COMPLEX MODEL VERSUS COMPLEX DATA IN AN APPLICATION OF PREDICTING MORTGAGE LOAN

Miss Yu Weng

An Independent Study Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Financial Engineering
Department of Banking and Finance
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2021

การเปรียบเทียบระหว่างแบบจำลองที่มีรูปแบบซับซ้อนกับแบบจำลองที่ใช้ข้อมูลที่มีความซับซ้อน ในการคาดการณ์สินเชื่ออสังหาริมทรัพย์

น.ส.หยู เหวง

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมการเงิน ภาควิชาการธนาคารและการเงิน
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Independent Study Title	COMPLEX MODEL VERSUS COMPLEX DATA IN AN APPLICATION OF PREDICTING MORTGAGE LOAN
By	Miss Yu Weng
Field of Study	Financial Engineering
Thesis Advisor	Dr. TANAWIT SAE SUE

Accepted by the FACULTY OF COMMERCE AND ACCOUNTANCY, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

INDEPENDENT STUDY COMMITTEE

.... Chairman
(Associate Professor Dr. THAISIRI WATEWAI)

.... Advisor
(Dr. TANAWIT SAE SUE)

.... Examiner
(Associate Professor Dr. SIRA SUCHINTABANDID)

หยู เหวง : การเปรียบเทียบระหว่างแบบจำลองที่มีรูปแบบซับซ้อนกับแบบจำลองที่ใช้ข้อมูลที่มีความซับซ้อนในการคาดการณ์สินเชื่ออสังหาริมทรัพย์. (COMPLEX MODEL VERSUS COMPLEX DATA IN AN APPLICATION OF PREDICTING MORTGAGE LOAN) อ.ที่ปรึกษาหลัก : ดร.ธนวิต แซ่ซือ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวัดประโยชน์ของการใช้แบบจำลองที่ซับซ้อนเปรียบเทียบกับการใช้ข้อมูลที่ซับซ้อน ในการประยุกต์ใช้คาดการณ์ความเสี่ยงด้านเครดิตสำหรับสินเชื่อจำนอง ในการศึกษาวิจัยนี้ แบบจำลองโครงข่ายระบบประสาทจะเป็นตัวแทนของแบบจำลองที่ซับซ้อน ในขณะที่แบบจำลองแบบการถดถอยเป็นตัวแทนของแบบจำลองอย่างง่ายที่ไม่ซับซ้อน ข้อมูลสองประเภท ได้แก่ ข้อมูลปกติและข้อมูลที่ซับซ้อน ได้ถูกนำมาใช้วิเคราะห์ โดยข้อมูลที่ซับซ้อนถูกสร้างมาจากการนำข้อมูลปกติมาผ่านเทคนิคการสกัดและการแปลงข้อมูลเพื่อให้ได้ค่าของข้อมูลที่ใกล้เคียงกับสิ่งที่ต้องการวัด ซึ่งตัวแปรที่สร้างมาเป็นส่วนหนึ่งของข้อมูลที่ซับซ้อนในงานวิจัยนี้คือ อัตราส่วนสินเชื่อต่อราคาบ้าน (Loan-to-value) และสัดส่วนค่าใช้จ่ายที่อยู่อาศัย (Housing expense ratio) ข้อมูลปกติมาจากชุดข้อมูลสินเชื่อครอบครัวเดี่ยวในสหรัฐอเมริกาตั้งแต่ปี ค.ศ. 2010 ถึงปีค.ศ. 2018 โดยมีแหล่งที่มาจากองค์กร Freddie Mac ผลจากการวิจัยโดยพิจารณาจากเมทริกซ์ความสับสนและความแม่นยำในการคาดการณ์ความเสี่ยงของสินเชื่อพบว่า ข้อมูลที่ซับซ้อนที่สร้างขึ้นสามารถช่วยให้แบบจำลองมีความแม่นยำเพิ่มขึ้น แต่ความแม่นยำไม่ได้เพิ่มขึ้นอย่างมากมายมหาศาล อีกทั้งประโยชน์ที่ได้รับเพิ่มเติมจากการใช้ข้อมูลที่ซับซ้อนในแบบจำลองที่ซับซ้อนมีเพียงเล็กน้อยเท่านั้น อย่างไรก็ตาม ผลลัพธ์จากการวิจัยชี้ให้เห็นว่า แบบจำลองที่ซับซ้อนมีประโยชน์อย่างมีนัยยะสำคัญในการคาดการณ์ความเสี่ยงของสินเชื่อ และสนับสนุนด้วยว่าแบบจำลองที่ซับซ้อนมีค่ามากกว่าข้อมูลที่ซับซ้อน

สาขาวิชา วิศวกรรมการเงิน ลายมือชื่อ
นิสิต ....
....
ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษา
หลัก ....

6484007226 : MAJOR FINANCIAL ENGINEERING

KEYWORD: Mortgage loan prediction, Logistic regression model, Neural network model, Loan-to-value, Housing expense ratio

Yu Weng : COMPLEX MODEL VERSUS COMPLEX DATA IN AN APPLICATION OF PREDICTING MORTGAGE LOAN . Advisor: Dr. TANAWIT SAE SUE

This research aims to measure the benefits of complex model versus those of meaningful information, through an application of credit risk prediction for mortgage loans. The neural network represents complex model and the regression model represents simple model. Two types of data are applied in this analysis: simple data and complex data. The complex data is obtained from the simple dataset using information extraction techniques and data transformation. The two specific variables constructed in our complex data are Loan-to-value and Housing Expense ratio. Applied to the monthly Single-Family Loan-Level Dataset of Freddie Mac from year 2010 to year 2018 in this experiment, the result of confusion matrix and accuracy metrics points out that the complex data constructed in this study can help model increase the accuracy, but it cannot have a huge boost. The added benefit of the complex data in both complex model and simple model is quite small. The result also points out that the complex model is more valuable than complex data.

Field of Study:	Financial Engineering	Student's Signature ....
Academic Year:	2021	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. TANAWIT SAE SUE as my supervisor, as well as Associate Professor Dr. THAISIRI WATEWAI and Associate Professor Dr. SIRA SUCHINTABANDID, for their guidance throughout this project.

Yu Weng

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I
INTRODUCTION

1.1 Background

Credit risk assessment plays an important role in maintaining the vigorous development of the financial market. Mortgage loans have been in existence for more than one hundred years. During the time, the mortgage loan system of various countries gradually became convergent, and then formed a banking system with mortgage as the core, and it was widely adopted by the banking industry in various countries. The mortgage loan system is mortgaged by real estate and takes into account the pledge guarantee. Furthermore, the mortgage loan system is related to both the banking industry and real estate, which may have a relatively large impact on the economy and makes it full of risks. It is precisely because of this the development and improvement of mortgage loan credit risk assessment has never stopped.

The economic crisis in 2008 further confirmed the importance of risk assessment. In 2000, the U.S. government encouraged Americans to buy houses, and the Federal Reserve had also adopted a loose monetary policy, constantly cutting interest rates, trying to use real estate to stimulate the U.S. economy and keep the U.S. economy prosperous. Commercial banks have also lowered standards to lend to people with low credit levels. In April 2007, the bankruptcy of New Century Finance, the second largest mortgage company in the United States, marked the beginning of the mortgage crisis in the United States. The crisis has led directly to the closure of more than 80 mortgage companies in the United States. It has also led to a sharp decline in major global capital markets, including the stock market, bond market, futures oil market and so on. The financial crisis has also warned other countries ‘mortgage market, such as China. There are many similarities between China's housing mortgage market and the US market. For example, the banking system generally underestimates the risk of housing mortgage loans, and the rapid rise of house prices has stimulated the vigorous development of personal housing mortgage loans (Zhang, 2010). How to estimate the risk correctly is a problem that needs to be focused on the world mortgage market.

