2015

MODELING THE PROBABILITY OF MORTGAGE DEFAULT VIA LOGISTIC REGRESSION AND SURVIVAL ANALYSIS

Qingfen Zhang
University of Rhode Island, jenniferzhang06@gmail.com

MODELING THE PROBABILITY OF MORTGAGE DEFAULT VIA LOGISTIC REGRESSION AND SURVIVAL ANALYSIS

BY

QINGFEN ZHANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

UNIVERSITY OF RHODE ISLAND

2015

MASTER OF SCIENCE THESIS

OF

QINGFEN ZHANG

APPROVED:

Thesis Committee:

Major Professor Natallia Katenka

Co Advisor Gavino Puggioni

Liliana Gonzalez

Orlando Merino

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2015

ABSTRACT

The goal of this thesis is to model and predict the probability of default (PD) for a mortgage portfolio. In order to achieve this goal, logistic regression and survival analysis methods are applied to a large dataset of mortgage portfolios recorded by one of the national banks. While logistic regression has been commonly used for modeling PD in the banking industry, survival analysis has not been explored extensively in the area. Here, survival analysis is offered as a competitive alternative to logistic regression.

The results of the final modeling for both methods show very similar fit in terms of the ROC with the survival model having slightly better performance than logistic regression in the training dataset and almost the same performance in the testing dataset. In term of prediction of defaulted and non-defaulted mortgage portfolios, the logistic regression model outperforms survival analysis in the training dataset, while survival model outperforms logistic regression in the testing dataset.

Overall, the results support that the survival analysis approach is competitive with the logistic regression approach traditionally used in the banking industry. In addition, the survival methodology offers a number of advantages useful for both credit risk management and capital management.

ACKNOWLEDGMENTS

I am writing the acknowledgment with sincere gratitude since I would never have been able to finish my thesis without the guidance of my committee members, help from my friends, and support from my family.

First and foremost, I wish to send my deepest thanks to my major advisor, Professor Natallia Katenka and my co-advisor, Professor Gavino Puggioni for the continuous support of my thesis development, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of analysis and writing of this thesis.

I would like to express my sincere thanks to Professor Liliana Gonzalez. She has been so supportive since the day I was accepted into this program. Thanks to her I got the teaching assistantship in the first year which encouraged me to quit my job to come to this program full time and led me to the field I really love.

Besides my advisor, I would like to thank the rest of my thesis committee, Professor Orlando Merino and Professor Tong Yu, for their insightful comments and questions.

Last but not least, I would like to thank my family for their support and encouragement. My in laws and parents have been helping us take care of the kids to give us more time working on the thesis. My husband, Xinkai, has been the best partner for all the years and I was lucky to have him working with me together on the rough road to the finish line of the thesis. My two lovely kids also support me for being such good kids. Even my two year old boy knows that mommy needs to write the thesis and doesn't have time playing with him all the time.

TABLE OF CONTENTS

ABSTRACT ii

ACKNOWLEDGMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES vi

LIST OF FIGURES viii

CHAPTER 1: INTRODUCTION 1

1.1 Risk Profile Review 4

CHAPTER 2: REVIEW OF CONCEPTS AND LITERATURE 12

2.1 Logistic Regression 12

2.2. Survival Analysis 15

CHAPTER 3: DATA AND PRELIMINARY ANALYSIS 21

3.1 Data Description 21

3.2 Univariate Analysis 22

3.3 Correlation Analysis 28

CHAPTER 4: ESTIMATION 31

4.1. Logistic Regression 31

4.1.1 Final Model Selection 31

4.1.2 Model Fit Statistics 32

4.1.3 Hosmer-Lemshow Goodness of Fit Test 34

4.1.4 Rank Ordering Testing 35

4.1.5 Residual Analysis 38

4.1.6. Model performance in both training dataset and testing dataset 40

4.1.7 Cross Validation 40

4.2. Survival Analysis 43

4.2.1 Probability of Density Function (pdf) 43

4.2.2 Survival curve 44
4.2.3 Hazard curve 46
4.2.4 Model Fit 46
4.3. Comparison and Summary 52
CHAPTER 5: CONCLUSION AND FUTURE STEPS 53
BIBLIOGRAPHY 57

LIST OF TABLES

Table 1: Summary statistics of the key variables in the training and testing dataset... 22

Table 2: Rule of thumb for interpreting the size of a correlation coefficient 28

Table 3: Correlation matrix of the key variables for the mortgage portfolio..... 29

Table 4: Logistic regression model estimation 31

Table 5: Testing on the null hypothesis that the coefficients equal to 0 for logistic regression model 33

Table 6: Model fit statistics for logistic regression model..... 34

Table 7: Hosmer and Lemeshow goodness-of-fit test for logistic regression..... 35

Table 8: Model performance testing for logistic regression 37

Table 9: Kolmogorov-Smirnow two sample test for logistic regression 38

Table 10: Model performance testing in both training and testing datasets for logistic regression 40

Table 11: The cross validation from 1000 runs for the logistic regression..... 41

Table 12: Survival model 1 – Model with the same variables as logistic regression model..... 47

Table 13: Testing the null hypothesis that coefficients equal to 0 for survival model 48

Table 14: Survival model 2 – Model by replacing month on book with unemployment rate..... 49

Table 15: Model fit statistics to compare survival model 1 and model 2 49

Table 16: Model performance testing for model 1 and model 2 in both training and testing datasets 50

Table 17: T-test on the difference equal to 0 52

Table 18: Model testing summary for both logistic regression model and survival model..... 52

LIST OF FIGURES

Figure 1: Mortgage Portfolio Default Rate 21

Figure 2: Default rate by origination FICO score 23

Figure 3: Default rate by current FICO score 24

Figure 4: Default rate by current LTV 25

Figure 5: Default rate by Case Shiller 1 year growth..... 26

Figure 6: Default rate by unemployment rate 26

Figure 7: Default rate by month on book 27

Figure 8: Scatter plot matrix of the key variables for the mortgage portfolio 30

Figure 9: ROC curve for the logistic regression model 36

Figure 10: Pearson residual and deviance residual for logistic regression 38

Figure 11: Distribution of the coefficients estimations from 1000 runs 42

Figure 12: Distribution of the AUC for logistic regression from 1000 runs..... 43

Figure 13: Distribution of the time to default for defaulted segment 44

Figure 14: Distribution of the time to default for whole population..... 44

Figure 15: Survival curve for defaulted segment..... 45

Figure 16: Survival curve for whole population 45

Figure 17: Hazard curve for defaulted segment (left) and whole population (right) ... 46

Figure 18: Distribution of the actual and predicted time to default 51

Figure 19: Distribution of difference between actual time to default and predicted time to default..... 52

CHAPTER 1: INTRODUCTION

Credit risk affects virtually every financial contract. Therefore the measurement, pricing, and the management of credit risk have received much attention from financial economists, bank supervisors and regulators, and financial market practitioners. Profits realized on loan products, such as credit cards and mortgage loans, depend heavily on whether customers pay interest regularly or miss payments and default on their loans. The latter is considered to be a credit risk which is the dominant source of risk for banks.

The key focus of the credit risk is to predict if a customer will default on her mortgage loan in the future, or to evaluate the probability of default (PD). The PD can be estimated based on the customers' credit bureau data, such as past credit activity, and their application data as well as their payment behavior for the loans on a book. A lower predicted probability of default means a better creditworthiness. For a loan origination, a bank generally sets a cut-off threshold and approves a credit to those customers that have the predicted probability of default (PD) less than the pre-defined threshold. For the ongoing credit risk management, the predicted probability will be combined with the other risk factors to determine the allowance of a loan loss reserve (ALLL), which in turn will be used to cover the losses when the loans default. The PD is not only important for effective risk and capital management, but also for the pricing of credit assets, bonds, loans and more sophisticated instruments such as derivatives.

The goal of this thesis is to predict the PD for a mortgage portfolio. A mortgage portfolio consists of all mortgage loans on a bank's book; and a mortgage loan is a loan secured by a real property through the use of a mortgage note, which serves as an evidence of the loan existence. A mortgage loan has a risk-based interest rate and is scheduled to amortize over a set period of time (called term), typically 15 or 30 years. All types of real property can be, and usually are, secured with a mortgage and bear an interest rate that is supposed to reflect the lender's risk. The lender's risk is based on the predicted PD and other risk parameters.

In order to predict PD, one needs to define the dependent variable on whether the mortgage loan defaults or not. The criterion that determines if a loan defaults varies on the product and the regulations. In what proposed next, the mortgage loan is flagged as default whenever one of the following conditions appears in the account's monthly data:

1) The payment has 180 days or more days past due.

2) There is a charge-off or a partial charge-off event for this account.

The bank maintains records of when each payment is due for every loan on the book. The due date information is used to populate the due date on a customer's mortgage bill or credit card bill. If a payment is delayed, the system will start calculating the accumulating days before the payment is recorded on the book, or the days past due (DPD). The bank will have a monitoring system to monitor the loans with the past due status. Different banks may have different response systems. For example, if a customer only misses one payment, it could just trigger the warning

process as the customer might just be on vacation and can forget to mail in the payment. In this situation, the bank may send a reminder to the customer. If the customer responds the reminder and pays in the following month, the number of days past due will be back to zero. However, if the customer keeps delaying the payment, the number of days past due will keep accumulating and when it exceeds a certain threshold (such as 90 days or 120 days), the bank will evaluate the loan and decides if any impairment is needed. The bank may request an appraisal of the underlying property and in the meantime, send letters to let the customer know that the property will be taken by the bank if the payment is still not received in some periods. In some situations, a customer may have temporary financial hardness, such as losing job or having a big medical bill to pay, and then the bank may choose to work with the customer to reduce the monthly payment either through extending the loan term or even taking some partial charge-off to further cut the bill. Charge-off means that the bank pays the loan from the bank's ALLL (reserve for the loan losses). Partial charge-off means that the bank pays part of the loan. This is one of the strategies to resolve a defaulted loan. In other situations, if the customer decides not to pay at all or there is no way the customer can keep the payment even if the payment is reduced, the bank starts the foreclosure process to recover the loan from the sale of the property.

Logistic regression has found wide acceptance as a model for the dependence of a binary response variable on a vector of explanatory variable (Strauss, 1992). It has been the most commonly used method in predicting PD (Stepanova and Thomas, 2002). Many methodologies have been investigated (Altman, 2010; Gurný, 2009;

Gurný 2010). Survival analysis is one of the alternatives to logistic regression that has recently been explored with application to different portfolios (Stepanova, 2000; Allen and Rose, 2006; Im et al, 2012). Originally, the methods of survival analysis have been developed and intensively applied in medical fields and specifically in life-and-death clinical trials. Recently, some banks have started exploring the application of survival analysis in predicting PD. If looking at the mortgage loan from a life cycle view, one can represent the time to mortgage default as a time to event (similarly to the time to death in a clinical trial) and model this time using survival analysis methods. In my thesis, I would like to apply both logistic regression and survival analysis methods to a large dataset of mortgage portfolios and compare the results in terms of prediction and interpretation.

1.1 Risk Profile Review

Many factors impact the default rates, such as FICO score, loan to value (LTV), month on book, etc. In what follows next, I will discuss the key factors in more details and explain how these factors will be tested using the mortgage portfolio data in later sections.

Industry (Mester, 1997; Brown et al, 2010) and academic researches (Altman and Saunders, 1997; Avery et al., 2003) suggest that mortgage default rate relies on FICO scores. A FICO score is a credit score developed by FICO, a company that specializes in what's known as "predictive analytics," which means they take information and analyze it to predict what's likely to happen. The FICO score is the best-known and the most widely used credit score model in the United States. It is used in about 90%

of consumer-lending decisions, according to a financial-services research firm CEB TowerGroup (Andriotis, 2015). Using mathematical models, the FICO score takes into account various factors in each of these five areas to determine credit risk: payment history, current level of indebtedness, types of credit used and length of credit history, and new credit. FICO company is not a credit reporting agency. In fact, to create credit scores, it takes information provided by one of the three major credit reporting agencies – Equifax, Experian or TransUnion. Both the Federal Home Loan Mortgage Corporation (Freddie Mac) and the Federal National Mortgage Corporation (Fannie Mae) have encouraged mortgage lenders to rely on credit scoring in order to increase consistency across underwriters (Mester, 1997).

