Research Report for Importance of Mortgage Downpayment as a Deterrent to Delinquency and Default as Observed in Black Knight (McDash) Servicing History

U.S. Department of Housing and Urban Development

April 2017

U.S. Department of Housing and Urban Development | Office of Policy Development and Research

Research Report for Importance of Mortgage Downpayment as a Deterrent to Delinquency and Default as Observed in Black Knight (McDash) Servicing History

U.S. Department of Housing and Urban Development

April 2017

Disclaimer

The contents of this report represent the views of the authors and do not necessarily reflect the views or policies of the U.S. Department of Housing and Urban Development or the U.S. government.

Preface

The cash downpayment serves the important function of mitigating mortgage credit risk. It also, however, represents a significant barrier to homeownership for many low- and moderate-income households. Although potential substitutes for various functions of the cash downpayment exist, such as mortgage insurance as a protection against lender loss, we still have much to learn about their effectiveness and how much cash downpayment is needed to mitigate various sources of credit risk.

This study aspires to advance our learning by using loan-level mortgage origination and performance data to examine the effectiveness of the cash downpayment as a deterrent to delinquency and also to serious default in terms of equivalent compensating credit enhancements. Consistent with previous empirical examination, the study finds that cash downpayments do mitigate risk; all else equal, lower proportional downpayments were associated with higher rates of delinquency and default. Compensating tradeoffs can provide an equivalent level of credit risk mitigation, however. The report quantifies the tradeoff between lower mortgage downpayments and other loan-level characteristics, such as higher credit score or lower debt-to-income ratio, that will produce an equivalent level of default risk and also how loan-level tradeoffs may differ across markets, depending on home price appreciation or unemployment trends.

This study underscores the importance of the cash downpayment and the potential for balancing different mortgage downpayment requirements with other relevant compensating factors to manage mortgage credit risk. It also makes an important contribution to the metrics for evaluating alternatives to current requirements that balance risk with broadening access to mortgage markets.

Contents

Introduction

The U.S. Department of Housing and Urban Development (HUD) engaged the Integrated Financial Engineering, Inc. (IFE) group to conduct an independent research project based on the Black Knight Financial Technology Solutions, LLC McDash™ loan data to determine the effectiveness of mortgage downpayments as a deterrent to delinquency and default. McDash data is a loan-level mortgage origination and performance data source maintained by Black Knight and collected from most major mortgage servicers. The project's objectives were to (1) form an analytical database containing performance and economic variables that facilitates empirical studies and (2) use econometric models to quantify the impact of mortgage downpayment on mortgage delinquency and default among different market segments and economic conditions.

Most existing literature applies the competing risk model in analyzing loan-level mortgage performance, which treats prepayments as endogenous and simultaneous with the default decision. Little has been published, however, on the use of the scorecard-type estimation approach to investigate the credit quality of individual mortgage loans. One reason scorecards are not in the public literature is because they are applied on a proprietary and confidential basis as a tool for competitive advantage by major mortgage banks and mortgage insurers and guarantors. Consider the well-known and popular FICO score, which is estimated with the scorecard approach. Even this much-used scorecard does not have public literature that reveals its structure and estimated coefficients. The point is that scorecard estimation of the type herein has a time-honored place in mortgage econometrics, just not a public one.

The Literature Review section focuses on empirical estimates of the effect of downpayments on delinquency and default. The Methodology and Model section presents the scorecard-type model and its dependent and independent variables. The Estimation Database section describes the methodology we applied to derive the estimation database, especially in dealing with outliers and duplicate loan observations. The McDash database does not identify proprietary data, such as borrowers or the address of the property. We devised an algorithm to identify multiple loans that may have been originated to support one specific housing transaction. The Empirical Results section presents the model results. We also computed compensating factors that show how much the combined loan-to-value must change for a unit change in other explanatory variables in order to hold the probability of delinquency or default constant. Our research implications and conclusions are presented in Implications and Conclusions section. Following the References section, various appendices describe in more detail the McDash data set, the steps taken in compiling the research sample, and the technique of our analytical models.

Literature Review

The literature reveals that extensive exploration has been dedicated to the factors that drive mortgage defaults. Default technically is the borrower's failure to abide by the mortgage contract provisions. The most common trigger of default is the failure to make required mortgage payments, and it is this aspect on which we focus this study. The term default is too vague, as it encompasses temporary or permanent delinquencies. As such, we need to be careful when comparing the empirical results of various studies, distinguishing initial delinquencies from foreclosures and settlement; and for foreclosures, distinguishing the start from the completion of foreclosure proceedings. The completion of foreclosure involves the "ultimate default," which is the transfer of the property to the lienholder or to third parties, and, beyond this, to the disposition of the property and the final accounting of losses.

The ample list of the sources cited in the References section at the end of this report describes the motivations for all default phases as the willingness-to-pay or the ability-to-pay factors. As a put option for mortgage buyers, rational defaults can occur when the market value of the collateral is lower than what is owed to the lender (Kau and Keenan, 1995; Vanderhoff, 1996). That is, rational borrowers become unwilling to continue paying when what they owe is more than the value of the purchase price, by an amount that compensates for negative consequences, like credit damage.

Adverse life events, such as loss of income or wealth, are cited as major factors for the inability to continue paying a mortgage. The upfront underwriting criteria and analysis, with indicators such as payment-to-income ratios, attempt to indicate default events based on a borrower's ability to pay (Arnone, Darbar, and Gambini, 2007), although underwriting indicators of how a borrower managed past obligations, measured by credit scores, may predict willingness to pay.

Classifying motivating factors into two categories is not very useful for our purposes, because a high original loan-to-value (LTV) ratio foreshadows a higher likelihood of the future house price to become lower than the mortgage debt, but it can reflect a lack of wealth, which, when given adverse life events, implies that the ability to pay is more likely to be a motivating factor for default.

One major factor of mortgage defaults is the adequacy of the collateral that backs the mortgage loan (Archer, Ling, and McGill, 1996). The major metric describing collateral risk is the LTV ratio. Most studies show that, as borrowers put up higher downpayments, the probability of default risk is lower.

Extensive research has been conducted regarding the effect of the relative amount of the downpayment, typically measured by the original LTV ratio, on mortgage performance. In this Literature Review section, we summarize the empirical relationships found in various studies. Because the focus is on the quantification of the deterrent effect of downpayments on delinquency and default, our literature review includes only those references that provide such empirical quantification, with the intention to compare those estimates with the McDash data. Different empirical approaches typically produce different empirical results, so we point out and classify the studies according to major differences we observed. For the following major differences, we accordingly provide estimates of the coefficient of LTV or combined loan-to-value ratios in this study—and basic information from each of the studies in table 1.

1. **Definition of default.** This study uses two measures of default, 90 days delinquent in the first 2 years after origination and the initiation of foreclosure proceedings during the loan's observed lifetime. We organize the references in table 1 according to how closely each reference matches these two measures of default. The table is arranged into two sections: the first is according to the delinquency measure; and the second is according to the foreclosure event, with our study being the first in each list. The closer the definition of default is to ours, the higher up the reference is located on the list. We list some references in this table more than once, because those have estimates using several definitions of default.

2. **Type of model and estimation technique.** Most models are competing risk models. We found few articles using the mortgage scorecard approach, so we cite several books on credit-scoring techniques: Siddiqi (2012); Thomas (2009); and Thomas, Edelman, and Crook (2002); however, we cite many articles using the competing risk model. Competing risk models estimate the relationship in the put-default option by simultaneously accounting for the call-prepayment option. Scorecard-type models estimate the default relationship by itself, with or without factors that are observed after loan origination. Our model is the latter one. Estimation methods include logistic regression; different estimators typically produce different estimates of the effects of downpayment on default. The competing risk approach deals with prepayments by estimating prepayment equations simultaneously with the default equations. The next columns list the modeling and estimation techniques and how prepayments are treated. We attempt to subgroup references into competing risk and scorecard models. For competing risk models, we show the LTV coefficient in the respective equation and ignore the simultaneous nature of each respective equation with the others.

3. **Other factors held constant.** If one factor affecting the default propensity is not included in a study, it is likely to affect the empirical results of the downpayment-default relationship, a situation known as the omitted relevant variable problem in the econometric literature. Most of the literature cited in the text includes the current LTV as an explanatory variable. Because our focus is on the origination LTV, we do not report the coefficients of this variable, but the inclusion of LTV as the variable is likely to affect estimated coefficients of the LTV. Most studies are on competing risk models, applying a proportional hazard estimation approach. The essence of the proportional hazard approach is to estimate the probability of delinquency or default in the current period conditional on the loan status during the previous period. For studies that have the origination LTV as an explanatory variable, the original LTV becomes less relevant and the current LTV becomes more relevant as time passes after origination. Therefore, it is not surprising that the magnitude of the effect of the origination LTV is considerably less in studies with LTV as an explanatory variable than in scorecard-type studies that do not use the proportional hazard methodology. In some sense, the current LTV also reflects the origination LTV, because its current value is due in large part to its first observation, the origination LTV.

   Another reason that the coefficient of the origination LTV in the proportional hazard models has a much smaller impact on delinquencies and defaults than in scorecard models is because the effect in the proportional hazard models is *per period*. In the scorecard approach, the effect is measured during the time period specified in the

definition of the dependent variable, such as 2 years or as a lifetime in our case. A comparable coefficient, then, would be to accumulate the per-period effect during the period specified in the scorecard dependent variable. Because of the multiple reasons cited previously, the proportional hazard and scorecard approaches produce different estimation coefficients of the origination LTV.

Table 2 lists other explanatory variables used in the respective equations, because any differences in the list of variables could affect the estimated LTV coefficient.

4. **How Explanatory Variables Are Constructed.** In particular, because default is a borrower put option, a number of studies (see table 2) attempt to measure the estimated value of the put option as an explanatory variable to indicate to what extent this option is in the money. These other studies are not directly comparable with ours, so we did not include them here unless they also have LTV as an explanatory variable. Other studies have applied indirect measures of LTV as an explanatory variable, for example, by simply allowing for a nonlinear relationship between LTV and default. We include these studies in this review.

Our study uses two measures of default, 90 days delinquent in the first 2 years after origination and the initiation of foreclosure proceedings during the loan's observed lifetime. The relevant papers are organized in table 1 based on these two measures. Note that any deviation from either the selected indicator of default or the time period for which these indicators are observed could cause the estimated LTV coefficient to differ.

For delinquency measurement, Hwang, Shu, and Van Order (2015) used a competing risk model to simulate the probability of 5-year 90-day delinquency. Competing risk models are estimated by multinomial logit, in which at least some measures of default and prepayment both are endogenous. In table 1, which cites the estimation method, we note three observations: (1) for cited references, multinomial logit is synonymous with the competing risk model; (2) the coefficient of the original LTV is not comparable with the binomial coefficient we estimate because of the simultaneous nature of the equations and the use of the current LTV as an explanatory variable; and (3) the variable, current LTV, in many of the studies is the LTV divided by 1 + house price appreciation (HPA) to reflect the house price appreciation, in which the numerator of the LTV is the amortized and possibly accrued mortgage balance—it is a different way to include the variable 1 + HPA.

Lam, Dunsky, and Kelly (2013) focused on 90-day delinquency in the first 7 years after origination. Their study incorporates the LTV variable by a spline function, which allows its effect to vary by ranges of the LTV. We report the coefficients according to the ranges covered by the various segments estimated. These are not the individual coefficients that are estimated for each segment, but rather the cumulative coefficients up to and including each segment. These values then become comparable with the coefficients of a continuous LTV variable, as in our study. For example, with linear splines, for LTVs in the range of the third and fourth knot points, $LTV_{spline3}$ and $LTV_{spline4}$, the contribution of LTV is equal to $\beta_1 \times LTV_{spline1} + \beta_2 \times (LTV_{spline2} - LTV_{spline1}) + \beta_3 \times (LTV_{spline3} - LTV_{spline2}) + \beta_4 \times (LTV - LTV_{spline3})$. If we take only significant coefficients into consideration, the Lam, Dunsky, and Kelly study demonstrates that the probability of 90-day delinquency increases monotonically with the origination LTV. They also concluded that the default probability is more sensitive to LTV changes for low FICO scores.

Table 1. Comparison of Different Loan-to-Value Coefficients

Reference	Default Measure	LTV Measure	Estimation Technique	Prepayment Treatment	Mortgage Data Coverage	Time Period	LTV Coefficient	
Delinquency Measurement								
McDash (Ours)	2-year 90-day delinquency	Combined LTV	Binomial logit	Counted as a "good" loan	Black Knight single-family and condo for-purchase	2000–2012	3.4502	
Hwang, Shu, and Van Order (2015)	5-year 90-day delinquency	Origination LTV	Multinomial logit	Endogenous	Freddie Mac single-family	1999–2012	2.24**	
Lam, Dunsky, and Kelly (2013)	7-year 90-day delinquency	Origination LTV splines	Multinomial logit	Endogenous	FHA single-family	1995–2008	GSE <70%: 0.188* 70–80%: 1.191 80–90%: 1.092 90–95%: 1.380 >95%: – 0.211*	FHA <80%: – 0.291* 80–90%: – 0.431* 90–95%: 2.813 >95%: – 0.350*
Calhoun and Deng (2002)	Lifetime 90-day delinquency	Origination LTV dummies	Multinomial logit	Endogenous	Fannie Mae, Freddie Mac 30-year single-family	1979–1993	30-year FRM <60%: – 1.348 60–70%: – 0.090 70–75%: 0.416 75–80%: 0.197 80–90%: 0.309 90–100%: 0.516	ARM <60%: – 1.310 60–70%: – 0.260 70–75%: 0.257 75–80%: 0.209 80–90%: 0.448 90–100%: 0.656
Kelly (2008)	Lifetime 90-day delinquency	Origination LTV	Multinomial logit	Endogenous	FHA single-family and condo for-purchase	1999–2006	National sample: – 0.056* MSA-level samples: – 0.028*	
IFE (2014)	Lifetime 90-day delinquency	Origination LTV dummies	Multinomial logit	Endogenous	FHA single-family	1975–2014	FRM30 90–95%: 0.0701 >95%: 0.0051 FRM15 90–95%: 0.0639 >95%: 0.0626 ARM 90–95%: 0.0734 >95%: 0.0829	
Freeman and Harden (2014)	Lifetime 30–90-day delinquency	Origination LTV	Multinomial logit	Endogenous	Community Advantage Panel Study for 30-year FRMs	1998–2009	– 0.02	
Deng, Quigley, and Van Order (1996)	Lifetime 30-day delinquency	Origination LTV	Multinomial logit	Endogenous	Freddie Mac single-family	1976–1983	2.491	

Garmaise (2015)	Lifetime 30-day delinquency	Origination LTV	Multinomial logit	Endogenous	Residential single-family	2004–2008	2.616
Deng, Quigley, and Van Order (2000)	Lifetime delinquency	Origination LTV dummies	Multinomial logit	Endogenous	Freddie Mac single-family	1976–1983	60–75%: 1.327 75–80%: 2.372 80–90%: 3.370 >90%: 3.333
Beem (2014)	Lifetime delinquency	Origination combined LTV	Binomial logit	Counted as a "good" loan	Fannie Mae, Freddie Mac 30-year single-family	1999–2011	0.1**
Foreclosure/Claim Measurement							
McDash (Ours)	Lifetime foreclosure	Combined LTV	Binomial logit	Counted as a "good" loan	Black Knight single-family and condo for-purchase first lien	2000–2012	4.3217
Lam, Dunsky, and Kelly (2013)	7-year foreclosure completion	Origination LTV splines	Multinomial logit	Endogenous	FHA single-family data	1995–2008	GSE <70%: – 1.994 70–80%: 1.461 80–90%: 0.520 90–95%: 0.711 >95%: 0.283* FHA <80%: – 1.724 80–90%: – 0.105 90–95%: 4.090 >95%: – 2.045
Freeman and Harden (2014)	Lifetime 90+ day or foreclosure	Origination LTV	Multinomial logit	Endogenous	Community Advantage Panel Study for 30-year FRMs	1998–2009	0.04
IFE (2014)	Lifetime claim	Origination LTV dummies	Multinomial logit	Endogenous	FHA single-family data	1975–2014	FRM30 90–95%: 0.1848 >95%: 0.2109 FRM15 90–95%: 0.3962 >95%: 0.6058 ARM 90–95%: 0.1459 >95%: 0.1429
Kelly (2008)	Lifetime claim	Origination LTV	Multinomial logit	Endogenous	FHA single-family and condo purchase money loans	1999–2006	National sample: – 0.057* MSA-level samples: 0.009*
Ghent and Kudlyak (2010)	Default that ends with the borrower vacating the home	Origination LTV & origination LTV dummies	Binomial logit	Counted as a "good" loan	Prime and nonprime private securitized loans, portfolio loans, and GSE loans	1997–2008	Origination LTV: 1.0** Origination LTV 80% dummy: 0.13**

ARM = adjustable-rate mortgage. FHA = Federal Housing Administration. FRM = fixed-rate mortgage. GSE = government-sponsored enterprise. IFE = Integrated Financial Engineering, Inc. LTV = loan-to-value. MSA = metropolitan statistical area.
* Statistically not significant.
** The coefficients are adjusted to reflect that the LTV ratios were measured as a decimal instead of as a percentage.

Table 2. Comparison of Explanatory Variables in the Literature

Variables	McDash (Ours)	Hwang, Shu, and Van Order (2015)	Lam, Dunsky, and Kelly (2013)	Calhoun and Deng (2002)	Kelly (2008)	IFE (2014)	Freeman and Harden (2014)	Deng, Quigley, and Van Order (1996)	Garmaise(2015)	Deng, Quigley, and Van Order (2000)	Beem (2014)	Ghent and Kudlyak (2010)
LTV at origination		v	v		v		v	v	v			v
CLTV	v										v	
CLTV > LTV dummy		v										
Current LTV		v				v						
Origination LTV dummies				v		v				v		
FICO score and MTM LTV Interaction			v									
Origination LTV = 80% dummy												v
Source of downpayment			v		v	v	v					
Loan size						v			v			
Loan grade	v											
FRM	v				v	v						v
Interest-only loan	v											v
Loan type missing	v											
Government loan	v											
Original term can be divided by 60	v											
Original term	v											
Single-family home	v				v							
Original credit score	v	v	v		v	v	v		v		v	v
Missing credit score dummy					v	v						
Prepayment penalty	v	v										
Primary residence	v			v								
Documentation type	v											
Jumbo loan	v											v
Missing payment status history	v											
Relative property value	v											
Relative loan size				v		v						
DTI ratio	v	v	v		v	v	v				v	
Yield curve slope	v		v			v						
30-year FRM rate	v											
Original interest rate											v	
CMT10/CMT1				v								
Has second loan	v		v									
Home price volatility						v						
Cumulative HPA	v				v	v						
Unemployment rate spread	v		v			v		v				

Mortgage rate spread	v	v				v						
Lifetime cumulative HPA	v											
Lifetime unemployment rate spread	v											
Lifetime mortgage rate spread	v											
Mortgage age		v	v	v		v						v
Mortgage age squared				v								
Mortgage age square root									v			
Seasonality			v	v		v						
Origination cohort			v									
Unpaid principal balance		v	v					v				
Spread at origination			v			v			v			
Spread at origination dummies												v
Burnout factor			v	v		v						
Credit burnout factor						v						
Census division			v									
Metropolitan location			v									
State laws			v			v						
Housing structure			v		v							
Borrower's ethnic/age/sex							v					
Borrower's marriage status							v	v		v		v
Borrower's education							v					
Borrower's employment status							v					
Household income							v	v				
Under water									v			
First-time buyer					v	v					v	
Origination year dummy		v		v							v	
State dummy		v										
Location dummy											v	
Loan purpose		v										v
Unemployment rate		v	v							v	v	
Relative unemployment rate						v						
Lagged unemployment rate												v
Mortgage premium value				v								
Probability of negative equity				v				v		v		
Probability of negative equity squared										v		v
Recourse												v
Fraction of contract value										v		
Fraction of contract value squared										v		
Spread in primary mortgage rate		v										
Number of housing units			v			v						
Reserve					v							
Builder					v							
Underserved area					v							
Refinance incentive						v						

CLTV = combined loan-to-value. CMT = Constant Maturity Treasury. DTI = debt-to-income. FRM = fixed-rate mortgage. HPA = house price appreciation. IFE = Integrated Financial Engineering, Inc. LTV = loan-to-value. MTM = Marked-to-market.

Note: The letter "v" in the data cells indicates that the analysis conducted in the study includes as a control the variable.

Calhoun and Deng (2002) focused on lifetime 90-day delinquency. They used origination LTV ratio dummies instead of a continuous variable, separating fixed-rate mortgage (FRM) and adjustable-rate mortgage (ARM) loans. In table 1, we show the results for both. Calhoun and

Deng observed high default probability for LTV from 70 to 75 percent, attributing it to the less stringent screening procedures for loans in this LTV range.

Kelly (2008) also focused on lifetime 90-day delinquency probability. Kelly observed negative LTV coefficients of origination LTV for the nation and for samples of metropolitan statistical areas (MSAs). The coefficients were not significant, however, because of limited LTV variation in the sample. They also found that traditional gift assistance for downpayments significantly increases the delinquency probability.

IFE (2014) estimated a competing-risk model for 90-day delinquencies. Origination LTV dummies were used in the regression. IFE found that loans with origination LTVs of more than 95 percent have a higher default probability compared with loans with LTVs from 90 to 95 percent for ARM loans. The opposite conclusion was made for FRM loans.

