financial products and risks (Demyanyk and Hasan, 2010; López and Pastor Sanz, 2015), the reasons for bank failures could be different (or not) from those previously observed. Better knowledge of bank failure determinants is also very important for regulators (in the Basel 3, 4 reforms perspective).

This paper is constructed as follows: part 2 introduces a literature review. The methodologies are presented more detail in part 3. Part 4 is about the data and variables. The final result is mentioned in part 5. Finally, conclusion and discussion are included in part 6.

2. Literature review on bank failure prediction

Taking into account the fact that bankruptcy prediction is an important and widely studied topic, we will concentrate our literature review on the prediction of bank failure using financial ratios.[1] Moreover, we will focus on the studies that implement at least one of the five techniques compared in this paper. The history of bankruptcy prediction originated from predicting the failure of businesses. The important contribution of Altman (1968) motivated researchers to use multivariate analysis to predict the bankruptcy of firms. He provided an original Z-score formula (1968) and showed its advantage by analyzing five main financial and economic aspects of a firm: the liquidity, size dimensions; operating efficiency and profitability of the assets, financial leverage as well as considering the capability of management in dealing with competitive conditions (total asset turnover). Sinkey (1975) employed discriminant analysis to predict bank failures.

In comparison, Martin (1977) and Ohlson (1980) employed logistic regression to predict failures of firms and bank. Martin (1977) attempted to predict the US commercial bank failure within 2 years during 1970 and 1976 by using 25 financial ratios of asset risk, liquidity, capital adequacy and earning. He suggested that logistic regression has a higher percentage of correctly classified than linear discriminant. Since these initial studies, empirical studies have been conducted to compare the prediction accuracy of these two approaches (Boyacioglu et al., 2009).

Nevertheless empirical studies do not demonstrate a clear advantage for one of the two main traditional techniques: discriminant analysis versus Logit and Probit models. But Canbas et al. (2005) on a sample of 40 privately owned Turkish commercial banks showed, using 49 ratios, that discriminant analysis obtains slightly better results than Probit and Tobit models. On the same vein, a recent study Chiaramonte et al. (2015) revealed, on a big sample of 3242 banks across 12 European countries, that Z-score is a good predictive model to identify banks in distress (better than the Probit model) and also has the great advantage of simple calculation. According to the empirical study by Lo (1986), the equivalence between LDA and LR may not be rejected.

However, in some standpoints, statistical techniques are no longer preferred in view of their relatively low accuracy (Ravi Kumar and Ravi, 2007). The attention to and confidence in machine learning has increased enormously during the past 5–10 years. Numerous studies suggest that intelligent techniques perform more effectively than traditional statistical techniques. The main difference between intelligent and statistical techniques is that statistical techniques usually require researchers to define the structures of the model a priori, and then to estimate parameters of the model to fit the data with observations, while with intelligent techniques the particular structure of the model is learned directly from the data (Wang et al., 2015). Moreover, the statistical analysis depends on strict assumptions (normal distribution, no correlations between independent variables), that can result in poor prediction accuracy.

Among several machine-learning methods, the artificial neural network seems to be the most favored tool in prediction issues. Ky (1991) was among the first to implement a neural network on 118 banks (59 failed and 59 non-failed banks) in Texas during 1985–1987, and indicated that the neural network performed more effectively than other methods (Discriminant Analysis, factor-logistic, k-NNs and Decision tree). Several studies (Miguel et al., 1993; Bell, 1997; Olmeda and Fernandez, 1997; Swicegood and Clark, 2001; Aktas et al., 2003; Wu and Wang, 2000) compare ANN and the classical statistical techniques (Discriminant Analysis and Logistic Model) to predict bank failure. They generally conclude in the superiority of the neural network approach. In their survey Vellido et al. (1999), also suggest that ANN is better than the logit model for predicting commercial bank failures. More recently Lee and Choi (2013) compared the prediction accuracy of neural networks and linear discriminant analysis on a sample of Korean companies. Their results indicated that the bankruptcy prediction accuracy using neural networks is greater than that of LDA. Finally,

[1] Readers can refer to Ravi Kumar and Ravi (2007) and Fethi and Pasiouras (2009) for a broader and more in depth literature review.

a meta-analysis performed by Adya and Collopy (1998) reveals that neural networks outperformed alternative approaches in 19 out of the 22 analyzed studies.

Unlike Neural networks and Support Vector Machines, the k-nearest neighbor algorithm is not implemented widely in finance. This technique is implemented widely in biological and transportation fields. This method, however, can function appreciatively and obtain high prediction accuracy. Min and Lee (2005) proposed support Vector Machines for bankruptcy prediction. Boyacioglu et al. (2009) examined ANNs, SVMs and multivariate statistical methods to predict the failure of 65 Turkish financial banks. 20 financial ratios belonging to 6 main groups were chosen: Capital adequacy, Asset quality, Management, Earning, Liquidity and the sensitivity to the market risk. Overall, the result proved that SVMs achieved the highest accuracy. They concluded that this method outperforms neural network, discriminant analysis and logit methods. SVM was also proved to work better than neural networks through the research of Chiaramonte et al. (2015) for a sample of 3242 EU banks. Park and Han (2002) used k-nearest neighbor for company bankruptcy prediction but we do not find empirical studies specifically dedicated to the use of k-nearest neighbor to predict bank failure.