While assessing credit risk, reliance on only the credit score is considered insufficient. Even before the mortgage meltdown, industry experts began to worry about the possibility of not fully capturing the credit risk embedded in mortgages. Reasons for concern before sub-prime mortgages began to default included rising loan to value (LTV) ratios, and a decreasing dependence on documentation of a borrower's assets, employment, and income (OCC, 2005). LTV is calculated as the loan amount divided by underlying property value. It is one of the key factors the bank check and monitor from credit risk perspective. As the property is used as a collateral, if the loan default, the bank can take the property and sale it to recover the loss. Therefore, when the property value is higher than the loan amount, the borrower has less motivation to default. The customer can decide to sell the property and payoff the loan with extra money of her own if she could not keep the monthly payment. Since it costs time and

money to sell the house, the bank normally will need 20% cushion for a mortgage loan origination. This is why most banks require 20% down payment when a customer applies a mortgage loan. This type of loan is called a prime loan. If the customer could not pay 20% down payment, a subprime loan could be applied for the amount that is lower than 20% down payment. The subprime loans normally have much higher interest rate than the interest rate for prime loans.

While the Equal Credit Opportunity Act (implemented by the Federal Reserve Board's Regulation B, also called fair lending) prohibits creditors from discriminating in any way during a credit transaction because of an applicant's demographic characteristics, such as race, religion, national origin, gender, marital status, or age, empirical research has shown that these factors do actually have predictive power of credit risk. A basic breakdown of borrowers into sub-prime and prime mortgages reveals some significant demographic distinctions. Sub-prime borrowers are disproportionately minorities, have less income, are older, and have fewer years of education and have significantly less financial sophistication (Lax, 2004). These demographic variables correlate quite well with FICO scores and LTV ratios, as borrowers in the sub-prime segment have both lower FICO scores and high LTV ratios than borrowers in the prime segment (Banasik et. al., 1996).

In order to calculate the LTV, the bank will need both the loan amount and the collateral value. The loan amount is easily captured on the book. In terms of the collateral value, there are multiple ways to get the house value. The most accurate way is to have a formal appraisal, which cost about $350-$500 for a single family house.

Another way is to update the house value based on the house price index as the property value is heavily impacted by the market and the house price index is a good indicator to reflect the house market in different locations. If the house market is going up, the house value will go up as well because the house can be sold at a higher price in a rising market. There are various types of house price indices and the Standard & Poor's Case Shiller (CS) home price index is one of the popular used indices among banks. The CS house price index is the repeated-sales house price index for the United States and it is the leading measure of U.S. residential real estate prices, tracking changes in the value of residential real estate both nationally as well as in the metropolitan regions. The composite and city indices are normalized to have a value of 100 in January 2000. Many banks subscribe the CS indices to manage their property secured residential portfolios, including mortgage, home equity loans and home equity lines, etc.

The house markets are very location oriented. The house with the same features can have very different values in different locations; this is why there is a saying "Location, Location, and Location" in house market. The fair lending prohibits the bank from discriminating the borrowers based on geographical information and hence no such information can be used in the model directly. The CS house price index is at the metropolitan region level and well captures the geographical information.

Based on this industry and academic research as well as interviews with business leaders, I have tested a number of explanatory variables. Ultimately, the final model

variables have been selected based on availability of data, predictive power, and business intuition.

The key risk drivers within mortgage can best be analyzed by examining the relationships among the following variables:

- Current Credit Score (Current FICO or FICO) and FICO score at the time account was booked (origination FICO),
- Month on books (MOB),
- House price index associated with the property's location (CS index), and
- The Loan to Value (LTV) based on original or on a derived adjustment considering the house price appreciation over the years from the origination LTV.

In addition to the factors described above, credit risk can depend on macroeconomic variables and factors. In economic downturns, the default probabilities increase and risk ratings deteriorate. The macroeconomic factors that are considered in this project include unemployment rate and CS index as described above, sourced primarily from U.S. federal government and Moody's economy.com.

1.2 **Methodologies**

Traditional risk assessment methods include discriminate analysis (DA) and logistic regression. Altman (1968) built a famous warning model of multi-variables, the Z-model by using multivariate discriminate analysis. Ohlson (1980) was the first one who used the logistic regression model to predict of financial risks. Wiginton (1980) was one of the first who applied a logistic regression model and discriminate analysis to credit rating and then compared the two methods. Wiginton showed that

the logistic regression model performed better than the discriminate analysis in terms of the proportion of individuals who were correctly classified. However, according to his findings even logistic regression failed to make a significantly high proportion of correct classifications to warrant the use of his model for unaided decision-making. Later, Tang (2002) tested the accuracy of the logistic regression model by sampling 5 listed companies with good financial conditions and 5 companies with bad conditions from the Shanghai and Shenzhen Securities markets and found that logistic regression could distinguish the company with good conditions from the company with bad conditions. Now logistic regression has become the main approach to the classification step in credit scoring and the most commonly used approach in credit risk management.

Numerous other statistical methods that attempted to fit more complex models with higher degrees of nonlinearity between the predictors and the response, such as support vector machines, neural networks, and Bayesian network classifiers, have also been investigated for credit scoring (Im et. al, 2012). The results do not always conclude which method is consistently better than the others. For example, Desai et a1 (1996) found that neural networks performed significantly better than linear discriminant analysis for predicting the 'bad' loans, whereas Yobas et a1 (2000) reported that the latter outperforms the former method. Furthermore, most of these studies only evaluated a limited number of classification techniques on one particular credit scoring data set. Hand (2006) argued that potential performance improvements attainable using more complex models were often offset by other sources of

uncertainty that were exacerbated by the added complexity. In addition, in the real business world, the choice of the best methodology also takes the cost and benefit into consideration. The increased complexity of these model methodologies may increase the implementation cost with only a marginal benefit; that is why logistic regression analysis has become the standard approach in banking industry.

Survival analysis is an area of statistics that deals with the analysis of survival data. The survival data can be collected in medical or reliability studies, for example, when a deteriorating system is monitored and the time until event of interest is recorded. The credit risk data is very similar to the survival data. The time until the loan gets to default in the credit risk data can be viewed as the time until the event of interest (e.g., death) in the survival data. In this interpretation, survival analysis can serve as a useful statistic tool for credit risk management. The idea of employing survival analysis for building credit-scoring models was first introduced by Narain (1992) and then developed further by Thomas et al. (1999). Narain (1992) applied the accelerated life exponential model to 24 months of loan data. The author showed that the proposed model estimated the number of failures at each failure time well. Then a scorecard was built using multiple regressions, and it was shown that a better credit-granting decision could be made if the score was supported by the estimated survival times. Thus, it was concluded by Narain (1992) that survival analysis could add a new dimension to the standard approach. However, the author did not make any comparison with alternative methods.

Even though the survival analysis has been introduced long time ago, it has not been thoroughly investigated and applied in the industry. The main purpose of this thesis is to apply both logistic regression and survival analysis methods to a large dataset of mortgage portfolios and to compare two methods in terms of data fit and prediction power. The long-term goal of this thesis is to learn both methodologies and their respective advantages and to be able to apply them effectively in my actual work.

The rest of this thesis is organized as follows. Chapter 2 introduces the basic concepts and the literature review on the methods used in this thesis. Chapter 3 describes the initial data analysis. Chapter 4 discusses the model results from both logistic regression and survival analysis methods and compares the model performances. Chapter 5 provides the final comments as well as the potential broad impact.

CHAPTER 2: REVIEW OF CONCEPTS AND LITERATURE

2.1 Logistic Regression

Logistic regression is a generalized linear model technique that allows one to predict discrete outcomes. The response variable in logistic regression is a Bernoulli variable that can take the value 1 with a probability of success θ, or the value 0 with probability of failure 1-θ. For credit risk analysis, let define a random variable D that takes values 1 and 0, where the value of 1 (D = 1) means the loan is default and 0 means the loan is not default. Then the probability of default is defined as the probability of success for the random variable D, that is θ=P(D=1). Although not as common and not discussed in this thesis, applications of logistic regression can be been extended to cases where the response variable has more than two categories known as a multinomial regression.

In logistic regression, the relationship between the response and the independent variables is described by the logit transformation of θ as follows:

$$\theta = \frac{e^{(\alpha + -\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}{1 + e^{(\alpha + -\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}},$$

where α is the intercept of the equation, β are the coefficients of the independent variables, and n is the number of independent variables.

An alternative form of the logistic regression equation is the following:

$$\text{Logit}\,[\theta(x)] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

or in terms of the credit risk variables:

$$\text{Logit}\,[P(D=1)] = \log\left[\frac{P(D=1)}{1-P(D=1)}\right] = \alpha + \beta_1 \text{MOB} + \beta_2 * \text{LTV}$$

$$+\beta_3 * \text{FICO} + \beta_4 * \text{CS},$$

where CS is the Case Shiller house price index, FICO score is the credit score, MOB is the number of months a loan in on book, and LTV is the loan to value calculated as the loan amount divided by collateral value. This logistic regression equation is going to be used to estimate the probability of default. More detailed description of the credit risk variables can be found in Chapter 1.1: Risk Profile Review.

Over the years, logistic regression has been the most commonly used methodology in credit risk modeling. For example, Kutty (1990) presented a logistic regression model for determining the default probability of developing countries debt. The study incorporated 79 countries' debts over a period of 19 years. The model predicted the default of the country's debt for Mexico, Brazil, and Argentina two years in advance. Westgaard et al (2001) applied a logistic regression model to predict the default probability based on financial variables.

Recently, Gurný (2013) estimated the PD of US banks using several statistical models, including logistic regression, probit model and linear discriminate analysis (LDA). In his work, the author analyzed a sample of 298 American commercial banks for model estimation which was collected during the financial crisis during the years 2007-2010. The stepwise selection was applied for logit and probit model. Based on the fit in the training data, logit model and probit model achieved a very similar explanatory power (96.30% for logit model and 95.85% for probit model in terms of pseudo R-square), even though the probit model had one extra indicator. The LDA model had a lower explanatory power (78.44%). LDA model predicted the response

outcome slightly better for non-default banks, but much worse for default banks. For out of sample analysis, the logit model outperformed among the three with average fit 80.4%, comparing with 62.2% from the probit model and 42.6% from the LDA model. The results of ROC analysis showed that the logit model also had the best performance with the area under the curve (AUC) of 96.48% compared to the probit model and the LDA model that had 82.28% and 83.52% , respectively. The AUC provides a simple figure of merit for the performance of the constructed classifier. Overall the results of these analyses confirmed that the logit model outperformed other models in application to both in training data and testing.

Earlier, Baesens et al (2003) also conducted a benchmarking study of various classification techniques on eight real-life credit scoring datasets originating, among others, from major Benelux (Belgium, The Netherlands and Luxembourg) and UK financial institutions. The techniques that were explored were logistic regression (LR), linear and quadratic discriminate analysis (LDA), and linear programming support vector machines (SVMs), neural networks naïve Bayes (NN), Decision trees and rules (DT) and K-nearest-neighbor classifiers (KNN). The performance criteria for classification were based on the AUC and the percentage of correctly classified (PCC) observations, which measured the proportion of correctly classified cases. Based on the eight datasets, the results indicated that different modeling techniques had different performance in different datasets. For example, the author found that linear SVM had the best performance for Australia portfolio while NN works the best on German portfolio in terms of both PCC and AUC. In general, it could be observed that the best

average rank was attributed to the NN classifier. However, the simpler, linear classification techniques such as LDA and LR also had a very good performance, which was in the majority of the cases not statistically different from that of the SVM and NN classifiers. Based on the research, Baesens et al (2003) concluded that the more complex models generally performed quiet similarly to logistic regression, in terms of predicting probability of default.

2.2. Survival Analysis

Survival analysis is one of the alternative approaches to logistic regression that have not been extensively explored; selected studies include Thomas et al., 1999; Stepanova et al., 2002; and Im, 2012. Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. In this thesis, the event of interest is the default of a mortgage loan.

The time to event can be measured, for instance, in days, weeks, months, or years. In this thesis, the time to default is recorded in months. Let T denote the time to default of a mortgage loan, and f(t) be the probability density function (pdf) and F(t) be the cumulative density function (cdf), or the probability that a loan will be less than or equal to any value t, F(t) = Pr {T <= t). Then the survival function can be defined by the following equation:

$$S(t) = P(T > t) = 1 - F(t).$$

For continuous survival data, the hazard function is a more popular characteristic than the pdf to describe the distributions. The hazard function is defined as a limit:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t | T \ge t)}{\Delta t},$$

which represents the instantaneous risk that an event occurs at time t.