Freeman and Harden (2014) used a lifetime 30-to-90-day delinquency definition. Similar to Kelly's (2008) result, a negative significant coefficient was observed; however, the estimated coefficients are close to zero. One possible reason for this observation is a limited LTV range in the sample. Of the loans, 90 percent have an LTV ratio of more than 90 percent. The Freeman and Harden research focused on the impact of downpayment sources on mortgage performance.

Deng, Quigley, and Van Order (1996) focused on lifetime 30-day delinquency. They observe that default rates for loans with LTV ratios of more than 95 percent are three or four times higher than default rates for loans with LTV ratios of 90 to 95 percent. The default rates for the latter loans are, in turn, about five times higher than loans with LTVs of less than 80 percent.

Garmaise (2015) also use lifetime 30-day delinquency. His research examines mainly the impact of borrowers' self-reported assets on delinquency probability.

Deng, Quigley, and Van Order (2000) focus on lifetime delinquency of any duration. They observe that the highest default probability was for loans with LTVs from 80 to 90 percent. They conclude that borrowers who chose high initial LTV loans are more likely to exercise options in the mortgage market—prepayment and also default. Their study indicates that the origination LTV reflects investor preferences for risk in the market.

The second set of studies focuses on foreclosure or claims. Foreclosure is the initiation of foreclosure proceedings, unless stated otherwise in table 1. The delinquency studies also had different definitions of the dependent variable in both the time period the delinquency is observed and the number of days delinquent, so the results are not entirely comparable. These default studies have more dramatic differences in how the dependent variable is defined, however, so the estimates are even less comparable with each other.

Lam, Dunsky, and Kelly (2013) analyzed 7-year foreclosure completions. The coefficients of their LTV spline function can also be converted to comparable values as discussed previously in the delinquency section.

Freeman and Harden (2014) estimated an equation for lifetime 90-day delinquency to foreclosure. They observed a positive coefficient of origination LTV; however, the estimated coefficients are close to zero. It is possible that they found a negative and statistically significant relationship between LTV and 30-to-90-day delinquency because of the narrow LTV range in their study.

IFE (2014) also estimated the lifetime claim probability (or, more precisely, log odds) with their competing risk model. They found loans with origination LTVs of more than 95 percent have a higher default probability compared with loans with LTVs from 90 to 95 percent. ARM loans with LTVs from 90 to 95 percent have a foreclosure probability similar to those with LTVs of more than 95 percent.

Kelly (2008) also estimated the lifetime claim probability. The coefficients of LTV were not significant for both the national and MSA samples, as was also the case for the 90-day delinquency equation.

Ghent and Kudlyak (2010) focused on the definition of default that is marked when a borrower vacates the home, and they used a scorecard-type approach similar to ours. They use a dummy variable of LTV80 to capture the high likelihood of a second mortgage being present for loans in which LTV equals exactly 80 percent. The positive coefficient of LTV80 indicated the higher default option value for these loans and has a strong effect on the probability of default. They find that lenders have less effective recourse in states that require lenders to go through a lengthy judicial foreclosure process, rather than a swift nonjudicial foreclosure process to obtain a deficiency judgment. The reported coefficients in table 1 are adjusted based on the magnitude of LTV used in the paper.

One important difference between our study and most of the others is the econometric technique. Ours is a "scorecard" approach similar to Beem (2014), which attempts to predict future relative performance based on the facts available at the time of origination, possibly controlling for the future state of the (local) economy. Scorecards have been estimated often but rarely published, because they are considered proprietary tools for competitive advantage. The primary modeling and estimation approach in the literature is the competing risk model estimated with multinomial logit. These are hazard-type models that attempt to explain the probability of (competing) performance outcomes along the life of the mortgage conditional upon its current status. The initial LTV becomes a less important factor, because the current LTV over time becomes the more important factor along this risk dimension. Compared with a scorecard model, then, the effect of origination LTV is less in the hazard-type models. Another difference is that the probabilities in the hazard-type models are per period, instead of during the time period specified in the definition of the delinquency or default, such as the first 2 years. As noted previously, most coefficients cited in table 1 are for the multinomial logit proportional hazard approach, making a direct comparison with the logit estimates of a scorecard model difficult.

Another contribution of the current analysis is its inclusion of different mortgage market segments. Most previous literature (table 1) focused on a specific segment of the mortgage market. Little discussion occurs among the performance of loans in different programs, such as Federal Housing Administration, Veterans Affairs, Fannie Mae, Freddie Mac, jumbo, and subprime markets. The use of the McDash data granted the opportunity to directly compare the loan credit quality among different mortgage markets and their associated underwriting rules.

Methodology and Model

This section describes the econometric models used to estimate the historical performance of for-purchase loans from fiscal year (FY) 2000 to FY 2012.

Model Specification

We specified binomial logistic models of the probability of default or foreclosure. The mathematical expression for the probability of loans to default or foreclose within a specific time period is given by

$$PD = \frac{e^{\alpha+X\beta}}{1+e^{\alpha+X\beta}}$$

Here, we use X to denote the vector of explanatory variables for the probability of default or foreclosure. We count prepaid loans as "good."

Discussion of Variables

This section describes some key variables in the model and the intuition behind specifications and use in the model.

Dependent Variables

We estimate bivariate logistic regressions for performance periods of the first 2 and 3 years after origination and lifetime and for 90 and 120 days delinquency and the initiation of foreclosure proceedings. After examining preliminary regression results for these nine different models and finding similar qualitative results, we selected the following two definitions to demonstrate the empirical results.

2-year 90-day delinquency. For the delinquency analysis, 90-day delinquency within 2 years is the dependent variable. Based on industry common practice, early delinquency within 2 years is an indicator of poor underwriting quality.

Lifetime foreclosure. For the default analysis, foreclosure within the loan's lifetime is the dependent variable.

We explain why borrowers do not pay for either 3 consecutive months in the first 2 years or when a foreclosure proceeding is initiated in a loan's observed lifetime. Even though cures may happen, we consider these two conditions to be significant indications of "bad" borrower behavior. Therefore, we want to discern the underlying factors, particularly the amount of the downpayment, leading to the decision to stop paying. With our definition of default, the variability in the duration of foreclosure proceedings and alternative loss mitigation efforts thereby did not need to be modeled. In the final analysis, it is primarily low or negative HPA that determines whether an ultimate default occurs and whether a loss is incurred; underwriting a borrower typically focuses on the borrower decision not to pay, without the complications accompanying the likelihood of loss in the final default.

Independent Variables

The McDash data set includes an extensive set of variables that we supplement with economic variables from Moody's Analytics. We group the available variables for this study into loan-

specific variables and macroeconomic variables. In this section, we define most of the variables and reserve the results for the Empirical Results section, in which we also discuss the remaining explanatory variables used in this study.

Loan-specific variables. All static loan-specific variables are generated at the origination of the loans.

Combined loan-to-value (CLTV) ratio. The sum of the origination primary and second lien loan amounts, divided by the property value is the calculation to obtain the CLTV ratio. Without the identity of the borrower and the specific house, only a second loan originated simultaneously with the primary loan is identified in the McDash data set. In the absence of second loans, CLTV is equal to the origination loan-to-value ratio.

Borrower credit score. Borrower credit or FICO scores at the loan level are an important predictor of default and prepayment behavior. We use a dummy variable to identify loans with missing credit scores.

Household debt-to-income (DTI) ratio. The DTI variable is likely subject to measurement error, because income, especially for low-documentation loans, is likely to be erroneously recorded. This variable also has a high degree of missing values, at 46.3 percent. The not-full-documentation loans have missing DTIs at a rate of 64.9 percent, and full-documentation loans have a rate of 11.1 percent. Extreme outliers also suggest measurement error. In addition to using the numerical variable, we created two dummy variables to reflect these situations; one variable indicates a missing value and the other indicates when we censored outlier values. We censored the observations with DTIs less than 5 percent and more than 70 percent, treating them as if they were missing, and created a separate DTI outlier indicator to account for this situation. The McDash data set does not include the total DTI, which includes the payment burden of all household debts.

Spread at origination (SATO). The spread between the mortgage note rate and the prevailing mortgage rate at the time of origination measures SATO, which is widely regarded as the lender surcharge for additional borrower risk characteristics not captured by standard underwriting hard data, such as a FICO score, LTV, DTI ratio, documentation level, and so on. A high SATO loan is generally riskier, compared with a similar loan with a low SATO (see table 2 for references using SATO-type variables).

Relative property value. The property value divided by the median house value of the corresponding MSA provides the relative property value. If the median house price at the MSA level is not available, then we substitute the state- or national-level median house price.

Grade B and C loans. Grade B and C loan dummy variables indicate grade-B or grade-C mortgages, which are generally believed to be of poorer credit quality than A-grade loans.

Prepayment penalty clause. Prepayment penalty loans are a special product type. After years of excluding prepayment penalty clauses, lenders reintroduced them to protect mortgage investors from multiple solicitations by mortgage brokers to refinance as soon as the market mortgage rate fell to less than the contract rate. This aggressive solicitation activity produces income for the brokers and a loss of value for investors. The prepayment penalty is a feature the borrower can choose when applying for a loan, creating a variable that may reflect self-selection behavior by a borrower. Loans with prepayment penalty clauses are riskier than those without such a clause in the contract, at least in part because borrowers are less likely to prepay, which means they may be exposed to adverse personal and economic events for a longer period of time.

Macroeconomic Variables. Macroeconomic variables are control variables. Some of these variables are not observed at the time of the underwriting decision but have a very important influence on whether delinquencies or defaults occur. Macroeconomic variables are typically explanatory variables in estimating a scorecard and then neutralized, that is, set at specific values when constructing an underwriting scorecard.

Our data consists of the monthly Federal Housing Finance Agency (FHFA) purchase-only house price index, along with the Core Based Statistical Area (CBSA) codes for metropolitan areas, metropolitan divisions, and state or national levels; monthly interest rates, including 1-year and 10-year Department of the Treasury rates and 30-year FRM rates; monthly unemployment rates, with CBSA codes for metropolitan areas, metropolitan divisions, and state or national levels; and the state-level census median housing prices for single-family houses and the national level for condos. All the economic data come from Moody's Analytics.[1] We now discuss the three sets of macroeconomic variables included as control variables.

House price appreciation (HPA). The house price appreciation variable, which measures house price movements after loan origination, is the main control variable in the possibility that a loan may become "under water" within the performance observation period. (IFE, 2014).

In our models, HPA is defined based on the dependent variable. For the dependent variable, 90 days delinquent within 2 years, the variable HPA_2yr is calculated as the cumulative HPA during the first 2 years, measured at the MSA level. If the HPA at the MSA level is not applicable, then we use the state- or national-level CBSA code. For each property, the measure is:

$$\text{HPA\_2yr} = \frac{HPI_{24}}{HPI_0} - 1$$

where HPI is the price index applicable to each property, based primarily on its MSA. The source is the FHFA purchase-only index.

For the dependent variable, foreclosure during a loan's lifetime, we calculate the variable "HPA_fore_life" as the cumulative HPA up to the initiation of the first time a foreclosure status occurs or, if there is no foreclosure, up to the end of the observation period, which, in our estimation data set, is a minimum of 5 years measured at the MSA level. If HPA at the MSA level is not applicable, then we use state- or national-level price index.

Although we intuitively prefer to measure HPA up to delinquency or default events, because that is the information borrowers have when making no-payment decisions, measuring the HPA in a period as short as 2 years appears to be more effective. In the lifetime scenario, a much higher probability exists for more extreme HPA to occur at different points in time than the cumulative HPA experienced at foreclosure. The HPA up to the point of foreclosure appears to provide a better fit in the lifetime scenario.

Yield curve slope. Expectations about future interest rates and differences in short-term and long-term borrowing rates, associated with the slope of the Treasury's yield curve, influence the choice between adjustable-rate mortgage and fixed-rate mortgage loans. We use the spread of the 10-year Constant Maturity Treasury (CMT) yield over the 1-year CMT yield to measure the slope of the Treasury yield curve.

[1] http://www.economy.com.

Estimation Database

In this section, we describe our methodology to identify and treat duplicate loan observations and summarize the outlier control process to derive the database used for regressions. Appendix A includes detailed analyses of the data steps.

The loan information raw data are from the McDash database. McDash loan data is a mortgage performance database maintained by Black Knight, collected from most major mortgage servicers. Before our analysis, we screened the loan data using two criteria: (1) the origination loan calendar year and (2) loan purpose. We first selected loans originated during calendar years 2000 through 2012, focusing on for-purchase loans. After this preliminary screening, the number of loan observations was reduced from 287,559,201 to 26,642,754.

Duplicate Observations

Approximately 26.5 million loans were originated from 2000 through 2012. We did not have borrower names or property addresses for those loans, but we suspect, in many cases, that two or more of the individually reported loans were in fact duplicates of the same mortgage transaction, and that these multiple reports were because of either (1) one or more servicing transfers, making them the same loan reported by multiple servicers, or (2) the fact that at least one loan was a second lien. Approximately 5 million loans had the same origination month, ZIP Code, property type identification, and original property value as at least one other loan observation. If the original loan amounts had less than a 10-percent difference, we assumed that these observations are because of servicing right transfers. A minimal difference in original loan amount may occur when the new servicer reports the amortized loan amount as of the servicing transfer date instead of the loan amount at the origination date. We combined the loan payment status histories for these loans and deleted one of the two loan observations.

If the difference in the original loan amount was more than 10 percent, we assumed that these loans are two different loans on the same property. The loan with the higher original loan amount was identified as the primary loan, and only the primary loan performance was used for analysis. If a foreclosure of the second lien triggered a foreclosure of the primary loan, this would appear in the performance status of the primary loan. Loans with the lower original loan amount were assumed to be second liens. We derived the CLTV by dividing the combined loan amount by the original property value. Loans with a CLTV of more than 125 percent were excluded from our analysis as outliers. About 3.5 million loans identified as second loans or duplicate servicer-transfer loans were excluded from the analysis data set. The flowchart describing this process is in appendix B.

After removing second and duplicate loans, we identified 23,189,377 loans as valid for use in our analysis.

Outlier Exclusion

Our research focuses on first mortgages for single-family homes and condominiums, excluding an additional 4,075,892 loans and reducing the sample size to 19,113,485 loans. We then screened the pared sample of loans for outliers, typically thought to be the result of transcription or data errors. Table 3 shows the data exclusion standards to omit outliers. Details regarding outlier identification and exclusion are in appendix C.

Table 3. Cutoff Levels for Outlier Exclusion

Variable Name	Data in Model
CLTV ratio	20% ≤ X ≤ 125%
Original credit score	300 ≤ X ≤ 850
DTI ratio	5% ≤ X ≤ 70%
Original property value	10,000 ≤ X ≤ 3,000,000
Original term	60 ≤ X ≤ 480
Loan payment history missing month	X ≤ 3
Original loan amount	10,000 ≤ X ≤ 2,000,000
Original interest rate	1% ≤ X ≤ 25%

CLTV = combined loan-to-value. DTI = debt-to-income.

After excluding all outliers, the total number of observations for the 2-year 90-day delinquency model is 14,360,593, or 75.13 percent of the loans before outlier exclusions. The total number of observations for the lifetime foreclosure model is 10,566,442, or 55.28 percent of the loans before outlier exclusions.

The distributions of the continuous loan origination variables in the final sample set are shown in table 4.

Table 4. Continuous Variable Distributions of Final Data Set

Variable	Mean	Median	Maximum	Quantiles						Minimum
				95%	90%	75%	25%	10%	5%	
CLTV ratio	0.8200	0.8	1.25	1	0.9921	0.9500	0.7619	0.6250	0.5065	0.2
Original credit score	714	722	850	801	791	767	670	624	598	300
DTI ratio	0.34	0.35	0.70	0.54	0.50	0.44	0.25	0.17	0.13	0.05
Original property value	284,626	210,000	3,000,000	725,000	550,000	350,000	135,000	90,000	72,000	10,000
Original term	352	360	480	360	360	360	360	360	240	60
Original loan amount	217,225	170,259	2,000,000	515,000	400,400	270,750	112,080	76,000	60,000	10,000
Original interest rate	0.0610	0.0613	0.161	0.0775	0.0725	0.0663	0.0550	0.05	0.0463	0.01

CLTV = combined loan-to-value. DTI = debt-to-income.

Empirical Results

In this section, we present the regression results and a series of analyses that provide insight into the extent to which downpayments reduce the probability of delinquency and default, the focus of this paper. Recall that these are for-purchase loans, single-family home and condominiums, performance of first mortgages, and from the cohorts 2000 to 2012.

Regression Results

Table 5 presents the estimated logistic regression coefficients and goodness-of-fit statistics for the delinquency model defined as 90 days delinquent within the first 2 years and for default defined as initiation of foreclosure proceedings during a loan's lifetime.

We focus on both factors, delinquency and foreclosure, used in automated underwriting decisions. In a separate analysis, we test whether a loan with early delinquency leads to a higher probability of falling into foreclosure during the remainder of its lifetime. The indicator has a coefficient of 2.52 in the foreclosure equation. This is not a surprising value, because the loan has to be 90 days delinquent to get to foreclosure status, although not all 90-day delinquencies result in foreclosure; however, this result suggests that an underwriting focus on the delinquency equation has merit in its own right, even if the underwriting focus is on foreclosure.[2]

We use the term probability of default (PD) to describe the outcome variable for the analysis conducted in this study. As mentioned in the literature review section, our definition often differs from prior studies. The term PD is used for both the delinquency and default equations, unless the context calls for distinguishing them. A positive coefficient indicates a higher PD.

FRMs have lower PDs than ARMs or balloon mortgages, which is usually thought to reflect adverse selection of ARMs by the more cash-strapped borrowers. Interest-only loans have higher PDs. Federal Housing Administration and Veterans Affairs loans have higher PDs than conventional loans, holding all the other factors constant. If the original term is not divisible by 60 months—that is, if it is not 15, 20, 25, or 30 years—the loan has a higher risk of delinquency. We suspect that loans with odd terms are mainly modified loans as a result of loss mitigation efforts. We expect such loans to have higher delinquency rates. In addition, the longer the original term, the higher the PDs, which likely reflects borrower self-selection: shorter terms require higher monthly payments, so borrowers with higher incomes and wealth may tend to select shorter term loans.

Single-family loans have a higher delinquency rate, but a slightly lower default rate, than condominiums. B and C loans exhibit a higher probability of default than prime loans, the reference category, and a lower probability of delinquency—although the magnitude is small in absolute value, and the estimated coefficient is not statistically significant at the 0.0001 probability level.

Government loans tend to have a higher delinquency risk and a higher default risk than conventional loans. For loans other than government or conventional loans, the delinquency risk is higher because of the relatively large monthly payments for jumbo loans. On the other hand, the default risk for other loans is lower than for conventional loans, because credit scores of conventional loan borrowers are quite good, enabling them to borrow without requirements for insurance.

[2] Furthermore, we experimented with 120-day delinquencies and a 3-year time horizon and found results that are similar to the 90-day results.

Table 5. Logistic Regression Statistics

	Delinquency		Foreclosure	
Variable	**Coefficient**	**Pr > ChiSq**	**Coefficient**	**Pr > ChiSq**
Intercept	1.1164	< 0.0001	– 5.7648	< 0.0001
CLTV ratio	3.3939	< 0.0001	4.4373	< 0.0001
Has second loan	– 0.0191	0.007	0.0369	< 0.0001
Original credit score	– 0.0127	< 0.0001	– 0.00884	< 0.0001
DTI ratio	1.4891	< 0.0001	0.4105	< 0.0001
DTI ratio is missing	0.3998	< 0.0001	0.5479	< 0.0001
DTI ratio is outlier	0.1608	< 0.0001	0.6734	< 0.0001
Spread at origination	37.9145	< 0.0001	10.9559	< 0.0001
Relative property value	– 0.0207	< 0.0001	– 0.2965	< 0.0001
Primary residence	– 0.0499	< 0.0001	– 0.3378	< 0.0001
Single-family home	0.0951	< 0.0001	– 0.0531	< 0.0001
B or C loan	– 0.0213	0.0016	0.5279	< 0.0001
Jumbo loan	0.138	< 0.0001	0.2018	< 0.0001
Full documentation	– 0.2835	< 0.0001	– 0.3898	< 0.0001
Unknown documentation	– 0.0658	< 0.0001	0.1682	< 0.0001
30-year FRM rate	13.5103	< 0.0001	76.9304	< 0.0001
FRM	– 0.0438	< 0.0001	– 0.4236	< 0.0001
Interest-only loan	0.6205	< 0.0001	0.5122	< 0.0001
Loan type missing	0.194	0.0003	– 0.4749	< 0.0001
Government loan	0.1977	< 0.0001	0.2281	0.0001
Other than government or conventional	0.1606	< 0.0001	– 0.4682	< 0.0001
Has prepayment penalty	0.2771	< 0.0001	0.0567	< 0.0001
Original term	0.00285	< 0.0001	0.00269	< 0.0001
Original term can be divided by 60	– 1.1346	< 0.0001	– 0.1819	< 0.0001
Loan source—correspondent	0.2324	< 0.0001	– 0.2061	< 0.0001
Loan source—whole sale	0.4825	< 0.0001	0.3176	< 0.0001
Loan source—unknown	0.4871	< 0.0001	0.3009	< 0.0001
Payment status history partially missing	0.0349	< 0.0001	0.1246	< 0.0001
Yield curve slope	– 4.0913	< 0.0001	10.0181	< 0.0001
2-year cumulative HPA	– 2.5273	< 0.0001		
2-year unemployment rate spread	11.5887	< 0.0001		
2-year mortgage rate spread	0.6658	0.0127		
Lifetime cumulative HPA until foreclosure			– 3.443	< 0.0001
Lifetime unemployment rate spread until foreclosure			15.6674	< 0.0001
Lifetime mortgage rate spread until foreclosure			43.4376	< 0.0001
Gini coefficient	0.713		0.701	

CLTV = combined loan-to-value. DTI = debt-to-income. FRM = fixed-rate mortgage. HPA = house price appreciation.

