

4-24-2009

Modeling Subprime Mortgage Delinquency, Default, Prepayment and Loss

Olgay Cangur
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds



Part of the [Systems Science Commons](#)

Let us know how access to this document benefits you.

Recommended Citation

Cangur, Olgay, "Modeling Subprime Mortgage Delinquency, Default, Prepayment and Loss" (2009).
Dissertations and Theses. Paper 5943.
<https://doi.org/10.15760/etd.7813>

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

MODELING SUBPRIME MORTGAGE DELINQUENCY, DEFAULT,
PREPAYMENT AND LOSS

by

OLGAY CANGUR

A dissertation submitted in partial fulfillment of the
requirements for the degree of


DOCTOR OF PHILOSOPHY
in
SYSTEMS SCIENCE

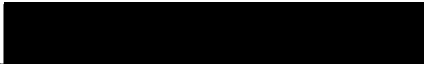
Portland State University
© 2009

DISSERTATION APPROVAL

The abstract and dissertation of Olgay Cangur for the Doctor of Philosophy in Systems Science were presented April 24, 2009, and accepted by the dissertation committee and the doctoral program.

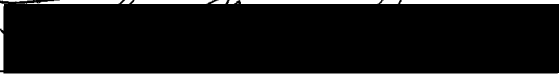
COMMITTEE APPROVALS:


Wayne W. Wakeland, Chair



Martin Zwick


Jong Sung Kim


Beverly Fuller


Timothy R. Anderson
Representative of the Office of Graduate Studies

DOCTORAL PROGRAM APPROVAL:


George G. Lendaris, Director
Systems Science Ph.D. Program

ABSTRACT

An abstract of the dissertation of Olgay Cangur for the Doctor of Philosophy in Systems Science presented April 24, 2009.

Title: Modeling Subprime Mortgage Delinquency, Default, Prepayment and Loss

The current financial environment presents significant challenges for the mortgage industry. Declining house prices have surfaced the importance of delinquency, loan default and loss predictions. Simple models of prepayment behavior are no longer applicable. Investors, originators, servicers and regulators are in need of more accurate predictions for their portfolios of interest.

This dissertation focuses on two topics relevant to modeling residential mortgages. The first topic provides a framework for modeling delinquencies, prepayments, defaults and losses that represents an enhancement over previous studies. A total of nine loan payment statuses are used (current, thirty-days delinquent, sixty-days delinquent, ninety-days delinquent, early foreclosure, late foreclosure, real estate owned, paid in full, and terminated with loss). This framework is compared to the previous framework discussed in the literature that used seven statuses.

The second topic applies reconstructability analysis (RA) to residential mortgage data in order to find new and interesting models. Many statistical methods are unable to reflect non-linearities and significant high-level interactions. RA is capable of doing both. The study explores the hypothesis that the inclusion of RA-suggested interaction terms would improve the accuracy of the logistic regression (LR) models used to forecast loan status changes within mortgage portfolios.

The first topic's result made two unique and important contributions to the mortgage management literature. First, it finds that the nine-state framework yields more accurate results than the seven-state framework. It also introduces a new state 'terminated with loss' that enables the framework to predict losses.

The second topic's results confirm the hypothesis that RA suggested interaction terms improve the performance of LR model. This is a useful contribution to the data mining literature since it enhances the performance of LR which is a widely used data mining methodology.

Dedication

I dedicate this dissertation to my wife, Laura, for her patience and loving support, to my son, Mehmet, and his future siblings yet to be named, to my parents, Cevher and Mehmet, for their excellent guidance and unconditional support, to my brother, Gokhan, for his invaluable mentorship and to my dog Muco.

I also dedicate this dissertation to Dawn for substituting my parents and giving me good advices in their absence, and to Emre for being more than a friend.

Acknowledgements

I would like to thank Dr. Wakeland, for his practical solutions to complicated issues and his effective approach to getting the dissertation done, Dr. Zwick, for constantly challenging me to strive for excellence and perfection, Dr. Kim, for teaching me the critical statistical knowledge I utilized as the backbone of this research, Dr. Fuller, for her superior empathy and understanding that helped me strengthen my confidence about this research, Dr. Anderson, for his inspirational vision and dedication to academic excellence, and Dr. Lendaris, for motivating me throughout the process.

I would like to acknowledge the importance of Portland State University resources, especially the library and the OCCAM (web based data mining program).

I would also like to thank Wilshire Credit Corporation for providing me the most comprehensive subprime mortgage data (Loan Performance Data), statistical tool (SAS version 9.1.3), database management tool (SQL Server 2005) and business knowledge.

Table of Contents

Acknowledgements	ii
Table of Contents	iii
List of Tables.....	iv
List of Figures.....	v
List of Definitions.....	vi
Chapter 1 - Introduction	1
1.1 Background and Significance	1
1.2 Specific Aims	5
1.3 Overview of Methods.....	5
Chapter 2 - Literature Review	8
2.1 Introduction	8
2.2 Prepayment Literature	11
Prepayment Literature Synthesis	19
2.3 Loan Default Literature	21
Loan Default Literature Synthesis.....	31
2.4 Loss Literature.....	33
Loss Literature Synthesis.....	39
2.5 Reconstructability Analysis, Monte Carlo Method, and Logistic Regression ...	41
Reconstructability Analysis.....	41
Monte Carlo Method	45
Logistic Regression	47
Chapter 3 – Methodology.....	49
3.1 Payment Model.....	49
3.2 Interaction Effects	58
3.3 Data	63
Chapter 4 – Results.....	65
4.1 Payment Model Results.....	65
Results of Empirical Testing of the 9-State Payment Model	65
Results of Formal Statistical Testing of Adding More States	75
4.2 Interaction Effects and Reconstructability Analysis Results	77
Chapter 5 –Discussion and Conclusions	82
5.1 9-State Payment Model Discussion.....	82
5.2 Interaction Effects and Reconstructability Analysis Discussion.....	88
5.3 Conclusions	93
References	94
Appendix A – Variables	100

List of Tables

Table 1 – Summary and comparison of existing and proposed payment models	50
Table 2 – Proposed 9-State payment model transition matrix	51
Table 3 – Transition matrix illustrating the mechanics of Monte Carlo method	53
Table 4 – Model framework comparison	56
Table 5 – 7-State model restriction tests construct compared to the 3-state model	57
Table 6 – 9-State model restriction tests construct compared to the 7-State model.....	57
Table 7 – Sample OCCAM output for RA.....	60
Table 8 – Origination Table Fields.....	63
Table 9 – Dynamic Historic Table Fields.....	64
Table 10 – Model comparison data	66
Table 11 – Testing the validity of splitting foreclosure into ‘Early and Late Foreclosure’ statuses	75
Table 12 – Testing the validity of splitting ‘Paid off’ status into ‘Terminated with Loss’ and ‘Paid in Full’ statuses.....	76
Table 13 – LR results	78
Table 14 – RA interaction tests	79
Table 15 – LR with interaction terms included	80
Table 16 – Model Comparison Framework.....	83

List of Figures

Figure 1 - Testing scheme for payment model	55
Figure 2 – 7-State and 9-State model predictions compared to the actual data for ‘Current’ loans	68
Figure 3 – 7-State and 9-State model predictions compared to the actual data for thirty day delinquent loans	69
Figure 4 - 7-State and 9-State model predictions compared to the actual data for sixty day delinquent loans	70
Figure 5 - 7-State and 9-State model predictions compared to the actual data for ninety plus day delinquent loans	71
Figure 6 - 7-State and 9-State model predictions compared to the actual data for foreclosure loans.....	72
Figure 7 - 7-State and 9-State model predictions compared to the actual data for REO loans.....	73
Figure 8 - 7-State and 9-State model predictions compared to the actual data for ‘Paid Off’ loans.....	74

List of Definitions

CLTV: Combined LTV. It is the sum of all existing loan amounts originated on the asset over the value of the asset.

FICO: Credit score of borrowers

Loan curing: Successful loan workout efforts that bring a delinquent loan back to paying status.

Loan workout: Servicer efforts to help borrowers make payments on the mortgage loan, such as payment plans, stipulated agreements, forbearance plans, modifications, reinstatement and so forth.

Loss mitigation: Servicer efforts to reduce the loss amount on a defaulted mortgage loan such as avoiding foreclosure sale, property preservation, deed-in-lieu, short sale and so forth.

Loss severity: The loss amount over the defaulted loan amount owed by the borrower.

LTV: Loan to value ratio is the ratio of loan amount to the value of the house.

Paid in Full: A loan terminated by a borrower paying the loan in full.

Paid off: A terminated loan either by borrower paying off the loan in full or with loss.

Realized loss: The final loss amount on a loan.

REO: Real estate owned by the servicer after the foreclosure sale.

Terminated with Loss: A loan terminated with loss. This status is usually after a liquidation process conducted by the servicer on the defaulted loan.

Chapter 1 - Introduction

1.1 Background and Significance

In 2006, residential home price appreciation (HPA) peaked with a trend that started in the mid 1990s. In early 2006, the average annual home price increased in the Pacific and the South Atlantic regions were 18% and 17%, respectively; while the Mid-Atlantic and the New England regions were 13% and 10%, respectively. Shortly thereafter, this steep climb turned into a catastrophic decline. Homebuyers' enthusiasm started to diminish with the declining HPA. Increased delinquencies and foreclosures led to massive losses on securitized deals that held millions of mortgages. Investors lost trillions of dollars on financial instruments that were once seen as reliable as US government Treasury bonds. By the end of 2008, US National Bureau of Economic Research declared that the US economy was in recession triggered by the mortgage crisis.

Most of the mortgage originators¹ were out of business by mid 2008, and important financial companies that provide the funding for mortgage businesses like Lehman Brothers, Bear Stearns, Merrill Lynch, Wachovia, Washington Mutual and many others were either bankrupt or sold by the third or fourth quarter of 2008.

¹ Originators lend money to the borrowers to purchase a new mortgage or refinance their existing mortgage. They collect similarly originated deals and create pools that are securitized into a financial security that is then traded by investors.

Consequently, servicers² are facing capacity issues due to higher delinquencies and foreclosures that are more labor intensive compared to loans that pay regularly. Delinquencies, foreclosure and loss projections are seen as very important to help optimize resource allocation. Models used for projection need to be revisited due to the dramatic shifts in borrowers' behavior following the shift in the economic conditions.

Published research in mortgage modeling falls into two broad categories: empirical studies and theoretical development. Empirical models use past behavior to draw inferences about key inter-relationships in order to create forecasting models. Theoretical models try to formulate the behavior of borrowers independent of specific regional, economic and sociologic conditions. These types of models provide general understanding that is applicable to a broad range of situations and conditions that borrowers might face.

However, there are drawbacks associated with both methods. Theoretical models try to formulate the borrowers' behavior mainly based on option theory. This theory assumes that borrowers try to maximize their wealth by hedging or mitigating their losses.

Theoretical models display the relationship between borrower behavior and economic conditions. However, their accuracy is questionable. Available theoretical frameworks

² Servicers collect the monthly payments from the borrower and remit them to the investor. They also take care of liquidating the house or any other service related issue regarding the mortgage loan.

perform poorly against real world (actual) data (Kau, Keenan and Kim (1993); Lekkas, Quigley and Van Order (1993); Capozza, Kazarian and Thomson (1998)).

On the other hand, empirical models use past behavior to predict future behavior assuming that economic conditions remain the same. This assumption works well when a strong housing trend is present, but not when markets are in a state of flux. Consequently, empirical models lack the flexibility to adjust to rapidly changing economic conditions beyond their boundaries. These models must be frequently revised and updated with the changing environment. Most of these empirical works are not more than complex trend analyses.

Theoretical models in mortgage literature are mostly based on option theory (Black and Scholes 1973). This work formulated the optimal strategy for valuing stock options. Subsequent researchers build mortgage valuation models using the option theory. One of the key assumptions of these models is that the borrower always uses an optimal strategy in order to maximize his/her wealth. However, this is not the case in real life. Borrowers face “frictions” that are difficulties in exercising optimal strategies for several reasons such as loss of job, no cash availability for an optimal refinancing opportunity, sentimental ownership towards the real estate and free-rent opportunity when the optimal strategy is to default and leave the house. Researchers strive to identify and implement these frictions in models in order to enhance the option theoretic framework and make it more realistic. Another important aspect of

these theoretical models is that the stochastic interest rates and home price forecasts are incorporated to generate an expected value for a specific mortgage.

As mentioned earlier, theoretical models lack accuracy due to the disconnect between actual borrower behavior and the assumptions made by the models. The use of such models in the mortgage industry is limited. This gap has been filled by empirical models, and has been proven to be reasonably accurate for making short-term predictions in a stable environment. The drawback of empirical models is that the assumptions need to change over time due to different economic, sociologic, and regional environmental conditions. These models are built for short-term prediction, for time periods much shorter than the average life of a mortgage loan.

The present research addresses these drawbacks by developing a comprehensive computer simulation model that forecasts the future payment statuses of a mortgage loan. Implied cash flows based on predicted future payment statuses enable researchers to price complex subprime securitizations³. The model uses a combination of theoretical and empirical frameworks to predict future delinquencies, future defaults and future prepayments. It utilizes logistic regression, Monte Carlo simulation, and a new system research methodology called reconstructability analysis. The following sections describe the research objectives and provide an overview of the methods.

³ Securitized deals are pooled mortgage loans, to create mortgage securities to be sold similar to bonds (Fabozzi 1992)

1.2 Specific Aims

This research carries out two inter-related studies in order to better forecast important aspects of subprime mortgages:

1) Prediction of delinquency transition and the frequency of prepayment⁴ and default:

This research develops and tests an enhanced model with additional loan statuses ('Early and Late Foreclosure', 'Terminated with Loss' and 'Paid in Full') compared to previously published models. The performance of the enhanced model is then compared against other models in the literature.

2) Identifying interaction effects between the key variables of the payment model:

This is an area that has been under-researched in the literature. A new methodology is used to identify significant interactions which then are introduced into predictive models and tested. Reconstructability analysis is used to detect and quantify statistically significant interactions that may be economically relevant.

1.3 Overview of Methods

1) 9-State Payment model: Delinquency prediction is an important aspect of mortgage analytics, because it serves as an early indicator of potential losses. Delinquency behavior is usually modeled by a Markov transition matrix, where each cell contains

⁴ Prepayment is the payment of the loan earlier than its maturity date. This action results in less total future cash flows for that loan.

the probability for each transition from different beginning and ending loan states for a given period. This period is usually one month while the numbers of states differ in the literature. The early literature indicates two states for a loan: 'Active' and 'Paid off'. Later, a three-state model added the "defaulted" state. The most recently reported model (De Franco 2002) has seven states: 'Current', '30-days delinquent', '60-days delinquent', '90+ days delinquent', 'Foreclosure', 'REO⁵' and 'Paid Off'. This present study includes an 'Early Foreclosure' status and a 'Late Foreclosure' instead of a single 'Foreclosure' status and also adds a 'Terminated with Loss' status. The 9-State model improves the understanding of the prepayment, delinquency, default and loss process. The accuracy of this new framework is compared to the 7-State model discussed in the literature. The measure of accuracy is the absolute error between the actual and predicted values. A hypothesis test is conducted on predictions from both models' results using the Mann-Whitney method in order to test whether the 9-State model performs significantly better than the 7-State model.

2) Interaction effects: The previous mortgage literature focuses on the use of statistical models to predict the behavior of the mortgage loans. These models use independent variables determined by the researcher to explain the behavior of the dependent variable. One downside of this type of approach is that these statistical models often ignore the possible interaction effects among the independent variables. The reconstructability analysis (RA) method in the information theoretic literature is a

⁵ REO: Real Estate Owned – Loans owned by the bank or the servicer that is in charge of liquidating the house.

robust technique that identifies such interactions and analyzes them to determine their significance. The web-based RA tool OCCAM⁶ is utilized for both the exploratory and the confirmatory analyses. Its results are tested to enhance the logistic regression models that predict the transitions. Incorporating interaction effects will not only improve the logistic regressions in the 9-state model presented in this research, but also the logistic regression analyses in general. A Chi-square test, at the model level, is utilized to confirm the hypothesis that the interaction terms explored using RA significantly improve the logistic regression.

Chapter 2 reviews the relevant literature including sections on mortgage loan prepayment, loan default, and loan loss. The pertinent modeling, simulation, Monte Carlo method, logistic regression and RA literature is also reviewed. Chapter 3 describes the research methodology, including the models and how their performance is tested and compared with existing models that are used in practice and discussed in the literature. Chapter 4 provides the results of both payment model and interaction effects research, and Chapter 5 discusses the conclusions drawn from the study, and presents the discussion including recommendations and identifies areas of future research.

⁶ OCCAM is a web-based reconstructability analysis program that can be accessed from <http://dmm.sysc.pdx.edu/occam/weboccam.cgi>

Chapter 2 - Literature Review

2.1 Introduction

A mortgage loan is an instrument for which cash flow is unknown because the borrower has an alternative to prepay or to default at any time. These uncertainties are called the prepayment risk⁷ and the default risk (Fabozzi, 1992). Researchers have applied both option theory and econometric techniques to estimate this uncertainty and forecast the performance to arrive at a value for a mortgage loan. Such forecasts are critical for investment decision-making.

Early literature started with determining prepayment risk for conventional mortgage loans, known as prime loans. These loans are assumed to have no default risk due to the underwriting guidelines by government agencies such as GNMA, FNMA and FHLMC. Thus, the literature concentrates on the borrower's behavior of selling the house, refinancing the mortgage, and the rare case of paying off the loan.

Following the introduction of high yield non-agency⁸ loans in the mortgage markets (subprime and Alt-A with higher credit risk), the default option of the mortgage borrower became a significant research interest. Researchers have investigated the

⁷ If loans in a deal pay in full before their maturity, the future cash flow of the deal degrades assuming everything else is held constant.

⁸ Loans that are not guaranteed by government agencies GNMA, FNMA and FHLMC. Subprime and Alternative-A loans fall into this category. These loans either have borrowers with impaired credit or somehow do not qualify to become prime loans that are guaranteed by the agency.

delinquencies and the probability of default related to these delinquencies. In addition, researchers have concentrated on the losses incurred on the mortgage loan.