Specifically, the survival probability is the probability that the loan i will "survive" beyond time period τ ($T_i \ge \tau$), which is $S(t_{i\tau}) = P(T_i > \tau)$, where i = 1, 2, …, I, is the number of loans; and τ = 1, 2, …, M, is the number of months. The hazard function at time t in this thesis is then defined as the probability of the default in time period τ ($T_i = \tau$) given that the loan did not default in any earlier time period ($T_i \ge \tau$). This definition implies that h(t) must be a "conditional probability": it is conditional on not having the event up to time τ (or conditional on surviving to time τ.) . Therefore, the probability of default could be expressed in the form of conditional probability$h(t_{i\tau}) = P(T_i = \tau | T_i \ge \tau)$. The greater the value of h(t), the higher the risk of the default by time τ. Each loan may have completely different hazard function as hazard is the characteristic of an individual. Unlike logistic regression, survival analysis models the distribution of the time to default, which then can be derived as the probability of default within some specified period of time.

To apply survival analysis in consumer credit modeling, we suppose that one or more further measurements are available for each individual, so that we have a vector of covariates, **X**, e.g., application characteristics such as current FICO score, current Loan to value, etc. In order to assess the relationship between the distribution of default time and these covariates, Cox (1972) proposed the following model:

$$h(t;\boldsymbol{X}) = e^{(X\beta)}h_0(t), \qquad (2.1)$$

where β is a vector of unknown parameters and h_0 is an unknown function giving the hazard for the standard set of conditions, when $\boldsymbol{X}$ = 0. It's called the proportional hazards (PH) model because the assumption is that the hazard of the individual with application characteristics X is proportional to some unknown baseline hazard. The vector of coefficients β is estimated using maximum likelihood.

PH models assume that the hazard functions are continuous. However, credit performance data are usually recorded only monthly so that several defaults at one time can be observed. These are tied default times, and the likelihood function must be modified because it is now unclear which individuals to include in the risk set at each default time t_1, t_2, t_3 ... The exact likelihood function has to include all possible orderings of tied defaults (Kalbfleisch and Prentice 1980), and hence is very difficult computationally. A number of approximations have been developed. One of these is achieved by replacing equation (2.1) by a discrete logistic model (Cox, 1972):

$$\frac{h(t;X)}{1-h(t;X)} = e^{(X\beta)}\frac{h_0(t)}{1-h_0(t)},$$

where $h(t,X) = P(t \le T < t+1 | T \ge t)$.

And then similarly to logistic regression, a logit link function can be used:

$$\text{Logit h(t)} = \log\left(\frac{\text{h(t)}}{1-\text{h(t)}}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

or in term of the credit risk variables:

$$Logit\ h(t) = \log\left(\frac{h(t)}{1-h(t)}\right) = \alpha + \beta_1 * MOB + \beta_2 * LTV + \beta_3 * \text{FICO} + \beta_4 * CS,$$

where CS is the Case Shiller house price index, FICO score is the credit score, and LTV is the loan to value calculated as loan amount divided by collateral value. More detailed description of the credit risk variables can be found in Chapter 1.1: Risk Profile Review.

Thomas et al. (1999) compared performance of exponential, Weibull and Cox's nonparametric models with logistic regression and found that survival-analysis methods were competitive with, and sometimes, superior to, the traditional logistic regression approach. The paper was developed based on personal loan data from a major UK financial institution. The data consisted of application information of 50,000 loans accepted between June 1994 and March 1997 together with their monthly performance description for the period up to July 1997. The monthly performance indicators were used to determine whether the loan was censored or defaulted, therefore, for each loan there was a survival time. In order to compare with standard credit scoring approaches, the data is also used to develop logistic regression model. The analysis and results suggested that proportional hazard models investigated in this sample were competitive with the logistic regression approach in identifying those loans who defaulted in the first year. The proportional hazard results for the second year with fewer defaults were not as encouraging and suggested that more sophisticated models might be appropriate. The survival analysis approach benefited more from a large sample of 'bads' than did the logistic regression approach. The poor performance under the second year criterion was also partly due to the fact that the ordering of risk of default did not change whatever the time period.

It was noted by Thomas et al. (1999) that there were several possible ways of improving the performance of the simplest survival-analysis models and Stepanova et al. (2002) explored three extensions of Cox's proportional hazard model. Another extension that could be used in survival analysis was to allow the coefficient to be time dependant as I allowed in this thesis. Due to its complexity, there are not many articles that apply survival analysis with time varying covariate in credit risk analysis. Im (2012) introduced a modification of the proportional hazards survival model that included a time-dependant variable in the model (Time-dependent proportional hazards TDPH) to capture temporal phenomena. The TDPH survival model represented the effects of dynamic economic conditions in a direct manner, without the need to identify a set of underlying macroeconomic factors that best characterizes the current state of the economy in terms of its impact on consumer credit risk and included them as additional predictor variables. The article was developed using a very large, real data set from a consumer credit company. The data consisted of the customers who were approved between January 2003 and July 2008 with monthly observations. The author tried TDPH model and compared with standard PH and logistic regression model by comparing the ROC curves and related performance measures based on 9-month default rates. The four models were the TDPH survival model, the standard PH survival model, a standard LR model, and an LR model with TDPH factor γ. For the LR model with TDPH factor γ, the standard LR model was fitted first, then the γ from TDPH estimation was included as an additional predictor variable. The article concluded that all four methods have somewhat similar

performance in terms of the ROC curves, however, the TDPH models did not perform better than the LR model in terms of the KS statistics. The similar performance of LR versus the standard PH method was consistent with what Stepanova and Thomas (2002) observed. Thus, inclusion of a time-dependency factor via TDPH modeling appeared to have potential benefit for the objective of scoring.

In practice, scoring a new customer using the TDPH model or the LR model with the TDPH factor γ would involve the forecast of the near-future γ values. As γ changes relatively smoothly for the most part, reasonable accurate extrapolation into the new-future is not infeasible. However, this will involve an additional model development in the real world and introduce more model risk. In addition, in the more recent Comprehensive Capital Analysis and Review (CCAR) effort that many banks are taking, the banks are required to forecast the expected losses for a much longer term which will face challenge of predicting the time-dependency factor γ for a longer term under such approach.

CHAPTER 3: DATA AND PRELIMINARY ANALYSIS

3.1 Data Description

The mortgage portfolio used in this thesis is a sample of 6106 distinct loan accounts that originated in 2004 and the information for each mortgage is collected monthly over period from January 2005 to May 2010 as long as it's on book. The observation is taken randomly every year for each mortgage loan. This means the number of months between the observation month and the default month is randomly distributed from 1 to 12 months. Based on this sampling method, there are a total of 20918 observations.

The rate of default for mortgage portfolio has been very low from January 2005 to June 2007, less than 0.5%, then increased to around 1% until January 2009, and then rapidly increased to as high as 2.5% in June 2009, during the well known sub-prime financial crisis.

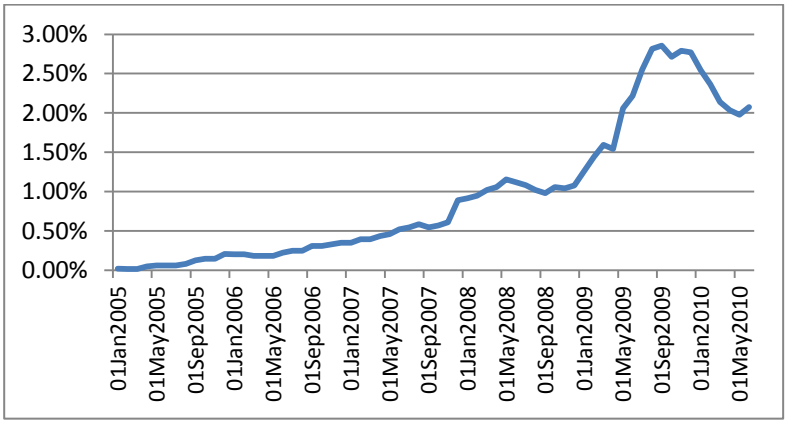


Figure 1: Mortgage Portfolio Default Rate

For the purpose of modeling, a random sample of 70% of the observations (14640) is selected; the rest of 30% (6278) of observations are used for testing.

The mean default rate in the training dataset is 0.7%. It is similar to the mean default rate in the testing dataset that is 0.8%. The key risk factors are also very similar. The average month on book in the training dataset and the testing dataset are almost the same (34.3 vs. 34). The average current FICO scores are 751 and 752 and the average current LTVs are 0.61 and 0.60 in the training data and the testing data, respectively. Therefore, both the training and the testing datasets have similar characteristics.

Table 1: Summary statistics of the key variables in the training and testing dataset

Variable	Training Data		Testing Data	
	Mean	**Std Dev**	**Mean**	**Std Dev**
Default rate	0.7%	0.1	0.8%	0.1
Current FICO	751	66	752	65
Current LTV	0.61	0.2	0.60	0.2
MOB	34.0	17	34.3	17
Origination FICO	736	56	737	57
Unemployment rate	6.27	2.2	6.28	2.2
Case Shiller	139	36	139	36

3.2 Univariate Analysis

As discussed earlier, FICO score is a very important factor that the majority of the credit industries use for risk management. The origination FICO score is the FICO score from the loan's application file. The origination FICO score does not change over the loan period, however, it defines the status of the customer's credit application which then may serve as a good indication for PD over the loan lifetime as shown in the Figure 2.

As shown in Figure 2, the loans with the origination FICO scores less than 660 have significantly higher default rate than the loans with the origination FICO scores

greater than or equal to 660. Generally, many banks have a credit policy that sets the lowest FICO score for which a loan application can be approved; but them almost all banks would have an exception policy according to which some loan applications that do not meet the credit requirements can also be approved. The lowest required FICO score can vary among the banks, ranging from 620 to 660; therefore, the loans that have the FICO score lower than 660 could be sometimes exceptionally approved.. There is a clear relationship between the origination FICO score and the rate of default rate as shown by Figure 2. The curve plotted in blue illustrates the relationship between the rate of default and the origination FICO score of all loans, whereas the curve plotted in red illustrates this relationship of only the loans with the origination FICO scores exceeding 660.

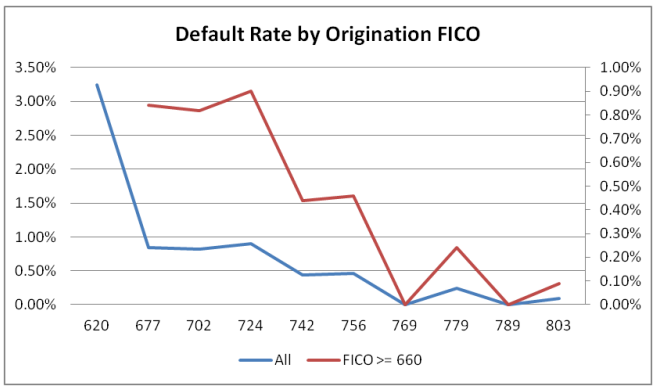


Figure 2: Default rate by origination FICO score

The FICO score is an indicator of a risk at a particular point in time. It changes as new information is added and as historical information ages. For example, past credit problems impact one's credit score less as time passes. Lenders request a current score when a new credit application is submitted, so they have the most recent information available.

As shown in Figure 3, similarly to the relationship between the rate of default and the origination FICO score, with the current FICO score increases, the rate of default decreases quickly for FICO scores below 660 and then remains relatively low for FICO scores above 750. After removing the loans with the current FICO scores less than 660, the rate of default shows a slightly different trend, it decreases for the current FICO scores less than 748, increases for the scores between 748 to 790, and then again decreases for the scores higher than 790. This observation suggests a difference in the modeling of PD of loans with the origination/current FICO scores below and above 660.