Origination credit scores have the expected sign, with a higher score presenting less risk and a higher magnitude for early delinquency than for lifetime default. This result suggests that the effect of the credit score measured at origination fades with the passage of time.

We suspect that loans with prepayment penalty clauses are riskier than those without such contractual clauses, at least in part, because borrowers of loans with prepayment penalty clauses are less likely to prepay, meaning they are exposed to the possibility of defaulting for a longer period of time.

Loans for the primary residence have lower PDs than those for second homes or investor properties. As expected, full-documentation loans have lower PDs than reduced-documentation loans, and jumbo loans have higher PDs. The loans with some missing payment status history have higher risk than loans with a complete status history. If the history of missing payments is longer than 3 months or after the first 3 months, we delete the loan observation.

Wholesale loans and unknown source loans have higher PDs than retail loans, the reference category. Correspondent loans are more likely to become delinquent, but are surprisingly less likely to default. During the subprime boom periods, many mortgage brokers actively solicited troubled borrowers to refinance into subprime loans. As a result, many loans that would have foreclosed were terminated as if they were fully prepaid, at least for the initial purchase loans.

Relatively higher priced houses within an MSA have lower PDs, possibly reflecting higher income and wealth of borrowers who can afford more expensive houses and may be less prone to cashflow issues.

Higher CLTVs have higher PDs.

Higher housing DTI ratios also have higher PDs. Furthermore, if the DTI ratios are missing or are outliers, the PDs will be higher relative to observations with DTI ratios. For example, the probabilities of default for a loan with missing DTI, a loan with outlier DTI, and a loan with the highest acceptable DTI are 2.61, 2.95, and 2.02 percent, respectively, given all other variables are set at median values in the data set.

The steeper the origination yield curve, the lower the delinquency rate, but higher the default rate. The steeper the origination yield curve generally suggests that the interest rate will rise rapidly in the future. ARM borrowers are likely to experience payment shocks at the end of the initial teaser periods. If a borrower's income does not rise at the same rate as the origination yield curve, the borrower may face the inability to pay and enter default. Such payment shocks, however, tend to occur several years after the origination, thus resulting in limited impact on the delinquency rate but stronger impact on the default rate.

Higher market rates at origination have higher PDs. The finding is consistent with the default option theory. When the market interest rate drops, previously originated loans would be priced at premium. In other words, borrowers would find that they are making higher than market rate payments. When refinancing is not generally feasible, such as the period after the 2008 financial crisis, borrowers making higher than market rate payments would have a higher incentive to default.

If a second lien exists, the primary loan has less delinquency risk, holding constant the LTV for primary loans without second loans or the CLTV for loans with second loans. The effect observed in this analysis may be the result of tighter underwriting requirements to qualify for second liens, so adverse selection exists for those who do not qualify for a second loan. Given lifetime results, it is probable that the concentration of initial delinquency risk on the smaller second component of mortgage loans in a short time after origination indirectly reduces the delinquency risk of loans with a second lien. Second loans mainly cover the risk above an 80 percent LTV; in fact, 95.33 percent of loans with second loans have a primary loan LTV of less than or equal to 80 percent. Private mortgage insurance is required for Fannie Mae and Freddie Mac to accept LTVs of more than 80 percent, and private mortgage insurance companies have complained about these "piggyback" mortgages because they reduce business volume and cause

them to be adversely selected. For the lifetime foreclosure rate, however, primary loans with second loans perform worse than those without second loans, most likely because a foreclosure on the more onerous second loan typically triggers a foreclosure on the first loan.

The higher the market mortgage rate SATO, the higher the PDs. Positive SATO stands for the lender surcharge for additional borrower risk, as described previously.

The next three variables—HPA, the unemployment rate, and market mortgage rate—are not known at origination, and, if they were included in the development of an automated underwriting system, these variables would be neutralized. These variables, however, strongly determine the absolute level of the PDs. They are independent of the credit quality of the mortgages, so their inclusion should not alter the effects of the fundamental underwriting factors, and in fact, are required to hold these important factors constant. The higher the HPA, the lower the PDs. The higher the unemployment rate spread from origination, the higher the PDs. The higher the SATO during the first 2 years for the delinquency equation and up to foreclosure or lifetime for the default equation, the higher the PDs.

Functional Form

Notice that the continuous variables are expressed in a linear functional form. We show in figures 1 and 2 diagnoses for the variables CLTV, FICO, and DTI that indicate whether a linear form is appropriate. (More diagnostic charts are in appendix D.)

Figure 1. Diagnostics Charts for Combined Loan-to-Value Ratio

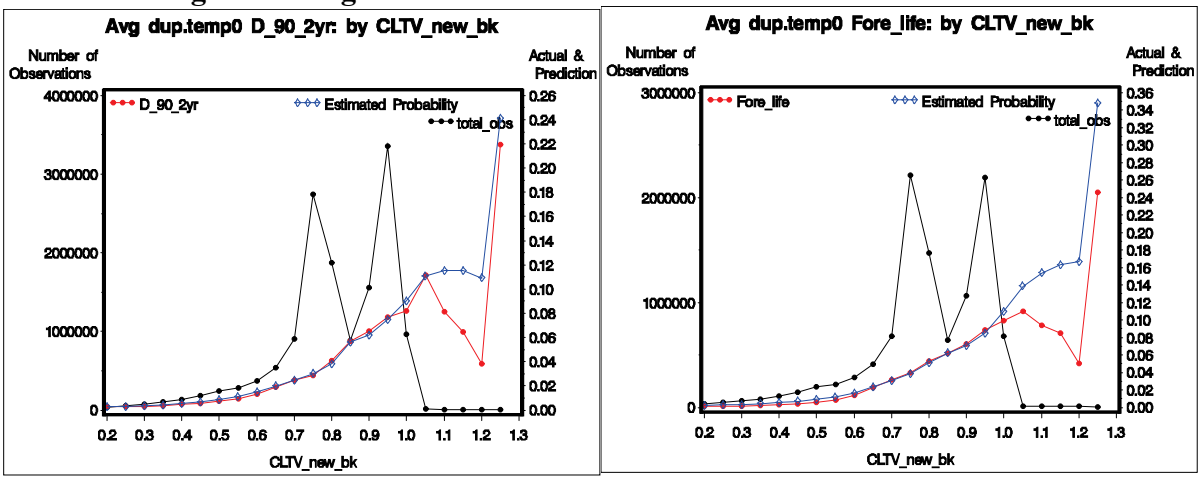


Figure 1 shows the predicted and actual PDs along the dimension of the respective variables, along with the distribution of the observations. It is more important to have the predicted and actual aligned over the range in which most observations are concentrated and not so much where data is thin. Figure 1 shows that the higher the CLTV, the higher the PD. When CLTV is higher than 100 percent, some abnormal behaviors are present in the raw data, but very few observations are in this range. Both models fit well up to about 110 percent, which is sufficient for analyses in the current market.

Figure 2. Diagnostics Charts for FICO Score

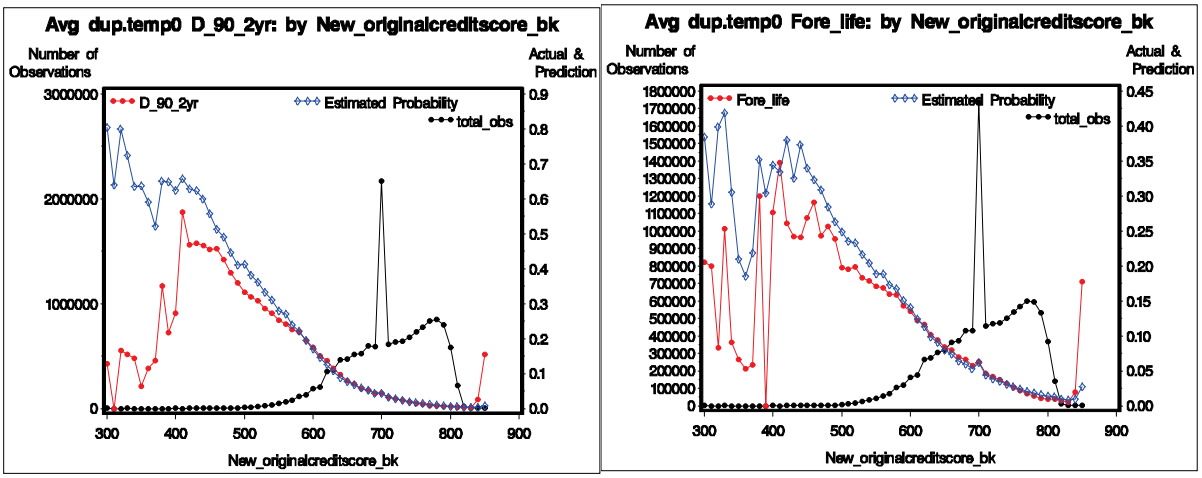
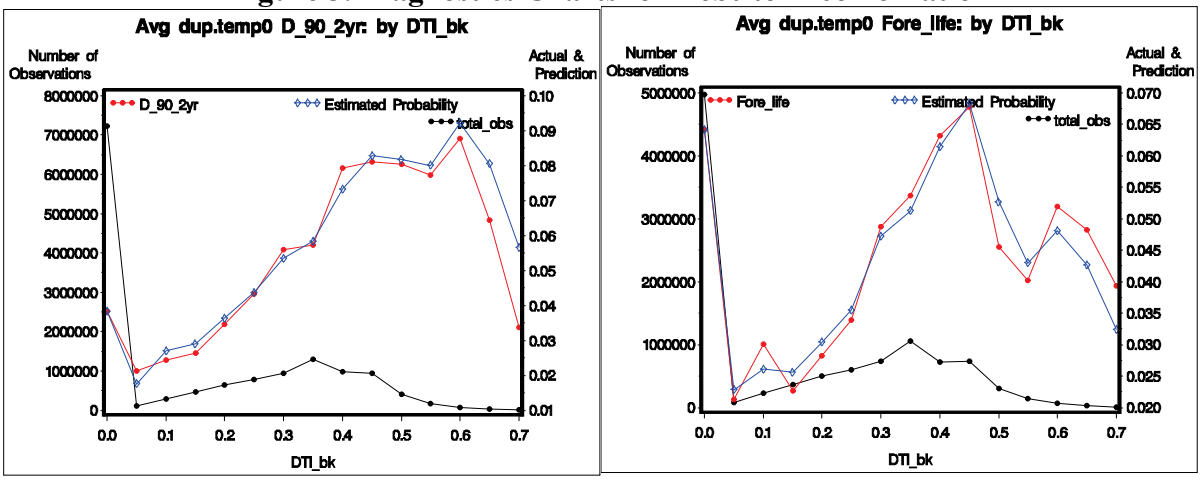


Figure 2 illustrates that the higher the credit score, the lower the PD. Again, some strange behaviors occur when a credit score is extremely high or low, but the sample data are thin in these regions and beyond the range for applicable policy relevance.

Figure 3 shows that the higher the DTI ratio, the higher the PD. A large population is left when DTI equals 0, because we let DTI equal 0 when it is missing or outside our specified relevant range of [0.05, 0.7]. In all the previous charts, the linear functional form appears to be appropriate. For the default equation, the predicted *pattern* is consistent with actual default rates, so a nonlinear form is not likely to adjust for a prediction that is considerably less than the actual.

Figure 3. Diagnostics Charts for Debt-to-Income Ratio



Compensating Factors for CLTV

In table 6, we show the percentage point changes in the CLTV that compensate for a one unit change of each continuous variable, holding other variables constant so that the log-odds ratio, and hence the PD, are the same. This is simply the variable's coefficient divided by negative one times the CLTV coefficient. Note that we use the percent sign (%) in the table, but its meaning is percentage points.

Table 6. Compensating Factors (Continuous Variables)

Variable	Delinquency (%)	Default (%)
Original credit score	0.37	0.20
DTI ratio	– 43.88	– 9.25
Spread at origination	– 1,117.14	– 491.36
Relative property value	0.61	6.68
30-year FRM rate	– 398.08	– 1,733.72
Original term	– 0.08	– 0.06
Yield curve slope	120.55	– 225.77
2-year cumulative HPA	74.47	
2-year unemployment rate spread	– 341.46	
2-year mortgage rate spread	– 19.62	
Lifetime cumulative HPA until foreclosure		77.59
Lifetime unemployment rate spread until foreclosure		– 353.08
Lifetime mortgage rate spread until foreclosure		– 978.92

% = percentage point. DTI = debt-to-income. FRM = fixed-rate mortgage. HPA = house price appreciation.

For example, in the default equation, if the credit score were to increase by 1.0 (for example, from 700 to 701), which decreases risk, the combined loan-to-value ratio needs to increase, which increases risk, by 0.20 percentage points to compensate in the sense of maintaining the same PD. In more practical terms, if the credit score were to *decrease* by 100 points (for example, from 680 to 580), the CLTV would have to decrease by 20 percentage points to compensate. Note that the linear form of the continuous variables makes the calculation of the compensating factor straightforward, but also note that this is not the reason for selecting the linear form, as discussed previously.

Another example highlights that the expression of the variable needs to be accounted for. Debt-to-income is measured as a decimal in our analysis. Converting it to percentage points requires the movement of DTI coefficient's decimal point to the left by two positions. For the delinquency equation, an increase in the DTI from 40 to 41, which increases risk, requires the reduction of CLTV by 0.42 percentage points to compensate for the increased risk. An increase of the DTI from 40 to 45 requires a decrease in CLTV of 2.19 percentage points to compensate for delinquencies and a decrease in CLTV of 0.46 percentage points to compensate for defaults. The finding demonstrates that DTI is more important for delinquencies but is less important for defaults.

The compensating factors regarding binary variables are shown in table 7. For example, to have the same probability of default as a prime loan, a B or C loan needs to have a CLTV that is 11.9 percentage points lower than the CLTV of an otherwise identical prime loan.

Another example is that for second and investor-owned homes, which perform worse than primary residences when all else is held constant, the CLTV needs to be about 7.61 percentage points lower to maintain the same PD. Conventional mortgage maximum loan-to-value ratios for second and investor-owned homes traditionally have been lower than those for primary

residences by 5 or more percentage points. Observations from prior research is consistent with our findings and is an example of how previous manual underwriting requirements, which were only vaguely based on empirical evidence, applied the compensating factors principle.

Table 7. Compensating Factors (Binary Variables)

Variable	Delinquency (%)	Default (%)
Prepayment penalty	– 8.16	– 1.28
DTI ratio is missing	– 11.78	– 12.35
DTI ratio is outlier	– 4.74	– 15.18
Primary residence	1.47	7.61
Single-family home	– 2.80	1.20
B or C loan	0.63	– 11.90
Jumbo loan	– 4.07	– 4.55
Full documentation	8.35	8.78
Unknown documentation	1.94	– 3.79
FRM	1.29	9.55
Interest-only loan	– 18.28	– 11.54
Loan type missing	– 5.72	10.70
Government loan	– 5.83	– 5.14
Other than government or conventional loan	– 4.73	10.55
Has prepayment penalty	– 8.16	– 1.28
Original term can be divided by 60	33.43	4.10
Loan source—correspondent	– 6.85	4.64
Loan source—whole sale	– 14.22	– 7.16
Loan source—unknown	– 14.35	– 6.78
Payment status history partially missing	– 1.03	– 2.81

DTI = debt-to-income. FRM = fixed-rate mortgage.

CLTV Analytics

In this section we present analytics to examine the CLTV-default relationship. We show these for the foreclosure equation, but they are similar for the delinquency equation. Charts for the delinquency equation are in appendix E, which show the PD-CLTV relationship when specific, interesting explanatory variables change. In the PD-CLTV relationship analyses, all other variables are set at their median values in the estimation data set.

Figure 4 shows that primary loans without second loans have lower PDs than primary loan with second loans, at the same levels of LTV or CLTV. This is because the borrowers with piggyback loans are likely to be riskier, and when a second loan defaults, the corresponding first loan could be triggered into default.

Figure 4. Effect of Second Loans on Default Probability

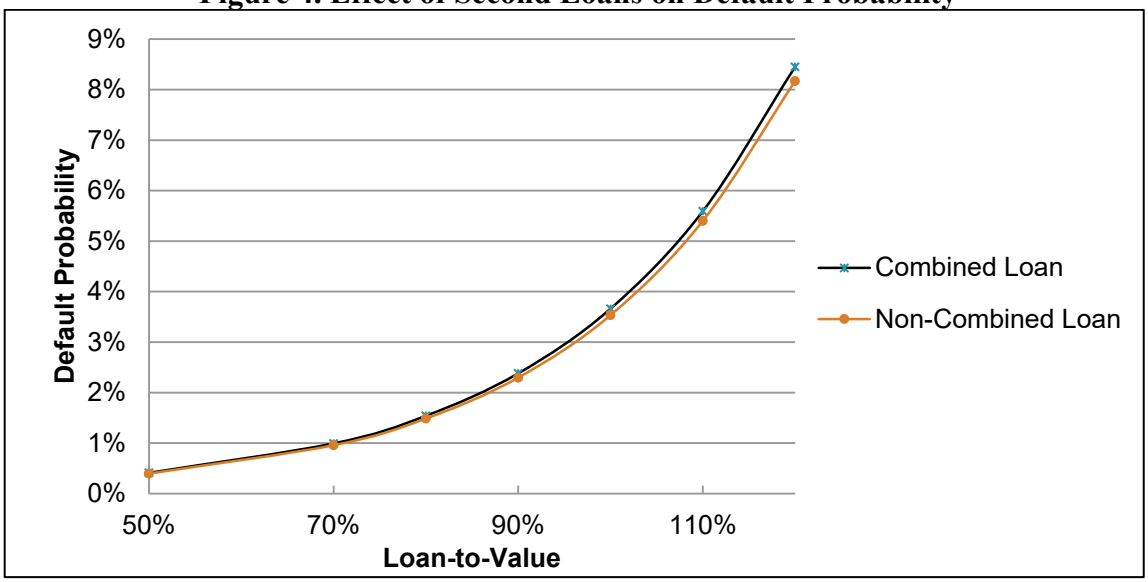


Figure 5 shows that very low FICO scores produce very high PDs, compared with high FICO scores. For a 90 percent CLTV loan, the PD for a FICO score of 500 is nearly 12 times higher than for a FICO score of 800. As might be expected, the delinquency relationship is even more exaggerated, as shown in figure 6. At a 90 percent CLTV, the probability of early delinquency is nearly 35 times higher at a FICO score of 500 compared with a FICO score of 800.