The Basel Committee on Banking Supervision, being composed of Banking Supervisions of thirteen countries, issued the Basel III agreement, which requires innovation and improvement of risk measurement methods, aiming to improve the management, internal control and risk prevention of each bank itself.

Since last century, many researchers have made a lot of efforts to find a better credit risk assessment method. Some researchers came up with the comprehensive financial ratio analysis method, which uses financial indicators to monitor credit risk of borrowers. Other researchers tried to use statistical methods to derive risk discrimination models. They developed many complex sampling methods, likelihood estimation methods and simulation methods (i.e. Bayesian rule, Gibbs sampling, Kalman filter, Monte-Carlo simulation), and tried to use these complex models to explain the complex relationship behind the data, such as logistic regression models, probit models, liner discriminant analysis and so on. But most of the traditional statistical methods need to be based on assumptions, or need to be based on a prior, which depends largely on the experience of researchers (Greenland, 2003). In the 1990s, the theory of neural networks based on cognitive science and mathematical methods developed, with high parallel computing ability, self-learning ability and fault-tolerant ability. Altman, Marco and Varetto (1995) applied the neural network method in financial crisis of Italian companies. Altman (1995) concluded that "neural network analysis method is not substantially superior to linear model in credit risk identification and prediction" in a comparative study of the neural network method and discriminant analysis method. Bensic, Sarlija and Susac (2005) pointed out that the neural network is obviously superior to the traditional statistical model. Therefore, there are always doubts about whether the neural network is making a significant contribution in the prediction model or just a fashion, whether the neural network can automatically dig out the hidden information behind the raw data, and whether the more meaningful information obtained from complex statistical methods can be input into the neural network as input data to obtain a more accurate classifier. Moreover, the neural network works randomly. To get a good neural network, it needs human debugging, which costs a lot of manpower and time.

1.2 Research Objectives

This research aims to measure the benefits of complex model versus those of meaningful information, through an application of credit risk prediction for mortgage loans. This study uses the neural network to represent complex model and use the regression model to represent simple model. Two types of data are applied in this analysis: simple data and complex data. The simple data obtained by doing standardization and data preprocessing on the raw data. The complex data is obtained from the simple data using information extraction techniques and data transformation.

This study first investigates whether a complex model, i.e. neural network, with simple data can outperform a logistic regression model with complex data. The result would suggest whether the complexity brought by the model or the data has more value. The study also investigates the added benefit of the complex data in a complex model by measuring the improvement of using the complex data in the complex model over simple data in the complex model. Similarly, the result would show whether the complex data brings any benefit to complex models and whether a complex model needs the power of complex data to improve its performance. At the end, the study will reveal whether an effort should be used for implementing complex model or extracting meaningful information or both.

1.3 Research Questions

In this study the first research question is whether the complex models or complex data is more valuable for prediction problems? In order to answer this question, the test dataset is used to compare the prediction performance of two models in experiment (the simple model with complex data and complex model with simple data). The second research question is how much the added benefit of the processed data can be brought in a complex model is measured by comparing the performance of a neural network with processed data and the neural network model with raw data. The monthly Single-Family Loan-Level Dataset of Freddie Mac is used to test the hypothesis.

The rest of this paper is organized as follow, in Chapter 2, focused review of the literature on various credit risk assessment methods; Data, models, and methods are highlighted in Chapter 3; and in Chapter 4, the work that has been done in this paper is summarized.

CHAPTER II
LITERATURE REVIEW

Due to the importance of loan credit risk evaluation, researchers are very focused on quantitative research. They have developed a variety of statistical models and data mining tools for credit assessment in recent decades. The modern data mining techniques have made a significant contribution to the field of the credit scoring models. The methods of credit risk assessment can be divided into the subjective qualitative analysis method, traditional statistical analysis methods and machine leaning techniques methods.

The subjective qualitative analysis method is the main method of early credit risk assessment. The application of this method mainly depends on the subjective judgment of experts. Factor analysis is a typical representative of subjective analysis methods, including five Cs method, five Ps method and five Ws method. However, the judgment result of this subjective evaluation method varies from person to person and the credibility is not high.

Subsequently, quantitative analysis method---traditional statistical analysis methods were developed. The main representative methods at this stage are linear discriminant model (LDA) (Fisher, 1936; Altman, 1968; Desai et al, 1996; Caouette et al, 1998; Hand et al, 1998), logistic regression model (LR) (William, 1995; Lee & Jung, 2000; Baesens et al, 2003), probit model (Grablowsky & Talley, 1981; Pindyck & Rubinfeld, 1997; Maddala, 2001) and k-nearest neighbor model (KNN) (Henley & Hand, 1996). Altman (1968) used the financial indicators of different characteristics of US bankrupt companies and non-bankrupt enterprises to construct a discriminant function to judge the financial risks of enterprises and used multivariate discriminant models to assess corporate financial risks earlier. Based on this, the Z credit scoring model was constructed. Reichert, Cho & Wagner (1983) proposed that the linear discriminant technique requires strict data, that is, it is required the data must be the strict normal distribution, so that its practical application is also limited. The logistic regression model was first proposed by Ohlson (1981). Through the investigation of bankrupt enterprises and non-bankrupt enterprises, the logit model is used to assess the probability of actual default of the enterprises, and it is found that the evaluation results are more accurate than the multivariate discriminant model. Logistic regression model is highly interpretable which has a clear form, and the parameters represent the effect of each feature on the output. However, LR model is essentially a linear classifier, so it does not deal with the problem of feature correlation well. It is also easy to underfit, resulting in a bad performance. The probit model is similar to the logistic model and can be considered as an extension of the latter.

With the development of computer science and data mining technology, more advanced statistical nonparametric models and machine leaning methods are used in the field of credit assessment, such as decision tree (Davis, Edelman & Gammerman, 1992), neural network models (Desai, Crook, & Overstreet, 1996; Malhotra & Malhotra, 2002), genetic programming models (Ong, Huang & Tzeng, 2005), random forest and support vector machines (SVM) (Vapnik, 1995; Huang, H. Chen, Hsu, W. Chen & S. Wu, 2004). Duan (2019) came up with a multilayer perceptron (MLP) model with three hidden layers trained by the back-propagation algorithm, and the Synthetic Minority Over-Sampling Technique (SMOTE) is used to improve the deep neural networks (DNN) prediction accuracy. Desai, Crook & Overstreet (1996) explored the ability of neural networks such as multilayer perceptions and modular neural networks, and traditional techniques such as linear discriminant analysis and logistic regression, in building credit scoring models in the credit different union environment. They concluded that neural networks outperform LDA, and logistic regression is as good as neural networks. Bensic, Sarlija & Susac (2005) compared the accuracy of best models extracted by different methodologies, such as logistic regression, neural networks, and classification and regression tree (CART) decision trees. Four different neural network algorithms are tested, including backpropagation, radial basis function network, probabilistic and learning vector quantization. The result shows that the highest total hit rate, and the lowest type I error are obtained by the probabilistic neural network.

In general, these artificial intelligence methods achieved better performance than traditional statistical methods. For conventional statistical classification techniques, an underlying probability model must be assumed in order to calculate the posterior probability upon which the classification decision is made. The more recently developed data mining techniques such as neural networks, genetic programming (GP) and support vector machines (SVM) can perform the classification task without this limitation. But no method is perfect. Although the SVM method avoids the "dimension disaster", it is difficult to implement this algorithm for large training sample set, and it is difficult to solve the multi-class problem. Neural network has strong nonlinear fitting ability, can map arbitrarily complex nonlinear relationships, and has simple learning rules and is convenient for computer implementation. But the most serious problem is the inability to explain his own reasoning process and reasoning basis. Another weakness is that when there is not enough data, neural network can't work.