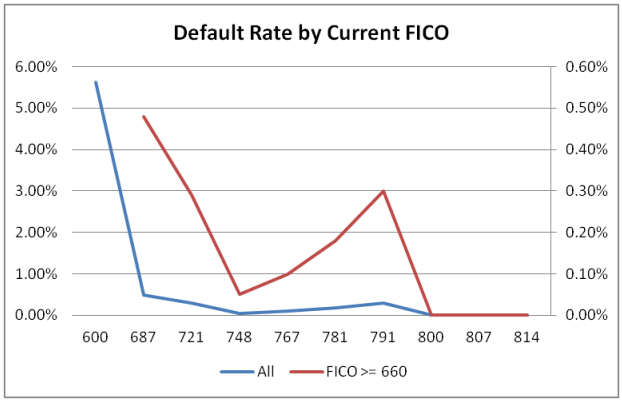


Figure 3: Default rate by current FICO score

As explained in Chapter 1.1, the current LTV is calculated as the current total loan amount divided by the current property value. The LTV is one more key factor that determines if a loan can be approved. In traditional residential mortgages and home equity loans there is an 80% rule, that is if the mortgage's LTV is more than 80%, the loan is most likely not approved or has to go through the exception review process. According to this 80% rule, a binary dummy variable that takes a value of 1, when the LTV is greater than 80%, is created and included in our initial modeling.

Figure 4 also shows that when the current LTV is greater than 80%, the default rate increases dramatically from below 0.4% to over 1%.

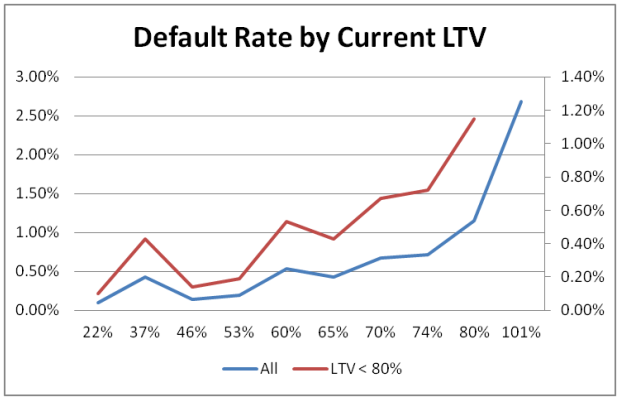


Figure 4: Default rate by current LTV

The property value is heavily impacted by the local house market which is reflected in CS index. Based on the data, there is no clear trend of the default rate and CS index directly. However, I observed that if CS one year growth rate is less than -12%, which means the house prices decreased 12% comparing with the price a year ago, the PD is significantly higher, as the blue line shows in Figure 5. The CS one year growth rate is calculated as the CS index today minus CS index a year ago and then divided by CS index a year ago. After removing the loans with CS 1 year growth less than -12%, the default rates generally decrease with the increased CS growth rate, which is the red line in Figure 5.

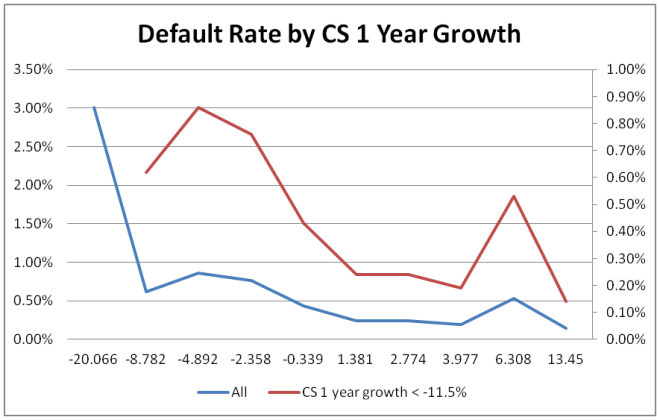


Figure 5: Default rate by Case Shiller 1 year growth

The unemployment rate is the most important macroeconomic factor that many bank tracks. When the unemployment rate is getting higher, more people lose their jobs from which many people get their main source of mortgage payment. The mortgage data analyzed in this thesis also confirmed that the higher the unemployment rate, the higher the probability of default as shown in Figure 6, especially after the unemployment rate reaches to around 6.5 - 7%, which will create the panic of the customers and then impact the confidence index.

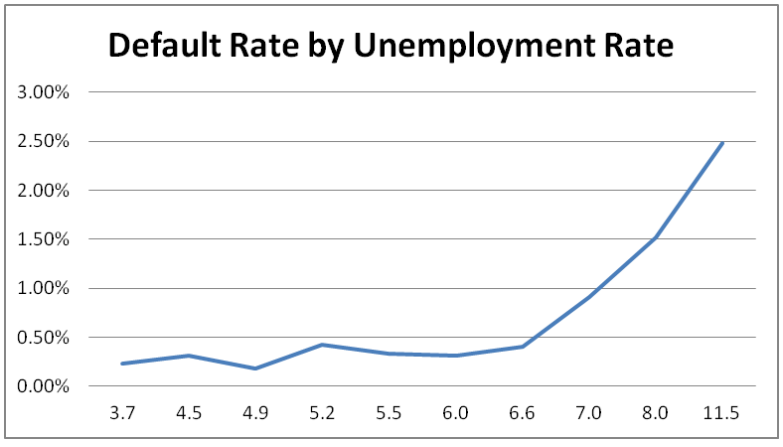


Figure 6: Default rate by unemployment rate

Month on book is also another key factor in the bank's monitoring process. The default rate is very low in the first year or two for mortgage loans. After three or five years (36 month to 60 month), the loan default rate may increase dramatically as seen in Figure 7.

Due to the specific history of the underlying data, the unemployment rate has been increasing along the month on book, therefore, the default rates have very similar trend with the two variables.

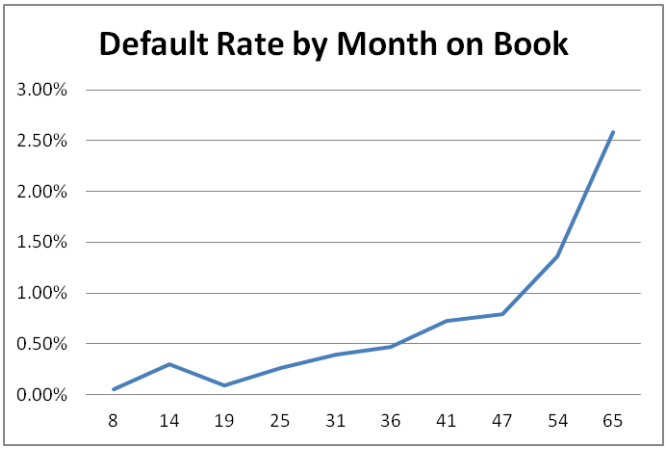


Figure 7: Default rate by month on book

Based on the univariate analysis, the following initial set of variables was selected:

- Current FICO score and dummy variables derived based on current FICO;
- Origination FICO score and dummy variables based on Origination FICO;
- Current LTV and dummy variable derived based on current LTV;
- CS growth rate;
- Unemployment rate;
- Month on book;

3.3 Correlation Analysis

Table 3 is a matrix of correlation coefficients for each pair of the most important variables. The purpose of this correlation analysis is to find the pairs of variables that are highly correlated and would require additional caution if included in the model.

The p-values are all less than 0.05, which means that the correlations among all variables are statistically significant. However, one should not confuse statistical significance with practical importance. If the sample size is large enough, even a weak correlation can be statistically significant.

In order to assess practical importance, one common computation is to square the correlation coefficient to get the coefficient of determination. This shows how much of the variation in one of the variables is associated with the variation in the other. For example, an r of 0.06273 between the current LTV and the month on book produces an R-square of only 0.39% (0. 06273 * 0. 06273 = 0.0039, or 0.39%). This means the knowledge of the month on book would account for only 0.39% of the variance in the current LTV, even though the p-value for their correlation is less than 0.05.

Hinkle et al (2003) proposed a rule of thumb for interpreting the size of a correlation coefficient (see Table 2). Note that the interpretation of correlation can also vary on the size of the data analyzed.

Table 2: Rule of thumb for interpreting the size of a correlation coefficient

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

The sample size in this thesis is large, so only the correlation higher than 0.6 – 0.8 is considered to be high enough for further consideration. This also corresponds to a range of 36% to 64% for the coefficient of determination. Therefore, only the correlation between the origination FICO score and the current FICO score are considered to be highly correlated and require extra caution if both of them are included in the model. The correlation between the month on book and the unemployment rate is 0.55766, which means the 31% of the variance in one variable can be explained by the variance in another variable. Even though it's not over 0.6, we will also need to be careful if both of the variables are to be included in the same model. The month on book and the CS one-year growth rate have a similar situation.

Table 3: Correlation matrix of the key variables for the mortgage portfolio

	Month on book	**Unemployment rate**	**Current LTV**	**Current FICO**	**Origination FICO**	**CS growth rate dummy**
Month on book	1	0.55766	0.06273	-0.03138	-0.0405	-0.59107
		<.0001	<.0001	0.0001	<.0001	<.0001
Unemployment rate	0.55766	1	0.33823	-0.06895	-0.06489	-0.50898
	<.0001		<.0001	<.0001	<.0001	<.0001
Current LTV	0.06273	0.33823	1	-0.20754	-0.20357	-0.25939
	<.0001	<.0001		<.0001	<.0001	<.0001
Current FICO	-0.03138	-0.06895	-0.20754	1	0.63529	0.01752
	0.0001	<.0001	<.0001		<.0001	0.0341
Origination FICO	-0.0405	-0.06489	-0.20357	0.63529	1	0.04269
	<.0001	<.0001	<.0001	<.0001		<.0001
CS growth rate dummy	-0.59107	-0.50898	-0.25939	0.01752	0.04269	1
	<.0001	<.0001	<.0001	0.0341	<.0001	

From Figure 8, the origination FICO score and the current FICO score have a positive relationship that both are higher or lower at the same time. When the month

on book increases, especially after the month on book is higher than 40 months, the unemployment rate also increases. Since the loans are originated in 2004, it's getting to 2007-2008 financial crisis period after 40 months, so the more time the loan is on book, the higher the unemployment rate. Similarly, the CS one year growth is higher when either the month on book or the unemployment rate is lower. There are no obvious relationships among the other factors.

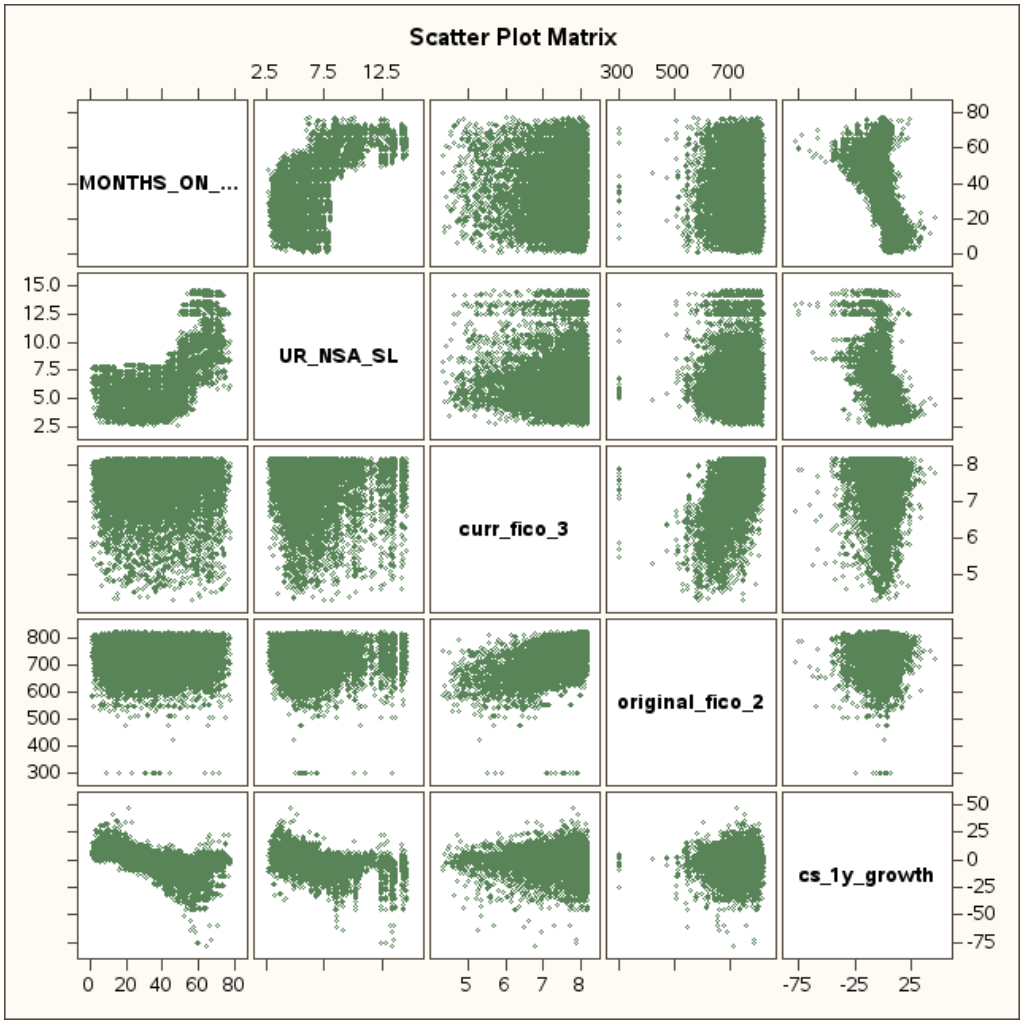


Figure 8: Scatter plot matrix of the key variables for the mortgage portfolio

CHAPTER 4: ESTIMATION

4.1. Logistic Regression

4.1.1 Final Model Selection

As discussed in Chapter 2, the logistic regression model applied in this thesis is as following:

$$\text{Logit}\,[\text{P(D}=1)] = \log\left[\frac{\text{P(D}=1)}{1-\text{P(D}=1)}\right] = \alpha + \beta_1 \text{MOB} + \beta_2 * \text{LTV}$$

$$+\beta_3 * \text{FICO} + \beta_4 * \text{CS}.$$

To fit the logistic regression model, the procedure *Proc Logistic* in SAS statistical software is applied and the estimation is based on the maximum likelihood function. I first run the logistic regression with all the initial set of factors based on univariate analysis detailed in Chapter 3.2. Table 4 shows the final model selected according to the stepwise regression and the review of the coefficients.