Finally some empirical studies compare the various predictions methods. Tam and Kiang (1992) compare discriminant analysis, Logit, k-nearest neighbor and artificial neural networks on bank failure prediction and find that the latter outperforms the other techniques. Martínez (1996) compares neural network back propagation methods with discriminant analysis, logit analysis and the k-nearest neighbor for a sample of Texan banks and concludes that the first set of methods outperforms. Zhao et al. (2009) compare Logit, ANN and k-NN. They find that ANN > Logit > k-NN when financial ratios rather than row data are used. These studies support neural networks as being the best methods of predicting bank failure. Serrano-Cinca and Gutierrez-Nieto (2011) compared 9 different methods to predict the bankruptcy of USA banks during the financial crisis, including Logistic Regression, Linear Discriminant Analysis, Support vectors Machines, k-nearest neighbor and Neural Networks. It can be concluded that no technique is clearly better than the others. Performance depends on the performance measure chosen; some techniques have more accuracy but less recall (1 minus Type II error rate).

Among numerous studies on predicting the bankruptcy of banks, history has shown that intelligent techniques (and specifically artificial neural networks) seem to work more effectively than statistical techniques. This study will execute both families of techniques in different methods and in a new attempt to make a comparison on two aspects: the accuracy and the importance of each ratio.

3. Methodology

3.1. Statistical techniques

Linear discriminant analysis and Logistic regression are popular methods for classifying objects based on their characteristics. These methods have been applied widely to predict the failure of firms and banks.

3.1.1. Linear discriminant analysis (LDA) and logistic regression (LR)

Logistic regression (LR) is a regression model where the outcome is categorical (in our case the bank is active or inactive). More technically, the model assumes a linear relationship between the logarithm of the odds ratio (ratio of probabilities, see equation below) and one or more independent variables (bank characteristics, x_j).

$$g(x) = ln\frac{P(y=1)}{P(y=0)} = \sum_{j=1}^{m} \beta_j x_j + \beta_0$$

Linear Discriminant Analysis (LDA) derives a linear combination of ratios which best discriminate between failed and non-failed firms. Observations are assigned to one of the two groups in some 'optimal' way, for example, so as to minimize the probability or cost of misclassification. Logistic is often preferred to LDA as it is more flexible in assumptions and types of data that can be analyzed.

Canbas et al. (2005) propose an integrated model that combines LDA and LR in order to help predict bank failure. They demonstrate that this combination improves the prediction accuracy. Serrano-Cinca and Gutierrez-Nieto (2011) combine LDA with Partial Least Square analysis in order to predict the failure of US banks during the 2008 financial crisis.

Although the LDA and LR have become the most commonly used in bankruptcy prediction, their inherent drawbacks of statistical assumptions such as linearity, normality and independence among variables have constrained their practical applications (Lee and Choi, 2013). To solve the limitation of a linear approach, intelligent techniques (in this paper considered as machine learning approaches) achieve a forward movement by introducing nonlinear separation between groups.

Several methods have been implemented to classify companies or financial institutions and predict bankruptcy or failure. In this paper, three machine learning algorithms are applied: k-Nearest Neighbors, Artificial Neural Network and Support Vector Machines. Neural network is a well-known model and is considered to be one of the most powerful tools in prediction even when their conceptions are not easy to be translated. These models are referred to as black box processed because the mechanism that transforms the input into the output is obfuscated by a figurative box. On the contrary, k-nearest neighbors is regarded as a lazy learning technique (meaning that generalization beyond the training data is delayed until a query is made to the system). The idea is to classify unlabeled examples by assigning to them the class of the majority of its neighbors. Support Vector Machines are in between since, being not overly-complex, it is possible to enter into the black box.

3.1.2. K-Nearest Neighbor (k-NN)

The k-Nearest Neighbor (k-NN) is an instance-based method, meaning that it assigns a new case to the majority class among the k-closest cases in the training set (Hand et al., 2001). In a brief description, nearest neighbor classifies by mapping the different characteristics of the dataset closely to different label groups, the given data with common features will then be placed in the same group. Each new case is classified based on the outcome of the majority of its neighbors.

Each bank is represented by a vector of its characteristics. Banks with similar characteristics tend to be placed closely together. The distance between each point to each group must be calculated in order to find out which group (active or inactive) the banks belong to. If the majority of neighbors of a given bank are classified as failed (active), this bank will be classified as failed (active).

There are three major decisions in the k-NN method: the set of stored cases, the distance metric used to compute the distance between cases, and the value of k (Weiss and Indurkhya, 1998).