Today, there are comprehensive models that forecast prepayment, delinquency, default and loss. These models operate by using various statistical modeling techniques that determine the risk of the mortgage investment. The forecast is run in conjunction with computer models that utilize simulation techniques such as Monte Carlo. Results are used for either valuation or by servicers to proactively manage their resources, such as human resources and call campaigns; also to help with pricing of mortgage servicing rights (PMSR).

Recently used methods include survival analysis, multinomial logistic regression, multiple regression, and Monte-Carlo simulation. Survival analysis is commonly used in prepayment literature. It is also utilized by the foreclosure and REO timeline researchers. Multinomial logistic regression is frequently used in the delinquency, default and loss literature due to its ability to generate probabilities for multiple outcomes. Monte Carlo simulation is used in pricing mortgage loans, especially when the models utilize option theory, sensitivity analyses and various future scenarios.

The literature review comprises four main areas:

1. Prepayment
2. Loan Default and Delinquency
3. Loss
4. Reconstructability Analysis, Monte Carlo Method and Logistic Regression

Section 2.2 presents the prepayments and expands on categories of prepayments; while Section 2.3 presents the literature for loan defaults and its relation with delinquencies and losses; Section 2.4 reviews research for loss and loss severities; while Section 2.5 discusses the literature of reconstructability analysis, Monte Carlo method and logistic regression.

2.2 Prepayment Literature

Prepayments are classified into three different categories by Hayre 2001:

- Home Sales
- Refinancing
- Curtailments and Full Payoffs

These categories play a critical role in the valuation of mortgage-backed securities. The various model projections are used by investors to manage their risks and plan their investment strategies.

One reason prepayment occurs is because of **housing turnover**. For instance, the borrower may decide to sell the property, which will terminate the contract resulting in a change in future cash flows. This behavior mostly depends on the following factors (Hayre 2001):

- Overall turnover rate: Percentage of all existing homes likely to be sold in a given period
- Relative mobility: Variability in the likelihood of moving due to different borrower demographics
- Seasoning: Variability in the likelihood of borrower moving after the time the loan was originated.

- **Lock-in Effect:** The effect of having a mortgage rate that is below the current mortgage rates. This reduces the likelihood of home sales.

Refinancing is another reason for prepayments. Borrowers choose to maximize their wealth by renewing their existing mortgage contract in various ways. Theoretically, this behavior should happen when the refinancing cost, together with the new mortgage value, is less than the existing mortgage value. This occurs with some delay on the borrowers' side; borrowers tend not to refinance optimally due to timing reasons in their refinancing decisions (Archer and Ling, 1993).

Refinancing can be looked at from the option theory point of view. It is the exercise of the call option that is implicit in the contract. The literature has many models that utilize the option theoretic approach. Most of these models endogenously generate interest rates and housing prices using Monte Carlo simulation techniques to value the refinancing option of the borrower. Since these models work under the assumption that the borrower will exercise the call option optimally, they tend to generate skewed results in adverse economic conditions (Kau, Keenan and Kim, 1993).

The key components regarding refinancing are (Hayre 2001):

- **Refinancing Incentive:** Borrower compares his/her mortgage rate to the available incentives in the market.

- **Burnout:** A decline in the refinancing rates even if no change occurs in the refinancing incentive.
- **Diversity in the borrower types:** Each borrower has unique characteristics that account for the refinancing.
- **Changes in the refinancing environment:** Regulatory, technological, market or borrower changes may affect the refinancing behavior.

Curtailment (partial payoff) and **full payoff** are additional options borrowers may choose when paying their mortgage debt. Even though they are less likely to be exercised, the rates might be significant when the loans are seasoned and the remaining balance is low. Another reason involves the demographics of the borrowers. For example, borrowers who are retired or close to retirement might be more likely to exercise full or partial payoff.

Hayre (2001) argues that since borrowers make use of mortgage interest payments as tax deductions, they are less likely to prepay fully without selling their homes or without refinancing. On the other hand, Chinloy (1993) indicates that the net present worth of future monthly mortgage payments is reduced when the borrower makes partial prepayments.

Key literature in prepayments follows:

Dunn and McConnell (1981) examine the impact of amortization, call, and prepayment features on the prices, risks and expected returns of GNMA securities⁹.

The amortization and prepayment features each have a positive effect on price, while the call feature has a negative impact. These features reduce the interest rate of GNMA securities, consequently decreasing their expected return.

Hall (1985) introduces the option theory to the prepayment literature. He suggests the idea that a mortgage can be seen as having options (payoff, default); and option theory can be useful to explain the behavior of mortgages. He also addresses the complications that arise while applying option theory to mortgage concepts. For example, non-optimal prepayment is a major issue and one of the key limitations of his model.

Schwartz and Torous (1989) implement a new valuation framework building on Dunn and McConnell's work (1981). Schwartz and Torous use a conditional probability of prepayment, rather than imposing an optimal value minimizing call condition to prepay. This probability depends on the state of the economy. To implement this idea,

⁹ GNMA - Government National Mortgage Association, a government-owned agency which buys mortgages from lending institutions, securitizes them, and then sells them to investors. Because the payments to investors are guaranteed by the full faith and credit of the U.S. Government, they return slightly less interest than other mortgage-backed securities.

they use the maximum-likelihood techniques (logit) to estimate a prepayment function in light of recent aggregate GNMA pool prepayment experiences.

Cunningham and Capone (1990) utilize multinomial logistic regression to analyze termination in adjustable- and fixed-rate mortgages. They conclude that interest-rate expectations affect fixed-rate mortgage terminations more strongly than adjustable-rate mortgage terminations. They also point out that loan-to-value ratio and debt-to-income ratio are the cornerstone determinants. They group variables into four categories: macroeconomic, mortgage related, property related, and borrower related.

Kang and Zenios (1992) discuss the development of prepayment models for pools of fixed-rate mortgages. They identify the key factors determining the prepayment rates as refinancing incentive, seasonal variations, seasoning of the mortgage pool, and the burnout effect. They build a model for each factor and calibrate their models using historical data. The multiplicative effects of each model determine the overall prepayment rate of the mortgage pool.

Schwartz and Torous (1992) investigate the interaction of prepayment and default decisions in the valuation process. Default decisions affect the timing of the cash flows in a mortgage pool, consequently affecting the value of the pool. They investigate the equilibrium valuation of the default insurance. They assume that the equilibrium insurance fee varies with the interest rate, and interest-rate volatility and

the value of the underlying collateral. The results indicated that default insurance is not properly priced.

Archer and Ling (1993) claim that residential mortgage borrowers behave sub-optimally with respect to their mortgage prepayment options. They develop a rational model of mortgage prepayment that incorporates non-optimal prepayment behavior. Their model handles the effects of interest-rate-motivated refinancing as well as non-interest-rate driven prepayment. Their paper also discusses the burnout effect within the mortgage pool.

Schwartz and Torous (1993) use a large data sample to assess the prepayment behavior of individual homeowners. Poisson regression is employed to estimate the parameters of a proportional hazards model for prepayment decision. Poisson regression handles time dependent covariates, multiple time scales, and non-proportionality better than the previous methods employed. They conclude that prepayments are affected by refinancing opportunities as well as regional differences.

McConnell and Singh (1994) introduce a dynamic programming model in which an individual mortgagor's decision to prepay is the feedback control variable. The mortgagor seeks to minimize the value of the mortgage subject to refinancing costs. The researchers use Monte Carlo method and conduct extensive sensitivity analyses to determine the robustness of this dynamic programming approach.

Stanton (1995) presents a new model of mortgage prepayments based on rational decisions by mortgage holders. The results indicate that mortgage holders act as though they face transaction costs that far exceed the explicit costs usually incurred on refinancing. These holders also wait an average of one year before refinancing even when it is optimal to do so earlier. In general, this article extends the option theoretic approach in order to better explain borrower frictions in prepayment.

Hakim (1997) estimates prepayment probabilities as a function of borrower characteristics, the loan characteristics, regional, and economical variables. He distinguishes induced prepayments from autonomous prepayments. Non-interest effects reveal the significance of the borrower's characteristics, property age and regional mobility rates on mortgage termination.

Yang, Buist and Megbolugbe (1996) introduce household income as the third stochastic variable in addition to the interest rate and house price appreciation (HPA). The presence of these variables allows consideration of consumption-theoretic determinants of mortgage termination. Also, the mortgage underwriting rules that restrict optimal prepayment is also explicitly modeled (such as prepayment penalties and due-on-sale clauses and so forth).

Lacour-Little and Chun (1999) investigate the effect of third party originators (TPOs). TPOs are mortgage brokers who have strong economic incentives to encourage

borrowers to refinance. The authors report the effect of TPOs on securities in terms of prepayments. Prepayments increase as the number of TPOs increase.

Deng, Quigley and Van Order (2000) present a unified model of competing risks of mortgage termination by prepayment and default. The model estimates these two probabilities jointly assuming they are dependent competing risks. They report that the borrowers with high loan-to-value (LTV) ratios loans are more likely to exercise their option of either prepaying or defaulting. The initial LTV ratio may reflect investor preferences for risk in the market for mortgages on owner-occupied housing. They conclude that a simple option model is not sufficient to explain the variability in actual prepayments and defaults.

Ambrose and Lacour-Little (2001) employ the risk methodology developed by Deng, Quigley and Van Order (2000). They conclude that loan age has a negative effect on prepayment risk for ARMs. This is consistent with the phenomenon that borrowers with high mobility and high propensity to refinance exit the pool early. They also note that loans with higher margins, higher spread compared to current fixed rates and loans originated by TPOs are more likely to prepay. Prepayments concentrate around the first and the second reset adjustments, subsequent adjustments did not produce significant effects.

Lacour-Little, Marschoun and Maxam (2002) emphasize the non-linear nature of the prepayment function. They also indicate use of non-parametric techniques in non-

linear and multivariate interaction conditions. The authors employed a kernel regression technique on loan level data to produce a non-parametric model of prepayment behavior. Kernel regression results indicate an $R^2 = 38.89\%$, an increase of nearly 5% in explaining the variability over the linear regression model.

Prepayment Literature Synthesis

Researchers focus on two important reasons for prepayments. The first reason is home sales, which depend on the housing turnover rate, mobility of borrowers, loan seasoning and lock-in effect¹⁰. The second reason is refinancing, which depends on the refinancing incentive, burnout effect, diversity of the borrower and the changing environment in the refinancing market. Important topics in the literature are listed below:

- Option theoretic framework is useful in formulating the behavior of the borrower. However, the borrower's non-optimal prepayment decision is not easy to incorporate. Researchers conclude that pure option theoretic models are not sufficient to explain the variability in prepayments.
- Seasonal variations, refinancing incentive, loan age, and the burnout effect are key components of empirical models.

¹⁰ The effect of having a mortgage rate that is below the current mortgage rates.

- The interaction between default and prepayment behavior is significant and is incorporated into the theoretic framework as noted in the literature as competing risks.
- Borrowers behave sub-optimally with respect to their mortgage prepayment.
- Interest rate, LTV, DTI¹¹, refinancing opportunities, property age, house prices and regional mobility rates affect the prepayment.
- Household income, interest rates and house prices as stochastic variables significantly explain prepayment behavior.

¹¹ Debt to income ratio of the borrower at the origination of the loan.

2.3 Loan Default Literature

Delinquency starts when the borrower misses a payment on his contract. Multiple missed payments eventually lead to the foreclosure process. The borrower can start to make the payments to eliminate the delinquency and the foreclosure, or the borrower may choose to not pay further on the loan resulting in default. It is usually defined as loans with four or more delinquent payments. Following the default, the house is liquidated and the loan amount is paid to the investor if the amount is higher than the net liquidation proceeds of the house.

One of the earliest studies on the default topic is by Morton (1975). He uses discriminant function analysis to determine the impact of the independent variables on current, delinquent and foreclosed mortgages. His findings indicate that borrowers with higher number of dependents are more likely to default. He also shows that three-family property, LTV, junior financing and existence of non-real estate debt are effective in predicting the likelihood of defaulting. The DTI is not significantly related to loan default or delinquency.

Vandell and Thibodeau (1985) formulate a theoretical model to investigate the behavior of default. Their theoretical model is based on optimizing borrower wealth. Their objective is to better understand the reasons for default beyond the equity related decisions of the borrower. Their hypotheses find the following effects significantly impact the default decision:

- Payment levels relative to income
- Current and expected housing market conditions
- Economic conditions
- Wealth
- Borrower characteristics proxying for variability in income or “crisis” events
- Transaction costs incurred upon default

They conclude that some of these variables dominate the equity effect on default and help explain the non-optimal default decision. Another important conclusion from this study is that non-equity effects could not be ignored.

Epperson, Kau, Keenan and Muller (1985) introduce option theory into the mortgage default literature, specifically the model by Black and Scholes (1973). Their simulation results show the sensitivity to the volatility of house prices and interest rates. They suggest that their research is an initial step towards modeling mortgage-backed securities from the option-theoretic perspective.

Cooperstein, Redburn and Meyers (1991) confirm the powerful influence of equity on mortgage defaults and the strong influence of interest rates on both defaults and prepayments. The interest rate effect explains the default behavior in periods of substantial economic fluctuations.

The irrational, non-wealth maximizing behavior of borrowers hinders the explanatory power of option-based pricing models. Kau, Keenan and Kim (1993) include transaction costs and sub-optimal termination behavior to overcome this problem. This is a step towards explaining the gap between empirical and theoretical approaches. Kau *et al.* show that transaction costs have a stronger influence on default than the influence of sub-optimal termination on default. However, both concepts are not as powerful as expected in explaining the irrational borrower behavior. They recommend future work to determine the difference between the actual default behavior and the predictions of option-based pricing of mortgages.

Hendershott and Schultz (1993) investigate the effect of negative equity. They binned LTVs into six categories and loan sizes into seven categories. Higher LTV means higher default rates and larger loan sizes are reported to have less defaults. They conclude that this is due to higher priced houses tending to more appreciate in prices. They also note that unemployment rate and the book value of borrower equity also are significant determinants of default as well as their interaction.

Jones (1993) models the role of moving, moral aversion, and deficiency costs in determining whether exercising the default option in home mortgage loans is rational. These factors can be considered as additional “frictions” regarding the borrower’s default decision.

Kau and Kim (1994) conclude that a rational individual might not exercise the put option (default) in the mortgage contract as soon as the anticipated cost of payments exceeds the house price. They indicate that there is considerable benefit if the house price increases in the near future. Defaulting at a later date might mitigate the consequences to the borrower. The cost to this strategy is the required monthly mortgage payment. Their important conclusion is that the observed delay in default, which is usually attributed to transaction costs, instead can be explained as an entirely rational choice in a dynamic environment.

Vandell (1995) discusses the usefulness of option-theoretic models in understanding the default behavior, but also points out their lack of accuracy. Consequently, he recommended the use of techniques such as Poisson regression used by Schwartz and Torous (1993) to enhance their model's explanatory power.

VanderHoff (1996) compares the adjustable-rate mortgage (ARM) defaults to the fixed-rate mortgage (FRM) defaults. His findings indicate that the ARM loans default more often than FRMs. The anticipated increase in interest rate has a larger impact on the ARM holders default decision compared to the anticipated payment increase. He concludes that ARM holders are less mobile than the FRM holders. His study supports the notion that defaults are not just due to negative equity in the house. He indicates that most of the observed defaults stopped paying when they actually had positive equity.

Deng, Quigley and Order (1996) emphasize the importance of current loan-to-value ratio in default modeling. They analyze the effect of down payment on default. They test the effect of unemployment, average annual house price change, household income and down payment rate. Their findings indicate a high sensitivity to LTV ratio.

Ambrose, Buttimer and Capone (1997) model the time between default and foreclosure, called the “free rent”. Their goal is to determine the value associated with such delay. They report that the probability of default increases as borrower expectations of delay between default and foreclosure increase. The probability of default decreases as borrowers expectations on the probability of the deficiency judgments increased. Lenders can alter borrower behavior by raising the transaction costs associated with the default. This would result in the reduction of the time between default and foreclosure. They conclude that the FHA/VA mortgage insurance could be lowered by actively seeking deficiency judgments similar to the conventional loans.

The role of age, LTV, rent-to-price ratio, trigger events and transaction costs are investigated using option-based modeling approach in Capozza, Kazarian, and Thomson (1997). They find a significant effect of trigger events and transaction costs on default. LTV, as usual, are the strongest reason for defaults. Their most notable finding is the negative correlation between rent-to-price ratio and defaults. They also indicate the significance of unemployment and divorce rates on defaults.

Unlike the previous default studies focus on foreclosures in one-step decision framework, Ambrose and Capone (1996) find that a foreclosure is one possible outcome of a default scenario. They note that the foreclosure is a separate event, conditioned upon both an initial default decision and subsequent changes in the economic environment. They state that mortgage servicers should understand the dynamic effect of key variables such as interest rates and house price appreciations during a default. This may lead to a significant understanding of the borrower behavior during the default period. Finally, they conclude that a servicer should identify which borrower is truly affected by a trigger event and offer them loss mitigating and foreclosure-avoiding options.

Capozza, Kazarian and Thomson (1998) introduce a new term called conditional probability of default. This is different from unconditional probability of default because it uses the most current data on the loan rather than just the origination data. The gap between empirical studies and the option-pricing methods arise from this notion. One important result is that variables important unconditionally, such as rental rate, interest rate reversion and interest rate volatility, are secondary in importance conditionally. They also conclude that trigger events, interest rate volatility and transactions costs have little effect and can be removed from the empirical models in order to reduce the risk of misspecification bias. Another distinct finding is that interest rate increase from origination rate reduces the probability of default since the option is “in the money” (meaning it is reasonable to continue making payments).

Their study also reinforces the idea that CLTV¹² is the key variable for predicting default.

Investigation on low-income neighborhoods by Van Order and Zorn (2000) reveal that both borrower income and neighborhood income is related to default. They also indicate that neighborhood income has stronger relationship with default than borrower income. High-income borrowers tend to default more; however their loss severities are lower. Typically, the relationship of LTV with default and severity is strongly inversely related.