Figure 5. Credit Score Effect on Default Probability

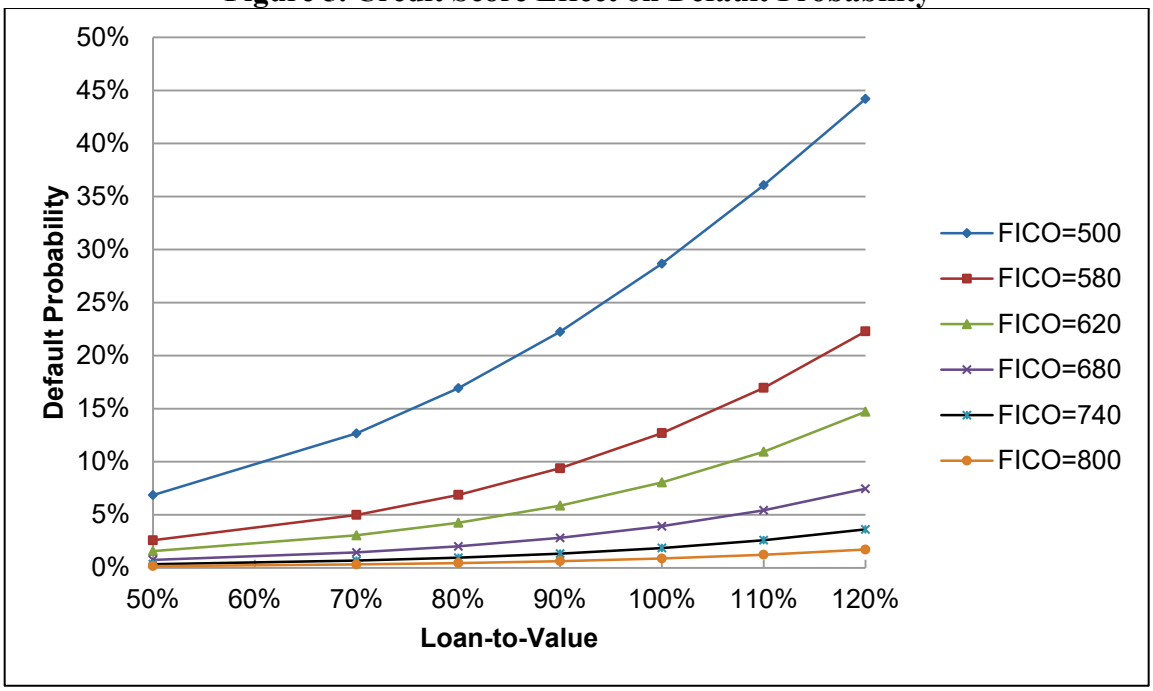


Figure 6. Credit Score Effect on Delinquency Probability

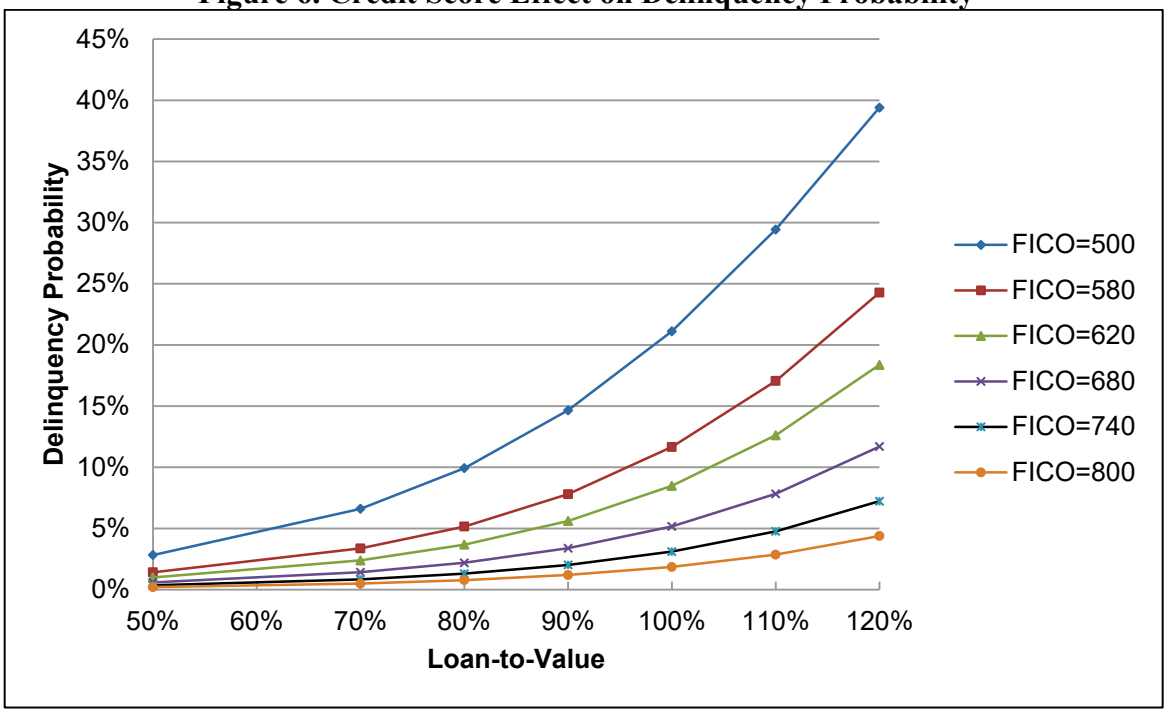
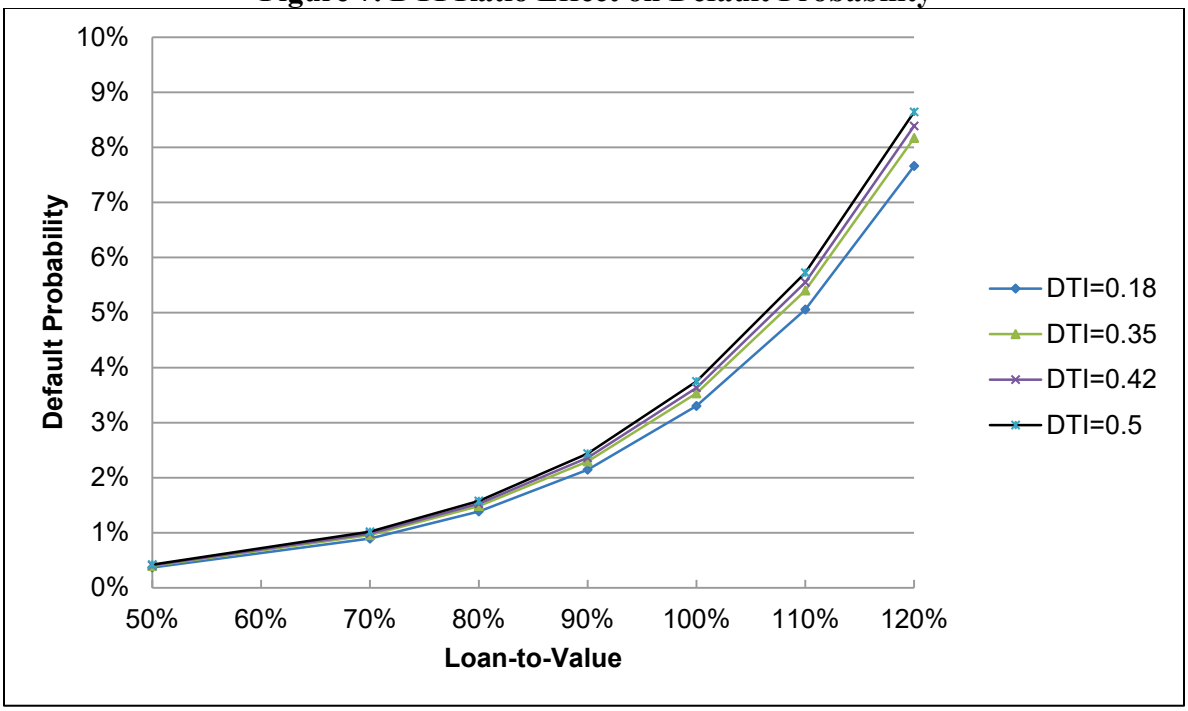


Figure 7 shows that higher DTI ratios produce higher PDs. For loans with 90 percent CLTV, the PD is 1.03 times higher for a 42 percent DTI ratio compared with a 35 percent DTI ratio. Again, this comparison shows the relatively weak effect DTI ratio has on estimated default probabilities.

Figure 7. DTI Ratio Effect on Default Probability



DTI = debt-to-income.

Figure 8 shows that higher-priced houses have lower PDs, likely because of stronger income and wealth of corresponding borrowers.

Figure 8. Relative Property Value Effect on Default Probability

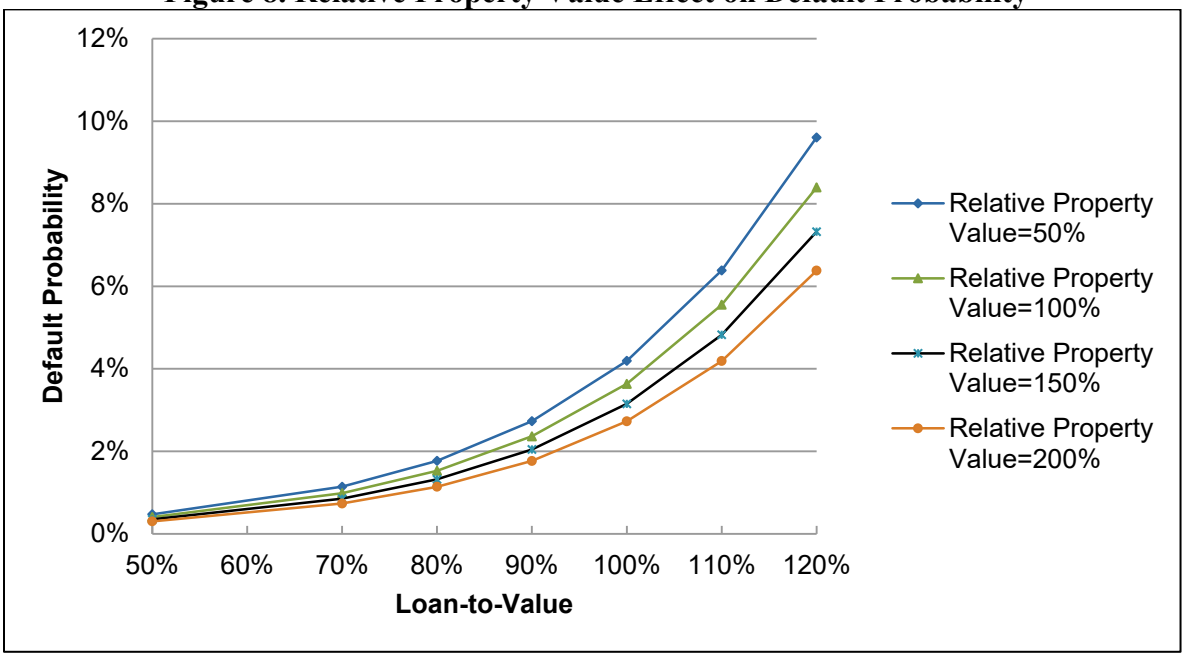


Figure 9 shows that B and C loans have significantly higher PDs.

Figure 9. B and C Loan Effect on Default Probability

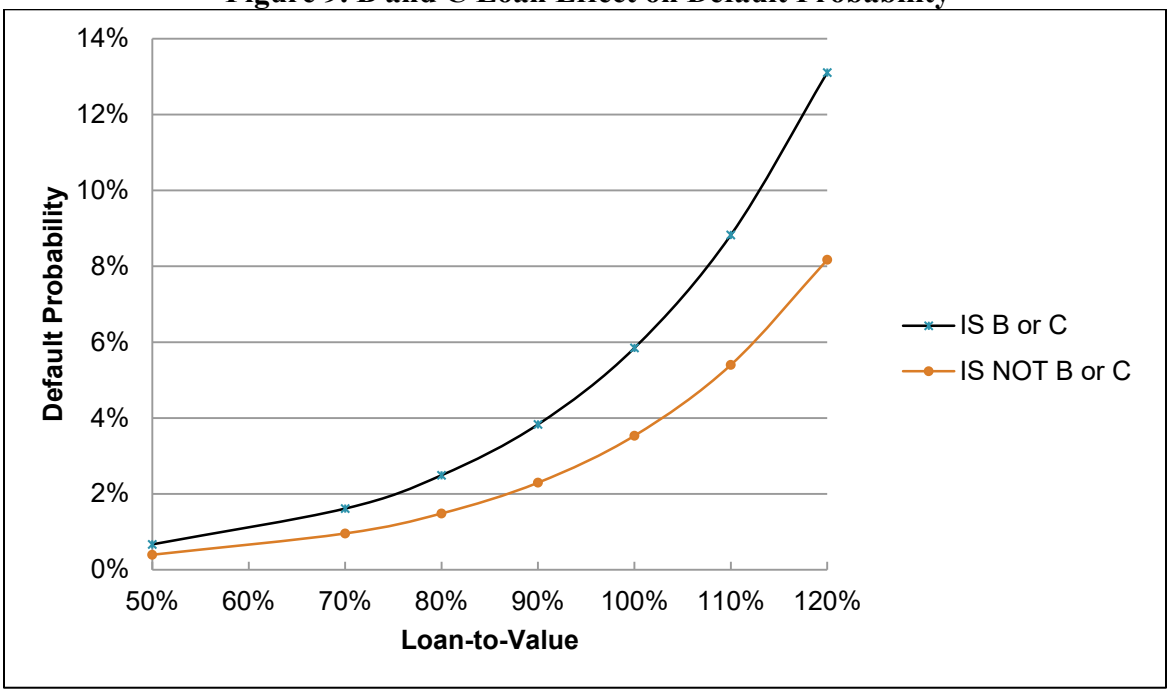


Figure 10 shows that jumbo loans have about a 20-percent higher probability of default rate than conforming loans.

Figure 10. Jumbo Loan Effect on Default Probability

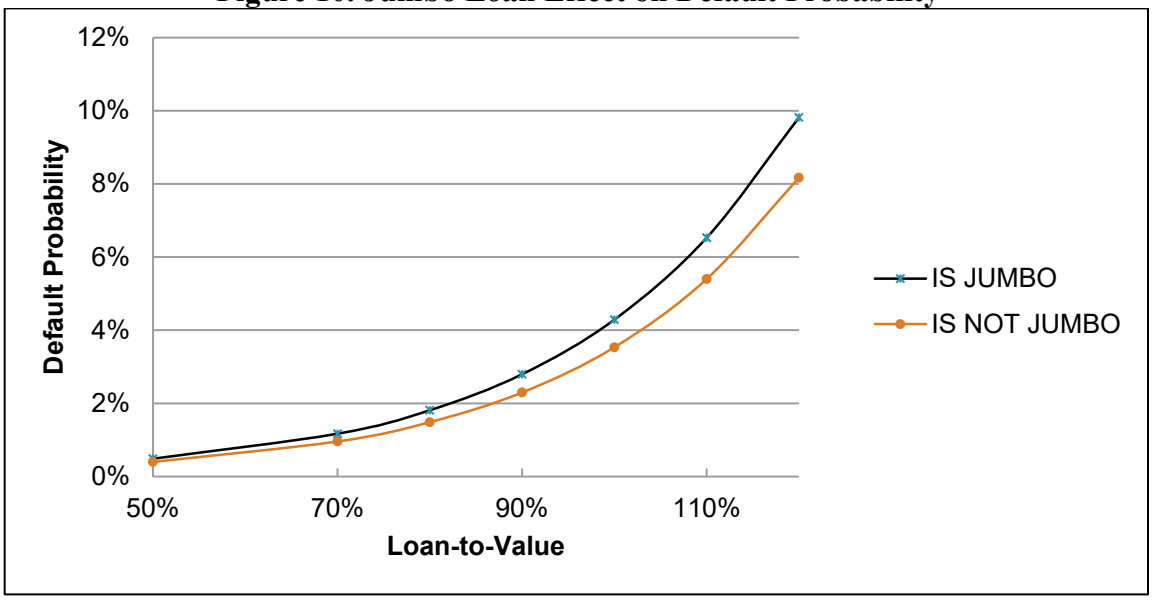


Figure 11 shows that full-documentation loans have lower PD rates. Full documentation is one indicator of good underwriting quality, so the PD of loans with full documentation is usually lower. Full-documentation underwriting can reduce PD about 30 percent.

Figure 11. Full-Documentation Effect on Default Probability

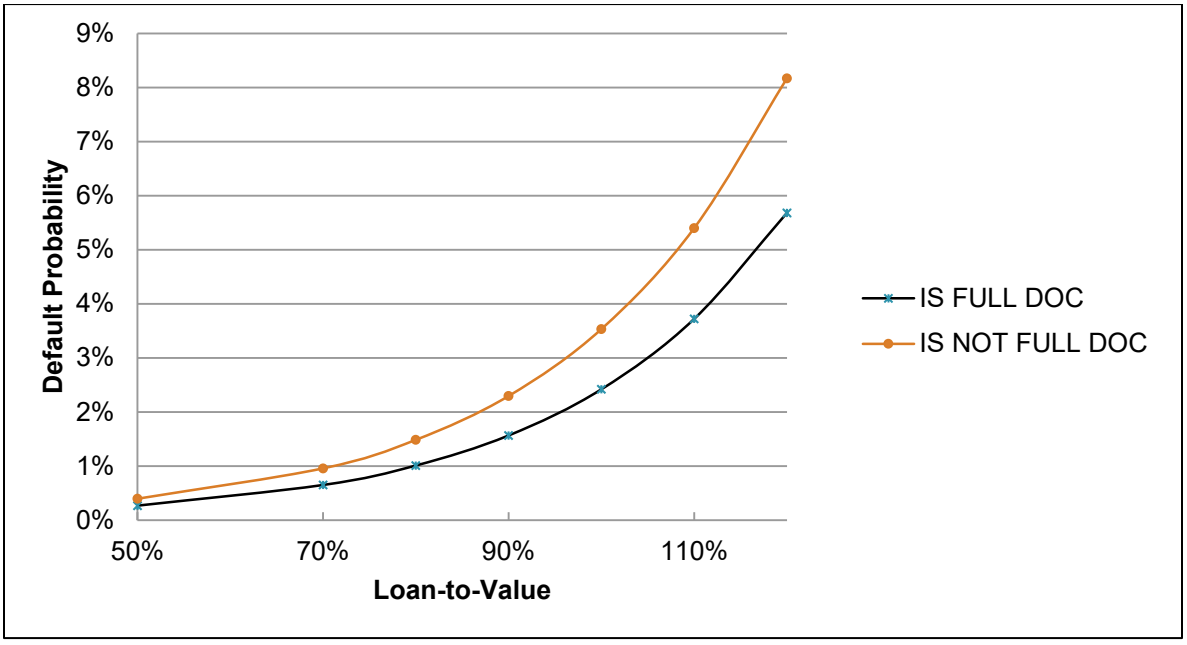
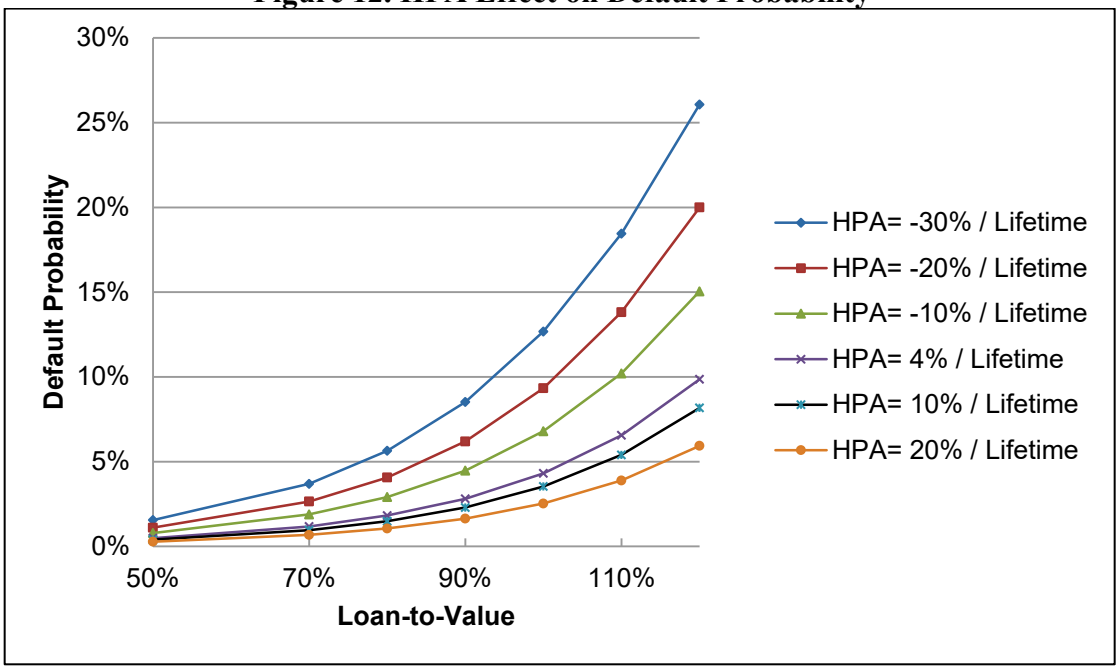


Figure 12 shows that house price appreciation has a significant effect on PDs. For a loan with 90 percent CLTV, the PD under a -30 percent HPA environment is about three times the PD under an average 4 percent HPA environment. The finding is only suggestive, because the measurement of lifetime HPA in our study is complicated by measuring only up to foreclosure for loans that go into foreclosure.

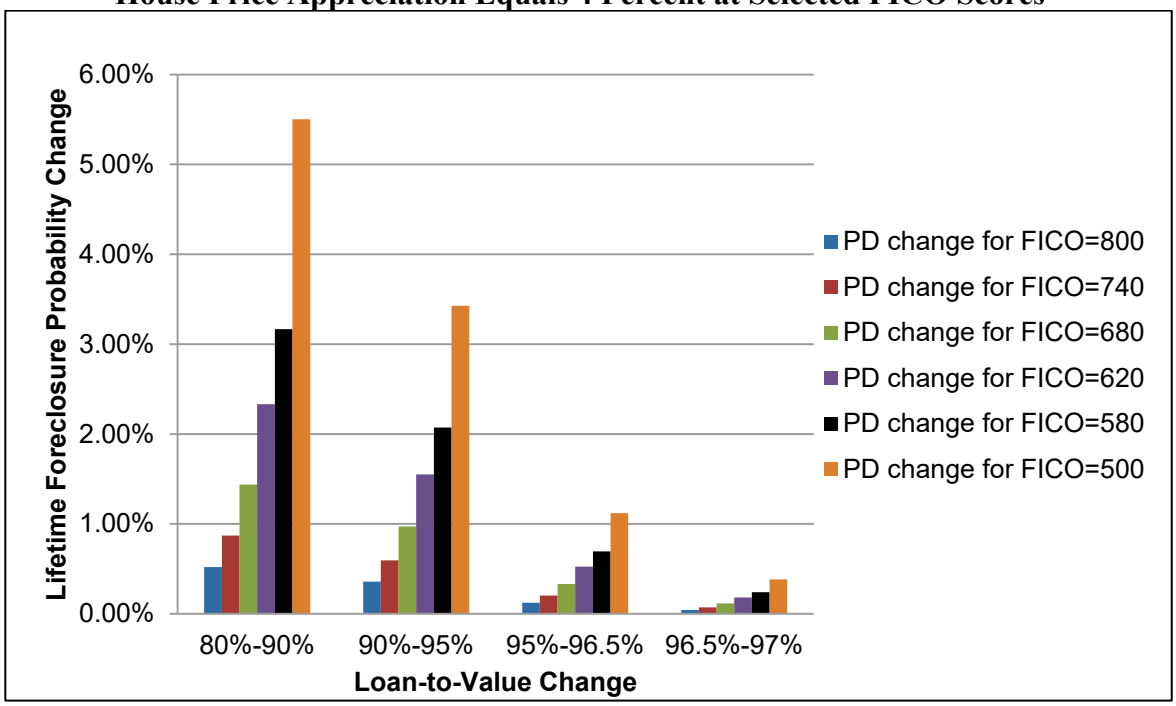
Figure 12. HPA Effect on Default Probability



HPA = house price appreciation.