In order to avoid the weaknesses of various classification methods and to make a full use of their advantages, researchers try to use hybrid classifiers. Some combined classifiers, which integrate two or more single classification methods, have shown higher correctness of predictability than any individual methods. Combined classifier research is currently flourishing in credit risk assessment. Shu. T. Luo, Bor. W. Cheng & H. Hsieh (2009) concluded the performance of a new classifier clustering-launched classification (CLC) which combines clustering and SVM is better than the benchmark SVM method, and the CLC method can classify data efficiently, and only needs one parameter. The neuro-fuzzy system introduced by Malhotra & D. K. Malhotra (2002), combined fuzzy systems and neural networks to get a better default probability prediction model. Hsieh (2005) derived a hybrid mining approach in the design of an effective credit scoring model, based on clustering and neural network techniques. He not only designed an accurate classifier, but clustering techniques were also successfully employed to preprocess the input samples for the purpose of indicating unrepresentative samples. Huang, C. Chen & J. Wang (2007) used three strategies to construct a hybrid SVM-based credit scoring models and found that the SVM classifier achieved an identical classificatory accuracy with relatively few input features. A multistage neural network ensemble learning model is proposed by Yu, Y. Wang & K. Lai (2008). Different from commonly used ‘‘one-member-one-vote’’ or ‘‘majority-rule’’ ensemble, the novel neural network ensemble aggregates the decision values from the different neural ensemble members, instead of their classification results directly.

Apart from investigating the most accurate model for assessment of default, some researchers focus on the influencing factors of default risk, such as loan to value (LTV) (Bian & Lin, 2018) and expected housing expense ratio. Kelly & Toole (2018) proposed the “double trigger” default model. They investigate whether the relationship between debt service, loan-to-value ratio and default can be informative in the calibration of macro-prudential limits. Qi & Yang (2008) show that loss given default can largely be explained by various characteristics associated with the loan, the underlying property, and the default, foreclosure, and settlement process. They also found that the current loan-to-value ratio is the single most important determinant.

This study differs from these recent papers as it uses estimation and statistical approaches to measure loan-to-value and expected housing expense ratio. In particular, the time series of LTV data of a loan is known initially but become unobservable at later time, so the unobservable LTV will be estimated by adjusting the initial LTV with the USA Housing Index and the current actual unpaid principal balance (UPB). Likewise, the unobservable expected housing expense ratio data will be estimated by adjusting its initial value with the Per Capita Personal Income, current UPB, and the remaining month of the loan. The estimated LTV and estimated housing expense ratio are treated as meaningful information in our study and are applied in the logistic model and neural network model to predict the outcome of them. There are two types of payments which are on-time payment and delayed payment.

CHAPTER III
METHODOLOGY

3.1 Overview

This research is based on the logistic regression model and neural network model. The outcome of a loan is classified into two categories: “on-time payment” and “delayed payment”. The objective of this study is to investigate:
(1) whether the complexity brought by the model or by the data has more value for predicting the loan payments.
(2) whether the complex data adds more benefit to the complex model for predicting the loan payments.

The dynamic, unobserved data, i.e. LTV and housing expense ratio is obtained by using data transformation techniques and represents an important part of the complex data. They will be discussed in detail later. The performance of neural network (complex model) with simple data is compared with logistic regression model (simple model) with complex data to explore the answer of problem (1). In addition, this study compares the performance of neural network with simple data and neural network with complex data to investigate problem (2). It also gives a conclusion about complex model and complex data.

3.2 Dataset description

The research considers the monthly Single-Family Loan-Level Dataset of Freddie Mac from 2010 to 2018, comprising of around 50000 loans in each year. Each entity’s repayment status in each due date is classified into two categories: “on-time payment” and “delayed payment”. In this case, a customer is assigned to the “on-time payment” class on a single due date if he pays on due date or delay no more than 90 days or otherwise is assigned to the “delayed payment” class if he had a payment in delay for 90 or more days. A loan treated as a default loan directly if a customer is assigned to “delayed payment” class in any single instalment due date or treat it as a “non-default” loan if the customer pays all instalments on time. The full list of variables in the original dataset is available in the Freddie Mac official website. It contains a total of 53 features. In this study, 28 original variables are used for analysis, and 2 extra processed variables (i.e., unobserved dynamic) were constructed. In addition, we incorporate the lag-1 information of 3 dynamic features into the dataset for the logistic regression. Initially, all 28 variables were taken into account in the two models, but after data cleaning and significance consideration only 23 out of 28 original variables were kept for further model training. Moreover, loans with loan terms less than 5 months are not considered.

Type	Feature Name
Observed Static	Xs1. Initial Credit Score
	Xs2. First Payment Date
	Xs3. First Time Homebuyer Flag
	Xs4. Maturity Date
	Xs6. Mortgage Insurance Percentage
	Xs7. Number of Units
	Xs8. Occupancy Status
	Xs9. Original Combined Loan-To-Value
	Xs10. Original Debt-To- Income Ratio
	Xs11. Original UPB
	Xs12. Original Loan-To-Value
	Xs13. Original Interest Rate
	Xs14. Channel
	Xs17. Property State
	Xs18. Property Type
	Xs20. Loan Sequence Number
	Xs21. Loan Purpose
	Xs22. Original Loan Term
	Xs23. Number of Borrowers
	Xs26. Super Conforming Flag
Observed Dynamic	Xd3. Current Actual UPB
	Xd2. Monthly Reporting Period
	Xd4. Current Loan Delinquency Status
	Xd5. Loan Age
	Xd6. Remaining Months to Legal Maturity
	Xd7. Repurchase Flag
	Xd9. Zero Balance Code
	Xd11. Current Interest Rate
Unobserved Dynamic (Processed Variable)	Xud1. Housing Expense Ratio (HER)
	Xud2. Loan to Value (LTV)

Table 1 List of Features

All training variables are divided into three different groups, i.e. “observed static”, “observed dynamic” and “unobserved dynamic”. A variable that has initial value and the value does not change over time is assigned into the “observed static” group (e.g. Channel, Property Type, Loan Purpose), then a variable whose initial value is known but the subsequent values fluctuate overtime is assigned into the “observed dynamic” group (e.g. Current Interest Rate, Current Actual UPB). Since the statistical methods are used to dig out changing values of HER and LTV, they are grouped into the “unobserved dynamic” group (see Table 1).

Let $w_{i,j}$ denotes the j^{th} observed static variable of loan i, $x_{i,k,t}$ denotes the k^{th} dynamic variable of loan i at time t, and $r_{i,v,t}$ denotes the v^{th} dynamic variable of loan i at time t.

Moreover, the current real UPB, property state and USA House Price Index are applied to estimate of current LTV. The quarterly Per Capita Income by State and the monthly total repayment are applied to estimate of monthly housing expense ratio. Before training a model, one hot encoding is required to convert all categorical variables into numerical value that could be provided to ML algorithms to do a better job in prediction. Variable Xs8, Xs14, Xs18, Xs21 are split into multiple columns whose number is the same as their number of categories. Data standardization and data scaling technologies are considered to applied in data preprocessing process (Table 2).