Loss mitigation is the process by which lenders attempt to minimize losses associated with foreclosure. Lenders and servicers lean towards adopting loss mitigation tactics rather than simply foreclosing on a defaulted loan. Ambrose and Buttimer (2000) formulate a mortgage-pricing model that fully specifies all possible borrower options embedded in the mortgage contract, such as reinstatement, forbearance¹³, and anti-deficiency judgment. They also determine the value of credit on borrower default behavior.

Giving the option of forbearance to a borrower increases the delinquency, but also increases reinstatements out of foreclosure, referred to as “cures”. Researchers show that creating an economic incentive such as waiving the default penalty can create an

¹² Current loan-to-value ratio. Since CLTV has the most current data compared to LTV which is an origination value.

¹³ A type of payment plan where the borrower accepts a stipulated agreement.

optimal cure condition for the borrowers in a stable economic environment. Ambrose and Buttimer (2000) also indicate the importance of the value of credit from borrowers' perspective, and recommend the use of future credit degradation to reinforce the impression that default is costly.

Ambrose, Capone, and Deng (2001) note the interaction effect of house-price-cycle-stage with the probability of negative equity. They employ a simulation using Monte Carlo method. As the housing prices enter into a significant downturn, the probability of negative equity and default relationship breaks down. This leads to more defaults even though the optimal default scenario is not reached.

A dynamic modeling approach by Calhoun and Deng (2002) uses multinomial logit to specify quarterly the conditional variables in their modeling environment. They analyze the different termination behavior of fixed-rate and adjustable-rate mortgages. The estimated impact of option theoretic variables on conditional probability of default is the same across both FRM and ARM borrowers. Any difference in behaviors of both types of borrowers can be explained by other fixed effects. These can be alternative motivations of each type for borrower. Age of the mortgage, year of origination, original LTV, and relative loan size also are indicated empirically significant in explaining default behavior.

Alexander *et al.* (2002) investigate the effect of TPO loans on default probabilities. They compare TPOs to retail originated loans. They observe that there is a significant difference between the two types of origination. This is due to TPOs being compensated for the origination but not held accountable for subsequent performance of the loan.

More recently in the literature, it is observed that the use of multinomial logistic regression modeling is becoming essential. Phillips and VanderHoff (2004) utilize the multinomial logistic regression model to determine possible outcomes of a defaulted scenario. Three outcomes are considered: the resumption of payments, termination by prepayment, and foreclosure. Findings indicate that the local area economics and housing market conditions affect the default resolution probabilities. They conclude that efficiency of default resolutions might be improved by legal and regulatory reforms. State specific legal statutes and regulations lead them to this conclusion.

The variables in their model are listed below:

- Mortgage value, equity, appreciation, income growth, age of the loan, defaulted time, workout flag indicating any workout option offered to the borrower, mortgage insurance flag, redemption period, tenancy flag, Texas and Florida indicator variables

LaCour-Little (2004) defines equity dilution as the additional debt secured on the house by a junior lien subsequent to the first loan origination. This is an important issue since it has a major impact on the equity of the borrower. Since junior liens are generally unobservable to the senior lien holder, predictions for default might be skewed. He estimates loans that are likely to have junior liens and examined their effect on default probabilities for senior lien holders.

Loan Default Literature Synthesis

Methods used in the default literature are discriminant function analysis, survival analysis, logistic regression, Monte Carlo simulation and optimization.

Important variables used in default research are: CLTV, transaction cost of default, borrower income, expected housing conditions, interest rate, existing borrower equity, down payment at origination, loan age, rent to price ratio, trigger events, unemployment rates, divorce rates, neighborhood income, year of origination and third party origination etc.

Utilizing the option theoretic framework, Kau and Kim (1994) concluded that borrowers could benefit from delaying their default when the HPA is in an increasing trend. Vandell (1995) determines that the option theoretic framework is not sufficient to explain the variability in defaults, and suggest using Poisson regression together with Monte Carlo simulations will do better.

Quigley, Deng and Order (1996) use three stochastic variables: interest rates, house prices, and household income. Ambrose, Buttimer and Capone (1998) model the value of free rent, which corresponds to the time between the default and the foreclosure sale. They conclude that the default is likely to happen when the borrower's expectation of free-rent-time increases.

Capozza, Kazarian and Thomson (1998) introduce the notion of conditional probability of default, bridging the gap between the empirical and theoretical studies of mortgage default research. They indicate that the unconditional probabilities are secondary in importance compared to the conditional probabilities. They determine that unconditional probabilities depend on the origination values of the variables of interest where as the conditional probability depends on the current values of those variables.

Ambrose, Capone and Deng (2001) note the interaction effect of house price cycle stage with the probability of negative equity. They conclude that as the house prices significantly degrade the relationship between the probability of negative equity and default breaks down.

DeFranco (2002) proposes a modeling framework and tests it against several classes of traditional mortgage prepayment and default models. His framework consists of multinomial logistic regression and Markov transition matrix for seven payment statuses. He concludes that his model is statistically and economically better than the previous models. He uses goodness of fit measures, statistical tests and out-of-sample forecasts.

Lacour-Little (2004) defines the equity dilution as the additional debt secured on the house by a junior lien subsequent to the first loan origination. He concludes that such loans are more prone to default.

2.4 Loss Literature

As mentioned earlier, a defaulted loan might incur losses. The loss literature concentrates on two key definitions:

- Frequency of loss: This is the probability of default given the current conditions.
- Loss severity: This is the loss amount relative to the defaulted balance given the condition of default. It depends on, but is not limited to, the factors listed below:
 - Current loan-to-value ratio
 - Default period
 - Age of the loan
 - Final default resolution
 - Cost of servicing

The literature on this topic is limited compared to the prepayment and loan default literature. The trend towards high-risk and high-yield products in 1990s and early 2000s increased the demand for predicting future losses on mortgage-backed securities. Subprime mortgage loans fit well into this definition and their loss behaviors are different from conventional prime mortgage loans. Earlier literature on prime loans ignores the probability of default or assumes it has no impact on the final valuation of the loan.

In recent literature, Capozza and Thomson (2005) outline a two-stage process for losses in subprime loans. The first stage is when the borrower stops making payments depending on the optimality of default. The second stage is the period when the lenders initiate the liquidation of the collateral through their mortgage servicer. Their study explored the role of following key characteristics:

- Borrower characteristics
- Collateral characteristics
- Judicial Process
- Trigger Events
- Option theory variables and loan terms

They conclude that the traditional approaches in academic literature that focus on option pricing methods are difficult to quantify. They have significant effect but little power to explain the variation in the default decision; however, borrower characteristics play an important role in determining the stopping boundary (frequency of default) and eventual total losses (loss severity). Property characteristics and legal requirements also have impact on the total loss. Surprisingly, trigger events such as unemployment and divorce did not appear to have a significant effect on the frequency of loss and loss severity.

One of the key papers in loss severity by Kau and Keenan (1999) is an important attempt to identify the severity of default. They propose that origination LTV is an important variable that determines the severity levels. This contradicts the findings of Lekkas *et al.* (1993); however, Kau and Keenan (1999) explain that the reason for this difference is the use of different severity measures. Lekkas *et al.* use severity rates and severity levels. They also report that as the interest rate of the mortgage contract increases, the probability of default increases; but the loss severity decreases. With imposed house price volatility both the severity and the probability of default increases. Seasoning only impacts the probability of default; the severity remains the same.

Kau and Keenan (1999) indicate that there are three reasons for why subprime borrowers end up with higher loss severities:

- Subprime borrowers are generally less skilled in property care and maintenance compared to prime borrowers
- They are generally less knowledgeable about property values and are likely to overpay for the property at the purchase time
- They may be buying properties that appreciate less and deteriorate faster

Ambrose and Capone (2000) investigate the hazard rates of repeated mortgage defaults, conditioned on reinstatement from an initial default. They conclude that the

two-year period for the subsequent default following the first default is riskier than periods more than two years for reinstated borrowers (payment plans, modifications and so forth). This is an important conclusion for current investors and servicers since the industry is working to avoid foreclosures.

The dynamics of borrower default and the conditions that result in foreclosures are gaining importance as mortgage lenders and servicers realize that loss mitigation efforts can reduce the incidence of foreclosures. These foreclosure forbearance programs are essential to lenders and servicers to reduce mortgage losses. Ambrose and Capone (2000) report that for 3,345 loans that are reinstated after the initial default, 22% defaulted, and none of them prepaid during the analysis period. They use financial, borrower and state specific characteristics to identify the number of months to the second default, given the first default.

Clauret (1990) examine the effect of LTV ratio at origination on the frequency of default and on the loss severity of a defaulted loan. LTV account for between 13% and 23% of the variability for years 1980 and 1983, respectively. This study is conducted on 204,706 loans from the Federal Housing Association (FHA) originated in between 1980 and 1983.

Another important finding in the literature is the relation of frequency of loss to the loss severity. A study by Crawford and Rosenblatt (1995) indicates that the variables

that increase severity are the same variables that reduce the probability of default. This indicates that default probability and loss severity are not independent.

They also indicate that the foreclosure decision of servicer does not obtain differences in severity across states and across market interest rates. This implies that the severity is not dependent on the servicer's performance in foreclosure process.

One of the most important studies in the loss frequency and loss severity field is by Lekkas, Quigley and Van Order (1993). This study uses the option theory to determine the optimal mortgage default from the perspective of the borrower. Their option theoretic model predictions (below) are tested:

- Loss severity should be independent of initial LTV
- Loss severity should be the same in regions with high default frequencies and should be independent of loan's origination year
- Loss severity should decrease with the age of the mortgage
- Loss severity should decrease as coupon rate minus the current interest rate decreases

The option-pricing model outcomes listed above are not consistent with the empirical data from years 1975-1990. The empirical data didn't support their theoretical

framework. Thus, they conclude that the borrower behavior is not consistent with the wealth maximization notion.

Overall, option-pricing models identify an optimal point where a borrower needs to default independently of region and initial LTV ratios. These hypotheses are rejected in several studies by Lekkas, Quigley and Van Order (1993), and Capone and Deng (1998). DeFranco (2002) indicate that these studies show the need for expanding the set of information used to predict severity beyond the existing option theoretic methods.

Other studies such as Smith and Lawrence (1993), Wilson (1995), Smith, Sanchez and Lawrence (1996) base their estimation on empirical methods for estimating the losses and loss severities. The variables in their studies are listed below:

Loan size, lender, LTV, property type, county, change in home prices, house price appreciation by state, indicator variables for being '30- or 60-days delinquent' in the last 12 months, logarithmic transformation of loan age, original and estimated current LTV, initial interest rate, maturity, borrower's age and occupation, average foreclosure time, number of months for right of redemption, indicator variable for judicial foreclosure, state level unemployment data, borrowers income.

Loss Literature Synthesis

Frequency of default and severity of loss are the key focus of researchers in the loss literature. Default frequency multiplied by the loss severity will give the expected loss on a single loan. This expected loss can be used for valuation purposes. It also will determine the cumulative loss within a pool of mortgage loans.

Capozza and Thomson (2005) conclude that option-pricing methods have significant effect but little power to explain the variation in the default decision. They also indicate that the borrower and property characteristics and legal requirements have impact on frequency and severity of loss. The trigger events do not affect the severity.

Kau and Keenan (1999) propose that with interest rate increases loss frequency increases, but the severity decreases. House price volatility increase results with an increase in loss severity and frequency. Seasoning only affects the severity of the loss where as frequency remains unchanged.

Ambrose and Capone (2000) research the hazard rates of repeated mortgage defaults, conditional on reinstating from an initial default episode. They conclude that the subsequent default for the reinstated borrowers (payment plans, modifications and so forth) has significantly greater risk than the first default in the first two years following the first default.

Clauret (1990) examines the effect of LTV ratio on the frequency of default and on the loss severity of a defaulted loan. LTV accounts for between 13% and 23% of the variability for years 1980-1983.

Crawford and Rosenblatt (1995) indicate that the variables that increase the severity are the variables that reduce the probability of default. This implies that default probability and loss severity are not independent.

DeFranco (2002) indicates that these studies show the need for expanding the set of information used to predict severity beyond the existing option theoretic methods.

2.5 Reconstructability Analysis, Monte Carlo Method, and Logistic Regression

Reconstructability Analysis

RA is a discrete multivariate modeling method. It includes both set-theoretic modeling of relationships and mappings, and information-theoretic modeling of frequency and probability distributions. System types can be both directed having input and output variables or neutral without input or output distinction. Zwick (2004) indicates that RA was developed by Klir (1986) together with Broekstra, Cavallo, Cellier, Conant, Jones, and Krippendorff (1986).

RA is a method for detecting and analyzing the structure of multivariate categorical data (Zwick 2004). The method is similar to log-linear analysis (Knoke & Burke 1980) of multi-way frequency tables in statistics. Where RA overlaps with log-linear analysis, the two methods are equivalent. There are, however, a number of aspects of RA methodology that are not present in log-linear analysis, and vice versa. One general difference is that RA is especially suited for exploratory as opposed to confirmatory modeling. The biggest difference from standard statistical methods is that RA works without the linearity and normality assumptions. It captures information within nonlinear relations and high-ordinality interactions between the specified variables. RA uses discrete variables, so continuous variables are discretized. Although discretization loses information, this may be compensated for by RA's ability to detect nonlinearities and interaction effects.

Set-theoretic RA is completely non-statistical and resembles logic design and machine learning found in the electrical and computer engineering literature (Zwick 2004).

Here is a very brief example illustrating how RA works. The data for this example consists of six variables. A,B,C,D,E are independent variables assumed to be predictive of the dependent variable F, which is a binary categorical variable that indicates whether a loan defaulted. The most complete model is when all the independent variables are used to explain the behavior of the dependent variable. In this case, it is the model ABCDEF. This model makes maximum use of the explanatory power within the data but it is maximally complex and may overfit the data. Conversely, there is the null model, ABCDE:F, which does not use any of the independent variables to predict F. It is minimally complex and does not explain the variation in the dependent variable F in any way.

RA uses information to quantify the explanatory power of a model, which is the reduction of uncertainty about the dependent variable provided by knowing the input variables. The null model has 0% information and thus no uncertainty reduction. The complete model has 100% information and maximum uncertainty reduction.

RA attempts to find the best model that falls in between the null and the complete model. The best model is determined based on a tradeoff between two criteria: the model complexity and the goodness of fit. This is accomplished by looking at various

criteria such as Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and statistical significance relative to a reference model.

Another important criterion is the percent correct measure of the model's accuracy of prediction. This indicates the ratio of the sum of true negatives and true positives over the sample population. However, the percent correct measure only is adequate when the dependent variable states occur equally in the sample population. For example, in the sample population where the transition is defined as current to 'Paid in Full', the number of 'Paid in Full' loans should be equal to the number of non-'Paid in Full' loans. Otherwise, more complex measures of predictive accuracy should be used such as Receiver Operating Characteristic Curve, and similarly the c-statistic.

RA is used to detect significant interaction effects. For example, a directed RA model of AB:AC:BC means that the variables A and B have independent effects on the dependent variable C. The first pair (AB) indicates the independent variables in the model, and following pairs (AC and BC) indicate the relationship type between an independent variable and the dependent variable. The model AB:AC:BC says that A and B separately predict the dependent variable C, these separate effects being integrated by the maximum entropy principle method. A higher-level model ABC indicates the independent relation of A and B with C as well as the interaction of A and B with C. In other words, it embraces the AB:AC:BC relation as well as the triadic interaction term. If the model AB:AC:BC is not significant (for example, relative to AB:C as a reference) and the model ABC is significant, the interaction of

variables A and B to explain C is significant even though the main effects of A and B are not significant. This is more fully explained in Zwick (2004) and Shervais, Zwick and Kramer (2005).

RA is used in the following areas, bio-medical data analysis, decision analysis, pattern recognition, syntactic constraints of languages, and dynamics of cellular automata.

Some papers that applied RA in the literature in ways relevant to this dissertation research are the following. In his study about the relationship between education, literacy and health, Carletti (2004) uses RA to detect interaction effects. He looks at four-way interactions and compares them to the immediate simple ancestor model that includes all possible three-way interactive relations except the four-way interaction itself. He notes that the test of significant difference between these models would identify whether the four-way interaction is significant. For example, the ABCD model is compared to the ABC:ABD:ACD:BCD model in order to detect a significant tetradic interaction effect that might be interpretable for understanding the relations between variables. A successful detection of statistically significant interaction effects using RA led Carletti to test these interactions in a regression model. He tests multiplicative and divisive forms and notes some significant interactions that improved the regression model's overall R^2 .

Mist (2007) wants to improve the feasibility of incorporating Chinese Medicine diagnosis by prescreening participants using questionnaires. He uses logistic regression models to predict Chinese Medicine diagnosis and enhances the accuracy of

these models with RA suggested interaction effects. He finds two interaction terms using RA which are significant in logistic regression models.

Shervais, Zwick and Kramer (2005) use RA as a tool for identifying gene-to-gene interactions in studies of human diseases. Detecting interactions with standard statistical methods is difficult when there is no significant main effect of the two individual genes. However, their interaction can be significant and can be detected by RA. RA is robust in these environments and compares favorably with other approaches including neural network modeling.

Monte Carlo Method

The Monte Carlo method was first introduced by Stanislaw Ulam during the Manhattan Project. Nicholas and Ulam (1949) use the Monte Carlo method for dealing with problems in mathematical physics. The method is a probabilistic approach to the study of differential equations.

Monte Carlo method utilizes random numbers to solve problems by brute force. The random numbers are either pre-generated or a pseudo random generator function is used. These random numbers in conjunction with the provided probability distribution functions create random values for the independent variables of interest. With the use of specified rules, the synthetic independent variable values generate an “outcome” sample, interpreted as the dependent variable distribution. Confidence intervals are

built using that distribution and results are deduced within the given confidence levels. The results are statistically tested for significance.

Efficiency of the Monte Carlo method depends on the speed of computing machines. There are specific software packages that utilize Monte Carlo method even in today's personal computers, providing random number generators that are essential for the Monte Carlo method.