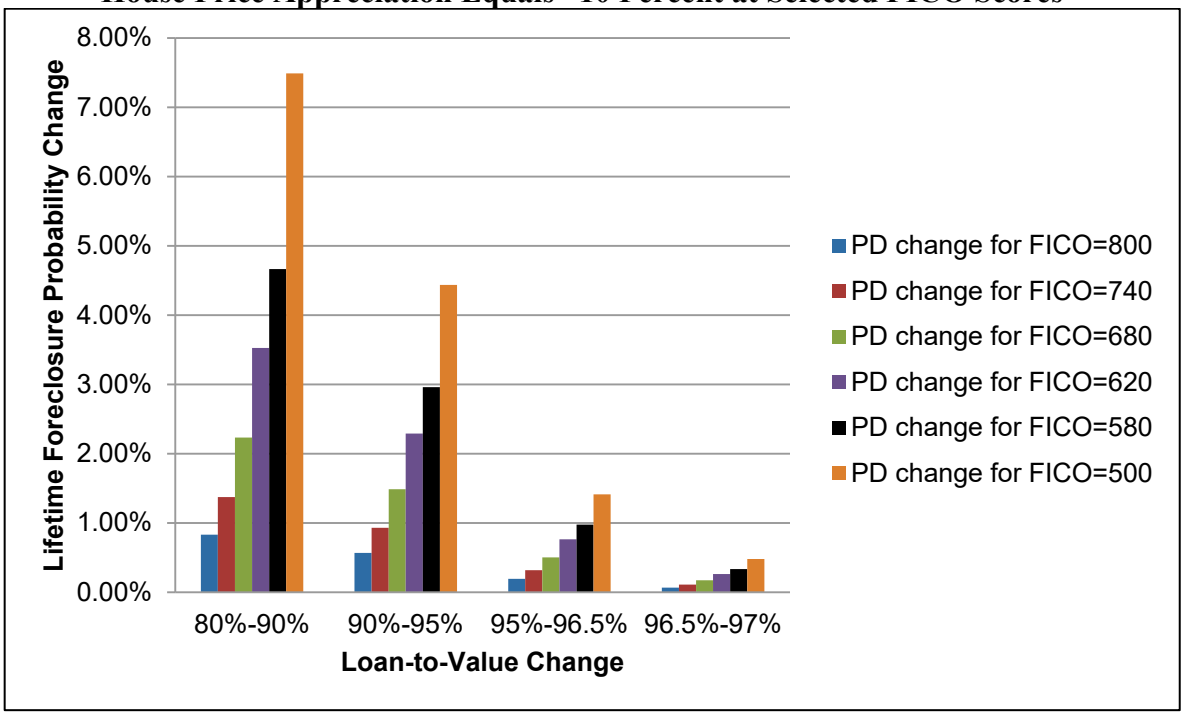
Figures 13, 14, and 15 now show the interaction of CLTV, FICO scores, and HPA. Policymakers might use these types of analyses when selecting combinations of maximum CLTV ratios and minimum FICO scores. To keep the discussion more tractable, we show only the probabilities of foreclosure (denoted as PD). The bars represent, for selected FICO score levels as indicated by the keys on the right side of the next three figures, the change in the PD by moving the maximum CLTV from 96.5 to 97.0 percent, for example. At a FICO score of 580 and assuming the house price drops 30 percent, the PD increases 0.3 percentage points if the maximum CLTV increases from 96.5 to 97.0 percent. This is one demonstration of the cost of relaxing the maximum allowed CLTV.

Figure 13. Lifetime Foreclosure Probability Change at House Price Appreciation Equals 4 Percent at Selected FICO Scores



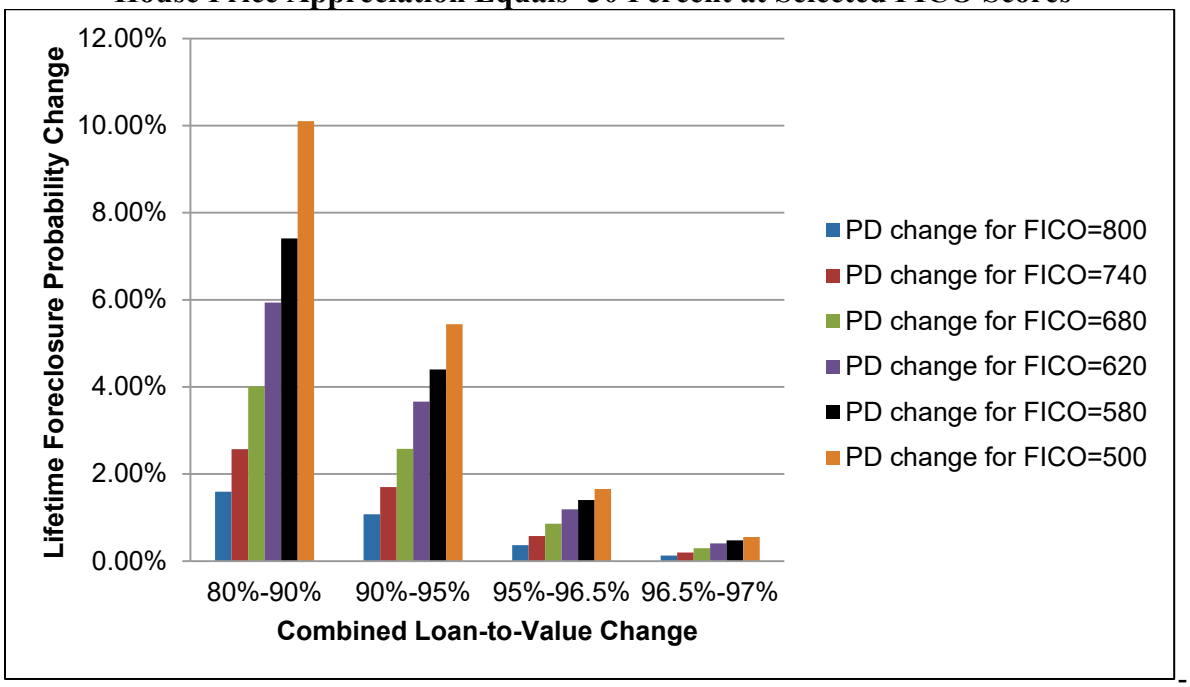
PD = probability of default.

Figure 14. Lifetime Foreclosure Probability Change at House Price Appreciation Equals –10 Percent at Selected FICO Scores



PD = probability of default.

Figure 15. Lifetime Foreclosure Probability Change at House Price Appreciation Equals -30 Percent at Selected FICO Scores



PD = probability of default.

Of course, the analysis of changing the maximum LTV or CLTV depends on the economic environment against which policymakers would want to protect. If private mortgage insurance is included, they may choose a benign environment such as a positive 4-percent HPA. As an alternative, policymakers may want to protect against a stressed environment when default losses may threaten solvency, such as a drop of 30 percent in housing prices.

They would, of course, attempt to balance the cost of higher PD rates with the increase of origination volume to include a mission to serve low- to moderate-income homebuyers . These are typical considerations when determining maximum CLTV ratios, and, often, the minimum acceptable FICO score may be established for different CLTV levels. For example, the minimum acceptable FICO score may be set at 580 to go to the maximum CLTV; however, the FICO score can go down to 500 if CLTV ratios are constrained to be no greater than 90 percent. This is an application of the compensating factors principle.

We observe that the magnitude of deterring PD is relatively high through the CLTV spectrum for ranges of credit scores in highly negative HPA environments and fairly low across HPA environments in the higher range of CLTV. The magnitude of deterrence to PD is relatively low through CLTV spectrum for ranges of credit scores in higher negative HPA environments and fairly low across HPA environments in the 90-to-97 range of CLTV.

In a similar way, these calculations often also include the selection of the maximum DTI. Note that, for the default equation, as discussed in the previous compensating factors section, the DTI has a relatively weak effect.

Implications and Conclusions

In this research, using a scorecard model and the McDash loan-level performance data, we quantified the impact of mortgage downpayments (in terms of the CLTV ratio) on mortgage defaults by controlling for other factors such as loan-specific elements and economic conditions during the performance period. Our definition of delinquency was 90-days past due within the first 2 years after origination; our definition of default was the initiation of foreclosure proceedings during a loan's lifetime. The estimation sample consisted of for-purchase mortgages and included government, conventional, and private-sector loans.

Our literature review focused on studies that quantified the effect of the initial downpayment on delinquency, default, or claims. Very few studies applied the scorecard-type approach. The literature instead was replete with studies using the competing risk model, which produces hazard-type estimates and a multinomial logit estimator that make it difficult to compare other estimates with ours. More reasons for the incompatibility of estimates include different definitions of delinquency and default, different estimation time periods along with the explanatory variables selected for inclusion, and the coverage of different mortgage market segments. We consequently drew no inferences regarding whether our estimates of the absolute and relative effects of CLTV on the probability of delinquency and default were consistent with previous literature.

We developed a useable estimation database from the loan performance records in the McDash loan database that is compiled from the major mortgage servicers Fannie Mae and Freddie Mac. Our major challenge was that the McDash database does not have borrower names or property addresses, which created uncertainty when considering multiple reporting incurred by servicing transfers and second liens. Devising an algorithm to address duplicate entries, we found that 8.58 percent of the loans examined during the 2000-to-2012 period were duplicitous due to servicing transfers and 6.24 percent to be second liens, distinguishing between the two types based on reported loan balances. The combined loan-to-value variable, which included the balance of the second lien when a second lien was identified, accounted for downpayments.

Although we did not create a typical mortgage scorecard, we produced analytics describing how important the CLTV is relative to other risk factors, such as the FICO credit score and the housing debt-to-income ratio. The first of these analytics derives from the time-honored tradition in manual underwriting of assessing creditworthiness by assessing compensating factors. For example, when a mortgage application reveals a blemished credit record, the application can still be accepted if accompanied by a greater downpayment to compensate for the credit reputation deficiency. The advent of mortgage scorecards made the evaluation of compensating factors implicit. On the other hand, we made the compensating factors explicit, as demonstrated from our estimated default models, focusing on the CLTV. We computed the amount of percentage points the CLTV has to decrease in order to offset the change in probability of default when some other explanatory variable increases. In addition, in the case of binary variables, the decrease in CLTV needed to offset the credit risk effect when they change from one to zero. For example, if the loan changes from a prime loan to a subprime loan, holding other factors constant, the CLTV needs to be lowered by 11.9 percentage points to offset the higher credit risk of the B or C loan.

We also showed the type of interaction among CLTV, FICO scores, and HPA rates that policymakers might use and showed tradeoffs when considering combinations of maximum LTV ratios and minimum FICO scores. Those are, in essence, the types of tradeoffs made within a scorecard model. Given a specified threshold for default rates and the relevant HPA, policymakers can choose different combinations of LTV ratios and FICO scores to manage mortgage credit risks. That is, the LTV ratio and FICO score can serve as compensating factor to each other to maintain a constant expected default risk.

Analytical inferences from estimated scorecard models evaluate where to set accept or reject cut points and various override rules, such as the minimum FICO score and the maximum DTI ratio.

References

Archer, Wayne, David Ling, and Gary McGill. 1996. "The Effect of Income and Collateral Constraints on Residential Mortgage Terminations," *Regional Science and Urban Economics* 26: 235–261.

Arnone, Marco, Salim M. Darbar, and Alessandro Gambini. 2007. "Banking Supervision: Quality and Governance." Working paper 07/82. Washington, DC: International Monetary Fund.

Beem, Richard H., Jr. 2014. "Residential Mortgage Delinquency Rates: The Determinants of Default," *Issues in Political Economy* 23: 59–75.

Calhoun, Charles A., and Yongheng Deng. 2002. "A Dynamic Analysis of Fixed- and Adjustable-Rate Mortgage Terminations," *The Journal of Real Estate Finance and Economics* 24 (1): 9–33.

Deng, Yongheng, John M. Quigley, and Robert Van Order. 1996. "Mortgage Default and Low Downpayment Loans: The Costs of Public Subsidy," *Regional Science and Urban Economics* 26 (3–4): 263–285.

———. 2000. "Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options," *Econometrica* 68 (2): 275–307.

Freeman, Allison, and Jeffrey J. Harden. 2014. "Affordable Homeownership: The Incidence and Effect of Downpayment Assistance," *Housing Policy Debate* 25 (2): 308–319.

Garmaise, Mark J. 2015. "Borrower Misreporting and Loan Performance," *The Journal of Finance* 70 (1): 449–484.

Ghent, Andra C., and Marianna Kudlyak. 2010. "Recourse and Residential Mortgage Default: Theory and Evidence from U.S. States." Working paper 09–10R. Richmond, VA: The Federal Reserve Bank of Richmond.

Hwang, Min, Binzi Shu, and Robert Van Order. 2015. Understanding the Underwriting in Prime Markets: the GSE Case. Working paper. Washington, DC: George Washington University.

Integrated Financial Engineering, Inc. (IFE). 2014. *Actuarial Review of the Federal Housing Administration Mutual Mortgage Insurance Fund Forward Loans for Fiscal Year 2014*. Washington, DC: U.S. Department of Housing and Urban Development.

Kau, James B., and Donald Keenan. 1995. "An Overview of Option-Theoretic Pricing of Mortgages," *Journal of Housing Research* 6: 217–244.

Kelly, Austin. 2008. "Skin in the Game: Zero Downpayment Mortgage Default," *Journal of Housing Research* 17 (2): 75–99.

Lam, Ken, Robert M. Dunsky, and Austin Kelly. 2013. Impacts of Down Payment Underwriting Standards on Loan Performance: Evidence from the GSEs and FHA Portfolios. Working paper 13–3. Washington, DC: Federal Housing Finance Agency.

Siddiqi, Naeem. 2012. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ: John Wiley & Sons, Inc.

Thomas, Lyn C. 2009. Consumer Credit Models: Pricing, Profit, and Portfolios. New York: Oxford University Press.

Thomas, Lyn C., David B. Edelman, and Jonathan N. Crook. 2002. Credit Scoring and Its Applications. Philadelphia: Society for Industrial and Applied Mathematics.

Vanderhoff, James. 1996. “Adjustable and Fixed Rate Mortgage Termination, Option Values and Local Market Condition: An Empirical Analysis,” *Real Estate Economics* 24 (3): 379–406.

Additional Reading

Ambrose, Brent W., and Richard J. Buttimer, Jr. 2000. "Embedded Options in the Mortgage Contract," *The Journal of Real Estate Finance and Economics* 21 (2): 95–112.

Campbell, Tim, and J. Kimball Dietrich. 1983. "The Determinants of Default on Conventional Residential Mortgages," *The Journal of Finance* 38 (5): 1569–1581.

Caplin, Andrew, Charles Freeman, and Joseph Tracy. 1997. "Collateral Damage: Refinancing Constraints and Regional Recessions," *Journal of Money, Credit and Banking* 29 (4): 497–516.

Deng, Yongheng. 1997. "Mortgage Termination: An Empirical Hazard Model with a Stochastic Term Structure," *The Journal of Real Estate Finance and Economics* 14 (3): 310–331.

Dunn, Kenneth B., and Chester Spatt. 2005. "The Effect of Refinancing Costs and Market Imperfections on the Optimal Call Strategy and the Pricing of Debt Contracts," *Real Estate Economics* 33 (4): 595–618.

Elmer, Peter, and Steve Seelip. 1999. "Insolvency, Trigger Events, and Consumer Risk Posture in the Theory of Single-Family Mortgage Default," *Journal of Housing Research* 10 (1): 1–25.

Findley, M. Chapman, and Dennis Capozza. 1977. "The Variable Rate Mortgage: An Option Theory Perspective," *Journal of Money, Credit and Banking* 9 (2): 356–364.

Foster, Charles, and Robert Van Order. 1984. "An Options-Based Model of Mortgage Default," *Housing Finance Review* 3 (4): 351–372.

Kau, James B., Donald Keenan, and Taewon Kim. 1994. "Default Probabilities for Mortgages," *Journal of Urban Economics* 35 (3): 278–296.

Kau, James B., and Taewon Kim. 1994. "Waiting to Default: The Value of Delay," *Journal of the American Real Estate and Urban Economics Association* 227: 195–207.

Pavlov, Andrey D. 2000. "Competing Risks of Mortgage Prepayments: Who Refinances, Who Moves, and Who Defaults?" *The Journal of Real Estate Finance and Economics* 23 (2): 185–212.

Peristiani, Stavros, Paul Bennett, Gordon Monsen, Richard Peach, and Jonathan Raiff. 1997. "Credit, Equity, and Mortgage Refinancings," *Economic Policy Review* 3 (2): 83–99.

Phillips, Richard A., Eric Rosenblatt, and James H. Vanderhoff. 1996. "The Probability of Fixed-And Adjustable-Rate Mortgage Termination," *The Journal of Real Estate Finance and Economics* 13 (2): 95–104.

Quercia, Roberto G., and Michael A. Stegman. 1992. "Residential Mortgage Default: A Review of the Literature," *Journal of Housing Research* 3 (2): 341–379.

Quigley, John M. 1987. "Interest Rate Variations, Mortgage Prepayments, and Household Mobility," *The Review of Economics and Statistics* 69 (4): 636–643.

Schwartz, Eduardo, and Walter Torous. 1993. "Mortgage Prepayment and Default Decisions: A Poisson Regression Approach," *Journal of the American Real Estate and Urban Economics Association* 21: 431–449.

Von Furstenburg, George. 1969. "Default Risk on FHA-Insured Home Mortgages as a Function of the Terms of Financing: A Quantitative Analysis," *The Journal of Finance* 24: 459–477.

Appendix A. Data Analysis

In this section, we discuss our data source and our data processing procedures, which include analysis of missing data, outliers, and variable construction.

The source of loan information data is from the McDash loan database. McDash data are the comprehensive mortgage performance data solutions that Black Knight Financial Technology Solutions, LLC developed and maintains. The McDash loan data are collected from major mortgage servicers and from Fannie Mae and Freddie Mac. We screened the loan data using two criteria: the origination loan calendar year and the loan purpose. We selected loans originating from 2000 through 2012 and focused on for-purchase loans. This initial screening reduced the number of loan observations from 287,559,201 to 26,642,754.

Data Quality

Missing Data

Table A1 shows the number of observations with missing values of individual variables. We first focused on four important variables: (1) original property value; (2) combined loan-to-value (CLTV) ratio; (3) original credit score, or FICO; and (4) housing debt-to-income (DTI) ratio. The database did not include the total DTI. Missing data in the variables seriously impair the ability to conduct empirical analyses.

CLTV is the key variable in this research. Table A1 shows more than 77 percent of the loans lack information on CLTV; hence, we computed the CLTV for individual loans using inferences on multiple mortgages on the same property. We present more detail about this procedure in this section.

Table A1. Loans With Missing Variable Data and Our Resolution

Variable	Total Observation: 26,642,754 Loans		
	Missing Observations	Percentage of Data Missing	Resolution
ZIP Code	785	<0.01	Not used in regression, but used to identify multiple loans
Property type	580	<0.01	Used missing dummy variable
Original loan amount	619	<0.01	Deleted observations with missing values
Original property value	151,545	0.57	Deleted observations with missing values
CLTV ratio	20,545,263	77.11	Imputed second lien balances
Original credit score	4,408,347	16.55	Used missing dummy variable
DTI housing ratio	12,338,426	46.31	Used missing dummy variable
Documentation type	6,724	0.03	Used missing dummy variable
Underwriting type	3,098,767	11.63	Used missing dummy variable
Teaser rate	26,579,262	99.76	Not used in regression

CLTV = combined loan-to-value. DTI = debt-to-income.

Data Distributions and Outliers

Outliers are data elements that appear to be erroneously inputted. This analysis begins by examining the distribution of the numeric data in each year. Table A2 shows the distribution of the seven numeric variables in our model, which include original interest rate, original loan amount, original property value, CLTV, original credit score, housing DTI ratio, and original term.

Table A2. Variables Distribution From 2000 Through 2012

Variables	Mean	Median	Maximum	Quantiles								Minimum
				99%	95%	90%	75%	25%	10%	5%	1%	
Original interest rate	6.06%	6.00%	75.00%	11.25%	8.74%	7.70%	6.63%	5.25%	4.38%	3.88%	3.13%	<0.01%
Original loan amount	196,638	155,000	90,205,897	785,527	463,200	368,000	245,531	100,000	63,000	44,950	21,480	(85,389)
Original property value	269,561	200,000	99,999,999	1,200,000	665,000	509,000	325,000	131,000	89,000	70,500	44,000	0
CLTV ratio	87	90	255	102	100	99	97	80	74	62	40	21
Original credit score	714	721	974	812	801	791	767	668	626	601	548	300
DTI ratio	37	37	99	99	63	52	45	26	18	13	5	1
Original term	344	360	999	381	360	360	360	360	342	180	180	1

CLTV = loan-to-value. DTI = debt-to-income.

We observe from table A2 that the values of all variables within 1 and 99 percent appear to be reasonable during the sample period, except for the housing DTI ratio. The analysis suggests that extreme outliers account for only a very small portion of the data set.

The housing DTI ratio has some particularly high values. For example, the 99th percentile has a value of 99, indicating that the monthly housing payment is 99 percent of the borrower's monthly income. It has been observed in the mortgage industry that the reported payment-to-income ratio is not particularly reliable across different submarkets, but values as extreme as these are most likely transcription errors. For estimation purposes, we use not only the variable DTI but also two dummy variables to indicate missing DTI and outliers. The new DTI variable is equal to zero (0) if the original DTI is missing or is an outlier. Otherwise, the new DTI is equal to the original DTI.