Feature Name	Preprocessing Method
Xs1. Initial Credit Score	Standardized
Xs3. First Time Homebuyer Flag	Change categorial to numeric
Xs4. Maturity Date	Minusing first payment date (Xs2) and scaling to yearly
Xs6. Mortgage Insurance Percentage	Convert to numeric
Xs7. Number of Units	Convert 1, and more than 1 to 0, 1
Xs8. Occupancy Status	One hot encoding on Multi-categorial Vars
Xs9. Original Combined Loan-To-Value	Standardized
Xs10. Original Debt-To- Income Ratio	Standardized
Xs11. Original UPB	Standardized
Xs12. Original Loan-To-Value	Standardized
Xs13. Original Interest Rate	Taking log and -1
Xs14. Channel	One hot encoding on Multi-categorial Vars
Xs18. Property Type	One hot encoding on Multi-categorial Vars
Xs21. Loan Purpose	One hot encoding on Multi-categorial Vars
Xs22. Original Loan Term	Standardized
Xs23. Number of Borrowers	Convert 1,2 to 0, 1
Xs26. Super Conforming Flag	Convert Y, Na to 1, 0
Xd3. Current Actual UPB	Standardized
Xd5. Loan Age	Standardized
Xd6. Remaining Months to Legal Maturity	Standardized
Xd7. Repurchase Flag	Change categorial to numeric
Xd9. Zero Balance Code	Convert non-Na, Na to -1, 0
Xd11. Current Interest Rate	Taking log and -1
Xud1. Housing Expense Ratio	Standardized
Xud2. Loan to Value (LTV)	Standardized

Table 2 Data Preprocessing

3.3 Simple model------Logistic regression model

Logistic regression (LR) is one of the most popular algorithms for solving bank loan default rate problems. Ordinal LR can be used to solve the classification problem. This study uses the ordinal logistic regression model as a representative of a simple model. Let $y_{i,t}$ denote the categorical outcome of loan i at time t, where $y_{i,t} = 1$ for delayed payment, $y_{i,t} = 0$ for on-time payment. In addition, we incorporate the lag-1 information of 3 dynamic features into the dataset.

Let $W_i = [w_{i,1}, \cdots w_{i,j}, \cdots\ w_{i,p}]'$, $X_{i,t} = [x_{i,1,t}, \cdots, x_{i,k,t} \cdots x_{i,q,t}]'$, and $R_{i,t} = [r_{i,1,t}, r_{i,2,t}\ ]'$ where p is the number of static variables; q is the number of observed dynamic variables. $R_{i,t}$ is the vector of processed variables.

Assume that $y_{i,t}$ is independent across i given $Y_{i,t}$. Consider the following logit model for outcome of each loan i at time t:

$$y_{i,t+1} = \begin{cases} 0, & if\ p(Y_{i,t}) \leq 0.5 \\ 1, & if\ p(Y_{i,t}) > 0.5 \end{cases}$$

while

$$p(Y_{i,t}) = \phi(Z_{i,t}) = \frac{e^{Z_{i,t}}}{1 + e^{Z_{i,t}}}$$

Here, the feature set consists of static and time-series variables.

1. Simple model with complex data

$$Z_{i,t} = \beta_W' W_i + \beta_X' Y_{i,t-1}$$

$$= \beta_0 + \beta_{w,1} w_{i,1} + \cdots + \beta_{w,p} w_{i,p} \qquad \boxed{\text{Static}}$$

$$+\beta_{X,1} x_{i,1,t} \ldots + \beta_{X,q} x_{i,q,t} + \beta_{X,3} x_{i,3,t-1} + \beta_{X,5} x_{i,5,t-1} + \beta_{X,6} x_{i,6,t-1} \qquad \boxed{\text{Observed dynamic}}$$

$$+ \beta_{R,1} r_{i,1,t} \ldots + \beta_{R,2} r_{i,q,t} \qquad \boxed{\text{Unobserved dynamic}}$$

where $i = 1, \cdots, n$, $t = 1, \cdots, T$

2. Simple model with simple data

$$Z_{i,t} = \beta_W' W_i + \beta_X' Y_{i,t-1}$$

$$= \beta_0 + \beta_{w,1} w_{i,1} + \cdots + \beta_{w,p} w_{i,p} \qquad \boxed{\text{Static}}$$

$$+\beta_{X,1} x_{i,1,t} \ldots + \beta_{X,q} x_{i,q,t} + \beta_{X,3} x_{i,3,t-1} + \beta_{X,5} x_{i,5,t-1} + \beta_{X,6} x_{i,6,t-1} \qquad \boxed{\text{Observed dynamic}}$$

where $i = 1, \cdots, n$, $t = 1, \cdots, T$

Then estimating the coefficients $\hat{\beta}$ by maximizing this likelihood function:

$$\ell(\beta) = \prod_{i:y_i=1} \mathrm{P}\big(X_{i,t}, R_{i,t}\big) \prod_{i':y_i{}'=0} \Big(1 - \mathrm{P}\big(X_{i,t}, R_{i,t}\big)\Big)$$

3.4 Complex model-------Neural network model

In this work, a neural network model represents complex models. The neural network can produce a system classifies the entity based on the input data through complex nonlinear transformation. A neural network optimizes certain parameters to get to the right output.

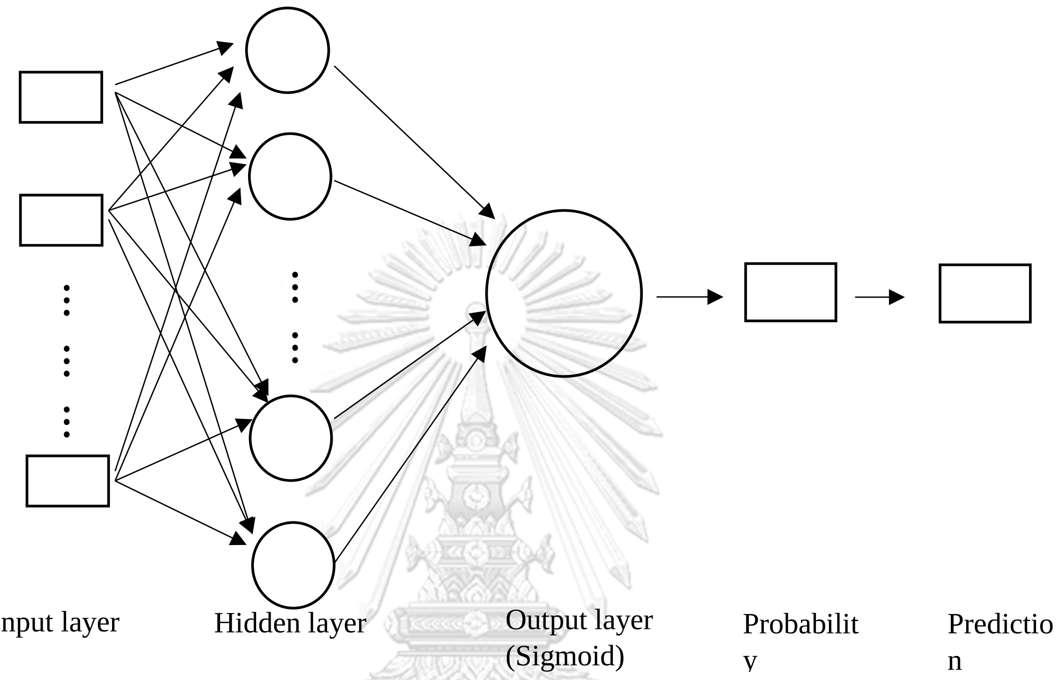


Figure 1 Neural network framework for observation l

As show in Figure 1, Neural networks flow from left to right. The 23 original features and 2 optionally unobserved features (25 columns from the input data frame) that arrive at the input neurons from the first row (first observation) of the input data frame. In this study, the network contains only one hidden layer. These 25 numbers are then multiplied by a set of weights $W^{[1]}$ (randomly initialized at first and later optimized). A tanh activation function is then applied on the result of this multiplication. This new set of numbers becomes the neurons in the hidden layer. These neurons are again multiplied by another set of weights $W^{[2]}$ (randomly initialized) with a sigmoid activation function applied to this result. The final result we obtain is a single number that lies between 0 and 1. Defined a threshold equal to 0.5 for rounding off this probability to 0 or 1. Neural network needs activation function (tanh and sigmoid activation functions in our case) to add non-linearity and enables it to learn complex features. While a neural network consists of a bunch multiplications and additions, which alone is linear, a linear classification model, even in high dimensions, will not be able to learn complex features as equal to the neural network models due to the non-linearity that the activation functions have added.