There are several ways of calculating this distance. Traditionally, the k-NN algorithm deploys Euclidean distance. If p and q are two vectors of characteristics (two banks), each of them has n features. The Euclidian distance between p and q is calculated as:

$$\text{Dist(p, q)}=\sqrt{(p_1 - q_1)^2+...+(p_n - q_n)^2}$$

To classify the bank as active or inactive, we should begin by assigning the number of neighbors, k. We can select any value of k to find the best grouping method. There are divergent hypotheses on selecting the 'best' k. Some researchers suppose that k should be the square root of number of features. However, others assume that k performs the best if it is between (2, 10). In this research, we experiment with various value of 'k' to find the optimal value.

3.1.3. Artificial neural networks

Taking advantages of computer potential, Artificial Neural Networks (ANNs) are inspired by biological neural networks. ANN is applied widely on a variety of tasks such as: computer vision, speech recognition, etc. ANN is a machine learning technique, which can simulate any relationship. Although ANN is not the only technique that can do this it is often preferred for the ability to obtain a solution in a reasonable time.

The idea is to learn from examples using several algorithms just as a human being learns new things. The advantages of ANNs are their flexible nonlinear modeling capability, strong adaptability, as well as their learning and massive parallel computing abilities (Ticknor, 2013). However, they cannot explain the causal relationship among variables, which restricts its application to managerial problems (Lee and Choi, 2013).

A fully connected network includes series of neuron layers. While each unit in the same layer cannot interconnect, each layer can. The connection between one unit in a given layer and another in the following layer is represented by a number call a weight, which can be positive or negative. There are two ways to transfer information: feed-forwarding and back-propagation. Feed-forwarding will forward information from input layer to output layer and this processing can lead to a wrong result. However, back-propagation can fix these errors by sending back the information to optimize the outcome.

When designing a multilayer network, the decision on choosing the number of hidden layers is very important. Lee et al. (2005) and Zhang et al. (1999) show that one hidden layer is sufficient for most classification problems. Meanwhile, Vasu and Ravi (2011) suggested choosing 2 hidden layers in order to be sure that the network architecture will be sufficiently complex to cope with the complexity of bank failure prediction. In our study, we applied for both 1 hidden layer and 2 hidden layers to examine which one performs better.

To eliminate the possibility of being linear, we use an activation function which creates a non-linear decision boundary. Various types of activation function exist, for example: sigmoid, tanh, rectified linear unit, leaky rectified linear unit or max out. We decided to use a sigmoid function since its characteristics are suitable for our output. It is the most widely used function.

3.1.4. Support Vector Machines (SVM)

The great advantage of Support Vector Machines (SVM) is that they combine the strengths of theory-driven conventional statistical methods and data-driven machine learning methods (Min and Lee, 2005). The method is based on the Vapnik's (1995) structural risk minimization principle. SVM is highly appreciated for successful applications in many fields such as bioinformatics, text, image recognition, etc. SVMs are supervised learning models that analyze data used for classification and regression analysis. This method is developed from Statistical Learning Theory (Boser et al., 1992). The basic idea is that input vectors (a vector represents the financial characteristics of a given bank) are non-linearly mapped to a very high dimension feature space. A linear decision surface is constructed in this feature space thus SVMs transform complex problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions.

Unlike numerous other methods which focus on whole training data, SVM pays attention to the most difficult to recognize data point based on the idea that if SVMs can figure out the toughest points, the others will be seen easily. The vectors most difficult to recognize are located close to the hyperplane separating active and failed banks, they are called support vectors. These points can be easily misclassified. The distance from the closest data points in each respective class to the hyperplane is called the margin. SVM will attempt to maximize these margins, so that the hyperplane is at the same distance from the 2 groups (failed and active banks). Intuitively, the more distant vectors are from the hyperplane, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

3.2. Implementation of statistical techniques

We use WEKA software to apply the listed statistical methods. WEKA is a collection of algorithms for data mining tasks. It can be run quickly for a big database. For more detailed information about the Weka package, the reader is referred to Witten and Frank (2005).

Firstly, 70% of the data from the 6-year period (5 year before failure plus the failure year) will be used on WEKA for training. The remaining 30% (5400 observations) will be used in order to test the prediction accuracy of each model. For each bank, we then checked whether it is correctly classified in the right group (active or inactive) at the right time (how many years before being inactive).

For the k-NN method, we tested several values of k to determine the appropriate value. As mentioned in part 3, there are several hypotheses on selecting k, for example: k should be the square root of total number of observation (in this paper are 43) or k should be between 1 and 10. We then tested both values of k from 1 to 10 and 43 For ANNs method, we set the hidden layer is equal to 1 or 2. The default training time is 500 on WEKA.

3.3. Data and variables

Initially, we collected over 5000 banks from Bankscope database. However, we set the condition that the number of inactive and active banks should be equivalent to test the performance of the machine learning approach. Finally, therefore, we select randomly a sample of 3000 banks including 1438 inactive and 1562 active banks. 6 year-periods include: year when banks go bankrupt and 5 years before being inactive was selected. Active banks were selected randomly with the criterion of being a US bank and still active until the first quarter of 2016.

After collecting and importing data in panel, we shuffled the order. Shuffling data is important to prevent bias learning process and predict more intelligently and in an integrated way. We then, divide the dataset into 2 subsets: a Training set (70% of data) and a Test set (30% of data). Theoretically, each method will learn 70% first training to create significant models. The remaining 30% will examine accuracy.