Monte Carlo method is applied in the mortgage industry for valuation of mortgage deals. The stochastic behavior of interest rates and housing price appreciations are generally modeled using the Monte Carlo method. Akesson and Lehoczky (2000) indicate the importance of the Monte Carlo method for pricing and hedging of complex, path dependent financial instruments. They develop a low discrepancy method to enhance the model predictability for mortgage backed securities valuation.

McConnell and Singh (1993), Kau and Keenan (1999), Kau and Kim (1994), Vandell (1995), Quigley, Deng and Order (1996), Ambrose, Buttimer and Capone (1998), Capozza, Kazarian and Thomson (1998) all utilize the Monte Carlo method to simulate the stochastic behavior of interest rates and house price appreciations. This provides an expected value of a mortgage loan given the different states of the environment.

Monte Carlo method is the essence of discrete event simulation (DES). In DES, each event has a frequency distribution or a timeline distribution. As the complexity of such systems increase, mathematical solutions to determine the system behavior (such as queuing theory) cannot be applied easily. Instead, the use of the Monte Carlo method enables a solution. The outcome is not a closed form solution to the problem but rather a computed numerical solution for which confidence intervals can be calculated for a given confidence level.

Logistic Regression

The origin of logistic regression goes back to 19th century. It was invented for the description of the growth of populations and the course of autocatalytic chemical reactions. It evolved through various papers and books from Aitchison and Brown (1957), Berkson (1980) to Hosmer and Lemeshow (1989). A clear explanation of logistic regression can be found in Tabachnick and Fidell (1996).

Logistic regression is a generalized linear model that utilizes the logit as its link function (Equation 1).

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i} \quad i = 1, \dots, n \quad \text{where } p_i = P(Y_i = 1) \quad [1]$$

There are multiple forms of logistic regression techniques:

- *Binomial Logistic Regression* is useful for a binary response variable. A binary variable only has two possible values, such as presence or absence of a particular disease.
- *Ordinal Logistic Regression* is used for an ordinal response variable. Ordinal variables are categorical variables that have three or more possible levels with a natural ordering, such as strongly disagree, disagree, neutral, agree, and strongly agree.
- *Multinomial Logistic Regression* is useful for a nominal response variable using an iterative-weighted least squares algorithm to obtain maximum likelihood estimates of the parameters. Multinomial logistic regression uses nominal variables that three or more categories with no natural ordering.

Binomial logistic regression is used when the dependent variable has two states. If dependent variable has more than two states that have a natural ordering, ordinal logistic regression is used, otherwise, multinomial logistic regression is preferred.

Chapter 3 – Methodology

This chapter includes three sections:

- **3.1 Payment model**: A new 9-State Markov transition model that is compared to the 7- state model proposed by DeFranco (2002) in terms of predictive accuracy.
- **3.2 Interaction effects**: A method for detecting (exploratory analysis) interaction effects using reconstructability analysis and using the interaction terms to enhance the predictive accuracy of the logistic regressions used in the payment model.
- **3.3 Data**: Description of the Loan Performance dataset, variables and tools used in the study.

3.1 Payment Model

This section discusses an extension to the model presented by DeFranco (2002). Table 1 summarizes the key models discussed in the literature review, contrasted with the enhanced 9-State model.

Model Name	Number of Loan States	Possible Loan Payment States
Combined Prepayment and Default (previous literature)	3	Active, Paid Off, Defaulted
7-State Model (DeFranco 2002)	7	Current, 30-days delinquent, 60-days delinquent, 90+ days delinquent, Foreclosure, REO, Paid Off
Proposed 9-State Model	9	Current, 30-days delinquent, 60-days delinquent, 90+ days delinquent, Early Foreclosure, Late Foreclosure REO, Paid in Full, Terminated with Loss

Table 1 – Summary and comparison of existing and proposed payment models

One of the drawbacks in DeFranco’s model is that there is only one termination status: ‘Paid Off’. Loans in that status could be ‘Terminated with Loss’ or ‘Paid in Full’. This separation will enable the modeling framework to identify expected losses for a mortgage loan. Furthermore, foreclosure is a long process that may take an average of 4-5 months, depending on the state and its corresponding legislation. In earlier stages of foreclosure, the probability of a loan curing is higher compared to the later stages of foreclosure due to heavy loan workout by the servicer. Thus, another drawback of DeFranco’s model is that it considers the behavior of ‘Late Foreclosure’ similar to ‘Early Foreclosure’. Table 2 shows all possible transitions between states. The probability for each transition is denoted as q_t^{x-y} where the t indicates that each probability could vary as a function of time (in this case it is one month transition), and the x-y indicate the starting and ending state. When state changes are not possible, the corresponding cell has a probability of zero.

From\To	Current	30 Days Delq	60 Days Delq	90+ Days Delq	Early Foreclosure	Late Foreclosure	REO	Paid in Full	Terminated with Losses
Current	q_t^{C-C}	q_t^{C-3}	0	0	0	0	0	q_t^{C-PIF}	0
30 Days Delq	q_t^{3-C}	q_t^{3-3}	q_t^{3-6}	0	0	0	0	q_t^{3-PIF}	0
60 Days Delq	q_t^{6-C}	q_t^{6-3}	q_t^{6-6}	q_t^{6-9}	q_t^{6-EF}	0	0	q_t^{6-PIF}	0
90+ Days Delq	q_t^{9-C}	q_t^{9-3}	q_t^{9-6}	q_t^{9-9}	q_t^{9-EF}	0	0	q_t^{9-PIF}	q_t^{9-L}
Early Foreclosure	q_t^{EF-C}	q_t^{EF-3}	q_t^{EF-6}	q_t^{EF-9}	q_t^{EF-EF}	q_t^{EF-LF}	q_t^{EF-R}	q_t^{EF-PIF}	q_t^{EF-L}
Late Foreclosure	q_t^{LF-C}	q_t^{LF-3}	q_t^{LF-6}	q_t^{LF-9}	0	q_t^{LF-LF}	q_t^{LF-R}	q_t^{LF-PIF}	q_t^{LF-L}
REO	0	0	0	0	0	0	q_t^{R-R}	0	q_t^{R-L}
Paid in Full	0	0	0	0	0	0	0	1	0
Terminated with Losses	0	0	0	0	0	0	0	0	1

Table 2 – Proposed 9-State payment model transition matrix

To determine the q values, a binomial logistic regression analysis is performed. The independent variables for these regression models are borrower, loan, and economic variables relevant to that particular state. The same analysis is conducted for the 7-State model using same variables.

For a simple prepayment model example where a loan is either prepaid or active, the binomial logistic regression form is shown in Equation 2.

$$\log [P(\text{Prepaid})/P(\text{Active})] = b_0 + b_1 * \text{CreditScore} + b_2 * \text{LoanAmount} + \dots + b_n * \text{LTV} \quad [2]$$

The probability of a loan prepaying can then be represented as shown in Equation 3.

$$P(\text{Prepaid}) = 1 / \{ 1 + \exp [-(b_0 + b_1 * \text{CreditScore} + b_2 * \text{LoanAmount} + \dots + b_n * \text{LTV})] \} \quad [3]$$

For variables with more than two nominal outcomes, the multinomial logistic regression is used. For example, if a loan is prepaid, active or defaulted as an outcome, then the multinomial logistic regression model is shown in Equation 4 and 5.

$$\log[P(\text{Prepaid})/P(\text{Active})] = b_{p0} + b_{p1}*\text{CreditScore} + b_{p2}*\text{LoanAmount} + \dots + b_{pn}*\text{LTV} \quad [4]$$

$$\log[P(\text{Defaulted})/P(\text{Active})] = b_{p0} + b_{p1}*\text{CreditScore} + b_{p2}*\text{LoanAmount} + \dots + b_{pn}*\text{LTV} \quad [5]$$

Active is the reference group and both equations are solved simultaneously using the least squares method to derive probabilities for each state of the dependent variable.

Multinomial logistic models are multi-equation models. A response variable with k categories will generate $k-1$ equations. For each of these $k-1$ equations, there is a binary logistic regression. Multinomial logistic regression simultaneously estimates the $k-1$ logits of each binary logistic regression.

The logistic regressions are then utilized by a discrete system simulation model using Monte Carlo method to simulate the status for the next month given the initial conditions of a loan. By repeating this process multiple times, the model outputs monthly statuses for a loan.

As an example, a '30-days delinquent' loan with given attributes, the multinomial logistic regression generates transition probabilities to all possible states. Table 3 displays the output of that operation.

Transition Matrix starting from 30-days delinquent	Current	30-days delq.	60-days delq	90+ days delq	Early FC ¹⁴	Late FC.	REO	Paid in Full	Terminated with Loss
Transition Probabilities	0.6	0.1	0.25	0	0.00	0	0	0.05	0
Cumulative Transition Probabilities	0.6	0.7	0.95	0.95	0.95	0.95	0.95	1	1
Range of cumulative Probabilities	0.0-0.6	0.6-0.7	0.7-0.95	0.95-0.95	0.95-0.95	0.95-0.95	0.95-0.95	0.95-1	1-1
Valid Transition Flag	1	1	1	0	0	0	0	1	0

Table 3 – Transition matrix illustrating the mechanics of Monte Carlo method

The simulation uses Monte Carlo method to generate random numbers between 0 and 1 using a uniform distribution. For example, Monte Carlo method assigns the number .563 randomly, the loan transitions into 'Current' status depending on the range of cumulative probabilities and valid ranges. Another run of Monte Carlo might assign the random number like .950. Then, the loan will transition to '60-days delinquent' since '90+ days delinquent', 'Early Foreclosure', 'Late Foreclosure' and 'Real Estate Owned' (REO) are not valid transitions when the transition starts from '30-days delinquent' status.

¹⁴ FC means foreclosure.

Both 7-State and 9-State models are tested for accuracy using the actual data that is separated from the initial training dataset. The measure of accuracy is the absolute error of the simulation model results from the real data. The ‘Early and Late Foreclosure’ states are aggregated into a single ‘Foreclosure’ state to enable comparison between two different models. ‘Terminated with Loss’ and ‘Paid in Full’ states also are aggregated into a single ‘Paid Off’ state. Table 4 displays the comparison framework. For each status, a statistical comparison test is conducted.

For example, if a loan portfolio shows 300 loans in foreclosure by the 15th period and the model predicts 250 loans, there is an absolute error of 50. The absolute error is calculated for each month for both models. The median absolute errors of both models are statistically tested using a Mann-Whitney U test to compare overall model performance. The null hypothesis is that the 9-State model is not different from the 7-State model in terms of estimating each status. The 9-State model may be considered an improvement upon the 7-State model, if the significance test indicates that the null hypothesis is rejected with $p < .05$. This predictive accuracy comparison methodology is explained in detail by Diebold and Mariano (1995).

Figure 1 displays the empirical methodology for testing the 9-State payment model’s performance compared to the 7-State model’s. The initial step is to collect model projections from both 7- and 9-State models. This is done by applying the model to a group of loans from the test data. The next step is to compare the predicted values to the actual for each month to compute absolute errors for each status for both models.

Lastly, median absolute errors are compared for each model using Mann-Whitney U test in order to confirm the hypothesis whether the 9-State model is significantly different and better than the 7-State model.

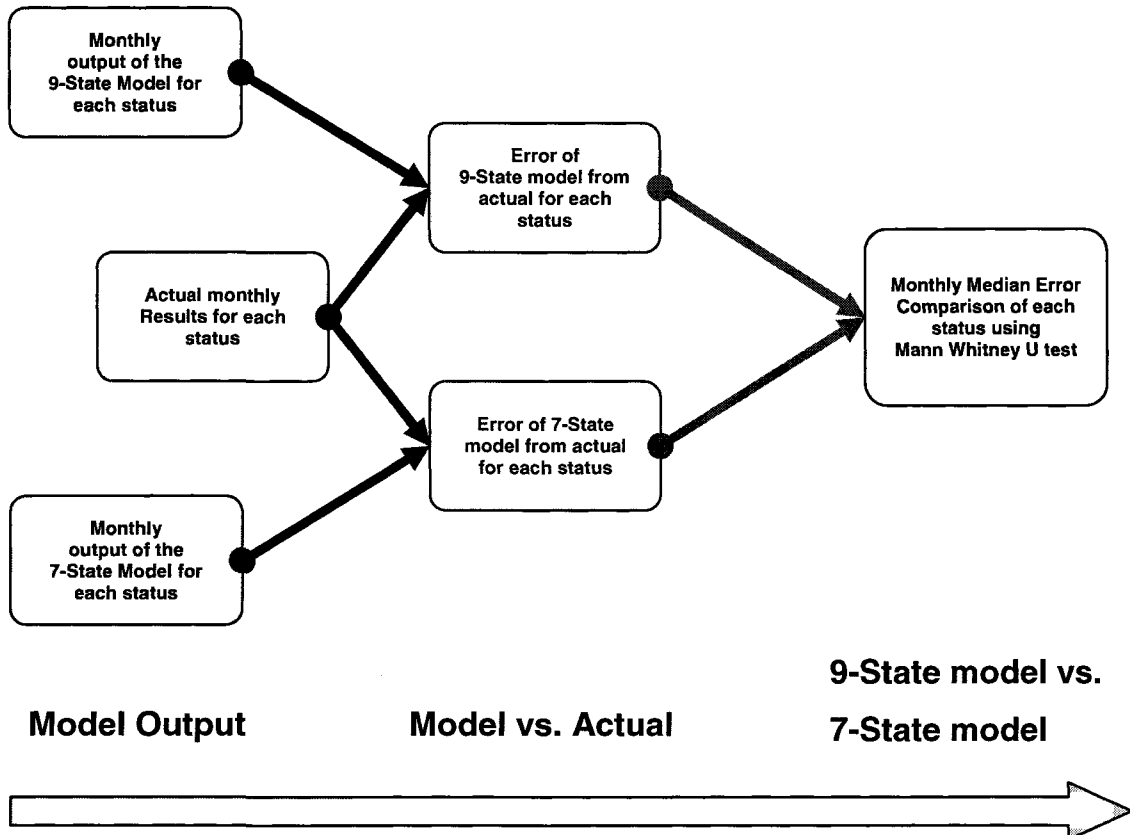


Figure 1 - Testing scheme for payment model

Formal statistical test of determining the difference between 9-State model vs. 7-State model goes back to DeFranco (2002). He tests the restrictions implied by the 3-state model and rejected these restrictions, concluding that the 7-State model is significantly impacting the accuracy of the predictions (Table 4).

The 3-state model implies that the probability of paying off is the same, regardless of the current payment status of the loan. In other words, the 3-state model forces the coefficients to be the same for all of the transitions to ‘Paid Off’ in the 7-State model. The Table 4 below shows the differences between these three frameworks.

Loan Payment Statuses by Model									
3 - State Model	Active					Default		Paid Off	
7 - State Model	Current	30	60	90+	FC		REO	Paid Off	
9 - State Model	Current	30	60	90+	Early FC	Late FC	REO	Terminated with Loss	Paid in Full

Table 4 – Model framework comparison

Table 5 shows the construct of the testing schema of DeFranco. The null hypothesis is that the 7-State model’s ‘Default’ and ‘Paid Off’ coefficients for each active status is the same as the 3-State model’s ‘Default’ and ‘Paid Off’ coefficients. Looking at each logistic regression and testing its significance will help reject the restriction of the 3-state model. Wald Chi-Square test is used to test the null hypothesis that the 7-State model is not different from the 3-state model for ‘Paid Off’, ‘Default’ and ‘Active’ statuses. He concludes for all statuses that the Wald Chi-square is significant, leading to the conclusion to reject the restrictions.

Logistic Regression	From Loan Status	To Loan Status		
		Active	Default	Paid Off
3-State Model	Active	q_t^{A-A}	q_t^{A-D}	q_t^{A-P}
7-State Model	Current	q_t^{C-A}	q_t^{C-D}	q_t^{C-P}
	30	q_t^{3-A}	q_t^{3-D}	q_t^{3-P}
	60	q_t^{6-A}	q_t^{6-D}	q_t^{6-P}
	90+	q_t^{9-A}	q_t^{9-D}	q_t^{9-P}
	Foreclosure	q_t^{F-A}	q_t^{F-D}	q_t^{F-P}

Table 5 – 7-State model restriction tests construct compared to the 3-state model

The difference between the 9-State model and the 7-State model in terms of active statuses is the separation of foreclosure state into ‘Early and Late Foreclosure’ states.

Table 6 displays the construct and use of Wald Chi-Square test for the logistic regression permitting testing the null hypothesis that use of two foreclosure states in 9-State model is not different from the single foreclosure state in 7-State model. The expected result is to see high Wald Chi-Square statistics for regressions specified below, having p-values smaller than the generally accepted $\alpha=.05$ confidence level.

Logistic Regression	From Loan Status	To Loan Status						
		Current	30	60	90+	Foreclosure	Default	Paid Off
7-State Model	Foreclosure	q_t^{F-C}	q_t^{F-3}	q_t^{F-6}	q_t^{F-9}	q_t^{F-F}	q_t^{F-D}	q_t^{F-P}
9-State Model	Early Foreclosure	q_t^{EF-C}	q_t^{EF-3}	q_t^{EF-6}	q_t^{EF-9}	q_t^{EF-F}	q_t^{EF-D}	q_t^{EF-P}
	Late Foreclosure	q_t^{LF-C}	q_t^{LF-3}	q_t^{LF-6}	q_t^{LF-9}	q_t^{LF-F}	q_t^{LF-D}	q_t^{LF-P}

Table 6 – 9-State model restriction tests construct compared to the 7-State model

3.2 Interaction Effects

In mortgage literature, different variables are used to model borrowers' behavior, such as credit score, unpaid balance (UPB), collateral value, LTV ratio, loan age (in months), property type, purpose of the mortgage, debt to income ratio, state and so forth. Economic conditions also are used, such as house price appreciation figures and prevailing interest rates. For example, holding all other variables constant, the correlation of LTV on default probability is assumed to be positive due to findings in the literature. However, this correlation does not always hold when the borrower has a high credit score (FICO). High credit score borrowers tend to maintain good credit histories and are less likely to default. This interaction effect between LTV and FICO is incorporated into earlier studies.