Table 4: Logistic regression model estimation

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	6.2115	0.857	52.5286	<.0001
MOB	1	0.0304	0.00694	19.133	<.0001
LTV	1	0.9295	0.361	6.6285	0.01
FICO	1	-1.7805	0.1075	274.4993	<.0001
CS growth rate dummy	1	-1.2205	0.2421	25.4138	<.0001

The model has four variables:

- x_1: MOB: number of month the account has been on book.
- x_2: LTV: current loan to value;

- x_3: FICO: current FICO score;
- x_4: CS growth rate dummy: Case Shiller 1 year growth greater than -12%;

The month on book and the current LTV are having positive sign which means the default rate is higher when the month on book and the LTV are higher. The current FICO score has a negative sign which means the default rate is lower when the FICO score is higher. And the CS index 1 year growth greater than -12% also has negative sign which means the default rate is lower for the segment with the CS index 1 year growth greater than -12% comparing with the segment of loans with the rate less than or equal to -12%. All the variables are statistically significant and make business sense. For example, the FICO score getting higher means the credit worthiness is better for a customer, and hence the probability of default will be smaller.

4.1.2 Model Fit Statistics

A Wald test is used to test the statistical significance of each coefficient (β) in the model. A Wald test calculates a Z statistic, which is:

$$Z = \hat{\beta}/SE$$

This z value is then squared, yielding a Wald statistic with a chi-square distribution.

However, several authors have identified problems with the use of the Wald statistic. Menard (1995) noted that for large coefficients, standard error is inflated, lowering the Wald statistic (chi-square) value. Agresti (1996) stated that the likelihood-ratio test is more reliable for small sample sizes than the Wald test. The likelihood-ratio test uses the ratio of the maximized value of the likelihood function

for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)]$$

This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistic to use when building a model through backward stepwise elimination.

Both the Wald test and the likelihood ratio test indicate that the coefficients for the model are statistically significant.

Table 5: Testing on the null hypothesis that the coefficients equal to 0 for logistic regression model

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	509.0738	4	<.0001
Wald	386.9448	4	<.0001

The deviance test is used instead of R^2 as the statistic for the overall fit of the logistic regression model. It is the fit of the observed values to the expected values. The bigger the difference (or "deviance") of the observed values from the expected values, the poorer the fit of the model. The maximum likelihood is a way of finding the smallest possible deviance between the observed and predicted values. The deviance is usually referred to as the "negative two log likelihood" (shown as "-2 Log L" in SAS). The deviance statistics is called -2LL by Cohen et al. (2003) and D by some other authors (Hosmer and Lemeshow, 1989), and it can be thought of as a chi-square value.

Akaike Information Criterion (AIC) and Schwarz Criterion (SC) are deviants of negative two times of the Log-Likelihood (-2 Log L) which penalizes the log-likelihood by the number of predictors in the model. AIC is calculated as AIC = -2 Log L + 2((k-1) + s), where k is the number of levels of the dependent variable and s is the number of predictors in the model. SC is defined as - 2 Log L + ((k-1) + s)*log(Σf_i), where f_i's are the frequency values of the ith observation, and k and s are defined as above. Like AIC, SC penalizes for the number of predictors in the model.

Table 6: Model fit statistics for logistic regression model

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1316.696	815.622
SC	1324.288	853.58
-2 Log L	1314.696	805.622

These three model fit testing statistics are used to choose among different candidate models with the smallest value as the best model. AIC, SC and deviance test indicate that the final model are better than the model with only intercept.

4.1.3 Hosmer-Lemshow Goodness of Fit Test

The Hosmer-Lemeshow test is a statistical test for the goodness of fit for the logistic regression model. The data are divided into approximately ten groups defined by increasing order of estimated risk. The observed and expected number of cases in each group is calculated and a Chi-squared statistic is calculated as follows:

$$\chi_{HL} = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{E_g(1-\frac{E_g}{n_g})},$$

with O_g, E_g and n_g be the observed events, expected events and number of observations for the gth risk decile group, and n be the number of groups. The test statistic follows a Chi-squared distribution with n-2 degrees of freedom. A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer >= 0.05) indicates a good logistic regression model fit. The P value is 0.1359, which means the model has a good fit.

Table 7: Hosmer and Lemeshow goodness-of-fit test for logistic regression

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
12.3577	8	0.1359

4.1.4 Rank Ordering Testing

Receiver Operating Characteristic (ROC) curve is a two-dimensional graph that visually depicts the performance and performance trade-off of a classification model (Fawcett, 2004). ROC curves are industry standard methods for comparing two or more scoring algorithms (Thomas et al, 2004). In a ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate (1-specificity) for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold.

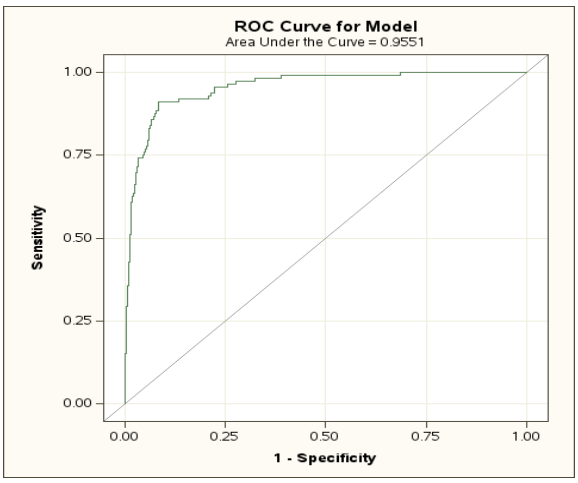


Figure 9: ROC curve for the logistic regression model

The area under the curve (AUC), also referred to as index of accuracy (A), or concordance index, c, in SAS, and it is an accepted traditional performance metric for a ROC curve. The AUC for the final model in training dataset is 0.9551.

A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value. It's the measure of the model power in terms of the rank ordering.

Another most widely used way to evaluate quality of a scorecard is the Gini coefficient besides ROC curve. The Gini coefficient had its first application in economics measuring the degree of inequality in income distribution and was calculated using the Lorenz curve (Kleiber, 2007). The Gini index has been brought into a lot of applications (Hand, 2005; Chatterjee et al, 2007), including credit scoring, where it is often referred as the accuracy ratio or power ratio. The Gini coefficient is used as a measure of how well a scorecard or variable is able to distinguish goods and bads. It is a rank ordering correlation coefficient and is exactly the same as the

Somer's D statistics provided by SAS, which is used to determine the strength and direction of relation between pairs of variables. Its values range from -1.0 (all pairs disagree) to 1.0 (all pairs agree). It is defined as (nc-nd)/t where nc is the number of pairs that are concordant, nd is the number of pairs that are discordant, and t is the number of total number of pairs with different responses.

Table 8: Model performance testing for logistic regression

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	95.5	**Somers' D (Gini)**	0.91
Percent Discordant	4.5	**Gamma**	0.91
Percent Tied	0	**Tau-a**	0.014
Pairs	1627136	**c**	0.955

Another common measure of discrimination used in credit scoring is Kolmogorov–Smirnov test (KS) statistic. Traditionally the KS statistic is used to compare an unknown, observed distribution to a known, theoretical distribution. The maximum distance between the cumulative distributions are calculated and measured against a critical value. If the distance is less than the critical value, there is a good chance that the distributions are the same.

In credit scoring, KS is often calculated as the maximum distance between the cumulative distribution of the predicted probability of defaults and the cumulative distribution of the predicted probability of the non defaults.

Let $F_{d,n1}(x)$ and $F_{non,n2}(x)$ be the empirical cumulative distribution functions of the default segment and the non-default segment respectively. The KS statistic in this case is $D_{n1,n2} = \sup |F_{d,n1}(x) - F_{non,n2}(x)|$, where sup is the supremum function and gives the max of the distance of the two distributions.

The null hypothesis that the two segments are from the same population will be rejected at level α if $D_{n1,n2} > c(\alpha)\sqrt{\frac{n1+n2}{n1*n2}}$. For example, in this analysis, $c(\alpha)\sqrt{\frac{n1+n2}{n1*n2}} = 1.36 * \sqrt{\frac{14640+6278}{14640*6278}} = 0.0205$ at level 0.05 ($c(\alpha)$ is 1.36), which is much smaller than 0.828, therefore, the null hypothesis will be rejected. In the credit world, the D value is more important than $c(\alpha)\sqrt{\frac{n1+n2}{n1*n2}}$. The D value ranges from 0 to 1 or 0 to 100 in percent format, with the higher D, the better distinguish the default and non-default segments, hence the better performance of the model. The D value is 0.828 for model in the training dataset.

Table 9: Kolmogorov-Smirnow two sample test for logistic regression

Kolmogorov-Smirnov Two-Sample Test		
D	**KS**	**Pr > KS**
0.828253	0.072166	<.0001

4.1.5 Residual Analysis

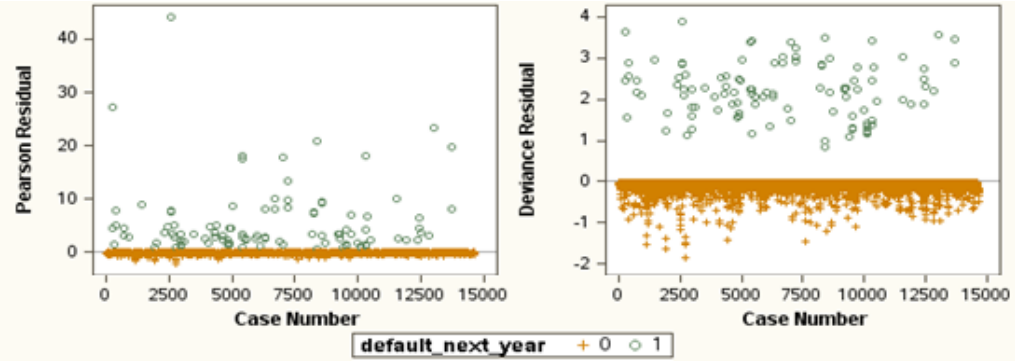


Figure 10: Pearson residual and deviance residual for logistic regression

The logistic regression function in SAS provides the Pearson and deviance residuals based on the diagnostics developed by Pregibon (1981). The Pearson and

deviance residuals are useful in identifying observations that are not explained well by the model. The Pearson residuals for the jth observation is:

$$\chi = \frac{\sqrt{w_j}(r_j - n_j\hat{p}_j)}{\sqrt{n_j\hat{p}_j\hat{q}_j}};$$

And the deviance residual for the jth is:

$$d_j = \begin{cases} -\sqrt{-2w_j n_j \log(\hat{q}_j)} & if\ r_j = 0 \\ \pm\sqrt{2w_j\left[w_j \log(\frac{r_j}{n_j\hat{p}_j}) + (n_j - r_j)\log(\frac{n_j - r_j}{n_j\hat{q}_j})\right]} & if\ 0 < r_j < n_j \\ \sqrt{-2w_j n_j \log(\hat{p}_j)} & if\ r_j = n_j \end{cases},$$

where r_j is the number of event response out of n_j trials for the jth observation; w_j is the weight of the jth observation; $\hat{p}_j$ is the estimate of π_j evaluated at $\hat{\beta}$, and $\hat{q}_j$= 1- $\hat{p}_j$; π_j is the probability of an event response for the jth observation given by $\pi_j = F(\alpha + \beta' x_j)$, where F(·) is the inverse link function; and $\hat{\beta}$ is the maximum likelihood estimate of $(\alpha, \beta_1, \ldots, \beta_s)'$.