Payment Status History Variables

Several variables that measure whether a loan is "good" or "bad" are derived from the payment history status variable in the delinquency history data set. The payment history status variable is a string variable containing 1,000 characters. Each character, in sequence, indicates the status of the loan during each month after loan origination. Thus, the payment history variable provides the status, and hence, the status transitions of each loan throughout its life, up to the current observation date. It is critical that the loan status history be complete to determine whether a loan is bad; that is, the loan has some specified negative event such as a missed mortgage payment during a time interval. We observed that many loans have missing performance statuses during their first several months. Table A3 shows the distribution of the first month when a nonmissing payment performance record appears in the data set since a loan's origination.

As table A3 shows, loan performance information has become more complete over time. In recent years, most of the first nonmissing months appear within 2 years of origination.

Table A3. Distribution of the First Month of Nonmissing Payment Performance Record

	Loan Percentage by Year				
Orig_ Year	First Nonmissing Payment History Month Equal to or Less Than 4 Months (%)	First Nonmissing Payment History Month From 5 Months to 2 Years (%)	First Nonmissing Payment Month From 2 to 5 Years (%)	First Nonmissing Payment History Month From 5 Years to Lifetime (%)	Missing Payment History (%)
2000	23.61	3.86	59.52	13.01	0
2001	28.49	4.25	60.74	6.52	0
2002	31.99	14.90	47.13	5.99	0
2003	36.05	43.84	18.65	1.46	0
2004	46.13	45.82	7.70	0.34	0
2005	76.84	16.94	6.18	0.04	<0.01
2006	79.19	19.08	1.72	0.00	<0.01
2007	88.59	10.98	0.42	0.01	0
2008	98.04	1.90	0.06	0.00	0
2009	97.40	2.60	0.00	0.00	0
2010	94.01	4.12	1.86	0.00	0
2011	96.27	2.95	0.78	0.00	0
2012	93.47	6.25	0.26	0.08	0

To improve the data's performance, we identified duplicate loans originating from servicing transfers and combined their loan performance history. Appendix B shows detailed methodology and results. After combining the performance histories of duplicitous loans, we deleted all loan observations with more than 3 months of missing performance information.

Variable Creation and Selection

Calculating CLTV and Managing Duplicate Loans for Servicing Transfers

The McDash database does not flag primary and secondary loans, nor does it dedupe multiple servicers reporting the same loans as a result of servicing transfers. We created edits to discern duplicitous reporting to create a more accurate and vetted data set. These edits work around the fact that borrowers' names and property addresses are not part of the data set. For duplicate loan records entered by multiple servicers, we combined the performance history of incomplete records to address missing payment history observations. For all loans, we created an original loan-to-value (LTV) ratio variable and a CLTV for all loans that would be identical to the LTV on loans without second liens. We used this information to determine whether loans with second loans perform better or worse than loans without second liens. Appendix B describes the methodology.

Delinquency Variables

To measure the impact of the underwriting process, we conducted analysis of early delinquency, and we apply the industry practice of 90 days delinquent in the initial 2 years. We also constructed a performance variable based on the initiation of foreclosure proceedings. Basing our analysis on the previous examination of the first nonmissing payment history month of each loan in 2 years and during the observable lifetime, we constructed the following variables:

- First delinquent month variables include the loan age in months during which a loan becomes 30, 60, 90, and 120 days delinquent for the first time. We also created variables indicating the first initiation of foreclosure proceedings, using the same method based on the payment history in the loan's lifetime.

- Number of delinquent months variables are derived by counting the number of months when each of the six delinquent statuses is observed during the first 2-year and 3-year periods after a loan's origination.
- Default episode month variables are defined in terms of beginning and ending months of delinquency. The first default episode beginning month is defined as the first month when a loan becomes 90 days delinquent, and the first default episode ending month is defined as the last month before the loan cures to the current payment status after the first default episode beginning month. The same algorithm is used to construct the second, third, fourth, and fifth default episode beginning and ending months.

Macroeconomic Variables

We prepared four macroeconomic variable data sets for model use. The first data set contained the monthly Federal Housing Finance Agency purchase-only house price index, with the Core Based Statistical Area (CBSA) codes for metropolitan areas, metropolitan divisions, and state- or national-levels. The second data set contained monthly interest rates, including 1-year LIBOR, 1-year Department of the Treasury rates, 10-year Treasury rates, and the 30-year fixed-rate mortgage (FRM) rates. The third data set contained monthly unemployment rates, with the CBSA codes for metropolitan area, metropolitan division, and state- or national-levels. The fourth data set contained the state-level census median housing prices for single-family houses and the national-level for condominiums. All the economic data come from Moody's Analytics.[3]

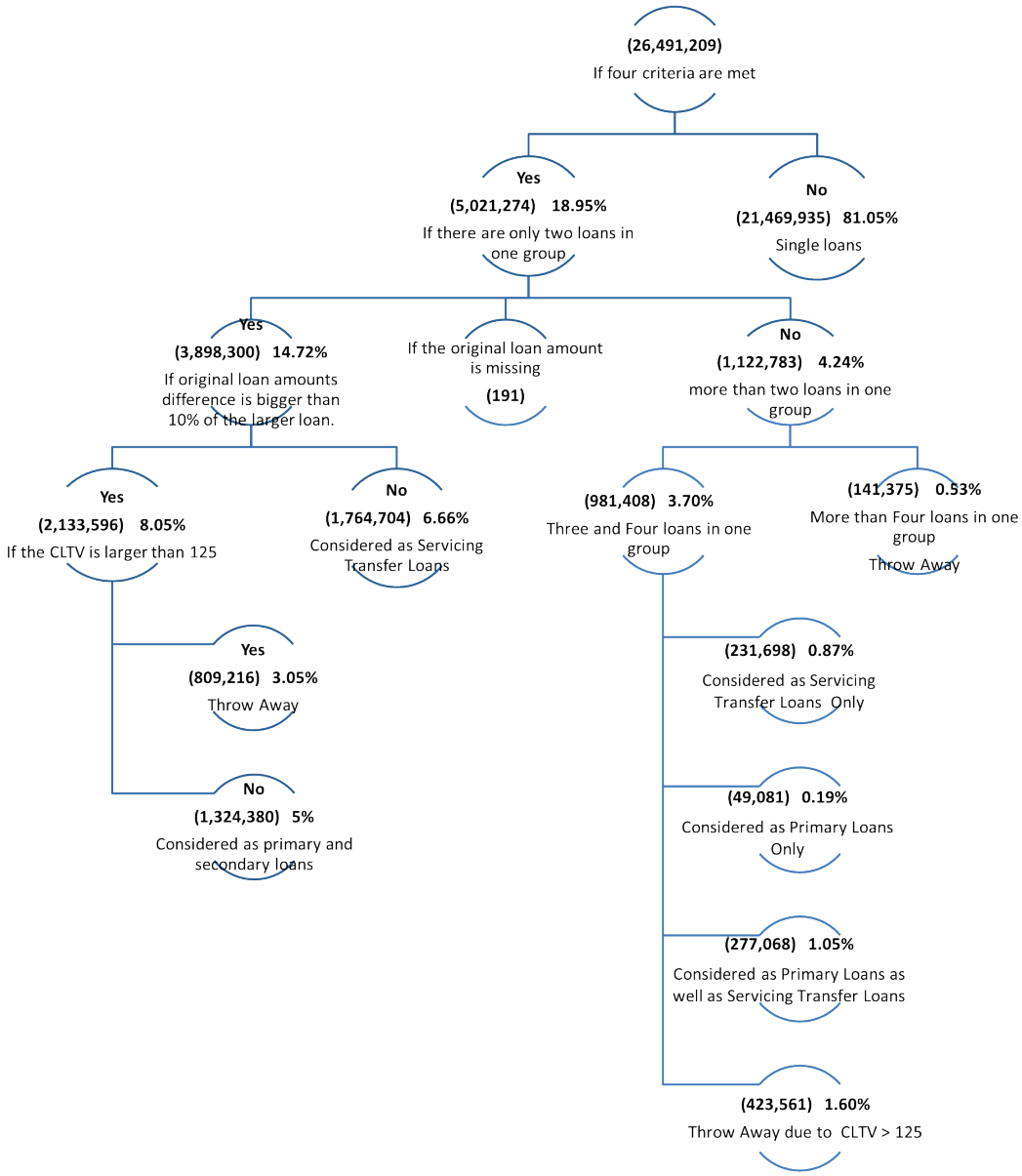
[3] http://www.economy.com.

Appendix B. Duplicate Observations

Figure B1 details the steps used to deal with suspected duplicate observations for loans on the same property. Approximately 26.5 million loans originated from 2000 through 2012. We did not have borrower names or property addresses, but we suspected, in many cases, that two or more of the individually reported loans were in fact on the same house, and that these multiple reports were because of either (1) one or more servicing transfers, making it the same loan reported by multiple servicers, or (2) at least one loan was a second lien. Approximately 5 million loans had the same origination month, ZIP Code, property type identification, and original property value as at least one other loan observation. These four criteria are cited in the first row of figure B1. If the original loan amounts had less than a 10-percent difference, we assumed that these observations are due to servicing transfers. A minimal difference in original loan amount may occur when the new servicer reports the amortized loan amount as of the servicing transfer date instead of the loan amount at the origination date. We combined the loan payment status histories for these loans and deleted one of the two loan observations.

If the difference in the original loan amount was larger than 10 percent, we assumed that these loans are two different loans on the same property. The loan with the higher original loan amount was identified as the primary loan, and we used only the primary loan performance for analysis; if a foreclosure of the second lien triggered a foreclosure of the first, this showed up in the performance status of the primary loan. Loans with lower original loan amounts were specified as second liens. We then derived the CLTV by dividing the combined loan amount by the original property value. Loans with a CLTV larger than 125 percent were excluded from our analysis as outliers. About 3.5 million loans identified as second loans or duplicate servicer-transfer loans were excluded from the analysis data set.

Figure B1. Observations Identified as Combined Loans or Servicing Right Transfer Loans



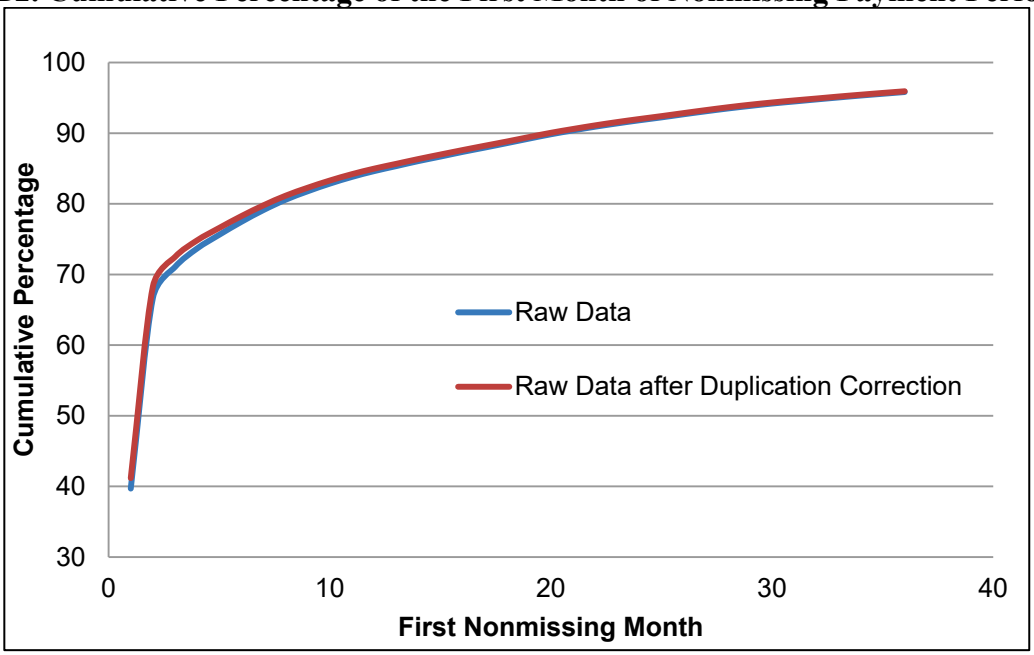
After removing second and duplicate loans, we identified 23,189,377 loans as valid that we used for our analysis. Table B1 replicates table A3 after the deduping procedure described previously.

Table B1. Distribution of the First Month of Nonmissing Payment Performance Record

	Loan Percentage by Year				
Origination Year	First Nonmissing Payment History Month Equal to or Less Than 4 Months (%)	First Nonmissing Payment History Month From 5 Months to 2 Years (%)	First Nonmissing Payment Month From 2 to 5 Years (%)	First Nonmissing Payment History Month From 5 Years to Lifetime (%)	Missing Payment History (%)
2000	23.98	3.84	59.79	12.38	0.00
2001	28.99	4.16	60.96	5.89	0.00
2002	32.63	14.78	47.23	5.36	0.00
2003	37.04	43.75	18.12	1.09	0.00
2004	47.47	45.52	6.73	0.27	0.00
2005	79.00	15.51	5.44	0.05	<0.01
2006	81.20	17.34	1.46	0.00	<0.01
2007	89.81	9.74	0.44	0.01	0.00
2008	98.08	1.86	0.06	0.00	0.00
2009	97.39	2.61	0.00	0.00	0.00
2010	95.35	4.11	0.55	0.00	0.00
2011	98.27	1.23	0.50	0.00	0.00
2012	94.47	5.26	0.26	0.01	0.00

Compared with table A3, the percentage of the first nonmissing payment history month that is equal to or less than 4 months increases, after combining the payment history of servicing transfer loans. Because we measured 90-day delinquencies, we excluded any loan observations with the first nonmissing payment history that is more than 4 months to avoid uncertain information.

Figure B2. Cumulative Percentage of the First Month of Nonmissing Payment Performance



From the previous chart and after applying the matching methodology, the accumulated percentage of the first month of nonmissing performance statuses is larger than before in earlier cohorts due to the combination of payment status histories among servicing transfer loans.

Appendix C. Outlier Exclusions

Table C1. Outlier Control Process

Outlier Condition for Lifetime Analysis	Number of Observations Excluded	Percentage	Cumulative Percentage	Outlier Condition for 2-Year Analysis	Number of Observations Excluded	Percentage	Cumulative Percentage
Any loan with missing payment history status more than 3 months	4,483,989	23.46	23.46	Any loan with missing payment history status more than 3 months	4,483,989	23.46	23.46
Any loan with original interest rate less than 1% or more than 25%	35	0.00	23.46	Any loan with original interest rate less than 1% or more than 25%	35	0.00	23.46
Any loan with original term less than 60 months or more than 480 months	5,204	0.03	23.49	Any loan with original term less than 60 months or more than 480 months	5,204	0.03	23.49
Any loan with original loan amount of more than $2,000,000 or less than $10,000	10,162	0.05	23.54	Any loan with original loan amount of more than $2,000,000 or less than $10,000	10,162	0.05	23.54
Any loan with original property value of more than $3,000,000 or less than $10,000	21,574	0.11	23.65	Any loan with original property value of more than $3,000,000 or less than $10,000	21,574	0.11	23.65
Any loan with LTV more than 125% or less than 20%	36,240	0.19	23.84	Any loan with LTV more than 125% or less than 20%	36,240	0.19	23.84
Any loan with CLTV more than 125% or less than 20%	2,547	0.01	23.86	Any loan with CLTV more than 125% or less than 20%	2,547	0.01	23.86

Any loan with original credit score higher than 850	2,394	0.01	23.87	Any loan with original credit score higher than 850	2,394	0.01	23.87
Any loan with missing CLTV	48	0.00	23.87	Any loan with missing CLTV	48	0.00	23.87
Any loan with origination year of 2010 or later is excluded for the lifetime analysis	3,300,676	17.27	41.14				
Any loan with payment status history of permanently missing is excluded for lifetime analysis	684,174	3.58	44.72				
				Any loan with payment status history of permanently missing within 2 years.	190,699	1.00	24.87

CLTV = combined loan-to-value. LTV = loan-to-value.

Appendix D. Diagnostics Charts

Here, we show the diagnostic charts for categorical and continuous variables that help to assess the appropriateness of the functional form of the variable. The left-hand charts are for 90-day delinquencies in the first 2 years, and the right-hand charts are for lifetime foreclosures.