Interactions between other variables (for example, unpaid balance and LTV, property type and collateral value and so forth) might help to further explain borrower's payment behavior. These possible interactions may be tested by including interaction variables in the regression models. The interaction variable is the product of the variables involved in the hypothesized interaction. Including all possible interaction terms can be computationally tedious to resolve in logistic regressions with more than ten independent variables.

Therefore, this research uses RA to study potential interactions between the variables of logistic regression equations that construct the transition models. RA can detect interaction effects, as described previously in the RA literature review.

The process for using RA in determining significant interaction effects consists of the following steps:

- Selecting a specific payment transition
 - '30-days delinquent' to 'Paid in Full' transition is selected since its one of the difficult transitions to predict
- Sampling observations and data manipulation
 - 20,000 observations are randomly selected where the DV states are equally sampled. This is done both for training and testing. Both of them had 10,000 'Paid in Full' and 10,000 non 'Paid in Full' loans.
- Discretizing continuous variables of interest (some of them are nominal with no need to discretize)
 - Out of 82 variables used in the logistic regression, 39 are binned (transformed from a continuous to a categorical variable)
 - All the continuous variables are binned into four bins with number of observations equally distributed in each bin
 - Optimal binning is chosen to be four equal bins after testing for two, three, four and five bins per continuous variable; four bins yielded the highest %correct(training) and %correct(test) results
- Determining the important variables in order to reduce the possible search space

- Based on the logistic regression results on the same training sample, the variables of RA model are limited to the ones that are significant in LR model. This ensures a significant reduction in computational complexity.
- Determining the best models based on BIC, AIC, statistical significance and high % correct results on test population.
 - In order to explore higher level interactions, OCCAM searches more complex models starting from the initial ‘best’ models. These yielded more interaction terms.
- Confirming the validity of the interaction effects on the test data.
 - This is to ensure whether the interaction term is significant in RA.
- Adding these new interaction terms to LR model and observing the training and testing percent correct in order to compare it to the LR model without interactions.
- Comparing the LR model with interactions to LR model without interactions using Chi-Square test.

A sample RA output (from the computer program OCCAM) is displayed as Table 7:

MODEL	Level	H	dDF	dLR	Alpha	Inf	%dH(DV)	dAIC	dBIC	%C(Training)
IV:DEHJP	4	10.92	143	18719.3	0	0.727	35.61	18433.3	17211.6	79.24
IV:DHIJP	4	10.93	143	18539.6	0	0.72	35.27	18253.6	17031.9	77.89
IV:CDHJP	4	10.95	71	17405.5	0	0.676	33.11	17263.5	16656.9	77.81
IV:DHJP	3	10.96	35	16927	0	0.658	32.2	16857	16557.9	77.59
IV:CDEHJP	5	10.91	287	19508.4	0	0.758	37.11	18934.3	16482.5	79.81
IV:DEIJP	4	10.93	191	18393	0	0.714	34.99	18011	16379.3	78.63
IV:DHJNP	4	10.96	71	17064.3	0	0.663	32.46	16922.3	16315.7	77.72
IV:DHJKP	4	10.96	71	17004.8	0	0.661	32.35	16862.8	16256.3	77.6
IV:CDHJNP	5	10.95	143	17579.1	0	0.683	33.44	17293.1	16071.4	78.06
IV:CDHJKP	5	10.95	143	17530.4	0	0.681	33.35	17244.4	16022.8	77.85
IV:CDHIJP	5	10.92	287	18989.5	0	0.738	36.13	18415.5	15963.7	78.48

Table 7 – Sample OCCAM output for RA

In Table 7 the “Model” column shows the alternative models with different independent variable combinations. The last letter, in this case, P, is reserved for the dependent variable; the other letters indicate independent variables. For example, IV:DEHJP means, a model with D, E, H and J as independent variables and P as the dependent variable. IV stands for the independent variables and a short indication of all the independent variables. “Inf” column indicates the amount of information in percentage compared to the full model, which has 100% information, but not shown in Table 7. “dBIC” and “dAIC” columns are information criteria that take into account the model simplicity and richness of information. When estimating model parameters using maximum likelihood estimation, it is possible to increase the likelihood by adding additional parameters, which may result in over-fitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. This penalty for additional parameters is stronger than that of the AIC. In Table 7, models are sorted by dBIC. These two criteria are used to evaluate model performance in terms of tradeoff between likelihood ratio and complexity.

The last column, “%correct (training)”, shows how well the model fit the training data. It is the ratio of the sum of true positives and false negatives to the sample size expressed as a percentage. Using these criteria, one can identify the best model. In Table 7, the model DEHJP will be the best fit with highest dBIC and a relatively high %correct (training).

Logistic regression model with interaction and logistic regression model without interaction are compared using the significance of the difference in likelihood ratios of both models at the overall model level. Difference of likelihood ratios and difference of degrees of freedom are calculated as:

$$\Delta\text{LikelihoodRatio} = \text{LikelihoodRatio}_{\text{LRint}} - \text{LikelihoodRatio}_{\text{LRnoint}} \quad [6]$$

$$\Delta\text{DF} = \text{DF}_{\text{LRint}} - \text{DF}_{\text{LRnoint}} \quad [7]$$

Chi-Square test is utilized to measure whether the difference between the two models is significant.

3.3 Data

The dataset used for this research is called the Loan Performance dataset. It consists of residential subprime mortgage data and it is the largest commercially available dataset in the industry. It provides both the origination information of mortgage loans and the monthly snapshots of payment and status information. It has information on 19 million loans over a 20-year period. This dataset is available to this researcher as an employee of Wilshire Credit Corporation, a wholly owned subsidiary of Bank of America. The company is supportive of the research.

Loan Performance Origination Data Table Schema	
Columns	Columns (continued)
Pool ID	Servicer Fee Rate
Deal No	Negative Amortization
Loan ID	Negative Amortization Limit
Property Zip	Index ID
State	Margin
Property Type	Periodic Interest Rate Cap
Number of units	Periodic Interest Rate Floor
Occupancy	Periodic Pay Cap
Origination Date	Periodic Pay Floor
Maturity	Maximum Lifetime Interest Rate
First Payment Date	Minimum Lifetime Interest Rate
Origination Amount	Rate Reset Frequency
Closing Balance	Pay Reset Frequency
Closing Interest Rate	First Rate Period
Sale Price	First Pay Period
Appraisal value	Amortization Term
Product Type	FICO Score
Term	Lien Position
Initial Rate	Credit Grade
Debt to Income Ratio	Prepayment Penalty
Loan Type	Prepayment Term
Purpose	First Rate Cap
Payment Frequency	Pledge Amount
Loan Source	Effective LTV
Buydown	First LTV
Documentation	Second LTV
Convertible Flag	Combined LTV
Pool Insurance	Servicer
Original LTV	Originator

Table 8 – Origination Table Fields

Table 8 shows the static (origination) information of the loans whereas Table 9 has the dynamic (monthly) information on payment, interest rate, status and so forth.

Loan Performance Dynamic Data Table Schema	
Columns	Columns (continued)
Pool ID	Exception
Deal No	Foreclosure Start Date
Loan ID	Foreclosure End Date
Last Interest Paid Date	Payoff Date
Balance	REO Date
Interest Rate	Investor Balance
Total Payment Due	Next Interest Rate
Scheduled Principal Payment	Loss Amt
Scheduled Monthly Payment	Net Pass Through Rate
MBA Method Delinquency Status	Month ID
OTS Method Delinquency Status	Pool Date
Delinquency History	

Table 9 – Dynamic Historic Table Fields

Only first lien and fixed loans are modeled. Data is processed and thoroughly cleaned to avoid missing data issues. Incomplete observations are removed (approximately 5% of the entire data). The variables used in this research are displayed in Appendix I.

Some variables shown in Appendix I might only apply to specific transitions. For example, foreclosure variables do not play a role in ‘Current’ to ‘Paid in Full’ transition.

Chapter 4 – Results

The results are presented in two sections. Section 4.1 gives the results for the 9-State payment model testing. Section 4.2 presents the results of the ‘30-days delinquent’ to ‘Paid in Full’ transition model, with regards to the comparison of LR, and LR with interactions effects.

4.1 Payment Model Results

Payment model testing is presented in two sections. The first part shows the empirical testing which compares the 7-State model predictions and 9-State model predictions to the actual data. This comparison emphasizes the improvement in prediction accuracy achieved by the 9-State model. The second part presents the formal statistical testing of new statuses introduced by the 9-State model. In other words, it focuses on the statistical validity of introducing new statuses to the payment model.

Results of Empirical Testing of the 9-State Payment Model

An empirical test is conducted for 40 months into the future, on 1,666 out-of-sample loans that are initially in foreclosure. The results are compared to the actual. This is a very stringent test of the model’s prediction capability.

Table 10, below, shows the results of 7-State and 9-State models, actual data and the comparison criterion which is the error from the actual results, with median absolute

Period	Current Loans			Absolute Error	
	Actual	7state	9state	7-State	9-State
1	50	52	71	2	21
2	62	68	88	6	26
3	92	85	98	7	6
4	89	86	102	3	13
5	88	81	114	7	26
6	99	85	110	14	11
7	99	86	110	13	11
8	104	89	107	15	3
9	101	86	110	15	9
10	106	83	104	23	2
11	117	84	101	33	16
12	115	84	97	31	18
13	111	81	102	30	9
14	107	79	98	28	9
15	100	78	94	22	6
16	99	81	89	18	10
17	104	78	83	26	21
18	91	77	85	14	6
19	94	77	83	17	11
20	96	72	78	24	18
21	92	71	81	21	11
22	84	65	80	19	4
23	74	57	74	17	-
24	76	62	70	14	6
25	74	58	68	16	6
26	79	56	68	23	11
27	89	53	63	36	26
28	82	53	61	29	21
29	72	53	62	19	10
30	75	49	53	26	22
31	73	42	51	31	22
32	66	47	52	19	14
33	69	49	50	20	19
34	65	44	46	21	19
35	66	41	47	25	19
36	56	39	47	17	9
37	62	40	40	22	22
38	65	42	39	23	26
39	67	38	37	29	30
40	67	34	43	33	24
Median Absolute Error				21	12

Table 10 – Model comparison data

error highlighted at the bottom. These are the results for loans that have not missed their monthly payments.

Discrepancies are calculated by period (monthly) using errors of model from the actual data. This is a common measure of forecast error in time series analysis. The median of these errors are tested for significant difference using the Mann-Whitney U test. The pairs are 7-State and 9-State errors.

Figure 2 displays from Table 10 the number of loans that are current in their payments for each period into the future. In this test, since all loans start from foreclosure there are no current loans initially. A few loans will transition from foreclosure to current in the first month (50 out of 1,666 in this case). As the number of periods increase, the transitions of loans in and out of current status determine the number of current loans for subsequent periods. As shown in Figure 2, this number peaks at 120, ten periods into the future and then declines slowly.

The solid, dotted and dashed lines indicate the actual data, the 7-State model results, and the 9-State model results, respectively. The 9-State model results are visually closer to the actual data on average compared to the 7-State model results.

The next step is to test whether the median absolute errors for both models are significantly different than each other. For current loans, the 9-State model median absolute error is 12, and the 7-State model median absolute error is 20.5, based on the 40 pairs of absolute errors in Table 10. Mann-Whitney U test results using these pairs indicate that the difference between the two median absolute error figures is significant ($p= 0.003$). This suggests that the 9-State model forecasts the number of current status loans better than the 7-State model.

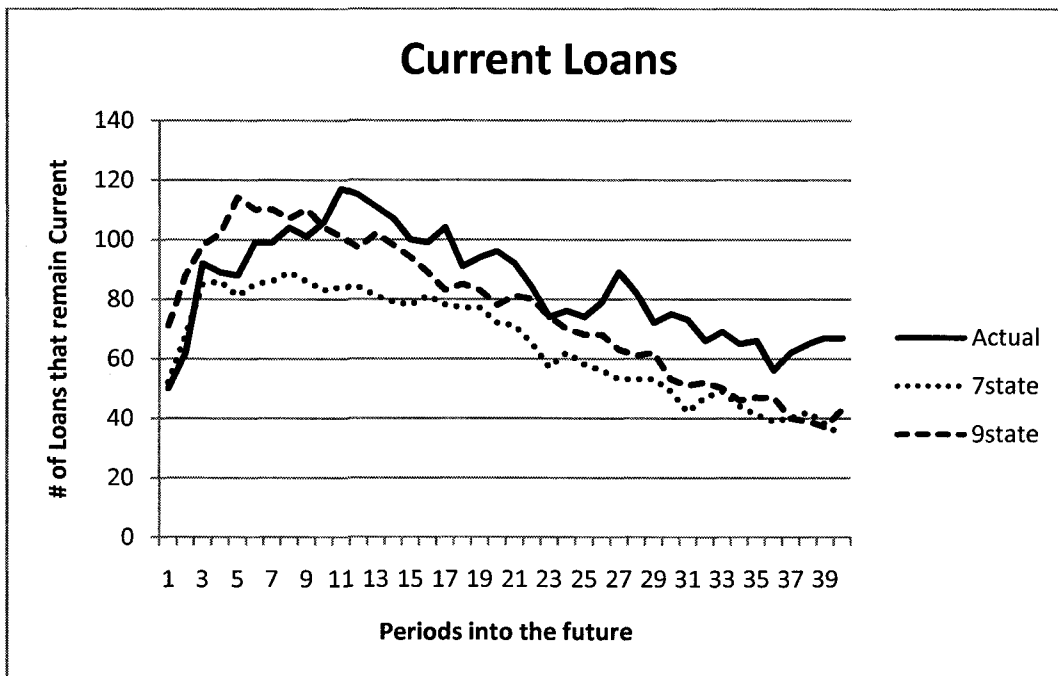


Figure 2 – 7-State and 9-State model predictions compared to the actual data for 'Current' loans

The next transition modeled is from foreclosure to 30-days delinquent. This is a less common transition, so the number of loans is smaller (Figure 3). The shape of the plot is similar to Figure 2. As shown in Figure 3, for 30-days delinquent loans, the 9-State model also performs visibly better than the 7-State model. The 9-State model median absolute error is 4 and the 7-State model median absolute error is 7.5. Mann-Whitney U test results indicate that the difference between the median absolute errors is significant ($p=0.001$). These results indicate that the 9-State model is better than the 7-State model for predicting the number of 30-days delinquent loans.

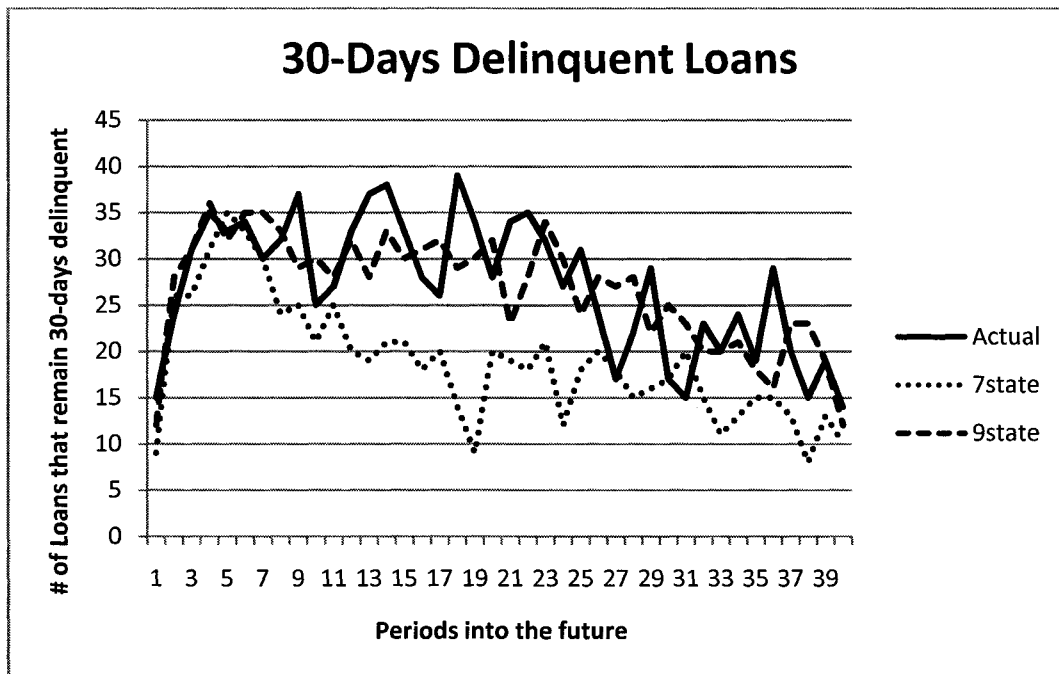


Figure 3 – 7-State and 9-State model predictions compared to the actual data for thirty day delinquent loans

For 60-days delinquent loans, the numbers are smaller still. The 9-State model predictions look similar to the 7-State model predictions (Figure 4). The 9-State model median absolute error is 4; while the 7-State model median absolute error is 6.5. Mann-Whitney U test results indicate that the difference between the median absolute errors is significant ($p=0.026$) indicating that the 9-State model performs better than 7-State in predicting 60-days delinquent status.