Once a prediction is obtained, then it will be compared with the true output value. To optimize the weights in order to make our predictions more accurate (right now the first input is being multiplied by random weights to give a random prediction), the calculate that shows how far off is the prediction from the actual value is necessary. Then the calculation of the gradients with respect to each weight using the loss needs to be done. The gradients tell the amount by which we need to increase or decrease each weight parameter in order to minimize the loss, this process is called backpropagation. All the weights in the network are updated as repeating the entire process with the other input samples. After all the input samples have been used to optimize weights, one epoch has passed. This process is repeated for multiple number of epochs till the loss stops decreasing or loss is small enough. Once all the training data has passed through this process, the final weights and deviations are used for testing. The one that has the highest probability will be selected as the prediction class.

Let $l^{(u)}$ be one input observation where u is the row number. Z is output from the input l, $z^{[1](u)}$ is the output from the uth neuron of the 1st layer. Here, the layer number in the superscript in square brackets and the neuron number in parenthesis. $a^{[1]}$ is the value of hidden layer which passed first activation function, n_x is the size of input layer, n_h is the size of hidden layer and n_y is the size of output layer.

The initial sizes of these weight matrices are:

$$W^{[1]} = (n_h\,, n_x)$$

$$b^{[1]} = (n_h\,, 1)$$

$$W^{[2]} = \left(n_y\,, n_h\right)$$

$$b^{[2]} = (n_y\,,1)$$

Training Algorithm:

1. Calculate the output Z for the first layer.

$$z^{[1](u)} = W^{[1]}\, l^{[1](u)} + b^{[1](u)}$$

2. Apply tanh activation function to get a.

$$a^{[1](u)} = \tanh\,(\,z^{[1](u)})$$

3. Calculate the value for the final output layer using the hidden layer values.

$$z^{[2](u)} = W^{[2]}\, a^{[1](u)} + b^{[2](u)}$$

4. Apply sigmoid activation function and obtain output probability.

$$\hat{y}^{(u)} = a^{[2](u)} = \sigma\left(z^{[2](u)}\right) = \frac{1}{1 + e^{-z^{[2](u)}}}$$

5. Use 0.5 threshold to round off this probability.

$$y_{prediction}^{(u)} = \begin{cases} 1 & \text{if } a^{[2](u)} > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

6. Compute the cost $\boldsymbol{J}$.

$$\boldsymbol{J} = -\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{u}=\mathbf{0}}^{\boldsymbol{m}}(\boldsymbol{y}^{(\boldsymbol{u})}\mathbf{log}\,(\hat{\boldsymbol{y}}^{(\boldsymbol{u})}) + (\mathbf{1} - \boldsymbol{y}^{(\boldsymbol{u})})\mathbf{log}(\mathbf{1} - \hat{\boldsymbol{y}}^{(\boldsymbol{u})}))$$

where m is the number of observations

7. Calculate the gradients term.

$$dZ^{[2]} = A^{[2]} - Y$$

$$dW^{[2]} = \frac{1}{m} dZ^{[2]} A^{[1]^T}$$

$$db^{[2]} = \frac{1}{m} \sum dZ^{[2]}$$

$$dZ^{[1]} = W^{[2]^T} * g^{[1]'} Z^{[1]}$$

where g is the activation function tanh ( )

$$dW^{[1]} = \frac{1}{m} dZ^{[1]} X^T$$

$$db^{[1]} = \frac{1}{m} \sum dZ^{[1]}$$

8. Update the weights, $\boldsymbol{\alpha}$ is learning rate.

$$W^{[2]} = W^{[2]} - \alpha * dW^{[2]}$$

$$b^{[2]} = b^{[2]} - \alpha * db^{[2]}$$

$$W^{[1]} = W^{[1]} - \alpha * dW^{[1]}$$

$$b^{[1]} = b^{[1]} - \alpha * db^{[1]}$$

9. Repeated multiple times until the cost is small enough or model converges.

3.5 Complex data extraction

Apart from the raw variables, three hidden variables are used in the models, i.e. LTV and Housing Expense ratio. The study uses different hidden information extraction techniques for each unobservable dynamic variable.

3.5.1 Estimated Loan-to-Value

As mentioned before, LTV can be found out using the current real UPB, property state and USA House Price Index. Since we know the property's state, even postal code, the USA State House Price Index is a good representor of the property value.
Approximate LTV at time t:

$$LTV_t = \frac{UPB_t}{Housing\ Index_t \times \frac{V_0}{House\ Index_0}}$$

where V_0 is the value of the property at the beginning of the loan contract.

3.5.2 Expected Housing Expense Ratio

The housing expense ratio is also referred to as the front-end ratio because it is a partial component of a borrower's total debt-to-income and may be considered first in the underwriting process for a mortgage loan. The estimated housing expense ratio aims to measure the current level of individual's obligation to pay housing debt, which should be a crucial factor in determining the insolvency risk of a loan. The estimated housing expense in this work is defined as the ratio between monthly debt obligation projected from the unpaid payment balance (UPB) and monthly income. It is expected that higher expected housing expense ratio is associated with higher insolvency risk and more likelihood of delayed payment.

For the income component in the housing expense ratio, we assumed that all borrowers from the same state share the same income changes in a month. Proportional change of the per capita personal income by state is a good representor of the percentage increase/decrease of the individual personal income. We first estimate I_0 , the initial monthly income by the following formula,

$$I_0 = \frac{UPB_0}{Expected\ loan\ duration \times DTI_0}$$

where

- DTI_0 is initial debt to be paid per month per monthly income which is observable and given from the data source.
- $Expected\ loan\ duration$ is the number of months since the note origination month of the mortgage to the maturity.
- UPB_0 is the original unpaid principal balance which equal to the total debt.

We then update the individual income by using the adjustment from Per Capita Personal Income at the loan's location:

$$I_t = I_0 \times \frac{Per\ Capita\ Personal\ Income\ _t}{Per\ Capita\ Personal\ \ Income_0}$$

Finally, the estimated Housing Expense ratio at time t is:

$$Expected\ Housing\ expense\ ratio_t = \frac{Current\ UPB_t}{remaining\ month\ of\ loan \times I_t}$$

3.6 Experimental design

The four model setups as shown in Table 3. We compare the simple model with complex data to complex model with simple data to answer the first research question and measuring the improvement of using the complex data in complex model over simple data in the complex model to answer second question. Three pairs of in-sample and out-of-sample datasets are considered in this study. The training data in each training data set is sampled from the data of each year in a moderate proportion 0.233, and the training data set is composed of 1000 data. The test dataset uses the same method, the proportion is 0.07, and the training data set is composed of 300 data. Finally, we obtained 3 pairs of in-sample and out-of-sample data in the experiment. Each out-of-sample data set is tested on all four types of models to get a general conclusion.

Type	Prediction Model	Input Data
Simple model / Simple data	Logistic regression	Static, Observed dynamic
Simple model / Complex data	Logistic regression	Static, Observed dynamic, Unobserved dynamic
Complex model / Simple data	Neural network	Static, Observed dynamic
Complex model / Complex data	Neural network	Static, Observed dynamic, Unobserved dynamic

Table 3 Experimental models

The performance of each model is compared based on the prediction accuracy. The confusion matrix is used to measure classifier performance. Since this is a binomial-class problem, we simplify treat one class as “positive” and the other classes as “negative”, and calculate：

- True Positive (TP): predicted to be default and it actually was default
- False Positive (FP): predicted to be non-default and it actually was default
- True Negative (TN): predicted to be non-default and it actually was non-default
- False Negative (FN): predicted to be default and it actually was non-default

We also compute the accuracy, precision, recall, and F1 values to measure the performance of models.