From bank financial statements we extract or construct 31 ratios. Zhao et al. (2009) demonstrate that the use of financial ratios, instead of raw accounting variables, significantly improves the performance of prediction techniques. Detailed accounting information is taken to forecast the status of banks and provide more adequate points of view. Ratios were selected by comparison with the lists of ratios used in previous empirical studies. Before 2007, these lists were presented in detail in the review by Ravi Kumar and Ravi (2007), after 2007 we take into account the lists presented in the papers referred in the literature review section.

Finally, the selected ratios cover: (i) loan quality, (ii) capital quality, (iii) operation efficiency, (iv) profitability and (v) liquidity. To shorten the name, we label each ratio by Z from Z1 to Z31. Each of these financial ratios is expected to have a strong influence on bank performances as well as possibly helping to predict the failure. For 31 ratios, we have dissimilar expectation signs on the bank's survival. Positive signs suggest that the higher the ratio the better the influence on the bank's survival. Negative signs indicate the contrary (Table 1).

4. Empirical results

4.1. Descriptive statistics

Table 2 presents the means and standard deviations of the 31 selected financial ratios for active and failed banks groups for one year before becoming inactive. As in Canbas et al. (2005) the last two columns present the F-test for the equality of means among the two groups and the significance levels. We find that **25 of the 31 ratios** have a significant different mean (for failed and non-failed banks) at a level than 5%. Hence, the null hypothesis that the two group means are equal is rejected at the 5% significance level for these ratios. We find that one year before failure, the loan quality is significantly lower for inactive banks (especially Z2 and Z4). Equity can be seen as a general buffer against risk and we observe that these banks have less equity whatever the measure of equity and the comparison point (assets, loans, liability) (see Z6, Z7, Z8, Z12, Z15). Operational efficiency is also lower for inactive banks compared with active banks (Z19, Z20). However, contrary to expectations, liquidity is higher for inactive banks (Z29, Z30, Z31), possibly because the banks in the sample became inactive for solvability problems rather than for liquidity problems. Note that our results are quite similar to those of Canbas et al. (2005). On a sample of 40 Turkish banks during the period 1997–2003, they find that the ratios that are the most different between failed and active banks are interest expenses on assets and interest income on interest expenses, equity/TA, liquid assets total assets, standard capital ratio. López and Pastor Sanz (2015) employed data from the FDIC between 2002 and 2012, their results state that failed banks are more concentrated in real estate loans and have more provisions.

4.2. Comparison of accuracy

To analyze in detail the predictive performance of each method, we use several indicators (see Powers (2011) for more details on these measures). Precision is the fraction of those predicted positive by the model that are actually positive. Recall, also referred to as sensitivity, is the fraction of those that are actually positive which were predicted positive. F-measure is the harmonic mean of precision and sensitivity. The value of F-measure ranges from 0 to 1. A value of 1 indicates perfect prediction. MCC (Matthews Correlation Coefficient) is the measurement of the quality of binary classification. This indicator was first introduced by the

Table 1
Expected sign of ratios on bank's survival.

Variables	Variables description	Expected sign
	Loan quality	
Z1	Loan Loss reserve/Gross Loans	Negative
Z2	Loan Loss provision/Net interest revenue	Negative
Z3	Impaired Loans/Gross Loans	Negative
Z4	Net charge off/Average Gross Loans	Negative
Z5	Impaired Loans/Equity	Negative
	Capital quality	
Z6	Tier 1 capital ratio	Positive
Z7	Total capital ratio	Positive
Z8	Equity/Total assets	Positive
Z9	Equity/Net Loans	Positive
Z10	Equity/Customer & short term funding	Positive
Z11	Equity/Liabilities	Positive
Z12	Capital funds/Total assets	Positive
Z13	Capital funds/Net loans	Positive
Z14	Capital funds/Deposit & Short term funding	Positive
Z15	Capital funds/Liabilities	Positive
	Operations	
Z16	Net interest margin	Positive
Z17	Net interest revenue/Average Assets	Positive
Z18	Other Operation income/Average Assets	Positive
Z19	Non-Interest expense/Average Assets	Negative
Z20	Pre-tax Operating Income/Average Assets	Positive
Z21	Non-Operating Items & taxes/Average Assets	Negative
	Profitability	
Z22	Return on Average Assets	Positive
Z23	Return on Average Equity	Positive
Z24	Inc. Net of Dist/Average Equity	Positive
Z25	Cost to Income Ratio	Negative
Z26	Recurring Earning Power	Positive
	Liquidity	
Z27	Net Loans/Total Asset	Negative
Z28	Net loans/Deposit & Short term funding	Negative
Z29	Net Loans/Total Deposit and Borrowing	Negative
Z30	Liquid Assets/Deposit & Short term Funding	Positive
Z31	Liquid Assets/Total Deposit & Borrowing	Positive

A positive sign indicates that when the ratio increases, the probability to fail decreases.

biochemist Brian Matthews in 1975. The value of MCC is between $[-1, 1]$. MMC = 1 indicates a prefect prediction; MMC = 0 indicates that the prediction is not better than random prediction; MMC = −1 indicates disagreement between prediction and observation. ROC Area (Receiver Operation Characteristic curve) is usually used for a binary classifier with the value between [0, 1]. This curve is created with y-axis is true positive rate, and x-axis is false positive rate. The closer to 1 the values of ROC are, the better the prediction. PRC Area (Precision/Recall plots): this indicator is used less frequently than others. However, Saito and Rehmsmeier (2015) suggested that PRC is more informative than a ROC plot when evaluating binary classifiers. PRC plots evaluate the fraction of true positives among positive predictions and hence can provide an accurate prediction of future classification performance.