Pregibon (1981) suggests using the index plots of several diagnostic statistics to identify influential observations and to quantify the effects on various aspects of the maximum likelihood fit. In general, the distributions of these diagnostic statistics are not known, so cutoff values cannot be given for determining when the values are large. However, the plots provide displays of the diagnostic values, allowing visual inspection and comparison of the values across observations. As shown in Figure 10, the model fits the non default segment better than it fits the default segment. This finding is in line with the business expectation as the default is a rare event hence it's hard to model.

4.1.6. Model performance in both training dataset and testing dataset

As discussed in Chapter 3.1, 70% random sample is taken for the model development and the rest of the 30% is used to check the model performance as out of sample testing. AUC is 0.949 in the testing dataset comparing with 0.955 in the training dataset. The model could not distinguish the default segment from the non-default segment in the testing dataset (KS 0.788) as well as it does in the training dataset (KS 0.828).

Table 10: Model performance testing in both training and testing datasets for logistic regression

Association of Predicted Probabilities and Observed Responses		
	Logistic Regression	
	Training Data	**Testing Data**
Percent Concordant	95.5	94.9
Somers' D (Gini)	0.91	0.898
c (AUC)	0.955	0.949
KS	0.828	0.788

4.1.7 Cross Validation

For model prediction, we would like an estimation method with low bias and low variance. There are many reasons for the bias and variances, such as model misspecification, data scarcity, over fitting, etc. Cross validation is one of the testing methods to check for the bias and variance. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.

There are several types of cross validation, including leave-p-out cross validation, leave-one-out cross validation, k-fold cross validation, and repeated random sub-

sampling validation which is the method used in this thesis. One round of such cross validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To follow the sampling selection, 70% of random sample is selected as the training dataset and the rest 30% for the testing dataset for each round. To reduce variability, 1000 rounds of cross-validation are performed, and the validation results are averaged over the rounds. The advantage of this method (over k-fold cross validation) is that the proportion of the training/testing split is not dependent on the number of iterations (folds). The disadvantage of this method is that some observations may never be selected in the testing subsample, whereas others may be selected more than once.

Based on the 1000 runs, the average coefficients as well as the standard deviations for every factor in the model are calculated as listed in Table 11. We can find that the coefficients for the model selected in this thesis are all reside in the 95% confident interval. For example, the coefficient for the month on book is 0.0304 and the average coefficient for this variable is 0.0306 with 95% confidence interval from 0.0229 to 0.0383.

Table 11: The cross validation from 1000 runs for the logistic regression

Model		**Cross Validation**			
Parameter	**Estimate**	**Mean**	**Std**	**95% CI Lower**	**95% CI Upper**
Intercept	6.2115	6.0081	0.4958	5.0363	6.9798
MOB	0.0304	0.0306	0.0039	0.0229	0.0383
LTV	0.9295	0.8593	0.2265	0.4154	1.3033
FICO	-1.7805	-1.7677	0.0604	-1.8860	-1.6493
CS growth rate dummy	-1.2205	-1.0722	0.1488	-1.3638	-0.7806

Figure 11 displays the distribution of the coefficients from the 1000 runs for each variable in the model. All of them are approximately normal distribution with the mean as shown in Table 11. The cross validation results show that the coefficients for the variables are stable for these factors and hence indicate small bias and variance.

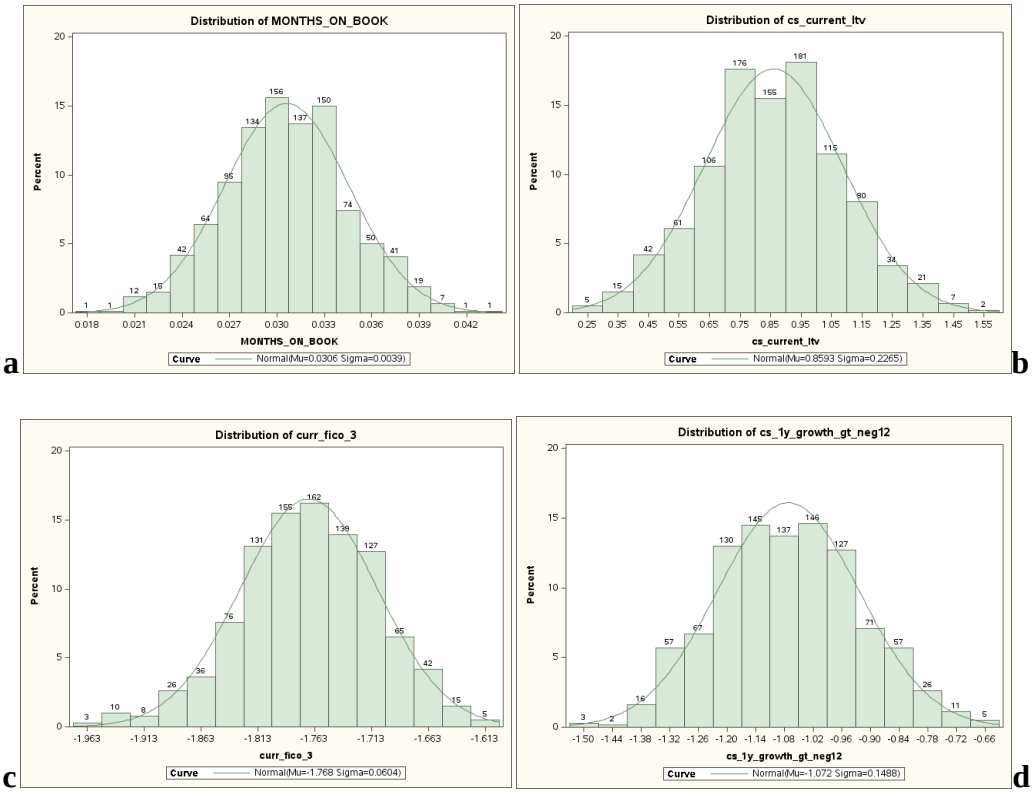
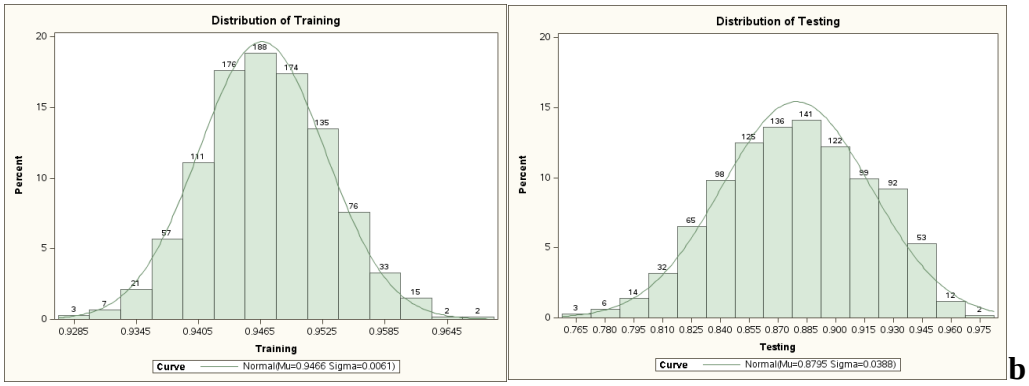


Figure 11: Distribution of the coefficients estimations from 1000 runs

The model normally performs better in the training dataset than in the testing dataset for the 1000 runs. As shown in Figure 12, the average AUC for the model in the training dataset is 0.9466 (graph a) while it's 0.8795 in the testing dataset (graph b).

a 

b

Figure 12: Distribution of the AUC for logistic regression from 1000 runs

4.2. Survival Analysis

Survival analysis models factors that influence the time to an event. Ordinary least squares estimation falls short because the residuals of survival analysis generally does not have a normal distributed and the model cannot handle censoring which is very common in survival data.

4.2.1 Probability of Density Function (pdf)

Density functions are essentially the histograms comprised of bins of vanishingly small widths. As indicated in Figure 13, the shorter survival times between 30 month and 60 months are more probable, indicating that the risk of the loan default in these periods is high.

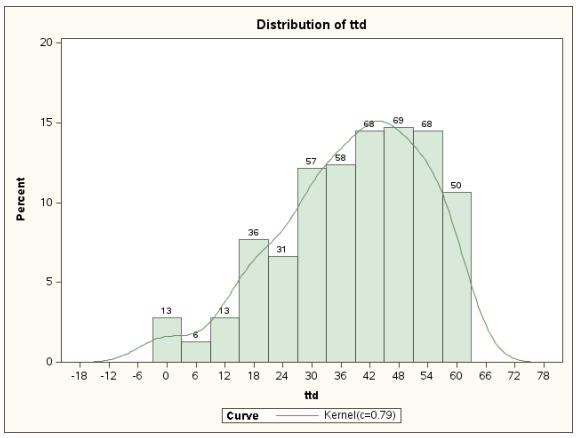


Figure 13: Distribution of the time to default for defaulted segment

Figure 13 is the pdf for all the defaulted loans, while Figure 14 is the pdf for all the loans in this thesis. We can see that there are a lot of loans censored around 67 to 78 months. This is due to the loans in the sample are originated in 2004 and majority of the loans have not defaulted before censoring.

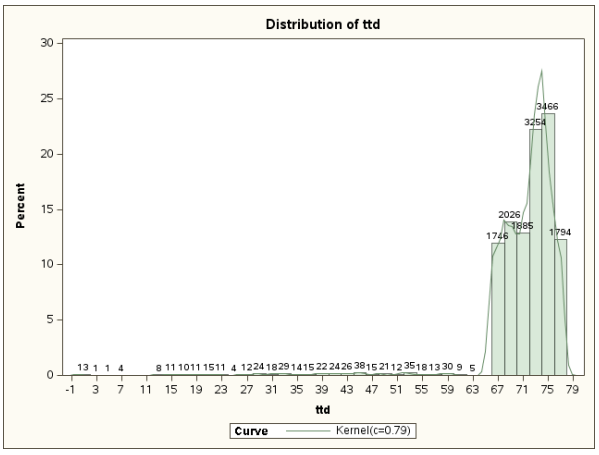


Figure 14: Distribution of the time to default for whole population

4.2.2 Survival curve

As discussed in more detail in Chapter 2 on survival analysis, a simple transformation of the cumulative distribution function produces the survival function, $S(t) = 1 - F(t)$. The survival function, $S(t)$, describes the probability of surviving past

time t, or Pr(T>t). For all the defaults in this datasets, we can see that majority of the defaults are defaulted within 60 months.

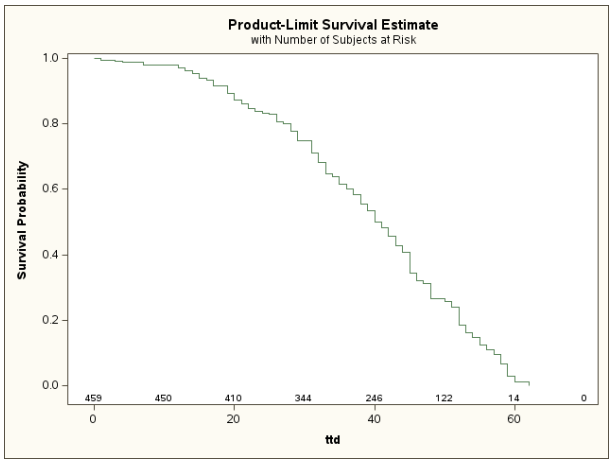


Figure 15: Survival curve for defaulted segment

From business perspective, a lot of defaults happen between 30 months to 60 months could partially due to the adjustable rate mortgage which generally have very low interest rate hence low monthly payment in the fixed interest rate term (36 months or 60 months).