F1-score is the harmonic mean of precision and recall:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is defined as the number of true positives over the number of true positives plus the number of false positives:

$$Precision = \frac{TP}{TP + FP}$$

Recall is defined as the number of true positives over the number of true positives plus the number of false negatives:

$$Recall = \frac{TP}{TP + FN}$$

F1-score is the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \times precision \times recall}{precision + recall}$$

The variation of the performance from each out-of-sample data set will be used to test the robustness of the results.

CHAPTER IV
RESULTS AND DISCUSSION

4.1 Sample Set

Although the amount of new loan data in each year is 50,000, considering it is changed after data cleaning such as discarding, conversion, standardization, a fixed ratio is applied on sample selection in every year. After experiments, it is determined that assigning the Tarin-Test selection ratio of the training set as 0.233 and the ratio of the test set as 0.02 is the best option, which can ensure that the training set data of the three sample groups is around 1000 and the test set data is around 300.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Sample Set 1										
Training	114	114	114	115	115	115	115	114	84	1000
Testing	35	34	34	35	34	35	35	34	25	301
Sample Set 2										
Training	114	114	114	115	115	115	115	114	84	1000
Testing	35	34	34	35	34	35	35	34	25	301
Sample Set 3										
Training	114	114	115	115	114	115	115	114	84	1000
Testing	34	34	35	35	34	35	35	34	25	301

Table 4 Data Sample

It can be clearly observed that the default rate in the training samples is quite low, the imbalance problem is serious in real credit data. In order to make the model to identify the default class and simplify the operation difficulty, the default data in the training set was duplicated 250 times, which significantly increased the default ratio. The change in default rate before and after replication as shown in Table 5.

No.	Default Rate	Default Rate after Duplicating
Training Set 1	0.37%	48.80%
Training Set 2	0.39%	50.37%
Training Set 3	0.46%	54.30%

Table 5 Default Rate

4.2 Unobserved Variables

The mean of approximated LTV in all experimental samples is 0.8984, the minimum value is -979.64 and the maximum value is 1682.60 with the standard deviation 10.19. The approximate HER has the mean value of -0.02, the minimum value of -569.04 and the maximum value of 2344.10 with the standard deviation 8.48. Both LTV and HER are most concentrated in sample 2 with small variance. Next, these normalized approximated data is ready to be plugged into the simple and the complex models.

4.3 Model Comparation

4.3.1 Logistic Regression Model

Logistic regression measures the relationship between the dependent variable and one or more independent variables(features) by estimating probabilities using the underlying logit function. In statistics, the logit function or the log-odds is the logarithm of the odds. Generalized linear model (GLM) is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The logistic regression model is an example of a broad class of models known as generalized linear models (GLM). In the R version 4.0.3 runtime environment, the glm() function, family = binomial('logit'), is used for our model training. In a significance test, coefficients marked as dot or stars means that their p value less than 0.05 can be considered significant. In the regression results, the coefficients of Xs8_3, Xs14_3, Xs18_3, Xs18_4, Xs18_5, Xs21_3, Xd9, Xd11 show NA, indicating that these variables might be a collinearity problem.

	With Unobserved Variables			Without Unobserved Variables		
Coefficients	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 1**	**Sample 2**	**Sample 3**
Intercept	3.809e+00	3.099e+10	5.839e+10	3.728e+00	-8.206e+10	2.249e+10
Xs1	-9.746e-01***	-5.996e-01***	- 7.183e-01***	-9.743e-01***	-5.930e-01***	-6.612e-01***
Xs3	8.649e-02***	-1.488e-01***	1.476e-01***	8.840e-02***	-1.412e-01***	1.438e-01***
Xs4	-5.394e+00***	-2.057e+00***	4.974e+00***	-5.399e+00***	-2.077e+00***	5.189e+00***
Xs6	-4.469e-01***	2.756e-02*	3.128e-01***	-4.487e-01***	3.078e-02*	3.211e-01***
Xs7	3.603e-02**	-1.620e-01***	2.465e-01***	3.541e-02**	-1.677e-01***	2.185e-01***
Xs8_1	4.782e-01***	7.362e-01***	-3.844e-01***	4.721e-01***	-1.677e-01***	-3.505e-01***
Xs8_2	5.123e-01***	1.535e+00***	1.212e+00***	5.067e-01***	1.537e+00***	1.130e+00***
Xs8_3	NA	NA	NA	NA	NA	NA
Xs9	-2.959e-01***	1.714e+00***	7.523e-01***	-2.947e-01***	1.690e+00***	7.018e-01***
Xs10	6.125e-01***	3.208e-01***	-5.904e-02***	6.127e-01***	3.122e-01***	2.147e-01***
Xs11	-1.876e+00***	2.066e+00***	-3.780e+00***	-1.884e+00***	2.115e+00***	-4.355e+00***
Xs12	1.123e-01*	-1.337e+00***	-7.147e-01***	1.113e-01*	-1.317e+00***	-6.717e-01***
Xs13	-3.204e-01***	1.187e-01***	2.622e-01***	-3.178e-01***	1.329e-01***	2.936e-01***
Xs14_1	-1.455e+00***	-4.231e-01***	3.031e-01***	-1.450e+00***	-3.982e-01***	2.890e-01***
Xs14_2	-8.892e-01***	-2.425e-01***	1.948e+00***	-8.873e-01***	-2.294e-01***	2.017e+00***
Xs14_3	NA	NA	NA	NA	NA	NA
Xs18_1	1.735e-01**	-3.099e+10	-5.839e+10	1.592e-01**	8.206e+10	-2.249e+10
Xs18_2	6.999e-03	-3.099e+10	-5.839e+10	-6.957e-04	8.2069e+10	-2.249e+10
Xs18_3	-1.274e+01	-3.099e+10	NA	-1.304e+01	8.206e+10	NA
Xs18_4	NA	-3.099e+10	5.839e+10	NA	8.206e+10	-2.249e+10
Xs18_5	NA	-3.099e+10	-5.839e+10	NA	8.206e+10	-2.249e+10
Xs21_1	-2.550e-01***	8.037e-01***	-1.518e+00***	-2.528e-01***	7.851e-01***	-1.565e+00***
Xs21_2	2.445e-03	4.842e-01***	-9.405e-01***	-1.072162e-03	4.770e-01***	-8.561e-01***
Xs21_3	NA	NA	NA	NA	NA	NA
Xs22	7.612e+00***	2.740e+00***	-6.042e+00***	7.608e+00***	2.720e+00***	-6.714e+00***
Xxs23	-4.367e-01***	-3.783e-01***	-7.392e-01***	-4.341e-01***	-3.772e-01***	-8.087e-01***
Xs26	-2.721e+00	-1.900e+01	-1.627e+01	-2.723e+00	-1.922e+01	-5.502e+00
Xd3	2.473e+02***	2.171e+02***	1.212e+01***	2.467e+02***	2.179e+02***	1.308e+01***
Xd3_tm1	-2.451e+02***	-2.186e+02***	-8.511e+00***	-2.445e+02***	-2.194e+02***	-8.827e+00***
Xd5	9.618e+00***	2.812e+00***	6.636e+00***	9.583e+00***	2.634e+00***	6.790e+00***
Xd5_tm1	-9.593e+00***	-2.852e+00***	-5.853e+00***	-9.557e+00***	-2.671e+00***	-5.893e+00***
Xd6	4.431e+02***	3.603e+00**	2.243e+00***	4.411e+02***	3.636e+00**	2.191e+00***
Xd6_tm1	-4.456e+02***	-4.359e+00***	-1.005e+00***	-4.435e+02***	-4.350e+00***	-5.245e-01*
Xd7	9.568e-01	8.439e+00	-2.543e+02	1.618e+00	1.019e+01	6.636e-01
Xd9	NA	NA	-2.639e+02	NA	NA	-1.619e+00
Xd11	NA	NA	-2.576e-02	NA	NA	-4.515e-02**
HER	1.679e-02	2.549e-02.	1.141e+00***	-	-	-
Ltv	1.418e-01***	-1.200e-01***	-6.619e-01***	-	-	-