4.2.1. Choice of parameters

Firstly, we made a decision on choosing the number of hidden layer for ANNs and the number of neighbors for the k-NN method. Regarding ANN methods, we test whether the number of hidden layers is 1 or 2. The result in Table 3 shows that for 1 or 2 hidden layers, the difference is small. Overall, with 1 and 2 hidden layers, ANNs can recall 74.4% and 75% respectively and the precision ratio is 75.7% and 75.7% respectively. Consequently, we may conclude as in Lee et al. (2005) and Zhang et al. (1999) that using 1 hidden layer is sufficient. Finally, we will use the result from ANNs with 2 hidden layers to compare with other methods.

For the k-NN method, we also implemented various values of *k* in order to try to find the 'best k'. The first assumption is that k should equal the square root of the total number of observations, which is 43. The second assumption is that k should be between 1 and 10. The first one brings only 72.9% precision, while the others obtained around 74% (see Table 4). We therefore state that the number of k-nearest neighbors should be between 1 and 10. In this case, we choose the k with the greatest precision which is k = 8 and denote it 8_NN.

4.2.2. Comparison of the five bank failure prediction methods

Following Vasu and Ravi (2011) we decompose the accuracy ratio into two dimensions: false positive (Type I error, the classifier

Table 2
Descriptive statics for the 31 financial ratios for active and failed banks – one year before being inactive.

Ratio	Inactive banks		Active banks		F	Sig.
	Mean	SD	Mean	SD		
Z1	1.60	0.944	1.45	0.913	20.058	0.000
Z2	13.45	24.780	4.09	8.065	200.135	0.000
Z3	2.03	3.015	2.00	2.115	0.115	0.735
Z4	0.53	1.077	0.19	0.654	110.766	0.000
Z5	14.88	24.632	12.55	14.892	10.005	0.002
Z6	13.17	4.952	15.14	6.743	81.813	0.000
Z7	14.53	4.827	16.40	6.713	75.745	0.000
Z8	10.09	3.213	11.10	3.398	70.150	0.000
Z9	16.61	8.880	18.29	13.153	16.502	0.000
Z10	12.26	4.964	13.41	7.119	25.630	0.000
Z11	11.40	4.295	12.71	5.188	56.090	0.000
Z12	10.37	3.221	11.48	3.313	85.988	0.000
Z13	17.08	9.010	18.90	13.129	19.327	0.000
Z14	12.63	5.148	13.87	7.186	29.358	0.000
Z15	11.73	4.326	13.14	5.132	66.062	0.000
Z16	3.95	1.088	3.76	1.436	16.040	0.000
Z17	3.52	0.967	3.35	1.213	17.886	0.000
Z18	1.00	1.190	1.12	1.474	6.184	0.013
Z19	3.52	1.810	3.18	1.605	30.231	0.000
Z20	1.00	1.663	1.30	0.998	38.089	0.000
Z21	−0.29	0.622	−0.24	0.572	4.716	0.030
Z22	0.67	1.238	0.98	0.780	67.328	0.000
Z23	7.28	11.764	8.83	6.525	20.391	0.000
Z24	1.88	11.626	5.27	5.961	103.154	0.000
Z25	69.85	32.613	67.83	14.711	4.909	0.027
Z26	1.46	1.416	1.48	1.260	0.081	0.777
Z27	65.16	13.438	65.97	13.338	2.763	0.097
Z28	78.41	20.623	78.06	17.342	0.263	0.608
Z29	73.60	15.244	75.59	15.583	12.394	0.000
Z30	9.72	10.426	7.71	8.077	35.414	0.000
Z31	9.24	9.080	7.50	7.843	31.809	0.000

Table presents the means and standard deviations (SD) of the 31 ratios (Z1 to Z31) used to compare active and inactive banks. The F-test (F) is used for comparison of means. The p-value for the F-test (Sig.) is given in the last column.

Table 3
The comparison of ANNs 1 hidden layer and 2 hidden layers.