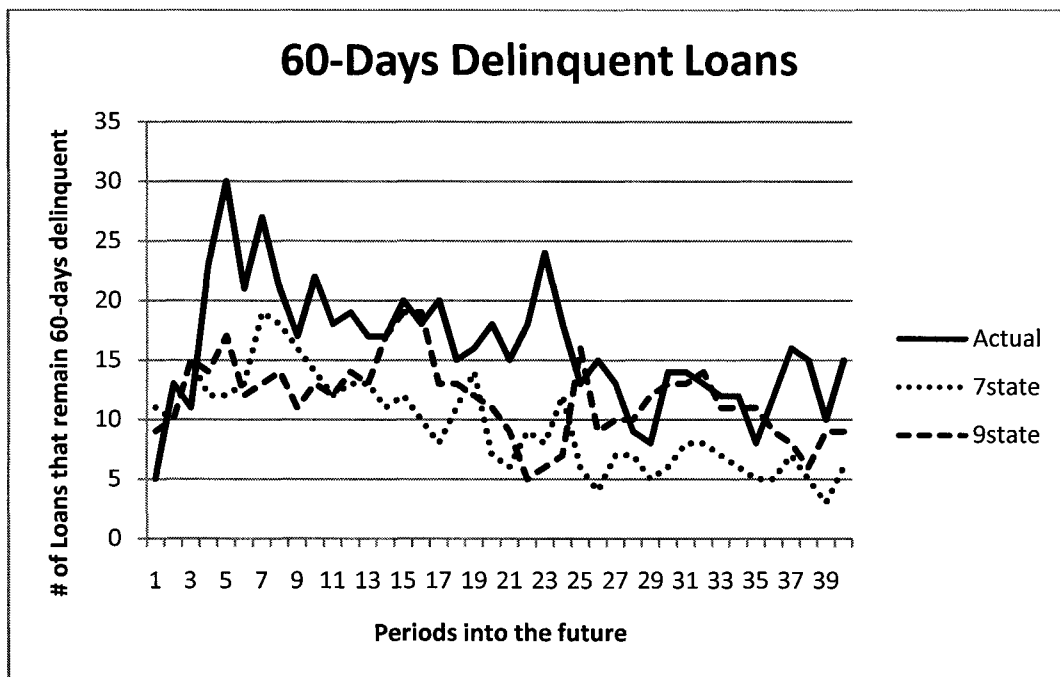


Figure 4 - 7-State and 9-State model predictions compared to the actual data for sixty day delinquent loans

For 90-days delinquent loans, the numbers are larger, as this is a more common transition. The shape of the plot is more exaggerated. The 7-State model performs better compared to the 9-State model (Figure 5). The 9-State model median absolute error is 39 and the 7-State model median absolute error is 36.5. However, Mann-Whitney U test results indicate that the difference between the median absolute errors is not significant with a p-value of 0.206. Both models' absolute error distributions are not statistically different than each other.

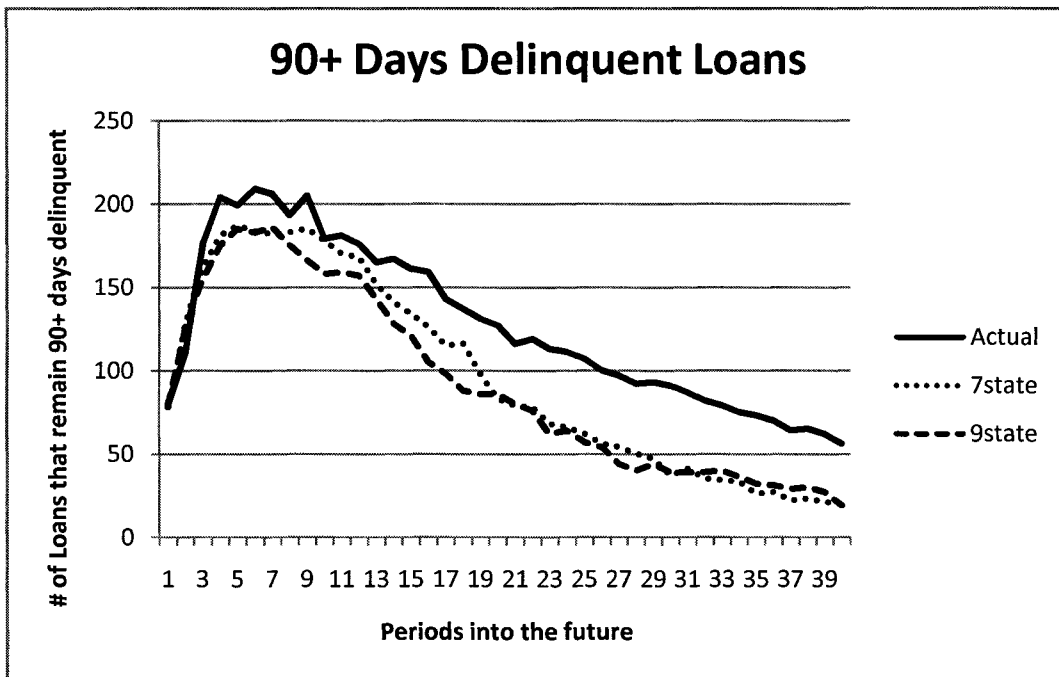


Figure 5 - 7-State and 9-State model predictions compared to the actual data for ninety plus day delinquent loans

Most of the loans that start in foreclosure remain in foreclosure for the next few periods, but over time, most of them eventually become 'Paid Off'. In terms of predicting the number of foreclosure loans, the 9-State model visibly performs better than the 7-State model (Figure 6). The 9-State model median absolute error is 21; while the 7-State model median absolute error is 26.5. However, Mann-Whitney U test results indicate that the difference between the median absolute errors is not significant ($p = 0.074$).

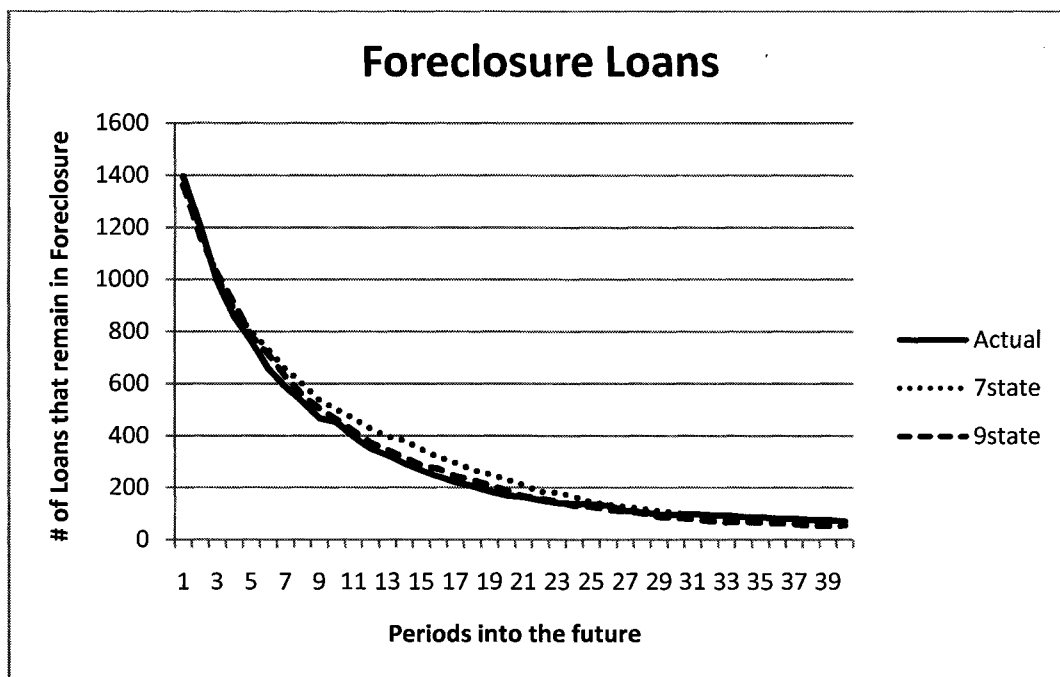


Figure 6 - 7-State and 9-State model predictions compared to the actual data for foreclosure loans

A larger number of loans in foreclosure transition to REO (bank owned) loans. The 9-State model also performs better than the 7-State model for predicting REO (Figure 7). The 9-State model median absolute error is 12.5 and the 7-State model median absolute error is 53.5. In this case, Mann-Whitney U test results indicate that the difference between the median absolute errors is significant ($p < 0.001$). The 9-State model outperforms the 7-State model in REO.

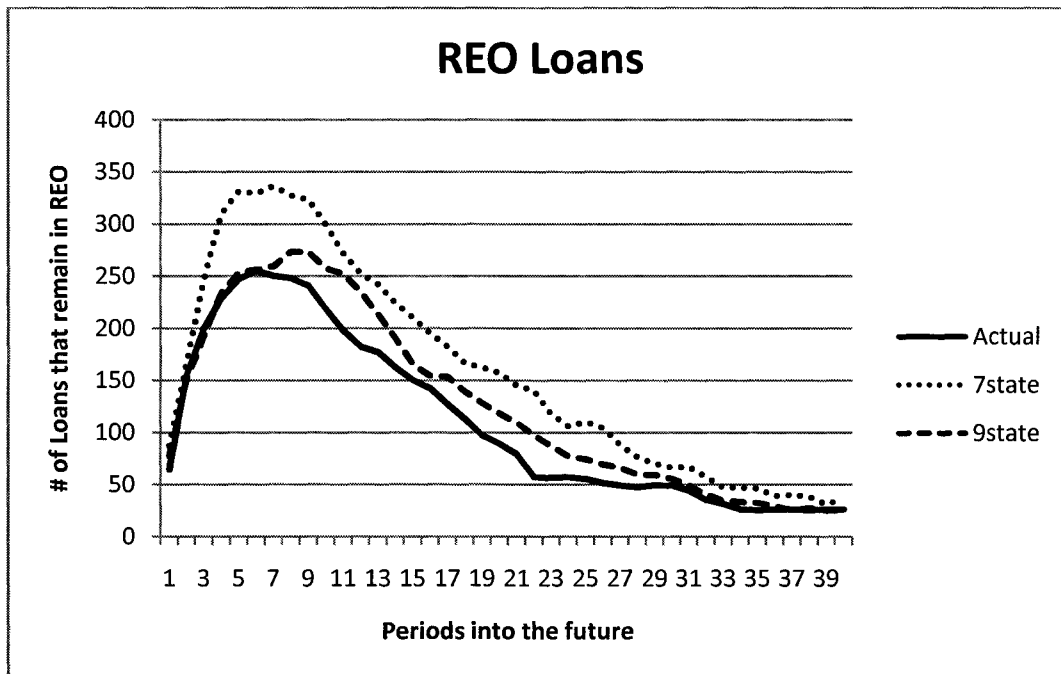


Figure 7 - 7-State and 9-State model predictions compared to the actual data for REO loans

For predicting 'Paid Off' loans over time, the 9-State model performs better on the first 20 periods than the 7-State model (Figure 7). The 9-State model median absolute error is 32 and the 7-State model median absolute error is 66.5. Mann-Whitney U test results indicate that the difference between the median absolute errors is significant ($p < 0.003$). Therefore, the 9-State model is significantly better than the 7-State model in predicting 'Paid off' loans.

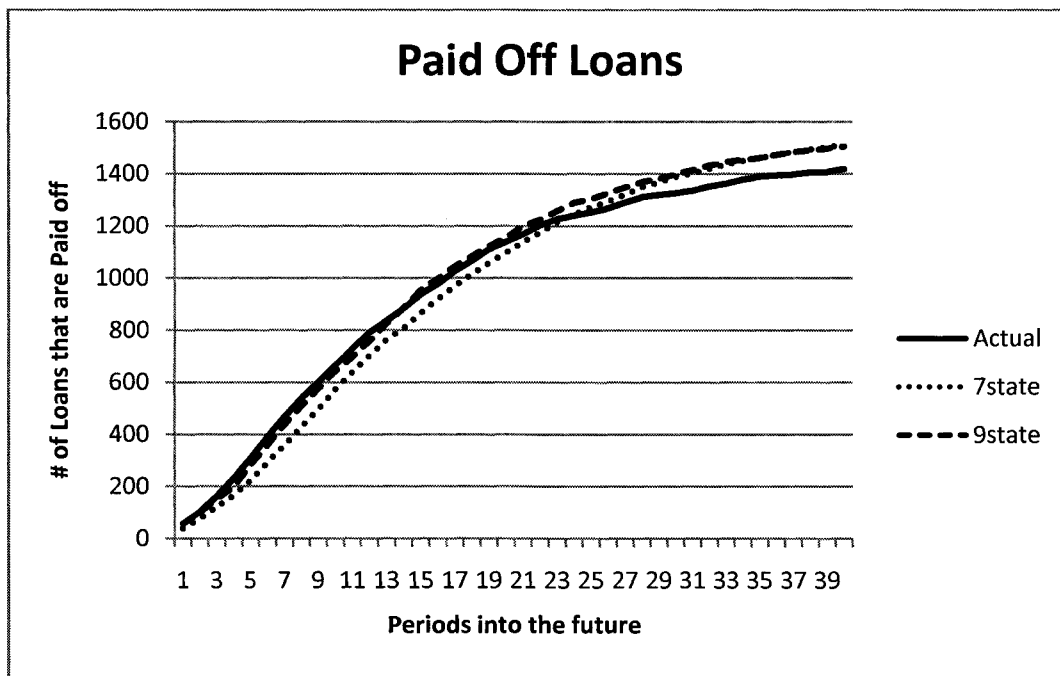


Figure 8 - 7-State and 9-State model predictions compared to the actual data for 'Paid Off' loans

Even though the preceding analysis provides compelling evidence that the 9-State model is superior to the 7-State model, additional tests are performed in the next section to verify the separation of 'Foreclosure' and 'Paid Off' statuses.

Results of Formal Statistical Testing of Adding More States

This subsection presents the results of formally testing the addition of two states to the Markov model. As mentioned in Chapter 3, the 9-State model should have significant Wald Chi-square test results for each individual LR models (for each state transition) to be valid.

Wald Chi-square test is chosen for its strictness compared to the log likelihood statistics and score statistics. It tests the null hypothesis that all the regression coefficients are equal to zero. Not rejecting the null hypothesis associated with any of the transitions introduced by the new statuses will tend to invalidate the usefulness of the additional statuses.

Table 11 shows the Wald Chi-square hypothesis test results for the two new foreclosure states.

Wald Chi-Square Results							
From\ To	Current	30	60	90+	Foreclosure	REO	Paid Off
Early FC	6768.4, df=19, p<.0001	1700.6, df=17, p<.0001	1222.0, df=14, p<.0001	3474.8, df=22, p<.0001	24227.9, df=38, p<.0001	8782.20, df=37, p<.0001	3181.8, df=18, p<.0001
Late FC	3580.8, df=3, p<.0001	1157.1, df=6, p<.0001	565.9, df=4, p<.0001	2065.0, df=11, p<.0001	4006.6, df=40, p<.0001	4675.86, df=17, p<.0001	1993.0, df=15, p<.0001

Table 11 – Testing the validity of splitting foreclosure into ‘Early and Late Foreclosure’ statuses

As shown in Table 11, the p-values for each transition show that all of the null hypotheses are rejected; indicating that splitting foreclosure into two statuses is valid from a statistical perspective.

Table 12 shows the hypothesis test results for the new 'Paid Off' statuses.

Wald Chi Square Results		
From\To	Loss	Paid Off
Current	N/A	1233.2, df=10, p<.0001
30	N/A	9001.2, df=18, p<.0001
60	N/A	4099.0, df=31, p<.0001
90+	956.5, df=12, p<.0001	2320.3, df=21, p<.0001
Early Foreclosure	757.9 df=14, p<.0001	3181.8, df=18 p<.0001
Late Foreclosure	806.2, df=9, p<.0001	1993.0, df=15 p<.0001

Table 12 – Testing the validity of splitting 'Paid off' status into 'Terminated with Loss' and 'Paid in Full' statuses

Similarly, Table 12 displays that splitting 'Paid Off' status into two new statuses, 'Terminated with Loss' and 'Paid in Full', yields statistically valid results. The null hypothesis implied by each transition is rejected with high confidence. This means that splitting "Paid Off" status into two new statuses is statistically valid.

4.2 Interaction Effects and Reconstructability Analysis Results

The hypothesis testing the usefulness of RA is:

- RA can suggest interactions that would significantly improve the performance of the LR.

'30-days delinquent' to 'Paid in Full' transition is chosen to test this hypotheses. Two sample populations are created; 20,000 for training and 20,000 for test. The dependent variable, binary variable indicating whether a loan is 'Paid in Full' or not, is equally distributed in both the training and test samples. Each continuous variable is binned into four equally numbered bins, which are better than two, three and five number of bins, by rank order in both the training and test populations. The binning results are compared based on overall model %correct (training) and %correct (test) but not separately for each variable in a univariate fashion. Stepwise regression is set to run at .05 significance for a variable to enter and stay in the model. Results are shown in Table 13.

The percentage of correct results are 64.26% on training population and 64.28% on the testing population.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi- Square	Pr > ChiSq
Intercept	1	4.9961	0.3776	175.0481	<.0001
app_value	1	1.68E-06	3.95E-07	18.0688	<.0001
fico	1	-0.00201	0.000266	57.1277	<.0001
cltv	1	0.0164	0.00144	129.8234	<.0001
Arm_rate	1	0.1267	0.016	62.8637	<.0001
remaining_upb	1	2.44E-06	8.40E-07	8.4094	0.0037
interest_pmts	1	-0.00096	0.000092	108.7677	<.0001
months_to_ppp	1	0.00192	0.00054	12.5847	0.0004
ever_30	1	0.00699	0.00293	5.681	0.0171
hpa_12_months	1	-3.8618	0.3028	162.6787	<.0001
hpa_o_months	1	-0.7743	0.1	59.9023	<.0001
unemployment_change_9	1	0.064	0.0163	15.5149	<.0001
unemployment_change_orig	1	-0.058	0.0119	23.8032	<.0001
mba_stat_1_c	1	0.2223	0.018	152.4563	<.0001
mba_stat_2_3	1	-0.1864	0.0202	85.3971	<.0001
mba_stat_2_6	1	-0.3166	0.0425	55.399	<.0001
mba_stat_3_c	1	0.4255	0.0631	45.4999	<.0001
mba_stat_3_3	1	0.2788	0.0646	18.6023	<.0001
mba_stat_3_6	1	0.19	0.0737	6.651	0.0099
prop_condo_flag	1	0.1139	0.0442	6.6268	0.01
prop_other_flag	1	-0.2211	0.0439	25.3779	<.0001
occp_owner_flag	1	0.0573	0.0246	5.4283	0.0198
ppp_2y_flag	1	0.1381	0.0514	7.2264	0.0072

Table 13 – LR results

In order to search faster, RA starts with the variables that are identified in the LR results shown in Table 13. One specific model performed better than LR, IV: AiBoZ: AkApZ: AsCaZ: CkZ, with a percent correct value of 64.27% on training and 64.40% on the test data, both slightly better than the LR results. This model consists of only 7 variables, which are cltv (Ai), arm rate (Ak), interest pmts (Ap), ever_30 (As), hpa_o_months (Bo), mba_stat_1_c (Ca), mba_stat_3_c (Ck) compared to the 22 variables used by LR model.