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 6 Estimation Result of Logistic Regression

With the unobservable variables, the tarin accuracy is better than the one without unobserved variables, and the test accuracy is quite close to the test accuracy of the simple model with simple data in each sample. The test accuracy is better than those without unobservable variables. Based on accuracy, it can be seen that the complex data does not add more value to the model's performance boost, as there is no significant change in the test accuracy. If we further use the AIC rule to judge their performance, we could find that simple models with complex data is better than simple models with simple data. Normally lower AIC values indicate a better-fit model. The reason why we cannot clearly see the better performance of complex data by using accuracy maybe because our sample size is not large enough. In general, in this experiment, all sample groups did get better results using simple models with complex data. Since it is necessary to consider the difficulty of data processing, the running ability of the computer, we only selected 3 sets of sample. In further experiments, the issue of increase the size of sample and the amount of data in each sample set and adjust training variables should be given attention, they may change the AIC value to more precise direction.

	With Unobserved Variables			**Without Unobserved Variables**		
Sample No.	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Train Accuracy	0.8448	0.7912	0.8398	0.8446	0.7879	0.8311
Test Accuracy	0.7455	0.8085	0.7017	0.7460	0.8076	0.7043
AIC	54,703	64,727	59,049	54,715	64,800	59,989

Table 7 Model Accuracy

4.3.2 Neural Network Model

A neural network optimizes certain parameters to get to the right output. To generate matrices with random parameters, first the size (number of neurons) of all the layers in the neural net must be obtained. The number of neurons is decided based on shape of the input and output matrices. The parameters are initialized based on random uniform distribution. The function initializeParameters ( ) takes as argument an input matrix and a list which contains the number of neurons in input layer, hidden layer, and output layer respectively. The function returns the trainable parameters W^1, b^1, W^2, b^2 those weights matrices are initialized randomly based on the layer sizes of the different layers as shown in Table 8.

Parameter	Initial Value for Complex Data	Initial Value for Simple Data
W^1	$(10\ ,35)$	$(10\ ,33)$
b^1	$(10\ ,1)$	$(10\ ,1)$
W^2	$(1\ ,10)$	$(1\ ,10)$
b^2	(1 ,1)	(1 ,1)

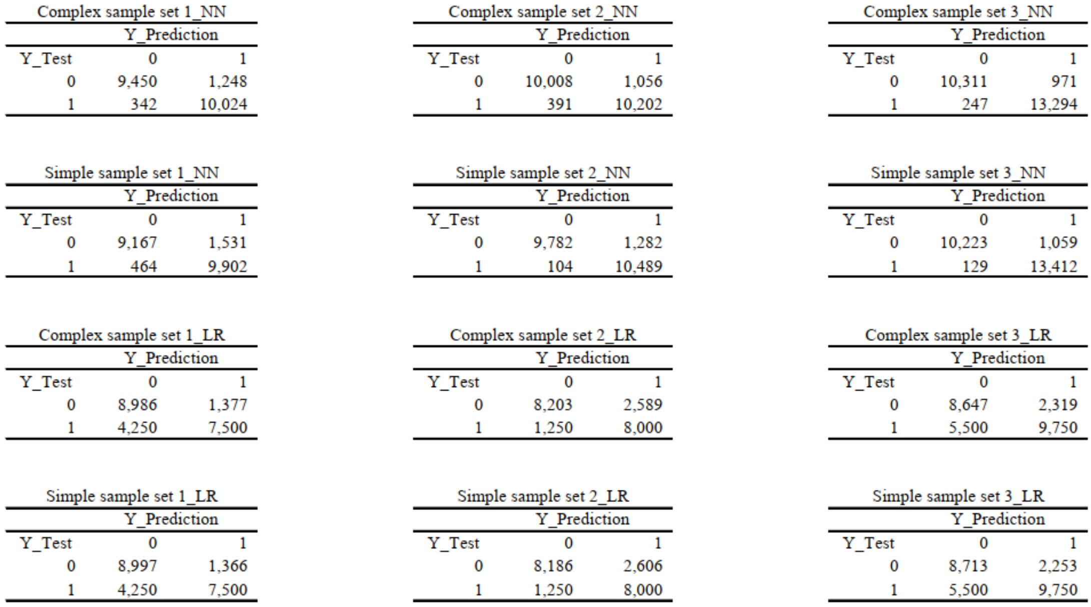
Table 8 Parameter Initial Value

In the model training process, set the epochs is 2000, the hidden neurons is 39 and the learning rate is 0.9. The loss will be print out after every 100 epochs. It can be seen that the cost has dropped significantly and quite low in the end of last iteration in each experimental group. At the same time, the sample sets with complex data have the lower cost. The lower the cost, the better the model training, this is also reflected on the prediction result in Table 11.

	Cost					
	Simple data set			Complex data set		
Iteration	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
100	0.4889	0.5128	0.4587	0.4880	0.5104	0.4598
200	0.4540	0.4778	0.4266	0.4491	0.4773	0.4210
300	0.4568	0.4543	0.4129	0.4296	0.4558	0.4020
400	0.4358	0.4704	0.3876	0.4281	0.4357	0.3810
500	0.4098	0.4318	0.3586	0.4002	0.4503	0.3581
600	0.3910	0.3942	0.3078	0.3785	0.4234	0.3047
700	0.3720	0.3636	0.3086	0.3720	0.3906	0.2727
800	0.3515	0.3410	0.2570	0.3613	0.3365	0.2484
900	0.3315	0.3165	0.2281	0.3438	0.3070	0.2265
1000	0.3118	0.2901	0.2090	0.3313	0.2850	0.2083
1100	0.2918	0.2560	0.1931	0.3080	0.2596	0.1917
1200	0.2756	0.2328	0.1795	0.2839	0.2307	0.1770
1300	0.2657	0.2127	0.1682	0.2644	0.2148	0.1637
1400	0.2715	0.2004	0.1576	0.2432	0.2039	0.1532
1500	0.2615	0.1903	0.1478	0.2173	0.1923	0.1441
1600	0.2467	0.1790	0.1389	0.1924	0.1761	0.1357
1700	0.2250	0.1698	0.1312	0.1783	0.1630	0.1284
1800	0.2042	0.1604	0.1251	0.1684	0.1571	0.1223
1900	0.1914	0.1515	0.1187	0.1596	0.1514	0.1173
2000	0.1829	0.1445	0.1132	0.1514	0.1462	0.1126

Table 9 Cost of Each Iteration

From the experimental results, complex model performs significantly better than simple model whether in complex or simple data conditions. Although training complex model takes more time, the performance benefits outweigh the costs. For neural network model, using the test set containing unobservable variables performs better than using a simple sample data set in accuracy and F1 score accuracy metrics. However, the difference between the prediction performance brought by the complex data and the simple data to the neural network is not as large as the difference between the performance of the logistic regression model and the neural network model under the same testing set condition. Also, for simple models, complex data not obviously improve model performance. So, it may not be worth the time and effort to mine complex data. Especially for complex models, the accuracy of the model itself is already high.