Figure 16 is the survival curve for the whole sample. The curve is very flat since default is a rare event and majority of the loans are censored instead of default.

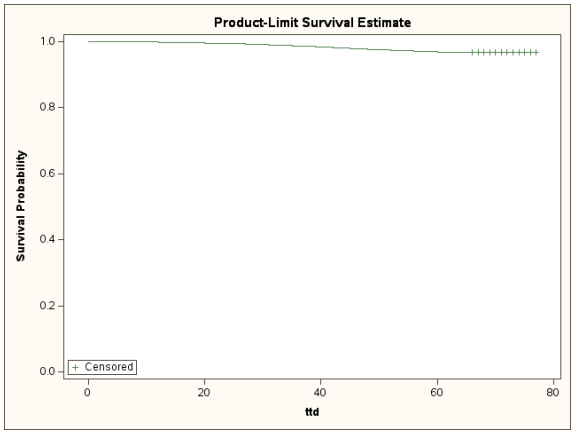
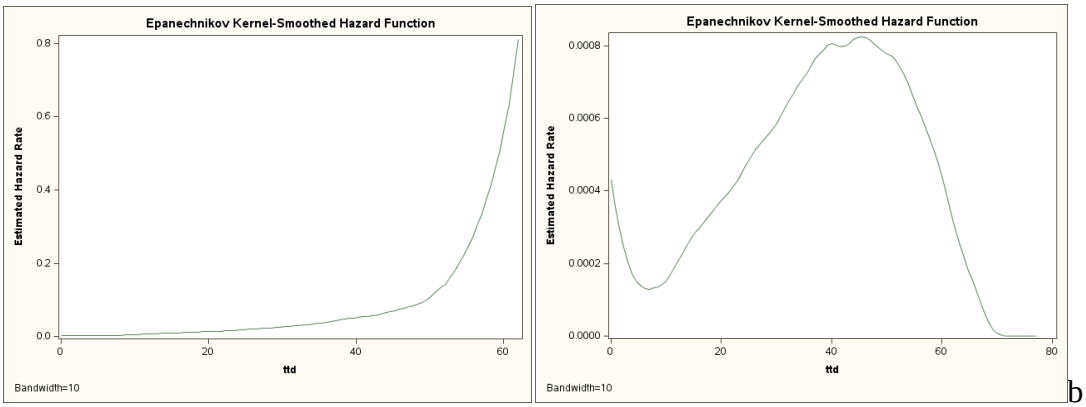


Figure 16: Survival curve for whole population

4.2.3 Hazard curve

The primary focus of survival analysis is typically to model the hazard rate (h(t)), which has the following relationship with the pdf and S(t), h(t) = f(t) / S(t). The hazard function describes the probability of the event occurring at time t (f(t)), conditional on the subject's survival up to that time t (S(t)). The hazard rate thus describes the instantaneous rate of failure at time t and ignores the accumulation of hazard up to time t. Figure 17 displays the graph of the hazard function for only the defaulted population (graph a) and the whole sample (graph b). The hazard of default increases steadily until 50 months and then increases dramatically afterwards. However, based on the hazard graph from the whole sample, the hazard of default has similar trend before 40-50 months, then the hazard drops precipitously from the 0.08% at around 50 months to 0.02% at around 60 months.



a b

Figure 17: Hazard curve for defaulted segment (left) and whole population (right)

4.2.4 Model Fit

4.2.4.1 Maximum likelihood from survival analysis

As discussed in more detail in Chapter 2, the model that is applied is as following:

$$Logit\ h(t) = \log\left(\frac{h(t)}{1-h(t)}\right) = \alpha + \beta_1 * MOB + \beta_2 * LTV + \beta_3 * \text{FICO} + \beta_4 * CS.$$

The model is estimated by *Proc PHreg* in SAS, which implements the regression method proposed by Cox (1972). PH in *Proc PHreg* stands for Proportional Hazard model. The hazard function is $h_i(t) = \lambda_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_n x_k)$. The reason the cox regression model is called proportional hazard model is because the hazard for any individual is a fixed proportion of the hazard for any other individual. If we take the ratio of the hazards for two individuals i and j, we will get:

$$\frac{h_i(t)}{h_j(t)} = \exp\{\big(\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_n(x_{in} - x_{jn}\big)\}.$$

We can see that $\lambda_0(t)$ cancels out of the numerator and denominator. Therefore, the *Proc PHreg* estimates the β coefficients of the proportional hazards model without having to specify the baseline hazard function $\lambda_0(t)$, which is partial maximum likelihood. Based on the univariate analysis, Table 12 is the survival model based on the same variables as in logistic regression.

Table 12: Survival model 1 – Model with the same variables as logistic regression model

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
MOB	1	0.01514	0.01081	1.9634	0.1611	1.015
LTV	1	1.42869	0.21425	44.4664	<.0001	4.173
FICO	1	-1.19313	0.04465	713.9675	<.0001	0.303
CS growth rate dummy	1	-0.61693	0.16525	13.937	0.0002	0.54

The hypothesis that each coefficient is 0 is tested by the following testing statistics (Table 13). The p-value is less than 0.0001 from both testing, so the null hypothesis is rejected and at least one of the coefficients is nonzero.

Table 13: Testing the null hypothesis that coefficients equal to 0 for survival model

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	788.7087	4	<.0001
Wald	903.1516	4	<.0001

Notice that in Table 12, there is no intercept estimate which is a characteristic feature of partial likelihood estimation. The last column, hazard ratio, is just exp(β). For dummy variable with value 1 and 0, the hazard ratio is the ratio of the estimated hazard for those with value 1 to the estimated hazard for those with a value of 0 by controlling for other covariates. For CS 1 year growth greater than -12% dummy variable, the hazard ratio is 0.54. This means the estimated hazard of default for accounts that with CS 1 year growth greater than -12% is only about 54% of the hazard for those with CS 1 year growth less than -12% if holding all the other variables the same.

For quantitative covariates, the estimated percent change in the hazard for each 1-unit increase in the covariate can be obtained by subtracting 1 from the hazard ratio and multiplying by 100. For the current FICO score, the hazard ratio is 0.303, which yields 100(0.303 – 1) = -69.7. As the FICO score is input as the raw FICO divided by 100, therefore, for each 100 FICO score increase, the hazard of default goes down by an estimated 69.7%.

Similarly, for 1 additional month on book, the hazard of default goes up by an estimated 1.5%. However, the p value for the month on book is greater than 0.05, which means this variable is not statistically significant. To ensure the best model to

be selected, I also tried to remove the month on book and use unemployment rate instead and get the following model. All the variables are now having p value less than 0.05.

Table 14: Survival model 2 – Model by replacing month on book with unemployment rate

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
Unemployment rate	1	0.15455	0.04136	13.9655	0.0002	1.167
LTV	1	1.11606	0.22415	24.7914	<.0001	3.053
FICO	1	-1.20129	0.04463	724.5922	<.0001	0.301
CS growth rate dummy	1	-0.39735	0.1746	5.1794	0.0229	0.672

4.2.4.2 Model Selection

As explained in Chapter 4.1, AIC, SC and log likelihood multiplied by -2 can be used to compare models with different sets of covariates. Even though these statistics cannot be used to construct a formal hypothesis test, the comparison could give us an indication with a smaller value meaning a better fit. As shown in Table 15, model 2 with the unemployment rate has slightly better fit comparing with model 1 which uses the same set of variables as the ones used in the logistic regression model.

Table 15: Model fit statistics to compare survival model 1 and model 2

Model Fit Statistics			
Criterion	**Without Covariates**	**Model 1**	**Model 2**
-2 LOG L	5135.466	4358.465	4346.758
AIC	5135.466	4366.465	4354.758
SC	5135.466	4383.067	4371.36

However, Table 16 indicates that model 1 has slightly better rank ordering power and distinguishing power comparing with model 2. The AUC for model 1 is 0.963 and

0.948 in the training and testing dataset respectively, while they are 0.961 and 0.944 for model 2.

Table 16: Model performance testing for model 1 and model 2 in both training and testing datasets

Association of Predicted Probabilities and Observed Responses				
	Model 1		Model 2	
	Training Data	Testing Data	Training Data	Testing Data
Percent Concordant	96.3	94.6	95.9	94.3
Somers' D (Gini)	0.925	0.897	0.921	0.888
c (AUC)	0.963	0.948	0.961	0.944
KS	0.8189	0.7987	0.8063	0.7963

The main purpose of the model used in the credit risk management is to predict which customer is more likely default so the bank could take actions to actively manage such accounts; therefore, model 1 is selected and Chapter 4.2 will compare this model with the logistic regression model.

4.2.4.3 Predicted Time to Default

One of the main outputs from survival model is the predicted time to event, which is time to default as in this thesis, for future accounts with specific covariates. Median survival times are often used in medical studies as a way to characterize the survival experience of a group of patients. The median survival time can be well estimated provided that the censoring is not too heavy (Ying, et al, 1995). Under heavy censoring, there may be a significant percentage of reflected intervals for which the median survival time cannot be estimated; this is because the probability that the estimated survival curve will not cross 0.5 can be substantial (Strawderman, et al 1997). In practice, the proportion of defaulted credits is very small and the proportion

of censored data will be very large in credit risk management data. This often introduces challenges for the time to default prediction. Lee, et al (2007) tried to tackle the heavy censoring issue by taking a lower quantiles prediction. However, it's prone to have relatively bigger tail errors based on the prediction from the lower quantiles, which is the limitation of using survival analysis for the time prediction based on the heavy censoring data.

Figure 18 shows the actual time to default and the predicted time to default for the defaulted loans. We can see that the predicted time to default have similar distribution but with a fat tail.

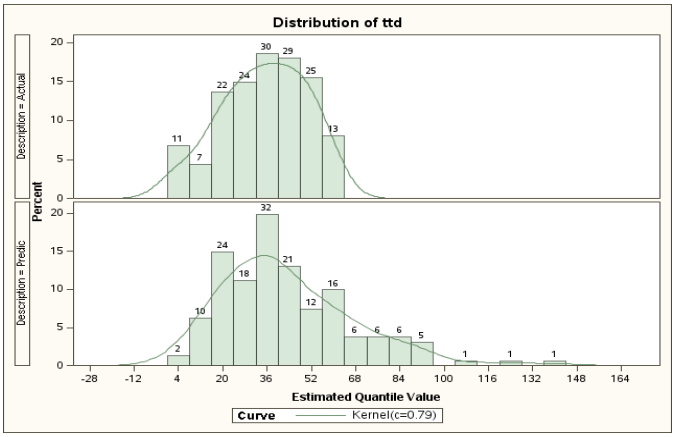


Figure 18: Distribution of the actual and predicted time to default

Another way to look at the prediction error is to calculate the delta as the predicted time to default minus the actual time to default. Figure 19 is the distribution of the delta. The distribution is asymmetrically distributed with mean equal to 0 based on the t-test (Table 17).

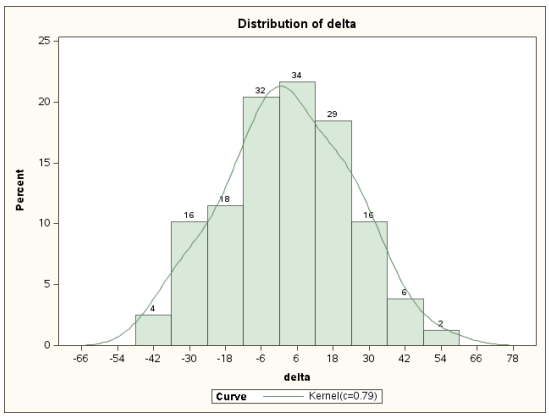


Figure 19: Distribution of difference between actual time to default and predicted time to default

Table 17: T-test on the difference equal to 0

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	1.599689	**Pr > |t|**	0.1117

4.3. Comparison and Summary

The results of the final modeling from both methods show very similar fit in terms of the ROC with the survival model having slightly better performance than logistic regression in the training dataset and almost the same performance in the testing dataset. In terms of prediction of defaulted and non-defaulted mortgage portfolios, the logistic regression model outperforms survival analysis in the training dataset, while survival model outperforms logistic regression in the testing dataset.