Method	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ANNs_1	0.758	0.741	0.734	0.493	0.771	0.739
ANNs_2	0.757	0.753	0.75	0.506	0.819	0.803

Table gives the accuracy measures for ANN with 1 hidden layer (ANNs_1) and two hidden layers (ANN_2). **Precision** is the fraction of those predicted positive that are actually positive. **Recall** is the fraction of those that are actually positive which were predicted positive. **F-measure** is the harmonic mean of precision and sensitivity. **MCC**: Matthews correlation coefficient. **ROC area**: Receiver Operation Characteristic curve. **PRC Area**: Precision/Recall plot.

misclassifies an actual active bank as a failed bank, FP in Table 5) and false negative (Type II error, the classifier misclassifies a failed bank as an active bank, 1-TP rate in Table 5). Note that for banks false negative is considered by banking regulators to be far more costly than false positive. As foreseeable from the literature review and from the previous results, Logistic and ANN obtain the best performance with the lowest values of type 1 and type 2 errors for both ratios. Table 5 highlighted that ANNs performed better than all the other methods whatever the performance measure (75.3% of TP rate and 25.9% of FP rate, which lead to 75.7% precision and 75.3% recall). As can be seen, k-NN and LR achieved similar results: around 74% precision. SVM and LDA obtain the lowest performance with only 71.6% and 72% precision respectively (TP). The distance among these results is not too significant, however we notice that the traditional logistic approach can predict more accurately than some of the machine learning approaches (as already observed by Zhao et al., 2009).

We then extract the result into years as in Table 6 to summarize the total errors of each method by year. The error here is defined when the active is classified as inactive and vice versa. ANNs make fewer errors than the others in the year that banks go inactive, and make the most errors 3 years before. Meanwhile, other methods can make errors evenly over the years. As noticed from the previous comment, SVMs make the most mistakes for most of every year. Surprisingly, the maximum number of incorrect classification occurred at the year or one year before failure.

We investigate to find out which method can recognize the failure when the other methods predict wrongly (Table 7). This

Table 4
Comparison of k-NNs with different number of neighbors.

Method	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
1_NN	0.731	0.731	0.73	0.459	0.728	0.668
2_NN	0.74	0.712	0.698	0.442	0.774	0.719
3_NN	0.741	0.74	0.739	0.477	0.791	0.743
4_NN	0.736	0.722	0.714	0.451	0.8	0.759
5_NN	0.736	0.736	0.734	0.469	0.804	0.768
6_NN	0.738	0.727	0.721	0.459	0.808	0.775
7_NN	0.741	0.739	0.737	0.476	0.81	0.78
8_NN	0.741	0.731	0.725	0.467	0.811	0.783
9_NN	0.741	0.739	0.737	0.476	0.81	0.783
10_NN	0.739	0.73	0.724	0.464	0.812	0.787
43_NN	0.729	0.724	0.72	0.448	0.806	0.798

Table gives the accuracy measures of k-NNs with a number of neighbors from 1 (1_NN) to 43 (43_NN). **Precision** is the fraction of those predicted positive that are actually positive. **Recall** is the fraction of those that are actually positive which were predicted positive. **F-measure** is the harmonic mean of precision and sensitivity. **MCC**: Matthews correlation coefficient. **ROC area**: Receiver Operation Characteristic curve. **PRC Area**: Precision/Recall plots.

Table 5
Performance of bank failure prediction methods.

Method	Confusion matrix		Precision	Recall	ROC Area	PRC Area
ANNs_2	2415	444	75.7%	75.3%	81.9%	80.3%
	892	1649				
8_NN	2455	404	74.1%	73.1%	81.1%	78.3%
	1048	1493				
LDA	2185	674	72.0%	72.0%	77.6%	75.8%
	836	1705				
LR	2235	624	73.9%	73.9%	79.6%	77.3%
	785	1756				
SVM	2121	738	71.6%	71.6%	71.5%	65.5%
	794	1747				

Table gives the accuracy measures for the five bank failure prediction techniques: ANN with two hidden layers (ANNs_2), k-NN with 8 neighbors (8_NN), Linear discriminant analysis (LDA), Logistic Regression (LR), Support Vector Machine, (SVM). **Precision** is the fraction of those predicted positive that are actually positive. **Recall** is the fraction of those that are actually positive which were predicted positive. **ROC area**: Receiver Operation Characteristic curve. **PRC Area**: Precision/Recall plots.

Table 6
Number of banks misclassified by year for the 5 predictions techniques.

Year	ANNs_2	KNN_8	LDA	Logistic	SVM	Total
0	199	252	285	262	295	1293
1	231	252	272	255	279	1290
2	220	258	248	212	225	1165
3	240	261	258	238	279	1279
4	234	213	218	221	219	1109
5	212	216	229	221	235	1118
Total	1336	1452	1510	1409	1532	

Table 7
Right when other methods are wrong.

Methods	ANNs_2	8_NN	LDA	Logistic	SVM
Total	469	92	3	14	12
Year 0	73	24	1	3	2
Year 1	85	24	0	2	1
Year 2	94	10	0	3	1
Year 3	82	15	0	0	1
Year 4	67	9	1	4	5
Year 5	68	10	1	2	2

Number of failed banks that one method can detect when all the others cannot.

criterion is important and not used widely to date. Our purpose is to observe how dominant the method is. Surprisingly, ANNs can recognize 469 instances while the other methods cannot. After ANNs, k-NNs can also predict 92 observations while the other methods cannot. However, LDA can recognize only 3 banks, which is very poor.