Figure D1. Diagnostics Charts for Loan Type

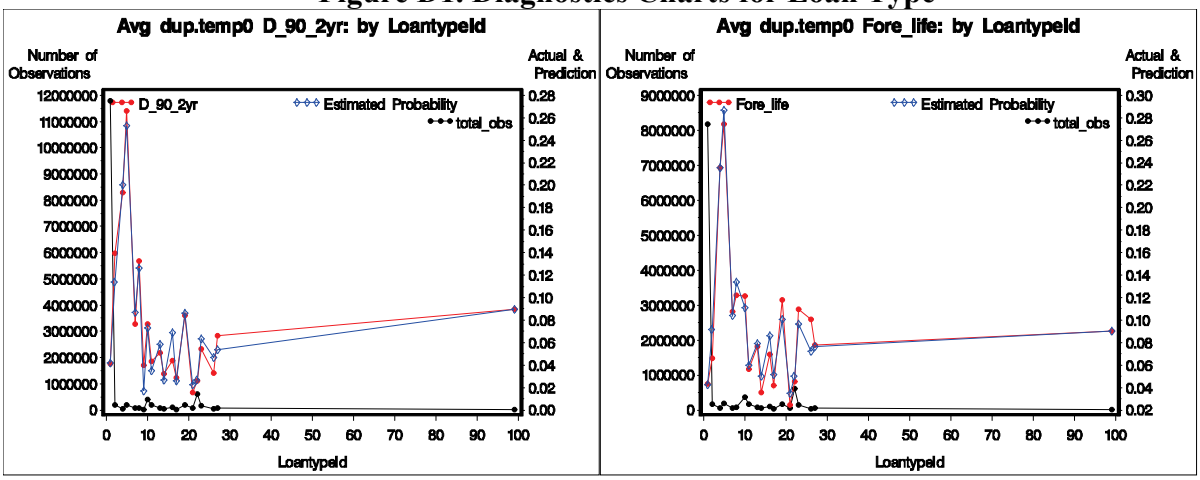


Figure D2. Diagnostics Charts for Product Type

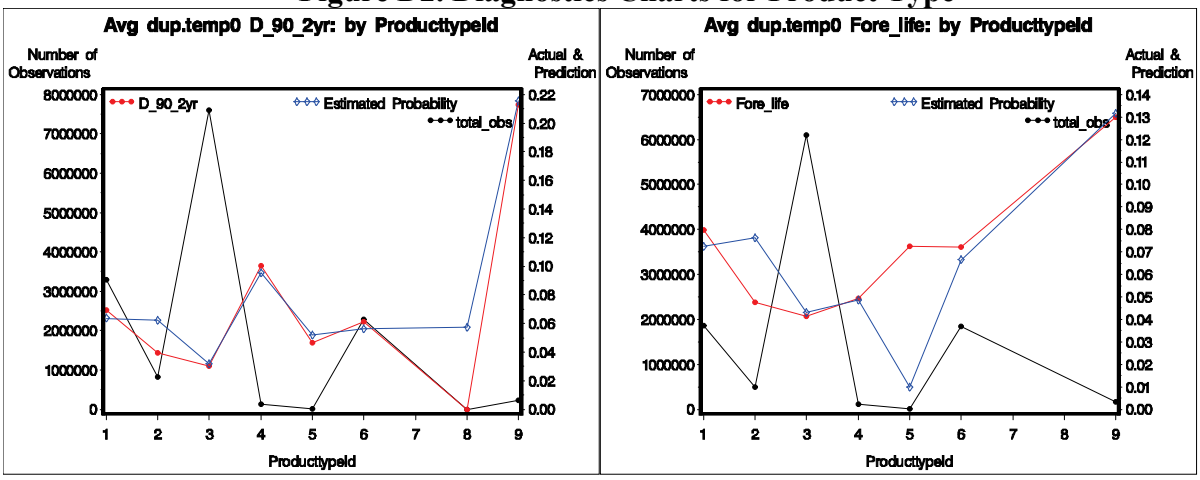


Figure D3. Diagnostics Charts for Original Term

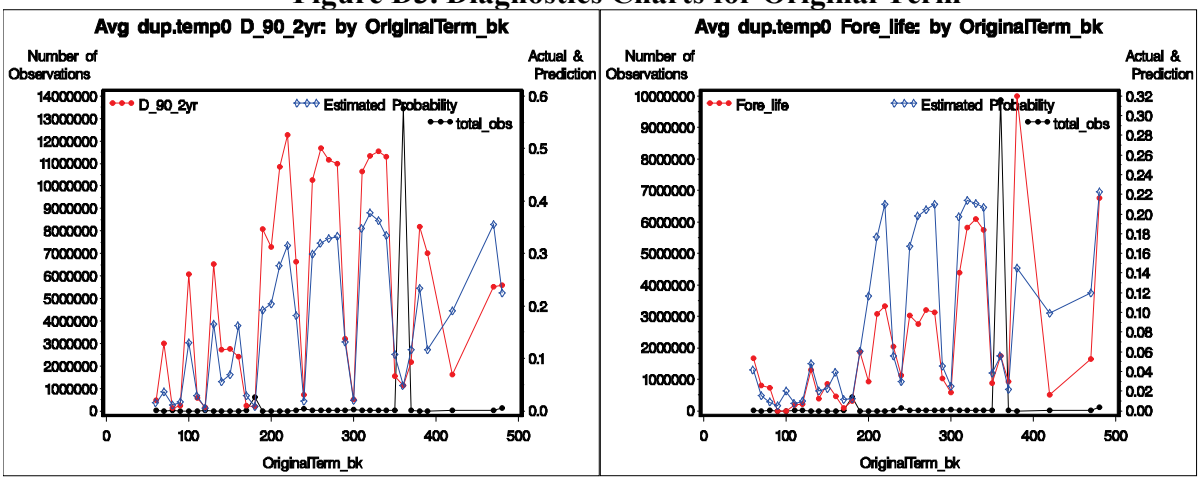


Figure D4. Diagnostics Charts for Payment and Interest Frequency

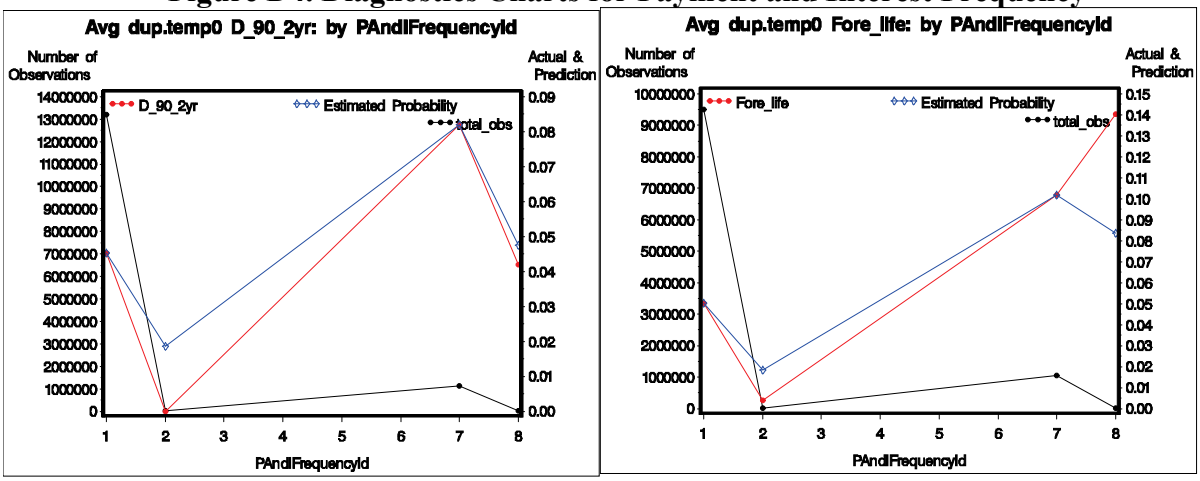


Figure D5. Diagnostics Charts for Occupancy

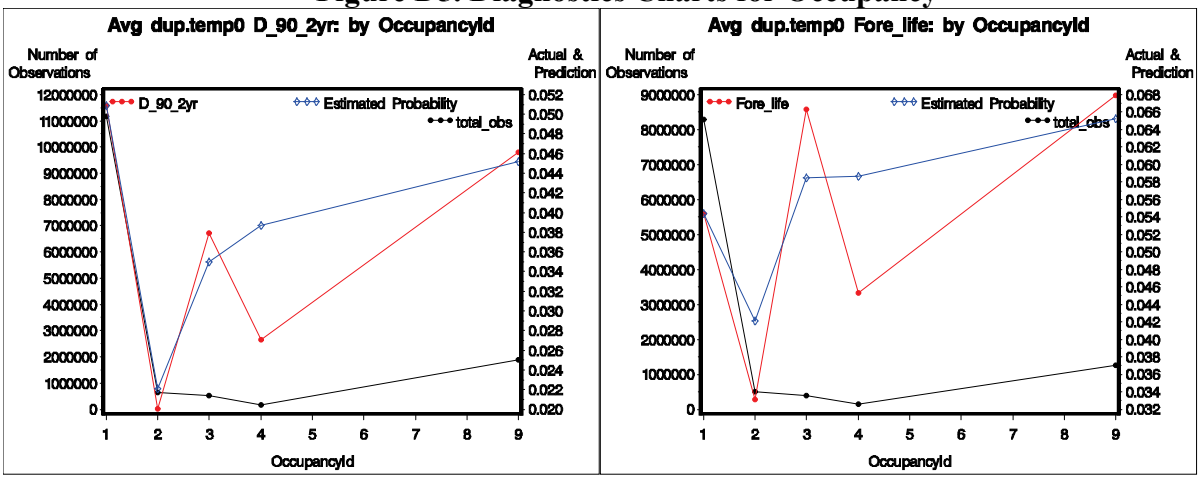


Figure D6. Diagnostics Charts for Prepayment Penalty Clause

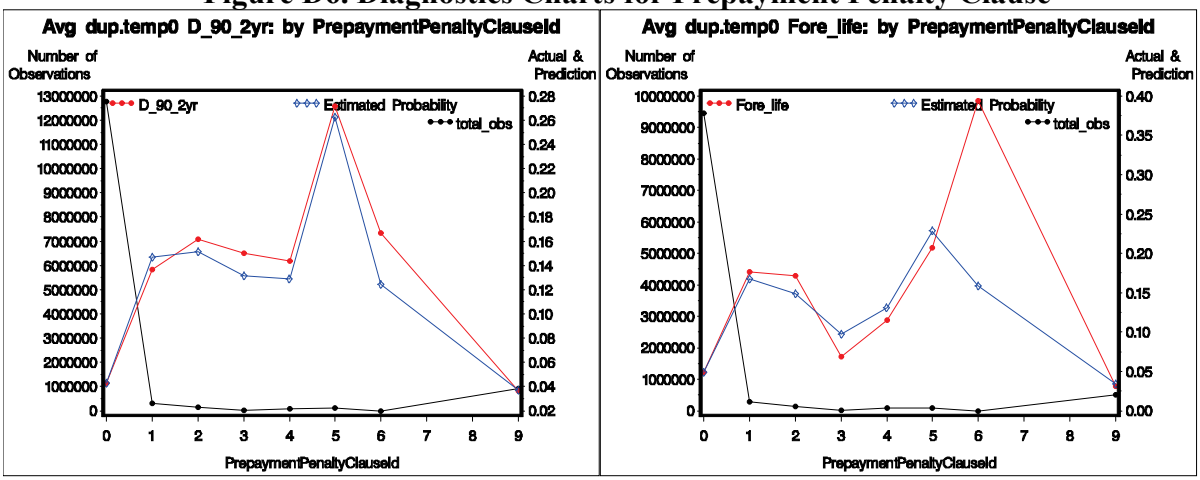


Figure D7. Diagnostics Charts for Documentation

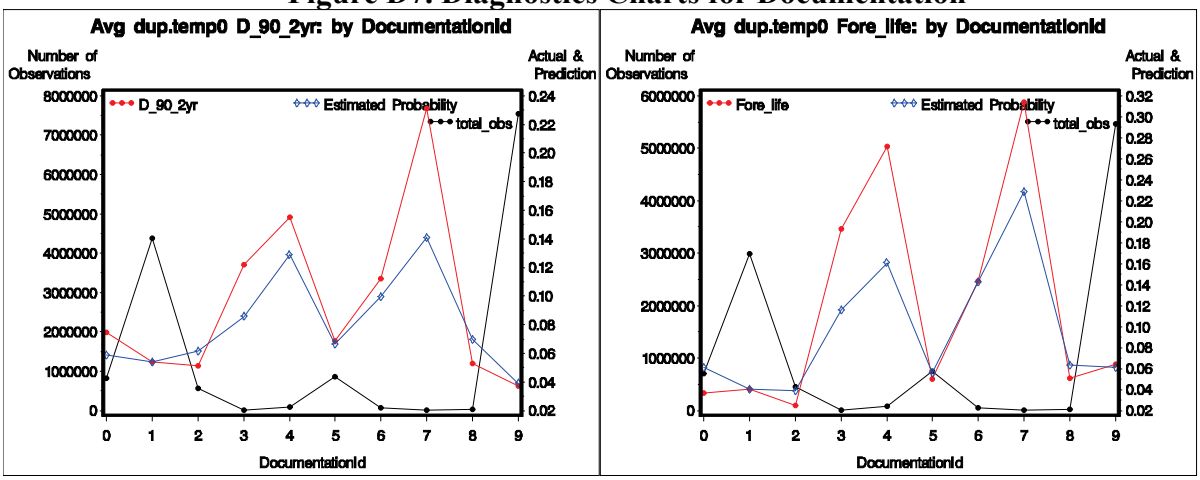


Figure D8. Diagnostics Charts for First Nonmissing Month

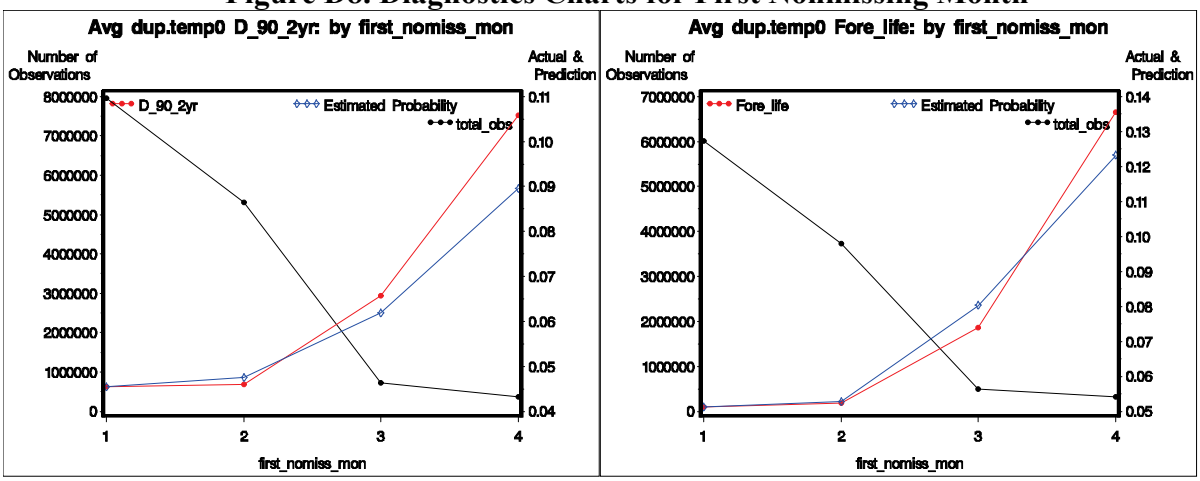


Figure D9. Diagnostics Charts for Original Credit Score

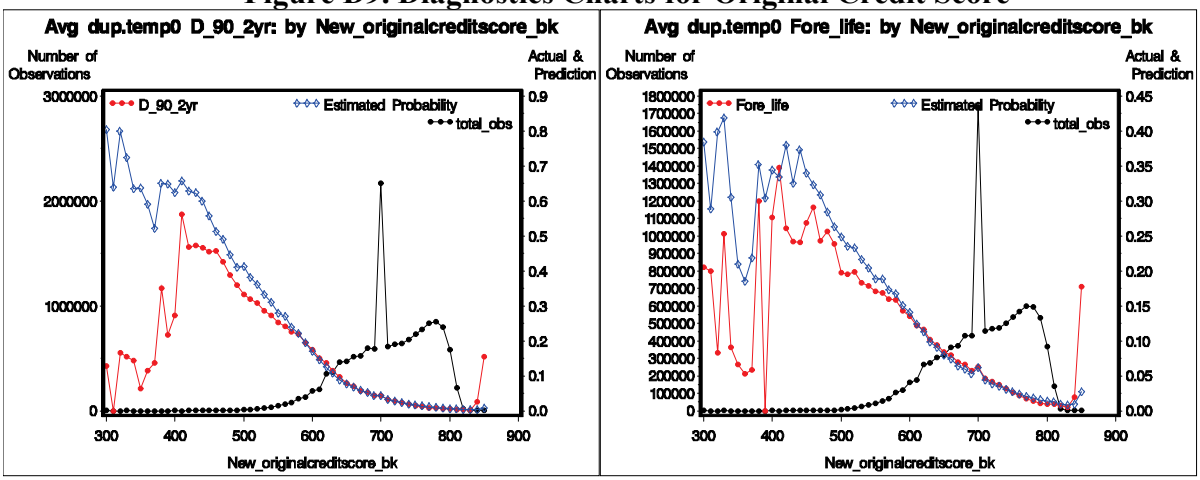


Figure D10. Diagnostics Charts for Relative Property Value

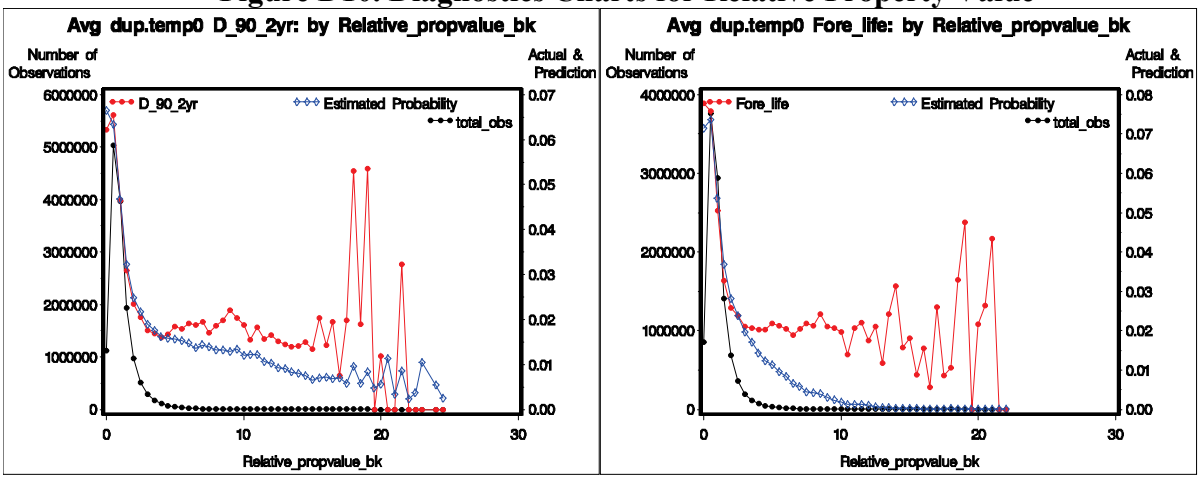


Figure D11. Diagnostics Charts for Combined Loan-to-Value

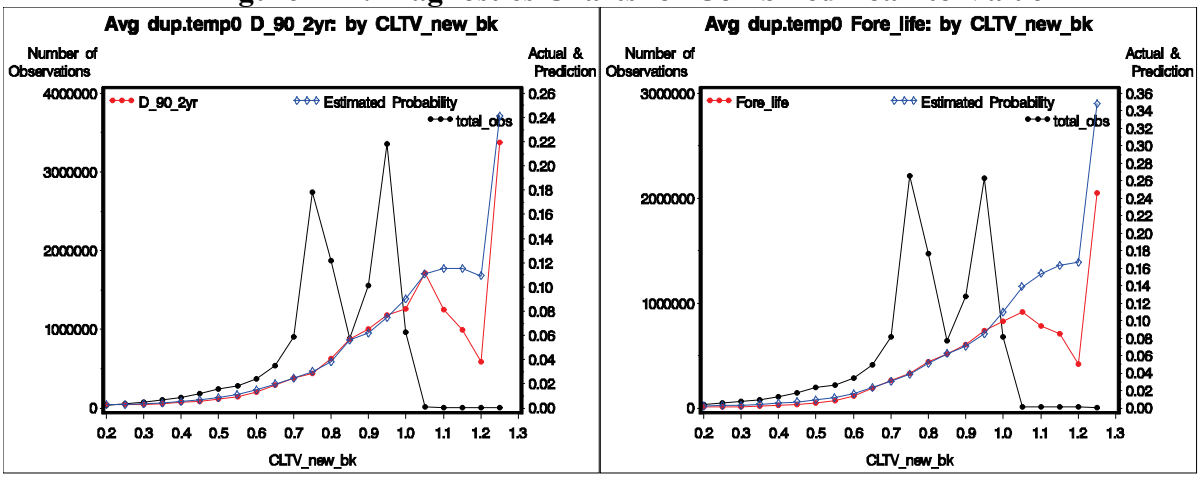


Figure D12. Diagnostics Charts for Yield Curve Slope

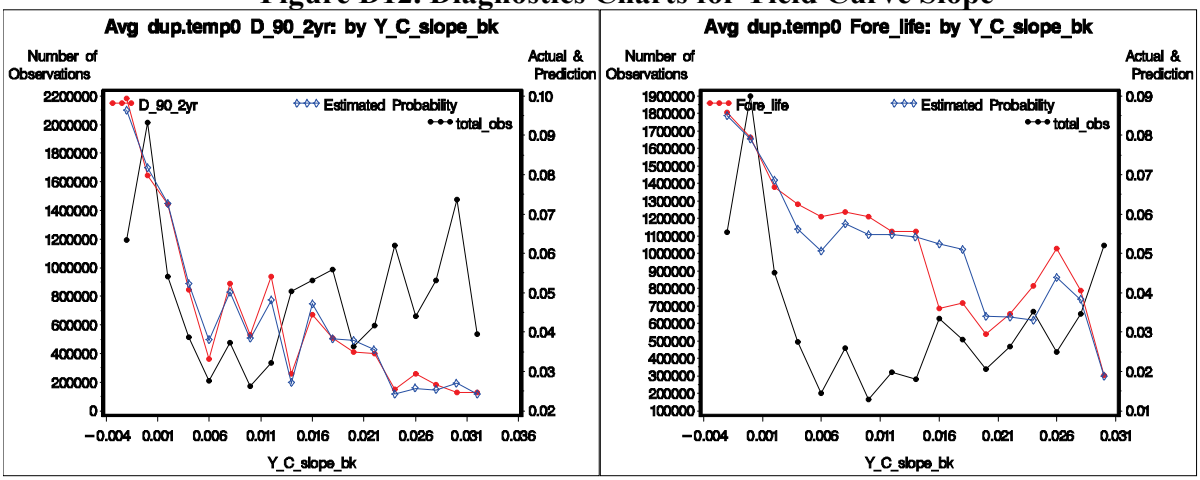


Figure D13. Diagnostics Charts for 30-Year Fixed-Rate Mortgage Rate

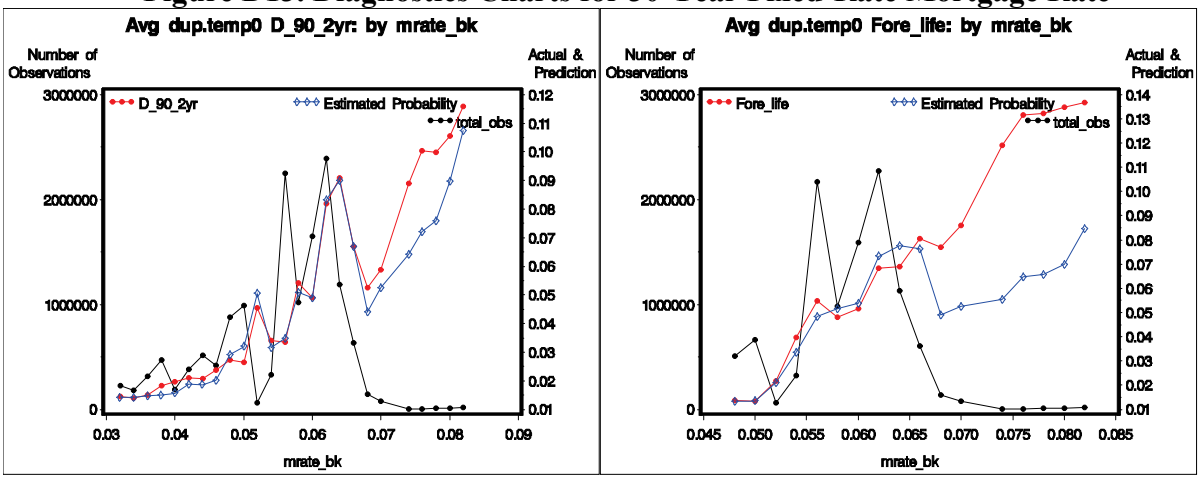
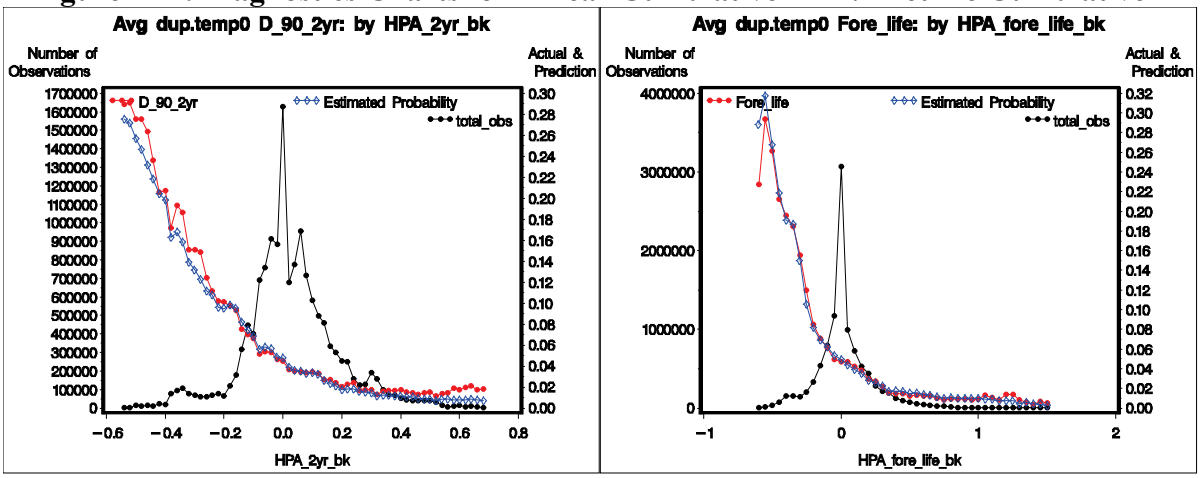
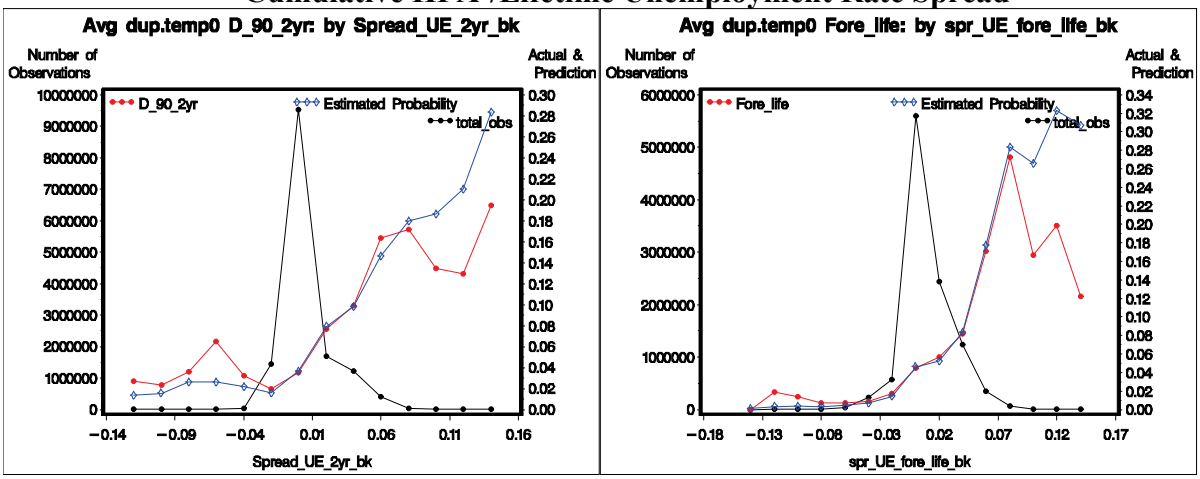