Complex sample set 1_NN

	Y_Prediction	
Y_Test	0	1
0	9,450	1,248
1	342	10,024

Complex sample set 2_NN

	Y_Prediction	
Y_Test	0	1
0	10,008	1,056
1	391	10,202

Complex sample set 3_NN

	Y_Prediction	
Y_Test	0	1
0	10,311	971
1	247	13,294

Simple sample set 1_NN

	Y_Prediction	
Y_Test	0	1
0	9,167	1,531
1	464	9,902

Simple sample set 2_NN

	Y_Prediction	
Y_Test	0	1
0	9,782	1,282
1	104	10,489

Simple sample set 3_NN

	Y_Prediction	
Y_Test	0	1
0	10,223	1,059
1	129	13,412

Complex sample set 1_LR

	Y_Prediction	
Y_Test	0	1
0	8,986	1,377
1	4,250	7,500

Complex sample set 2_LR

	Y_Prediction	
Y_Test	0	1
0	8,203	2,589
1	1,250	8,000

Complex sample set 3_LR

	Y_Prediction	
Y_Test	0	1
0	8,647	2,319
1	5,500	9,750

Simple sample set 1_LR

	Y_Prediction	
Y_Test	0	1
0	8,997	1,366
1	4,250	7,500

Simple sample set 2_LR

	Y_Prediction	
Y_Test	0	1
0	8,186	2,606
1	1,250	8,000

Simple sample set 3_LR

	Y_Prediction	
Y_Test	0	1
0	8,713	2,253
1	5,500	9,750

Table 10 Confusion Matrix

Before training the model, the default rate in the three training sets were increased to a level of 50% for solving the imbalance problem. There might be a bias in the estimation of the intercept due to the oversized training sample sets especially for the use of prediction default probability. So, normally the threshold of deciding whether the loan is default must be adjusted carefully (less than 0.5) to match with the bias. However, in this study, we do not predict exact probabilities, and the predicted class is less sensitive to this bias. By choosing to keep the threshold 0.5, the bias would be offset from not adjusting the intercept.

	NN					
	Simple data			**Complex data**		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Accuracy	90.53 %	93.60 %	90.75 %	95.21 %	93.32 %	95.09 %
Precision	95.52 %	99.02 %	94.37 %	99.05 %	96.31 %	98.18 %
Recall	86.61 %	89.11 %	88.88 %	92.68 %	90.62 %	93.19 %
F1 Score	90.85 %	93.80 %	95.76 %	92.65 %	93.38 %	95.62 %
	LR					
	Simple data			**Complex data**		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
Accuracy	74.60 %	80.76 %	70.43 %	74.55 %	80.85 %	70.17 %
Precision	63.83 %	86.49 %	63.93 %	63.83 %	86.49 %	63.93 %
Recall	84.59 %	75.43 %	81.23%	84.49 %	75.55 %	80.79%
F1 Score	72.76 %	80.58 %	71.55 %	72.72 %	80.65 %	71.38%

Table 11 Accuracy Metrics

CHAPTER IV
CONCLUSION

The objective of this study is to investigate whether a complex model or complex data has more value for prediction problems. It also investigates whether there is the added benefit of the processed data in a complex model. This study uses the monthly Single-Family Loan-Level Dataset of Freddie Mac to test the hypothesis. The logistic regression model is used to represent a simple model and the neural network model is used to represent a complex model. The complex data, namely LTV and HER, were extracted based on some structural models. At the end, the study compares the performance of all experimental models by using the confusion matrix and accuracy metrics.

From this experiment, a significant conclusion is that the prediction performance of all combination model are good, neural networks are especially good. Even though the logistic model does not perform badly, the overall performance of the neural network model is much better than the logistic model. After a closer comparison: complex model with simple data is better than simple model with complex data. So, the experimental results tell us that the complexity brought by the model has more value for predicting the loan payments, this is also the answer to the first research question. The study also pointed out that our two processed variables can help model increase the accuracy, but it cannot have a huge boost. The processed data adds a little value in the complex model base on those two unobserved variables in this study. If there is no better model selection, we can consider investing in mining complex data, but if there is a lot of room for model improvement, it is recommended to work on complex models first.

For the neural network, because the relationship between the internal data itself has been automatically extracted sufficiently complex via the computer, so there is no need to spend a lot of effort to do work on the data. The epochs number, the hidden neurons and the learning rate are more important and valuable even if it takes a lot of time to try out. In future research, more research should be carried out on the selection of the three parameter values that mentioned above. We can draw the curve that training error and test error variating with the number of epochs to find out the point where the test error starts to increase.

This study uses real data. The biggest problem with real default data sets is imbalance. Here, based on the purpose of simple and convenient operation, the default observation data in the training dataset are duplicated, but this changes the real data distribution. So which method can better solve the non-equilibrium problem is also worthy of our further study. Second, it should be considered to expand the number of data in each sample set to ensure obtained are more a general result.

REFERENCES

Arundina, T., Omar, M. A., & Kartiwi, M. (2015). The Predictive Accuracy of Sukuk Ratings; Multinomial Logistic and Neural Network Inferences. *Pacific-Basin Finance Journal*, *34*, 273–292.

Aron, J., & Muellbauer, J. (2016). Modelling and forecasting mortgage delinquency and foreclosure in the UK. *Journal of Urban Economics*, *94*, 32–53.

Bian, X., Lin, Z., & Liu, Y. (2018). House price, loan-to-value ratio and credit risk. *Journal of Banking & Finance*, *92*, 1–12.

Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modeling Small Business Credit Scoring by Using Logistic Regression, Neural Networks, and Decision Trees. *Journal of Intelligent Systems in Accounting Finance & Management*, *13*(3), 1–19.

Chen, M.-C., & Huang, S.-H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, *24*(4), 433–441.

Davis, R. H., Edelman, D. B., & Gammerman, A. J. (1992). Machine learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, *4*(1), 43–51.

Desai, V. S., Crook, J. N., & Overstreet Jr., G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37.

Freddie Mac Single Family Loan-Level Dataset. (2017). Available from http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856.

Hoffmann, F., Baesens, B., Martens, J., Put, F., & Vanthienen, J. (2002). Comparing a genetic fuzzy and a neuro fuzzy classifier for credit Scoring. *International Journal of Intelligent Systems*, *17*(11), 1067–1083.

Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications*, *28*(4), 655–665.

Henley, W. E., & Hand, D. J. (1996). A k-nearest-neighbor classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society*, *45*(1), 77–95.

Jing, D. (2019). Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. *Journal of the Franklin Institute*, *3564716*(8), 4716–4731.

Kelly, R., & O'Toole, C. (2018). Mortgage default, lending conditions and macroprudential policy: Loan-level evidence from UK buy-to-lets. *Journal of Financial Stability*, *36*, 322–335.

Kim, C.-J., & Nelson, C. R. (1999). *State-space models with regime switch*. London: The MIT Press.

Luo, S.-T., Cheng, B.-W., & Hsieh, C.-H. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. *Expert Systems with Applications*, *36*(4), 7562–7566.

Lee, ian-S., & Chen, I.-F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *28*(4), 743–752.
Lai, K. K., Yu, L., Wang, S., & Zhou, L. (2006). Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model. *ICANN*.

Liu, Z., & Xiong, Z. (2017). Master Thesis. Hunan University.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research*, *136*(1), 190–211.

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega*, *31*(2), 83–96.

Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications*, *29*(1), 41–47.

Reichert, A. K., Cho, C.-C., & Wagner, G. M. (1983). An examination of the conceptual issues involved in developing credit scoring models. *Journal of Business & Economic Statistics*, *1*(2), 101–114.

Rajaratnam, K. (2009). *A Simplified Approach to modeling the credit-risk of CMO.*

Wang, Y., Wang, S., & Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, *13*(6), 820–831.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11-12), 1131-1152.

Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, *34*(2), 1434–1444.

VITA

NAME	YU WENG
DATE OF BIRTH	
PLACE OF BIRTH	Kunming, Yunnan, China