Further RA search using dAIC, dBIC, significance and %correct (test) as the model selection criteria revealed eight “interesting” interaction effects: AiAp, AiBo, AkAp, ApBo, BoCa, ArBo, AsCa, and AsBo. These are the most commonly identified interactions from the best models of each RA run. These interactions are then tested in RA for significance (Table 14) and also tested in LR for significance within the stepwise selection procedure (Table 15).

MODEL	Level	H	dDF	dLR	Alpha	Inf	%dH(DV)	dAIC	dBIC	%C(Data)	%C(Test)
IV: AiApZ	0	14.1474	15	467.35	0.000	0.0271	1.6856	437.3486	318.7962	56.51	56.705
IV: AiZ: ApZ	1	14.1486	6	433.65	0.000	0.0251	1.564	421.6455	374.2246	56.395	56.81
Significance Test			9	33.70	0.000			Interaction AiApZ is significantly better than AiZ: ApZ			
IV: AiBoZ	0	14.1455	15	519.46	0.000	0.0301	1.8735	489.4569	370.9045	56.725	56.78
IV: AiZ: BoZ	1	14.1489	6	426.02	0.000	0.0247	1.5366	414.0226	366.6017	55.94	56.61
Significance Test			9	93.43	0.000			Interaction AiBoZ is significantly better than AiZ: BoZ			
IV: AkApZ	0	14.1448	15	539.59	0.000	0.0312	1.9461	509.5855	391.0331	56.76	57.265
IV: AkZ: ApZ	1	14.1468	6	483.57	0.000	0.028	1.7441	471.5728	424.1519	56.445	57.065
Significance Test			9	56.01	0.000			Interaction AkApZ is significantly better than AkZ: ApZ			
IV: ApBoZ	0	14.139	15	699.82	0.000	0.0405	2.5241	669.8191	551.2668	58.15	58.915
IV: ApZ: BoZ	1	14.1405	6	659.53	0.000	0.0382	2.3787	647.5253	600.1044	57.75	58.535
Significance Test			9	40.29	0.000			Interaction ApBoZ is significantly better than ApZ: BoZ			
IV: BoCaZ	0	14.1231	7	1,140.20	0.000	0.066	4.1124	1126.1993	1070.8748	59.685	60.47
IV: BoZ: CaZ	1	14.1233	4	1,136.43	0.000	0.0658	4.0988	1128.4307	1096.8167	59.685	60.47
Significance Test			3	3.77	0.288			Interaction BoCaZ is not significantly better than BoZ: CaZ			
IV: ArBoZ	0	14.1446	15	544.51	0.000	0.0315	1.9639	514.5143	395.962	56.835	57.135
IV: ArZ: BoZ	1	14.1461	6	502.63	0.000	0.0291	1.8128	490.6281	443.2072	56.835	57.135
Significance Test			9	41.89	0.000			Interaction ArBoZ is significantly better than ArZ: BoZ			
IV: AsCaZ	0	14.1293	7	968.93	0.000	0.0561	3.4947	954.93	899.6056	60.04	60.29
IV: AsZ: CaZ	1	14.1298	4	956.30	0.000	0.0554	3.4491	948.3006	916.6866	59.99	60.235
Significance Test			3	12.63	0.006			Interaction AsCaZ is significantly better than AsZ: CaZ			
IV: AsBoZ	0	14.106	15	1,615.22	0.000	0.0935	5.8257	1585.2238	1466.6715	61.245	61.805
IV: AsZ: BoZ	1	14.1076	6	1,570.19	0.000	0.0909	5.6633	1558.1904	1510.7694	60.84	61.755
Significance Test			9	45.03	0.000			Interaction AsBoZ is significantly better than AsZ: BoZ			

Table 14 – RA interaction tests

The confirmatory RA test results indicate that all interactions suggested except BoCa (hpa_o_months * mba_stat_1_c), are significant interactions. BoCa has been selected as an “interesting” term because it is often identified in the more complex exploratory RA models. The BoCa interaction did turn out to be significant in the LR results displayed in Table 15. This interesting result is discussed further in Chapter 5.

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi Square	Pr > ChiSq	RA Code
Intercept	1	1.1186	0.7395	2.2879	0.1304	
interest_pmts	1	-0.00162	0.000157	105.7106	<.0001	Ap
cltv	1	0.0511	0.00557	84.2365	<.0001	Al
mba_stat_2_c	1	0.1644	0.0202	65.9593	<.0001	Cf
ever_30*mba_stat_1_c	1	-0.0214	0.00277	59.7712	<.0001	AsCa
fico	1	-0.002	0.000282	50.1831	<.0001	Ag
months_to_ppp	1	0.0131	0.00193	45.9248	<.0001	Ar
mba_stat_3_c	1	0.2438	0.0368	43.927	<.0001	Ck
ever_30	1	0.0969	0.015	41.5647	<.0001	As
cltv*hpa_o_months	1	-0.028	0.00435	41.3678	<.0001	AiBo
months_to*hpa_o_months	1	-0.00937	0.00147	40.7376	<.0001	ArBo
ever_30*hpa_o_months	1	-0.0558	0.0107	27.4504	<.0001	AsBo
prop_other_flag	1	-0.2263	0.0441	26.3734	<.0001	E
Arm_rate*interest_pmts	1	0.000108	0.000023	22.8717	<.0001	AkAp
hpa_o_months	1	1.697	0.3711	20.9066	<.0001	Bo
hpa_12_months	1	-6.4925	1.4357	20.4506	<.0001	Bn
app_value	1	1.74E-06	4.04E-07	18.4876	<.0001	L
unemployment_change_orig	1	-0.0485	0.0121	16.1109	<.0001	Bt
doc_full_flag	1	0.1953	0.0499	15.3283	<.0001	R
remaining_upb	1	0.000013	3.67E-06	11.8295	0.0006	An
doc_low_flag	1	0.1761	0.0516	11.66	0.0006	S
unemployment_change_9	1	0.056	0.0164	11.6445	0.0006	Br
mba_stat_2_6	1	-0.1401	0.0416	11.3305	0.0008	Ch
seasoning_by_orig_Da	1	-0.0046	0.00148	9.6839	0.0019	Am
mba_stat_3_3	1	0.1138	0.0367	9.6001	0.0019	Cl
hpa_o_mon*mba_stat_1_c	1	0.2401	0.0781	9.455	0.0021	BoCa
orig_amt	1	-9.37E-06	3.55E-06	6.9632	0.0083	K
prop_condo_flag	1	0.1158	0.0445	6.7834	0.0092	B
ppp_2y_flag	1	0.1318	0.0517	6.5081	0.0107	Ac
doc_nodoc_flag	1	0.1977	0.0776	6.4978	0.0108	T
occp_owner_flag	1	0.0579	0.0247	5.4836	0.0192	G
hpa_9_months	1	3.7027	1.8059	4.2037	0.0403	Bm
Arm_rate	1	0.0493	0.0249	3.9321	0.0474	Ak

Table 15 – LR with interaction terms included

Out of eight interactions determined by RA model, the six highlighted in Table 15 are found to be significant in the best fit LR model with all eight included. These interactions are meaningful from the mortgage business perspective. The LR model with these six interactions achieves 64.68% correct on training population and 64.95% on testing population. These results are better than the LR model without interactions by .42% and .67%, respectively, for the training and test populations.

The % correct values on the test data for LR and LR with interactions tend to support the hypothesis, but further testing is needed to determine the significance of this results.

The Chi-square test considers two parameters: the difference in likelihood ratio and the difference in the degrees of freedom. It also requires that the models of comparison are hierarchically related. The null hypothesis is that the initial LR model without interactions and LR model with interaction effects are not significantly different. Here are the results:

$$H_0: \Delta LR = 0$$

$$\Delta LR = \text{LikelihoodRatio}_{LR_{\text{interaction}}} - \text{LikelihoodRatio}_{LR_{\text{no_interaction}}}$$

$$\Delta LR = 2472.7133 - 2273.2040 = 199.5093$$

$$\Delta DF = DF_{LR_{\text{interaction}}} - DF_{LR_{\text{no_interaction}}}$$

$$\Delta DF = 33 - 22 = 11$$

$$\text{Chi-Square}(199.5093, 11) < .0001$$

The null hypothesis can be clearly rejected, as the LR model with RA-suggested interaction effects is significantly better than the LR model with no interactions.

Chapter 5 –Discussion and Conclusions

This chapter consists of two sub-sections. Section 5.1 and 5.2 summarize the research results, state the contributions, discuss the implications, list the limitations, and identify the possible future research related to the 9-State payment model and interaction effects results, respectively. Section 5.3 summarizes the overall significance of the study.

5.1 9-State Payment Model Discussion

In the first phase of the research, a 9-State mortgage payment transition model is presented and its forecast accuracy is compared to a 7-State model introduced by DeFranco (2002). DeFranco points out that his 7-State model could be further improved by adding and testing more payment statuses. This dissertation introduces new statuses, 'Early Foreclosure' and 'Late Foreclosure' to replace 'Foreclosure.' A loan that is early in the foreclosure process is observed to have higher cure rates compared to the later stages. Thus, the foreclosure status is split in order to incorporate this observation and improve predictive accuracy. It also introduces 'Terminated with Loss' and 'Paid in Full' statuses to replace 'Paid Off' status. The increasing loss exposure risk on mortgage portfolios requires the separation of 'Paid Off' status in order to enable the model to distinguish loss loans from the fully paid off loans.

In order to make the comparison fair, the models are estimated using the same estimation methodology (logistic regression) and the same training and test data. The 9-State model is significantly better in predicting five statuses, 'Current', '30-days delinquent', '60- days delinquent', 'REO' and 'Paid off'. The 7-State model and 9-State model are statistically indifferent when predicting '90- days delinquent' and 'Foreclosure' statuses.

In addition to these status prediction tests, the “restrictions” of the 7-State model compared to the 9-State model also are formally tested. Looking at the Table 16 below, based on DeFranco (2002), “restriction” of 3-State model over the 7-State model is the use of 'Active' status, instead of the five separate statuses 'Current', '30', '60', '90+', 'Foreclosure'. In order to justify these additional states, the “restriction” should be rejected. In other words, addition of more states should be statistically tested using Wald Chi-square and the null hypotheses that each new logistic regressions coefficient is not different from zero should be rejected. In the 7-State model, there are two “restrictions” implied over the 9-State model, 'foreclosure' and 'Paid Off' statuses.

Loan Payment Statuses by Model									
3 - State Model	Active					Default		Paid Off	
7 - State Model	Current	30	60	90+	FC		REO	Paid Off	
9 - State Model	Current	30	60	90+	Early FC	Late FC	REO	Terminated with Loss	Paid in Full

Table 16 – Model Comparison Framework

This dissertation tests the removal of two restrictions: 1) having a single 'Foreclosure' status instead of 'Early and Late Foreclosure', and 2) having a single 'Paid Off' status instead of 'Terminated with Loss' and 'Paid in Full' statuses. A Wald Chi-square test is used to formally validate the increase in the number of statuses to nine.

The main academic contribution of the 9-State model is the resulting improvement in accuracy. It provides better forecasts for delinquency, default, loss and prepayment.

Another important academic contribution is the introduction of 'Terminated with Loss' status. Even though loss projections are not presented in the results section, the model framework enables the user to predict the frequency of loss.

The 9-State model also contributes to the current business environment. It is used in Wilshire Credit Corporation¹⁵ to predict future statuses of mortgage loans. This information then is used to anticipate near future resource needs (employees to handle loan processing needs). This model also is utilized to price servicing rights of a portfolio, since the price of servicing rights directly depends on the status of the various loans in the portfolio. A more crucial potential application is pricing of wholly-owned mortgage loans which is a multi- billion dollar business that features high yields when assets are correctly priced.

¹⁵ Wilshire Credit Corporation is a wholly-owned subsidiary of Bank of America.

Improved accuracy and having an additional loss status has four important implications:

1. The model becomes the state-of-the-art in the default, prepayment and loss literature. It is stated by DeFranco (2002) as possible future work to add new statuses and to test the validity of these new statuses. This is a valuable academic implication.
2. The model enables mortgage investors, mortgage servicers and rating companies to better predict delinquencies, defaults, prepayments and losses for pricing the mortgages, pricing the servicing rights, and assigning bond ratings as well as generating accurate cash flows into the future for better management of mortgage related assets. This is a valuable business implication.
3. This model, also, is useful to identify the recidivism rate of loan modifications¹⁶ currently supported by government under different economic stability programs.¹⁷ The model can better predict the recidivism rate of such borrowers in order to optimize an exit strategy for the loan. This is an important economic implication.

¹⁶ Modification means change of term, interest rate or unpaid balance to help the borrower make future payments.

¹⁷ FHA program for refinancing, ASF Fast-track program for modification, TARP program for repurchasing troubled assets etc.

4. This model can be used to rank servicers in a more accurate fashion. Business models rank servicers generating expected results for a portfolio of loans on a historical basis. Then the expected results get compared to the actual results. So, servicers can be ranked based on how well they perform against the expected results. Again, an important business implication.

For future predictability, the model needs to have an accurate house price index, unemployment and interest rate projections. These three variables are exogenous to the model. This is a limitation because the 9-State model accuracy is limited to the accuracy of the projections of these economic variables.

The 9-State model uses borrowers credit scores based on the origination date of the loan. This is a limitation since credit can change through time: yet it is a very important indicator of determining the likelihood of prepaying or defaulting based on Ambrose and Buttimer (2000). This limitation can be resolved only when credit score data is available monthly.

Another important limitation to the 9-State model is the changes in the government and/or business regulations. For example, a recently initiated government program named TARP¹⁸ can drastically lower the accuracy of the predictions of the model. There are also regulations from government induced business alliances such as HOPE

¹⁸ Troubled Asset Relief Program launched by government in last quarter of 2008.

NOW Alliance and ASF¹⁹ Fast Track Modification Program. The most recent one established is the Homeownership Affordability and Stability Plan by the government as of February 17th, 2009. Such programs can initiate an easy way of refinancing, or a foreclosure moratorium which alters the basic dynamics of the payment model fundamentally. Even a slight disturbance in any of the transitions within the payment model will yield inaccurate projections of the future.

There are three specific future research related areas to the 9-State model. The first is to complement the loss *frequency* predictions with models that predict the severity of loss, in order to determine the overall expected loss on a portfolio.

A second future research project could involve the separation of 'REO', and '90-days delinquent' into more statuses. One must ask, however, 'How many more statuses can be involved?' A statistician might answer 'As long as the Wald Chi-square tests are significant, more statuses can be usefully included in the model'. A practical business person might answer 'As long as the forecast accuracy increases and the model remains intuitively interpretable.' There is no statistical test or a goodness of fit measure established to answer the question of parsimony versus accuracy. Since models similar to the 9-State model are actively used in the business world, accuracy seems to be valued over parsimony in such business environments.

¹⁹ American Securitization Forum.

A third potential future research topic would be to incorporate more economic and geographic variables into the logistic regressions of the payment model. Such variables might be, but not limited to, home purchase supply and demand within a given geographic area and regional house affordability. Adding such variables may increase accuracy.

5.2 Interaction Effects and Reconstructability Analysis Discussion

The second phase of the research presents the use of reconstructability analysis as a methodology to detect interaction effects. Interaction effects in mortgage prepayment, default and loss literature are studied and emphasized by Hendershott and Schultz (1993), Ambrose, Capone, and Deng (2001), Lacour-Little, Marschoun and Maxam (2002). In this research, RA-suggests interaction effects are hypothesized to enhance the predictability of logistic regressions.

RA is a method for analyzing multivariate categorical data (Zwick and Johnson 2004) and so is LR. However, LR has the advantage of being able to process continuous data whereas RA can more effectively utilize interaction terms and non-linear relationships. Combinations of these methodologies are expected to yield better results. Carletti (2004) detects interaction effects using RA and utilized this knowledge to enhance the regression analysis conducted in his research. The present research extends this methodology to enhance the logistic regression models in the mortgage payment model framework.

The transition from '30-days delinquent' to 'Paid in Full' is selected to investigate the above hypotheses because it is both useful and difficult to predict. All training and test data are the same for each methodology.

First, the LR model is trained with no interactions. The resulting model achieves a 64.28% correct result on the test data using 22 independent variables. Eight interactions are derived from multiple RA runs as the most “interesting” interaction effects based on different overall model selection criteria such as dBIC, dAIC and % correct on test data. A logistic regression model that includes the eight RA-suggested interactions then is trained and tested. This model scores 64.95% correct on the test data using 32 variables including the six of the eight interactions suggested by RA.

A confirmatory interaction significance test is also conducted using RA for each of the eight suggested interactions; the results indicates that seven of the eight interactions are significant except for BoCa ($hpa_o_months * mba_stat_1_c$). Interestingly, this interaction is significant in the LR model with interactions.

The term BoCa represents the interaction effect of variables, house price appreciation from the loan origination and a binary variable that indicates whether the loan is current last months. It is coded as $hpa_o_months * mba_stat_1_c$ in the logistic regression model. Even though, the confirmatory test in RA indicates no significant effect on its own, it is significant in the logistic regression model that incorporates RA-suggested interaction effects. This suggested that the interaction term might be

significant in the absence of either Bo (hpa_o_months) or Ca (mba_stat_1_c) which can be justified by looking at Tables 13 and Tables 15, respectively. Ca (mba_stat_1_c) is present in LR without interaction, but it is absent in the LR with interactions.

The hypothesis is tested using Chi-square tests to identify significant differences in the likelihood ratio of compared models. The LR model with RA-suggested interactions outperformed the LR model with no interactions.