Table 18: Model testing summary for both logistic regression model and survival model

Association of Predicted Probabilities and Observed Responses				
	Logistic Regression		**Survival Model**	
	Training	**Testing**	**Training**	**Testing**
Percent Concordant	95.5	94.9	96.3	94.6
Somers' D (Gini)	0.91	0.898	0.925	0.897
c (AUC)	0.955	0.949	0.963	0.948
KS	0.828	0.788	0.8189	0.7987

CHAPTER 5: CONCLUSION AND FUTURE STEPS

5.1. Conclusion

I have developed a logistic regression and a survival model with time varying covariates for modeling default behavior of a mortgage portfolio.

Using a very large set of real data from a bank's mortgage portfolio, the logistic regression model has similar performance in terms of rank ordering power. For survival model, I implemented a hazard model with time varying covariates in predicting the time-to-default and then predicted the non-default and default using the same time frame. The analysis supports that survival analysis models are competitive with the industry standard logistic regression approaches.

As discussed earlier, many more complex models have been investigated in different articles; however, none of them becomes the common practice in the real world. This is partially due to the fact that the flexibility attainable using more complex models leads sometime to poor predicting performance. Moreover, the cost of implementation of such models is higher than the potential value added by them. Survival analysis is an alternative to logistic regression that is still reasonably simple. My thesis confirmed that for the sole purpose of predicting probability of default within a single specific period, survival modeling has little advantage over logistic regression model. This is consistent with the findings from Stepanova and Thomas (2003). However, survival analysis methodology offers a number of advantages that will be very useful for both credit risk management and capital management. First, it

provides a consistent method of predicting probability of default within arbitrary different periods of time. With logistic regression model, in order to get the prediction for different time window, different models have to be built with perhaps different data structures. Second, survival analysis can take into consideration the most recent data. In contrast, for logistic regression, if the probability of default within 24 months, the latest 24 months of data will not be able to be used as we will have to have at least 24 months of performance window in order for us to observe the actual defaults. Third, as Stepanova and Thomas (2001) illustrated, another use of the survival probability can be used to calculate the expected profit from a loan. The article introduced the idea of expected profit from a loan which can be calculated as the sum of the present values of the installments each multiplied by the probability of receiving it (the customer's survival probability), less the loan amount. In this case, the profit from a loan can be estimated, which then can be used in the profitability management. Last but not the least, the survival analysis provide more complete information on the predicted time to default distribution. Even with certain limitation due to the heavy censoring, the knowledge obtained from the predicted distribution of T can be useful in the broader context of profit modeling. There are also limitations in using survival methods for this type of data. The first limitation is from the application of the survival analysis in the banking industry perspective. All the models built in the banking industry need to be understood by the business users so they can better manage the business based on the model outputs. Logistic regression is built for binary data (in our application are, default or not) and it is usually estimated using maximum

likelihood. Both the coefficient of the parameters and the probability of default can be interpreted in a straightforward fashion. In contrast, the survival analysis model, especially the PH hazard model, models the hazard rate. The hazard rate is more difficult to interpret from a business point of view, and the estimation is carried using partial maximum likelihood without having to define the base hazard, and it is less common among practitioners. This could be one of the reasons why logistic regression is still the prevailing method in the industry for default analysis. Another potential limitation is from the cost-benefit perspective. One of the main usages of the default probability is for the reserve calculation. In order to calculate the reserves, when exactly the loan will default does not matter too much as long as we know what is the probability of the loan will default in the next year. The logistic regression model gives the predicted probability of default directly for the next 12 months; however, one would need additional calculations in order to get the probability of default from survival analysis modeling. This increases the implementation cost without adding too much value.

In summary, when the default modeling gets more attention in broader areas, such as profitability management, which is the directions that banking industry is heading, the additional benefits from the survival analysis modeling can be leveraged.

5.2. Future Steps

This thesis is carried out on a sample of the mortgage loans originated in 2004. It will be helpful to test the model on the mortgage loans out of this sample when data are available. As the default event is rare event, it would be of interest to investigate

whether generalized extreme value regression for binary rare event data can give improvement in prediction.

BIBLIOGRAPHY

Agresti, A. "*An introduction to categorical data analysis*", (1996), New York: Wiley

Andriotis, Annamaria. "Millions more to see their FICO scores", 2015, *http://blogs.wsj.com/totalreturn/2015/01/12/millions-more-to-see-their-fico-scores/*

Allen, LN and Rose, LC. "Financial survival analysis of defaulted debtors", *Journal of the Operational Research Society*, (2006), 57, 630-636

Altman, E.I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy" *Journal of Finance*, (1968), 23(4): 589-609

Altman, Edward I. "The Z-Metrics Methodology for Estimating Company Credit Ratings and Default Risk Probabilities", *RiskMetrics Group*, (2010)

Altman, Edward I. and Saunders, Anthony. "Credit Risk Measurement: Developments over the last 20 Years", *Journal of Banking & Finance*, (1997), 21(11-12), pp. 1721-1742

Avery, Robert B; Bostic, Raphael W.; Calem , Paul S. and Canner , Glenn B. "Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files", *Real Estate Economics*, (2000), 28(3), pp. 523-547.

Avery, Robert B.; Calem, Paul S.; and Canner , Glenn B. "An Overview of Consumer Data and Credit Reporting", *Board of Governors Federal Reserve Bulletin*, (February, 2003) pp. 47-73

Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J and Vanthienen J. "Benchmarking state-of-the-art classification algorithms for credit scoring", *Journal of the Operational Research Society*, (2003), 54: 627–635

Banasik, J.L. et. al., "Does Scoring a Subpopulation Make a Difference*?" The International Review of Retail, Distribution and Consumer Research*, Vol. 6, No. 2 (1996), pp. 180-195

Brown, Meta; Haughwout , Andrew F.; Lee, Donghoon; and Van der Klaauw, Wilbert. "The Financial Crisis at the Kitchen Table: Trends in Household Debt and Credit". *FRB of New York Staff Report* (2010), No. 480

Cohen, J.; Cohen, P.; West, S.G and Aiken, L.S. "*Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*", (2003), 3rd edition, Mahwah, NJ: Lawrence Erlbaum Associates,

Cox, D. R. "Regression models and life-tables (with discussion)" *Journal of Royal Statistics Society*, (1972), B, 34, 187-220

Crook JN, Edelman DB and Thomas LC. "Recent developments in consumer credit risk management", *European Journal of Operational Research*, (2007), 183: 1447–1465

Desai VS, Crook JN and Overstreet Jr GA. "A comparison of neural networks and linear scoring models in the credit union environment", *European Journal of Operational Research*, (1996), 95: 24-37

Fawcett, T. "ROC graphs: Notes and practical considerations for researchers", *Machine Learning*, (2004), 31

Gurný, Petr and Gurný, Martin. “Estimation of PD of Financial Institutions within Linear Discriminant Analysis”, *Mathematical Methods in Economics*, (2009)

Gurný, Petr and Gurný, Martin. “Comparison of the Credit Scoring Models on PD Estimation of US Banks”, *Mathematical Methods in Economics* (2010)

Gurný, Petr and Gurný, Martin. “Comparison of Credit Scoring Models on Probability of Default Estimation for US Banks”, *Prague Economic Papers*, (2013)

Hand DJ. “Good Practice in Retail Credit Scorecard Assessment”, *The Journal of the Operational Research Society,* Vol. 56, No. 9 (Sep., 2005), pp. 1109-1117

Hand DJ. “Classifier technology and the illusion of progress”, *Statistic Science*, (2006), 1: 1–14

Hand DJ and Henley WE. “Statistical classification methods in consumer credit scoring: A review”, *Journal of Royal Statistics Society Series A,* (1997), 160:523–541

Hinkle DE, Wiersma W, Jurs SG. “*Applied Statistics for the Behavioral Sciences*”, (2003), 5th ed. Boston: Houghton Mifflin

Hosmer, D. W. and Lemeshow S. “*Applied Logistic Regression*”, (1989), New York: John Wiley & Sons, Inc

Im. J-K, Apley, DW, Qi. C and Shan. X, “A time-dependent proportional hazards survival model for credit risk analysis”, *Journal of the Operational Research Society* (2012), 63, 306-321

Kalbfleisch, J. D. and Prentice, R. L. “*The Statistical Analysis of Failure Time Data*”, (1980), New York: Wiley

Kleiber, Christian. "The Lorenz curve in economics and econometrics", *Center of Business and Economics (WWZ), University of Basel,* (2007), Working Paper

Kutty, Gopalan. "Logistic regression and probability of default of developing countries debt", *Applied Economics*, (1990), Vol. 22 Issue 12, pp. 1649-1660

Lax, Howard, "Subprime Lending: An Investigation of Economic Efficiency," *Housing Policy Debate,* (2004), 15:3

Lee, Myoung-jae; Häkkinen, Unto and Rosenqvist, Gunnar. "Finding the Best Treatment under Heavy Censoring and Hidden Bias", *Journal of the Royal Statistical Society. Series A,* Vol. 170, No. 1 (2007), pp. 133-147

Menard, S. "Applied logistic regression analysis", *Sage University Paper Series on Quantitative Applications in the Social Sciences*, (1995), 07–106, Thousand Oaks, CA: Sage

Mester, Loretta. "What's the Point of Credit Scoring?" *Philadelphia Federal Reserve Business Review* (September/October 1997)

Narain, B. "Survival analysis and the credit granting decision", *Credit Scoring and Credit Control*, OUP, Oxford, U.K., (1992) 109-121.

OCC (Office of the Comptroller of the Currency), Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, Office of Thrift Supervision, National Credit Union Administration, "*Credit Risk Management Guidance for Home Equity Lending*", (May 2005)

Ohlson, J.A. "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research,* (1980), 18(1): 109-31

Pregibon, D. "Logistic regression diagnostics", *Annals of Statistics*, (1981), 9, 705-724

Rosenberg E and Gleit A. "Quantitative methods in credit management: A survey", *Operations Research*, (1994), 42: 589–613

Chatterjee, Satyajit; Corbae, Dean; Nakajima, Makoto and Ríos-Rull, José-Víctor. "A Quantitative Theory of Unsecured Consumer Credit with Risk of Default", *Econometrica*, Vol. 75, No. 6 (Nov., 2007), pp. 1525-1589

Stepanova, Maria and Thomas, Lyn, "Survival analysis methods for personal loan data", *Operations Research*, Vol. 50, No. 2, (March-April, 2002), pp. 277-289

Strauss, David. "The many faces of logistic regression", *The American Statistician*, Vol. 46, No. 4 (Nov., 1992), pp. 321-327

Strawderman, Robert L.; Parzen, Michael I.; and Wells, Martin T. "Accurate Confidence Limits for Quantiles under Random Censoring", *Biometrics*, Vol. 53, No. 4 (Dec., 1997), pp. 1399-1415

Tang, Y.Y. "Application of warning model of financial crisis in credit risk management", *Shanghai Finance*, (2002) 2, 12-14 (Chinese)

Thomas LC. "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers" *International Journal of Forecast*, (2000), 16: 149–172

Thomas LC, Edelman DB and Crook JN. "Credit Scoring and its Applications", *SIAM: Philadelphia, PA*, (2002)

Thomas LC, Edelman DB and Crook JN. "Readings in Credit Scoring: Recent Developments, Advances, and Aims" *Oxford University Press: Oxford, UK*, (2004)

Thomas, L. C., J. Banasik, J. N. Crook. "Not if but when loans default", *Journal of Operational Research Society*, (1999), 50 1185-1190

Thomas LC, Oliver RW and Hand DJ. "A survey of the issues in consumer credit modeling research" *Journal of the Operational Research Society*, (2005), 56:1006–1015

Westgaarda, Sjur and Nico van der Wijstb. "Default probabilities in a corporate bank portfolio: a logistic model approach", *European Journal of Operational Research*, (December 2001), olume 135, Issue 2, 1, Pages 338–349

Wiginton, J. "A Note on the Comparison of logit and discriminant model of Consumer Credit Behavior" *Journal of Financial and Quantitative Analysis*, (1980), 15, 757–770

Ying, Z., Jung, S. H., and Wei, L. J. "Survival analysis with median regression models", *Journal of the American Statistical Association*, (1995), 90, 178–184

Yobas MB, Crook JN and Ross P. "Credit scoring using neural and evolutionary techniques" *IMA Journal of Math Application Business Industry*, (2000), 11: 111-125