5. Conclusion

This paper proposes an empirical study on the prediction of bank failure through 2 approaches: machine learning and two traditional statistical approaches. We observed firstly that machine learning, ANNs and k-NN methods perform more effectively than traditional methods. However, the difference in prediction accuracy between ANNs and k-NN methods and the traditional logistic regression method is not very big. In addition, we observed that SVM does not perform better than traditional methods. Nevertheless, ANN and k-nearest neighbor demonstrate their remarkable ability when they can detect the failure correctly but the other methods cannot.

All 31 ratios are important to predict bank failures. Each group has at least one significant ratio that affects the survival of the banks. Among them, three groups play a more important role, namely operation efficiency, profitability and liquidity. Notably, the ratios Z3 (Impaired Loans/Gross Loans), Z6 (Tier 1 capital ratio), Z12 (Capital funds/Total assets), Z18 (Other Operation Income/Average Assets), Z17 (Net interest revenue/Average Assets), Z21 (Non Operation Items&taxes/Average Assets), Z22 (Return on Average Assets), Z25 (Cost to income ratio) Z27 (Net Loans/Total Asset), Z28 (Net loans/Deposit &Short Term funding) and Z29 (Net Loans/Total Deposit& Borrowing) are more relevant than the others.

Our results have important institutional and policy implications. In effect, banks and bank supervisors developed early warning systems to prevent individual bank failure and banking crisis. Sahajwala and Van den Bergh (2000) provide an overview of the different approaches that are being used or developed in this field. Our study can help banks and bank supervisors to design such early warning systems because it shows that the traditional logistic regression models perform quite well and they can be complemented by machine learning techniques (ANNs and k-NN) to detect the most difficult cases. Moreover, these methods are based on ratios analysis and our study provides some information on the financial ratios that could help to better predict bank failures.

The limitation of this study is that we emphasize accounting information and ignore bank market data. Moreover, we could not determine the role of each ratio in machine learning techniques.

References

Adya, M., Collopy, F., 1998. How effective are neural networks at forecasting and prediction?: A review and evaluation. Int. J. Forecast. 17 (5–6), 488–495.

Aktas, R., Doganay, M., Yildiz, B., 2003. Predicting the financial failure: a comparison of statistical methods and neural networks. Ankara Univ. J. SBF 58, 1–24.

Altman, E.I., 1968. Financial ratios: discriminant analysis and the prediction of corporate bankruptcy. J. Finance 23, 589–609.

Beaver, W.H., 1966. Financial ratios as predictors of failures. Emp. Res. Account. 4, 71–111.

Bell, T.B., 1997. Neural nets or logit model?: A comparison of each model's ability to predict commercial bank failures. Int. J. Intell. Syst. Account. Finance Manag. 6, 249–264.

Boser, B.E., Guyon, I., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Pittsburgh, ACM. Proceedings of the Fifth Annual Workshop of Computational Learning Theory 5. pp. 144–152.

Boyacioglu, M.A., Kara, Y., Baykan, O.K., 2009. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: a comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey. Expert Syst. Appl. 36, 3355–3366.

Canbas, S., Cabuk, A., Kilic, S.B., 2005. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: the Turkish case. Eur. J. Oper. Res. 166, 528–546.

Chiaramonte, L., Croci, E., Poli, F., 2015. Should we trust the Z-score? Evidence from the european banking industry. Global Finance J. 28, 111–131.

Demyanyk, Y., Hasan, I., 2010. Financial crises and bank failures: a review of prediction methods. Omega 38, 315–324.

Fethi, M.D., Pasiouras, F., 2009. Assessing Bank Performance with Operational Research and Artificial Intelligence Techniques: A Survey. University of Bath School of Management (Working Paper Series).

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT Press, Cambridge, MA.

Ky, T., 1991. Neural network models and the prediction of bank bankruptcy. Omega 19 (5), 429–445.

López, F.J., Pastor Sanz, I.I., 2015. Bankruptcy visualization and prediction using neural networks: a study of U.S. commercial banks. Expert Syst. Appl. 42, 2857–2869.

Lantz, B., 2013. Machine Learning with R. Packt Publishing Ltd.

Lee, S., Choi, W.S., 2013. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. Expert Syst. Appl. 40, 2941–2946.

Lee, K., Booth, D., Alam, P., 2005. A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. Expert Syst. Appl. 29, 1–16.

Lo, A.W., 1986. Logit versus discriminant analysis. a specification test and application to corporate bankruptcies. J. Econometr. 31 (2), 151–178.

Martínez, I., 1996. Forecasting company failure: neural approach versus discriminant analysis: an application to Spanish insurance companies. In: Sierra Molina, G., Bonsón Ponte, E. (Eds.), Intelligent Systems in Accounting and Finance, pp. 169–185 Huelva.

Martin, D., 1977. Early warning of bank failure: a logit regression approach. J. Bank. Finance 1 (3), 249–276.

Min, J.H., Lee, Y.C., 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst. Appl. 28, 603–614.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. 18, 109–131.

Olmeda, I., Fernandez, E., 1997. Hybrid classifiers for financial multicriteria decision making: the case of bankruptcy prediction. Computational Econ. 10, 317–335.