Overall, the hypothesis tests and accuracy comparisons indicate that RA can suggest interactions that significantly will improve the performance of logistic regressions of the payment model. Even though this improvement might be small in terms of percentage increase in accuracy (.67%), it might imply a significant amount of savings. For example, out of 10,000 mortgage loans which are 30-days delinquent, if the model predicts with a .67% increase in accuracy, it means that 67 more loans will be predicted accurately. Assuming an average balance of a mortgage loan is around \$250,000, the accuracy will apply to a total balance of approximately \$16,750,000. From an investor perspective, more accuracy on such a balance is important.

The RA and the interaction terms research leads to two important discussions that are:

- Missing independent variable states in the test data that don't get scored
- Computational limitations of OCCAM and SAS

As the complexity increases in OCCAM, more variables or interaction terms get introduced increasing the predictability. However, the difference between %correct (training) and %correct (test) increases dramatically as the complexity increases. This is primarily due to the observations in the test data that OCCAM is not able to generate their probabilities. This is because the RA method only can generate probabilities for combinations of independent variable states that exist in the training data. When such combinations do not exist, OCCAM assigns the probabilities of the independence model.

This issue is called over-fitting. An over-fit model is too specialized for the data it is trained with such that it performs poorly against a random out-of-sample population. Consequently, as OCCAM introduces more variables, the RA model becomes over-fit; in other words, performs poorly on the out-of-sample population.

One way to resolve this issue will be to enable OCCAM to intelligently remove one variable at a time in order to obtain simpler models, and thereby generate probabilities for these observations. This may significantly increase the predictive power of RA-models identified by OCCAM. This is an important future work.

Another important consideration is the computational time and the memory resources required by both OCCAM (web-based RA tool) and SAS (the statistical software used) to evaluate RA and LR models, respectively. Even though LR took less than a minute to find the optimal model with no interaction effects, it is impossible to

conduct a search for all possible models that include interaction effects of two or more variables (factorial design) in SAS due to memory issues. The OCCAM algorithm, however, is able to start searching for models (in a way that can potentially include all possible models) without an initial set up. It faces computational difficulty regarding both the computation time and the memory space because the search space grows exponentially with the increasing model complexity in terms of the number of variables and the number of interactions. Some OCCAM runs stopped with no error notification, and some were aborted due to low memory. Given a starting list of variables (from LR), OCCAM is able to do exploratory searches that include interactions of two or more variables, and is useful in identifying important interactions effects. SAS is not able to set up the factorial design to do such an exploratory search.

As mentioned earlier, OCCAM can generate high level models with interactions, but as the number of variables increase and model complexity increases, the number of computations needed exceeds OCCAM's capacities. This can be addressed partially through hardware (e.g. parallel computing or cutting edge systems) and partially through software improvements (better implementations). This also is an important area for future work.

5.3 Conclusions

This research tests a 9-State mortgage payment model in comparison to the 7-State model presented in the literature. The 9-State model adds 'Early Foreclosure' and 'Late Foreclosure' statuses replacing the single 'Foreclosure' status, and 'Terminated with Loss' and 'Paid in Full' statuses replacing the single 'Paid Off' status. The addition of these states is tested formally using Wald Chi-square tests. The tests validate the addition of these states from a statistical perspective. The results of 9-State model also are empirically tested against the 7-State model using a large sample of loans that are initially in foreclosure. These results indicate that 9-State model has significantly improves accuracy over 7-State model on five of the seven loan payment statuses over a period of 40 months into the future.

This research also demonstrates that reconstructability analysis is helpful in detecting and suggesting interactions effects that can be used in the logistic regression models within the payment model framework. It suggests six new interaction effects to the logistic regression model, regarding the transition from '30-days delinquent' to 'Paid in Full'. It improves the accuracy (measured in percent correct) by 0.67%. This increase is significant at the model level based on the Chi-square test conducted on the difference of likelihood ratios and difference of degrees of freedom.

References

- Aitchison, J., Brown, J. A. C. (1957). The Lognormal Distribution: With Special Reference to Its Uses in Economics. Cambridge, Cambridge University Press.
- Akesson, F., Lehoczky, J. P. (2000). "Path Generation for Quasi-Monte Carlo Simulation of Mortgage-Backed Securities." Management Science **46**(9): 1171-1187.
- Alexander, W. P., Grimshaw, S. D., McQueen, G. R., Slade, B. A. (2002). "Some Loans Are More Equal than Others: Third Party Originations and Defaults in the Subprime Mortgage Industry." Real Estate Economics **30**(4): 667-697.
- Ambrose, B. W., Capone, C. A. (1996). "Modeling the Conditional Probability of Foreclosure in the Context of Single-Family Mortgage Default Resolution." Real Estate Economics **26**(3): 391-429.
- Ambrose, B. W., Capone, C. A. (2000). "The Hazard Rates of First and Second Defaults." Journal of Real Estate Finance and Economics **20**(3): 275-293.
- Ambrose, B. W., Capone, C. A., Deng Y. (2001). "Optimal Put Exercise: An Empirical Examination of Conditions for Mortgage Foreclosure." Journal of Real Estate Finance and Economics **23**(2): 213-234.
- Ambrose, B. W., LaCour-Little, M. (2001). "Prepayment Risk in Adjustable Rate Mortgages Subject to Initial Year Discounts: Some New Evidence." Real Estate Economics **29**(2): 305-327.
- Ambrose, B. W., Buttimer, R. J. (2000). "Embedded Options in the Mortgage Contract." Journal of Real Estate Finance and Economics **21**(2): 95-111.
- Ambrose, B. W., Buttimer, R. J., Capone, C. A. (1997). "Pricing Mortgage Default and Foreclosure Delay." Journal of Money, Credit, and Banking **29**(3): 314-325.
- Archer, W. R., Ling, D. C. (1993). "Pricing Mortgage-Backed Securities: Integrating Optimal Call and Empirical Models of Prepayment." Journal of the American Real Estate and Urban Economics Association **21**(4): 373-404.
- Berkson, J. (1980). "Minimum Chi-Square, Not Maximum Likelihood!" Annals of Mathematical Statistics **8**: 457-487.
- Black, F., Scholes, M. (1973). "The Pricing of Options and Corporate Liabilities." The Journal of Political Economy **81**(3): 637-654.

Calhoun, C.A., Deng, Y. (2002). "A Dynamic Analysis of Fixed- And Adjustable-Rate Mortgage Terminations." Journal of Real Estate Finance and Economics **24**(1/2): 9-33.

Capozza, D. R., Kazarian, D., Thomson, T. A (1998). "The Conditional Probability of Mortgage Default." Real Estate Economics **26**(3): 359-389.

Capozza, D. R., Kazarian, D., Thomson, T. A. (1997). "Mortgage Default in Local Markets." Real Estate Economics **25**(4): 631-655.

Capozza, D. R., Thomson, T. A (2005). "Optimal Stopping and Losses on Subprime Mortgages." The Journal of Real Estate Finance and Economics **30**(2): 115-131.

Carletti, R. (2004). A Study of the Relationship between Education, Literacy, and Health. Portland, Portland State University. **Ph.D.:** 118.

Chinloy, P. (1993). "Elective Mortgage Prepayment: Termination and Curtailment." Journal of the American Real Estate and Urban Economics Association **21**(3): 313-332.

Clauretje, T. M. (1990). "A Note on Mortgage Risk: Default vs. Loss Rates." AREUEA Journal **18**(2): 202-206.

Cooperstein, R. L., Redburn, F. S., Meyers, H. G. (1991). "Modeling Mortgage Terminations in Turbulent Times." AREUEA Journal **19**(4): 473-494.

Crawford, G. W., Rosenblatt, E. (1995). "Efficient Mortgage Default Option Exercise: Evidence from Loss Severity." The Journal of Real Estate Research **10**(5): 543-555.

Cunningham, D.F., Capone, C. A. (1990). "The Relative Termination Experience of Adjustable to Fixed-Rate Mortgages." The Journal of Finance **16**(5): 1687-1703.

DeFranco, R. G. (2002). Modeling Residential Mortgage Termination and Severity using Loan Level Data. Economics. Berkeley, University of California. **Ph.D.:** 110.

Deng, Y., Quigley, J. M., Van Order, R (2000). "Mortgage Terminations, Heterogeneity and the Exercise of Mortgage options." Econometrica **68**(2): 275-307.

- Deng, Y., Quigley, J. M., Van Order, R., Mac, F. (1996). "Mortgage default and low down payment loans: The costs of public subsidy." Regional Science and Urban Economics **26**: 263-285.
- Diebold, F. X., Mariano, R. S. (1995). "Comparing Predictive Accuracy." Journal of Business & Economic Statistics **13**(3): 253-263.
- Eckhardt, R. (1987). "Stan Ulam, John von Neumann, and the Monte Carlo method." Los Alamos Science **Special Issue**(15): 131-137.
- Epperson, J. F., Kau, J. B., Keenan, D. C., Muller, W. J. (1985). "Pricing Default Risk in Mortgages." AREUEA Journal **13**(3): 261-272.
- Fabozzi, F. (1997). The Handbook of Fixed Income Securities, McGraw-Hill.
- Fabozzi, F., Modigliani, F. (1992). Mortgage & Mortgage Backed Securities Markets. Boston, Massachusetts, Harvard Business School Press.
- Hakim, S. R. (1997). "Autonomous and Financial Mortgage Prepayment." Journal of Real Estate Research **13**(1): 1-16.
- Hall, A. R. (1985). "Valuing the Mortgage Borrower's Prepayment Option." AREUEA Journal **13**(3): 229-247.
- Hayre, L. (2001). Guide to Mortgage-Backed and Asset-Backed Securities. , John Wiley & Sons, Inc.
- Hendershott, P. H., Schultz, W. R. (1993). "Equity and Non-equity Determinants of FHA Single-Family Mortgage Foreclosures in the 1980s." Journal of the American Real Estate and Urban Economics Association **21**(4): 405-430.
- Hosmer, D. W., Lemeshow, S. (1989). Applied Logistic Regression. New York, Wiley.
- Hosseini, J. C., Harmon, R. R., and Zwick, M. (1991). "An Information Theoretic Framework for Exploratory Multivariate Market Segmentation Research." Decision Sciences **22**: 663-677.
- Jones, L. D. (1993). "Deficiency Judgments and the Exercise of the Default Option in Home Mortgage Loans." Journal of Law and Economics **36**(1): 115-138.
- Kang, P., Zenios, S. A. (1992). "Complete Prepayment Models for Mortgage-Backed Securities." Management Science **38**(11): 1665-1685.

Kau, J. B., Keenan, D. C. (1999). "Patterns of Rational Default." Regional Science and Urban Economics **29**: 765-785.

Kau, J. B., Keenan, D. C., Kim, T. (1993). "Transaction Costs, Suboptimal Termination and Default Probabilities." Journal of the American Real Estate and Urban Economics Association **21**(3): 247-263.

Kau, J. B., Kim, T. (1994). "Waiting to Default: The Value of Delay." Journal of the American Real Estate and Urban Economics Association **22**(3): 539-551.

Klir, G. (1986). "Reconstructability Analysis: An Offspring of Ashby's Constraint Theory." Systems Research **3**(4): 267-271.

Knoke, D., Burke, P. J. (1980). Log Linear Models. Quantitative Applications in the Social Sciences. Sage, Beverly Hills. **20**.

Krippendorff, K. (1986). Information Theory: Structural Models for Qualitative Data. Quantitative Applications in the Social Sciences. Sage, Beverly Hills. **62**.

LaCour-Little, M. (2004). "Equity Dilution: An Alternative Perspective on Mortgage Default." Real Estate Economics **32**(3): 359-384.

LaCour-Little, M., Chun, G.H. (1999). "Third Party Originators and Mortgage Prepayment Risk: An Agency Problem?" Journal of Real Estate Research **17**(1/2): 55-70.

LaCour-Little, M., Marschoun, M. A., Maxam, C. L., Stein, H. (2002). "Improving Parametric Mortgage Prepayment Models with Non-Parametric Kernel Regression." Journal of Real Estate Review **24**(3): 299-327.

Lekkas, V., Quigley, J. M., Van Order, R. (1993). "Loan Loss Severity and Optimal Mortgage Default." Journal of the American Real Estate and Urban Economics Association **21**(4): 353-371.

LeRoy, S. F. (1996). "Mortgage Valuation Under Optimal Prepayment." The Review of Financial Studies **9**(3): 817-844.

McConnell, J. J., Singh, M. (1993). "Valuation and Analysis of Collateralized Mortgage Obligations." Management Science **39**(6): 692-709.

Metropolis, N., Ulam, S. (1949). "The Monte Carlo method." Journal of the American Statistical Association **44**(247): 335-341.

- Mist, S. (2007). Prediction of Traditional Chinese Medicine Diagnosis from Psychosocial Questionnaires. Systems Science. Portland, Portland State University. **Ph.D.:** 219.
- Morton, T. G. (1975). "A Discriminant Function Analysis of Residential Mortgage Delinquency and Foreclosure." Real Estate Economics **3**(1): 73-88.
- Phillips, R. A., VanderHoff, J. H. (2004). "The Conditional Probability of Foreclosure: An Empirical Analysis of Conventional Mortgage Loan Defaults." Real Estate Economics **32**(4): 571-587.
- Schwartz, E. S., Torous, W. N (1992). "Prepayment, Default, and the Valuation of Mortgage Pass-through Securities." Journal of Business **65**(2): 221-239.
- Schwartz, E. S., Torous, W. N. (1989). "Prepayment and the Valuation of Mortgage-Backed Securities." The Journal of Finance **19**(2): 375-392.
- Schwartz, E. S., Torous, W. N. (1993). "Mortgage Prepayment and Default Decisions: A Poisson Regression Approach." Journal of the American Real Estate and Urban Economics Association **21**(4): 431-449.
- Shervais, S., Zwick, M., and Kramer, P. (2005). "Reconstructability Analysis As A Tool For Identifying Gene-Gene Interactions In Studies Of Human Diseases." Presented at IEEE Systems, Man, and Cybernetics meeting **Oct. 10-12**.
- Stanton, R. (1995). "Rational Prepayment and the Valuation of Mortgage-Backed Securities." The Review of Financial Studies **8**(3): 677-708.
- Stokes, A., Armstrong, L. (2006). Some Delinquency Measures Tick Upwards in Latest MBA National Delinquency Survey. Washington D.C., Mortgage Bankers Association.
- Tabachnick, B. G., L. S. Fidell (1996). Using multivariate statistics. New York, Harper Collins.
- Van Order R., Zorn, P. M. (2000). "Income, Location and Default: Some Implications for Community Lending." Real Estate Economics **28**(3): 385-404.
- Vandell, K. D. (1995). "How Ruthless Is Mortgage Default? A Review and Synthesis of the Evidence." Journal of Housing Research **6**(2): 245-264.
- Vandell, K. D., Thibodeau, T. (1985). "Estimation of Mortgage Defaults Using Disaggregate Loan History Data." AREUEA Journal **13**(3): 292-316.

VanderHoff, J. (1996). "Adjustable and Fixed Rate Mortgage Termination, Option Values and Local Market Conditions: An Empirical Analysis." Real Estate Economics **24**(3): 379-406.

Wright, A., Ricciardi, T., and Zwick, M. (2005). Application of Information-Theoretic Data Mining Techniques in a National Ambulatory Practice Outcomes Research Network. Presented at the American Medical Informatics Association annual symposium, Washington DC.

Yang, T. T., Buist, H., Megbolugbe, I. F. (1998). "An Analysis of the Ex Ante Probabilities of Mortgage Prepayment and Default." Real Estate Economics **24**(4): 671-678.

Zwick M., Johnson. M. S. (2004). "State-based reconstructability analysis." Kybernetes **33**(5-6): 1041-1052.

Zwick M., Shu. H. (1996). "Set-Theoretic Reconstructability of Elementary Cellular Automata." Advances in Systems Science and Applications **Special Issue 1**, 31-36.

Zwick, M. (2001). Wholes and Parts in General Systems Methodology. The Character Concept in Evolutionary Biology. G. Wagner, Academic Press.

Zwick, M. (2004). "An Overview of Reconstructability Analysis." Kybernetes **33**(5/6): 877-905.

Appendix A – Variables

prop_sfr_flag	seasoning_by_orig_Date	hpa_3_months
prop_condo_flag	pmts_owed	hpa_6_months
prop_2_4_units_flag	pmts_made	hpa_9_months
prop_pud_flag	remaining_upb	hpa_12_months
prop_other_flag	principal_pmts	hpa_o_months
prop_noinfo_flag	interest_pmts	unemployment_change_3
occp_owner_flag	unemployment	unemployment_change_6
occp_second_home_flag	months_to_ppp	unemployment_change_9
occp_investor_flag	ever_30	unemployment_change_12
occp_other_flag	ever_60	unemployment_change_orig
orig_amt	ever_90	mba_stat_1_c
app_value	state	mba_stat_1_3
term	judicial_state	mba_stat_1_6
init_rate	fc_time	mba_stat_1_9
purp_purchase_flag	fc_cost	mba_stat_1_F
purp_cashrefi_flag	bk_time	mba_stat_2_c
purp_nocashrefi_flag	bk_cost	mba_stat_2_3
purp_other_flag	delinq_tax_rate	mba_stat_2_6
doc_full_flag	fc_trans_tax_perc	mba_stat_2_9
doc_low_flag	reo_time	mba_stat_2_F
doc_nodoc_flag	reo_variable_expense	mba_stat_3_c
ppp_flag	reo_eviction_cost	mba_stat_3_3
ppp_flag_missing	reo_trashout	mba_stat_3_6
ppp_2y_flag	reo_fixed_closing_cost	mba_stat_3_9
ppp_3y_flag	reo_broker_fee	mba_stat_3_F
ppp_5y_flag	reo_trans_tax_perc	int_rate_chg_3
ppp_1y_flag	arm_interest_rate_difference	int_rate_chg_6
fico	fixed_interest_rate_difference	int_rate_chg_9
ltv	prob_neg_equity	int_rate_chg_12
cltv	term_120_flag	int_rate_chg_3_arm
Fixed_rate	term_180_flag	int_rate_chg_6_arm
fixed_fees	term_240_flag	int_rate_chg_9_arm
Arm_rate	term_300_flag	int_rate_chg_12_arm
Arm_fees_points	term_360_flag	months_in_foreclosure
Arm_margin	term_480_flag	months_in_reo