Park, C.-S., Han, I., 2002. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. Expert Syst. Appl. 23 (3), 255–264.

Powers, D., 2011. Evaluation: from precision, recall and F-Measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2 (1), 37–63.

Ravi Kumar, P., Ravi, V., 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. Eur. J. Oper. Res. 180, 1–28.

Sahajwala, R., Van den Bergh, P., 2000. Supervisory Risk Assessment and Early Warning Systems. Basel Committee on Banking Supervision Working Papers. 53p.

Saito, T., Rehmsmeier, M., 2015. The precision – recall plots is more informative than the ROC Plot when evaluating binary classifiers on Imbalanced datasets. Plos One J. 11/024, 1–22. http://dx.doi.org/10.1371/journal.pone.0118432.

Serrano-Cinca, C., Gutierrez-Nieto, B., 2011. Partial Least Square Discriminant Analysis (PLS-DA) for Bankruptcy Prediction. CEB Working Paper. (n° 11/024).

Sinkey, J.F., 1975. A multivariate analysis of the characteristics of problem banks. J. Finance 30, 21–36.

Swicegood, P., Clark, J.A., 2001. Off-site monitoring for predicting bank under performance: a comparison of neural networks, discriminant analysis and professional human judgement. Int. J. Intell. Syst. Account. Finance Manag. 10, 169–186.
Tam, K.Y., Kiang, M., 1992. Predicting bank failures: a neural network approach. Decis. Sci. 23, 926–947.
Ticknor, J., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Syst. Appl. 40 (14), 5501–5506.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Vasu, M., Ravi, V., 2011. Bankruptcy prediction in banks by principal component analysis threshold accepting trained wavelet neural network hybrid. In: Proceedings of the 7th International Conference on Data Mining. Las Vegas, July, 18–21.
Vellido, A., Lisboa, P., Vaughan, J., 1999. Neural networks in business: a survey of applications (1992–1998). Expert Syst. Appl. 17, 51–70.
Wang, G., Ma, J., Yang, S., 2015. Improved boosting based on feature selection for corporate bankruptcy prediction. Expert Syst. Appl. 41, 2353–2361.
Weiss, S.M., Indurkhya, N., 1998. Predictive Data Mining: A Pratical Guide. Morgan Kaufmann, San Francisco, CA.
Witten, I.H., Frank, E., 2005. Data Mining: Pratical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufamn, San Fransisco, CA.
Wu, C., Wang, X.M., 2000. A neural network approach for analyzing small business lending decisions. Rev. Quant. Finance Account. 15, 259–276.
Zhang, H., Hu, M.Y., Patuwo, B.E., Indro, D.C., 1999. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. Eur. J. Operational Res. 116, 16–32.
Zhao, H., Sinha, A.P., Ge, W., 2009. Effects of feature construction on classification performance: an empirical study in bank failure prediction. Expert Syst. Appl. 36 (2), 2633–2644.

ANNÉE 2018

VU :

La Directeur de Thèse

Jean-Laurent VIVIANI

VU :

La Responsable de l'école Doctorale

VU pour autorisation de soutenance

Rennes, le

Le président de l'Université de Rennes 1

David ALIS

VU après soutenance pour autorisation de publication

Le président de Jury

Mots clés : Banque, faillite, extraction de texte, apprentissage automatique

Résumé : La thèse se compose de six chapitres. Chaque chapitre peut être lu indépendamment des autres, mais les six chapitres partagent le thème général de la thèse : L'utilisation de techniques d'apprentissage automatique pour prédire, expliquer et prévenir les défaillances des banques. Le chapitre 1 résume les motivations et les contributions de la thèse. Le chapitre 2 présente la revue de la littérature scientifique. Le chapitre 3 compare la précision de deux approches qui tentent de prédire la défaillance des banques : les techniques statistiques traditionnelles et les techniques d'apprentissage automatique. Le chapitre 4 examine examen des pertes matérielles publiés par la Federal Deposit Insurance Corporation sur les banques américaines en faillite de 2008 à 2015 à l'aide de techniques de text mining. Le chapitre 5 examine l'efficacité de la provision pour pertes sur prêts des grandes banques américaines par le biais de l'analyse des enveloppes de données et des réseaux de neurones. Le chapitre 6 commente les principaux résultats et discute des orientations pour les recherches futures.

Keywords: Bank, Failure, Text Mining, Machine Learning

Abstract: The thesis consists of six chapters. Each chapter can be read independently of the others, but all six chapters share the thesis's overall topic: *Using Machine learning techniques to predict, explain and prevent the failure of banks.* Chapter 1 summarises the motivation and contribution of the thesis. Chapter 2 introduces a global review of the literature. Chapter 3 compares the accuracy of two approaches: traditional statistical techniques and machine learning techniques, which attempt to predict bank failure. Chapter 4 investigates the material loss review published by the Federal Deposit Insurance Corporation on the U.S. failed banks from 2008 to 2015 using text mining techniques. Chapter 5 investigates the efficiency of loan loss provision of large US banks via Data Envelopment Analysis and Neural networks. Chapter 6 remarks the main finding and discusses directions for future research.