Figure D14. Diagnostics Charts for 2-Year Cumulative HPA/Lifetime Cumulative HPA



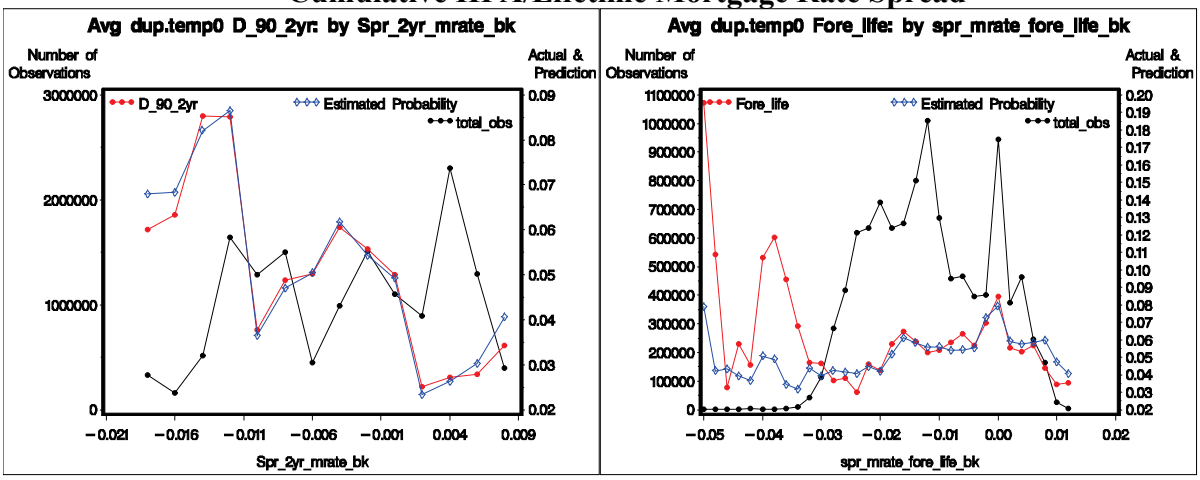
HPA = house price appreciation.

Figure D15. Diagnostics Charts for 2-Year Cumulative HPA†/Lifetime Unemployment Rate Spread



HPA = house price appreciation.

Figure D16. Diagnostics Charts for 2-Year Cumulative HPA/Lifetime Mortgage Rate Spread



HPA = house price appreciation.

Figure D17. Diagnostics Charts for Debt-to-Income Ratio

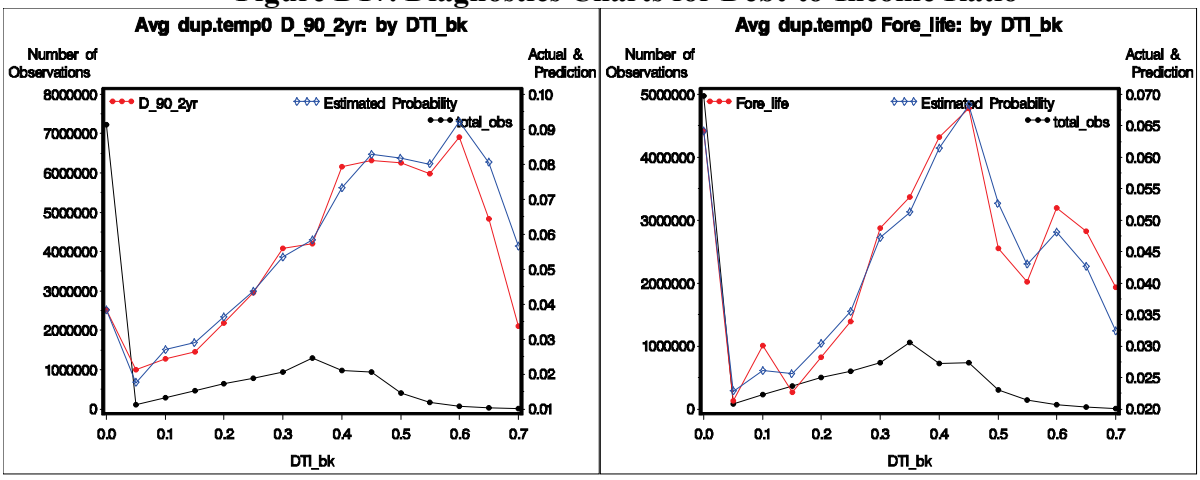
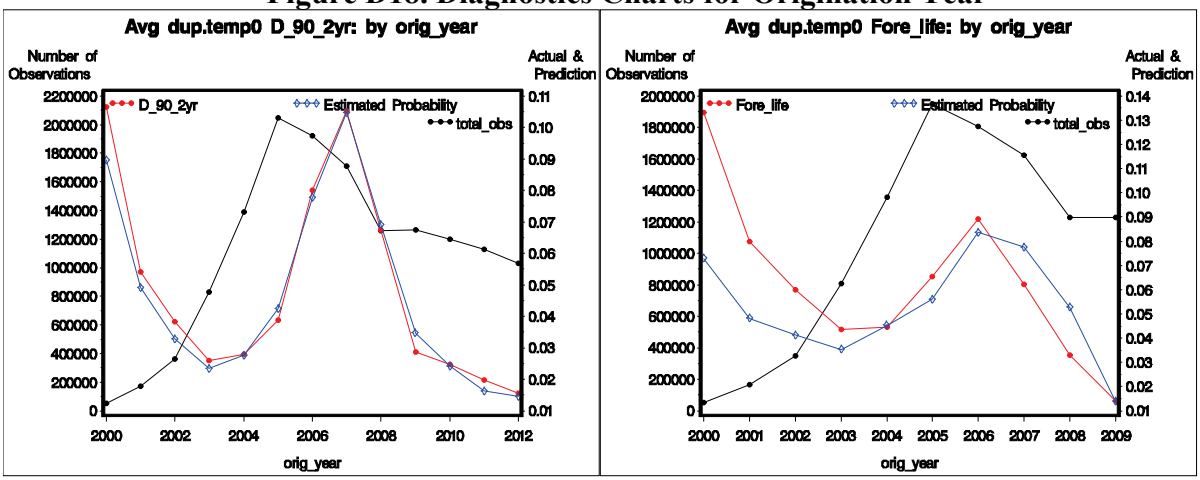


Figure D18. Diagnostics Charts for Origination Year



Appendix E. CLTV Analytics for the Delinquency Model

In this appendix, we present analytics to illuminate the combined loan-to-value (CLTV)-2-year 90-day delinquency relationship, to illustrate the delinquency equation. These analytics show the delinquency probability-CLTV relationship when specific, interesting explanatory variables change. In these analyses, all other variables are set at their median values in the data set.

Figure E1. Combined Loan Effect on Delinquency Probability

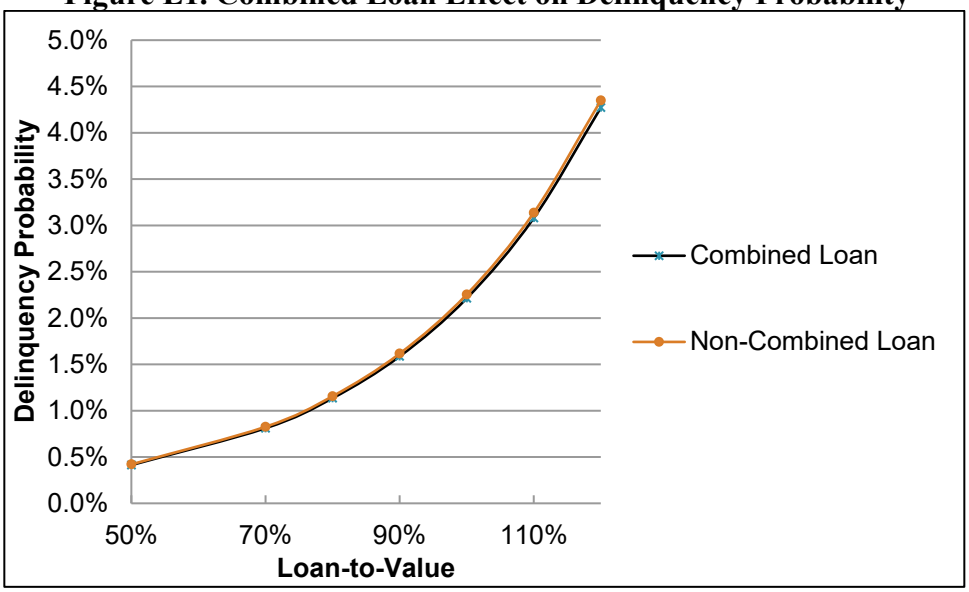


Figure E2. Jumbo Loan Effect on Delinquency Probability

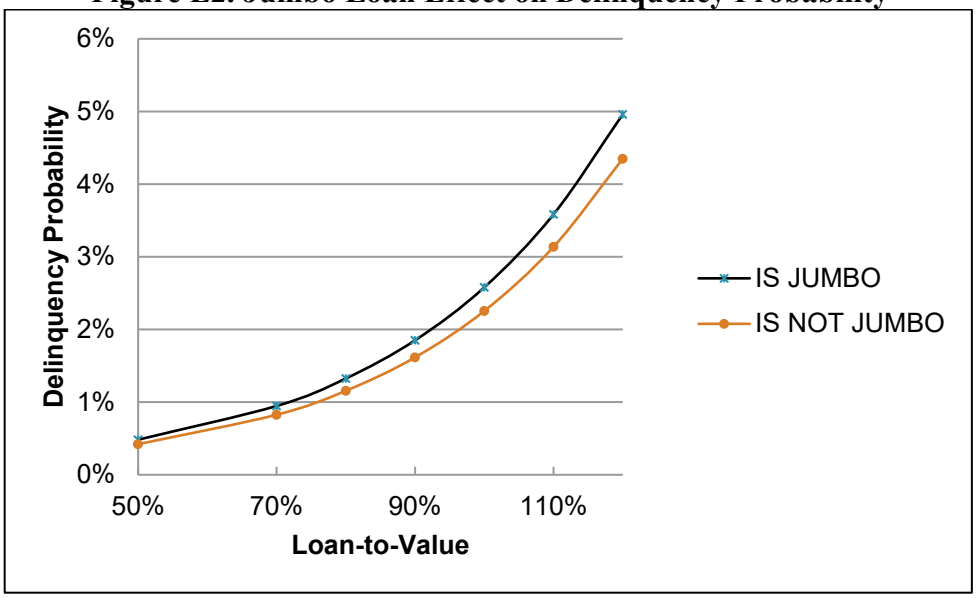


Figure E3. Full Documentation Effect on Delinquency Probability

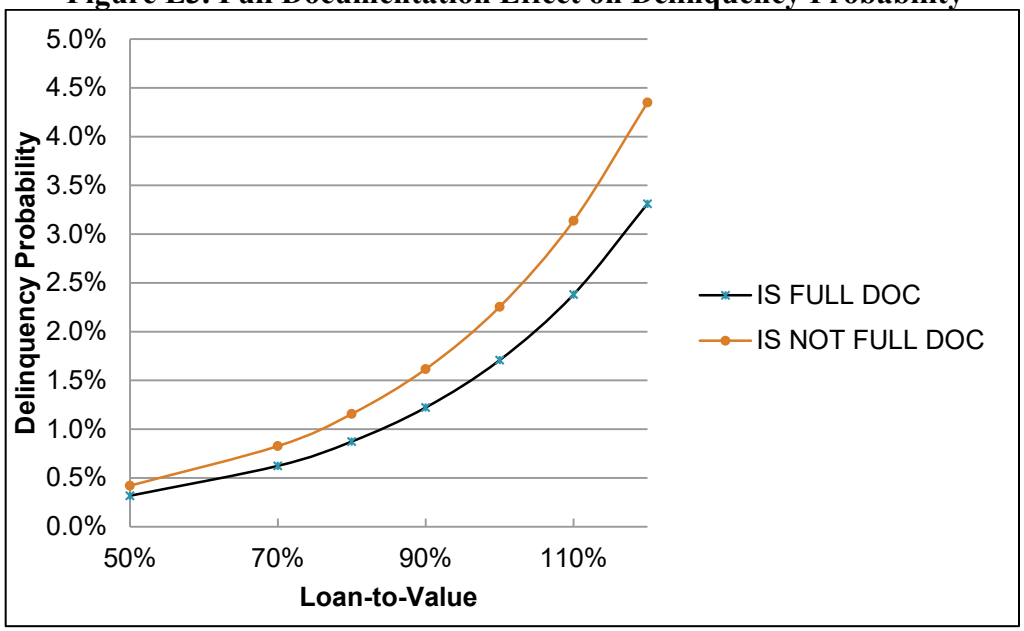


Figure E4. B and C Loan Effect on Delinquency Probability

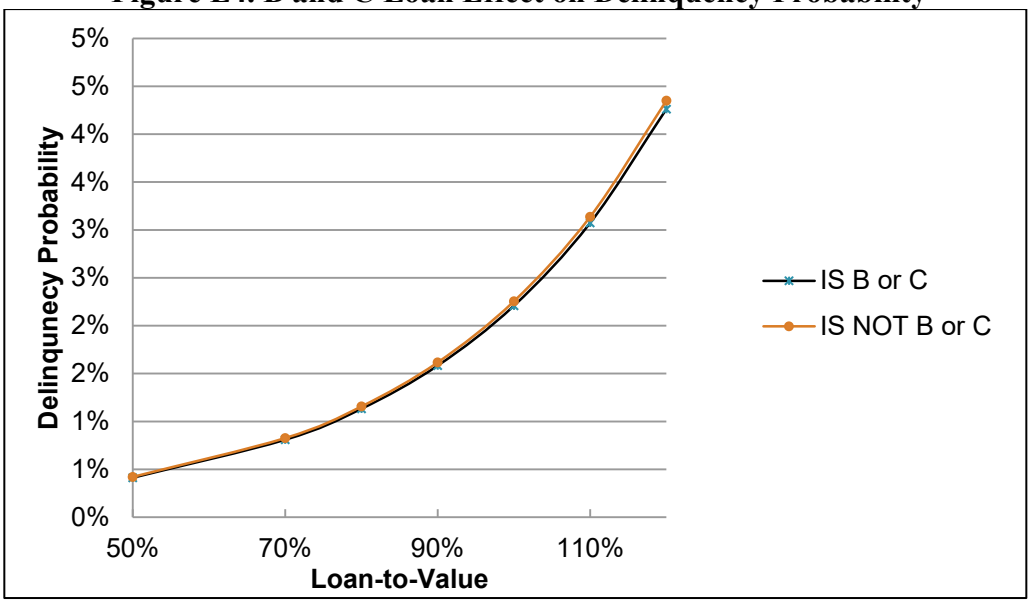


Figure E5. Relative Property Value Effect on Delinquency Probability

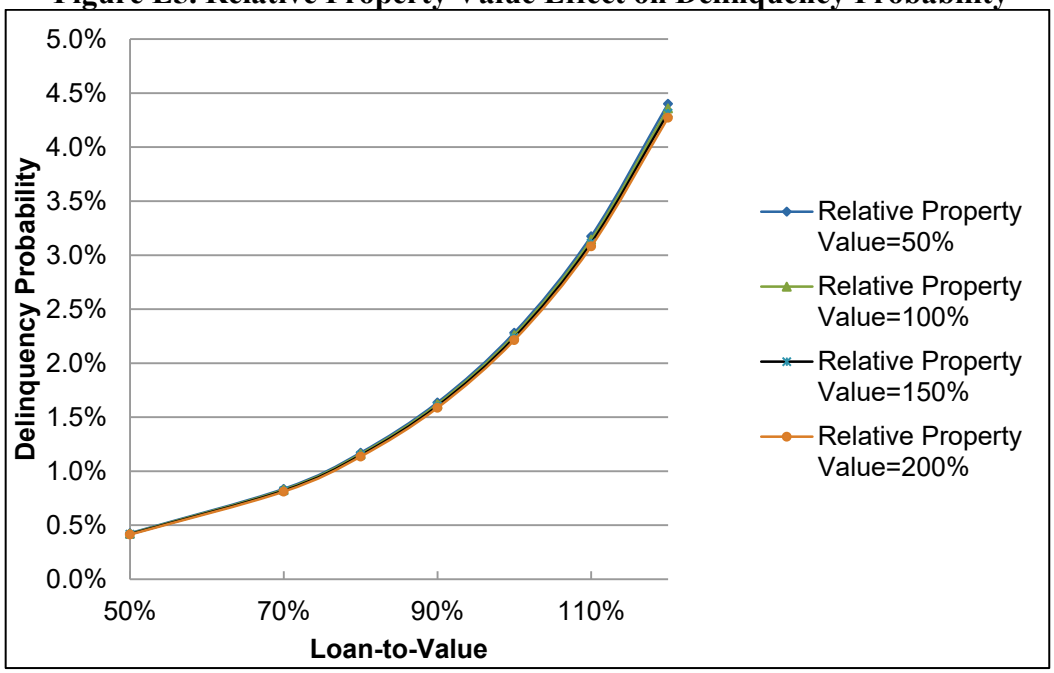
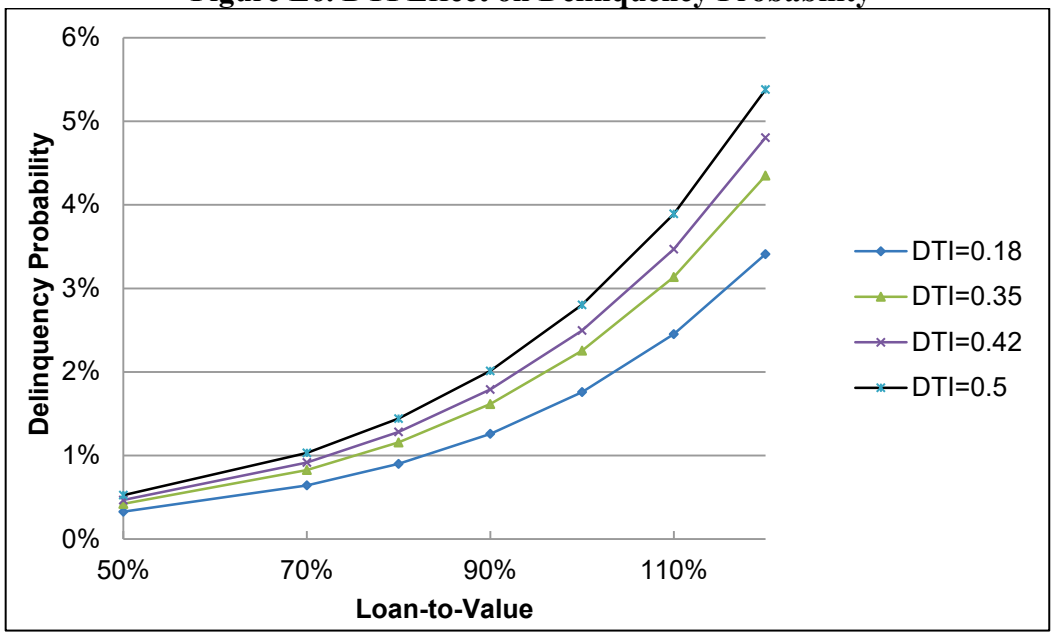


Figure E6. DTI Effect on Delinquency Probability



DTI = debt-to-income.

Figure E7. FICO Effect on Delinquency Probability

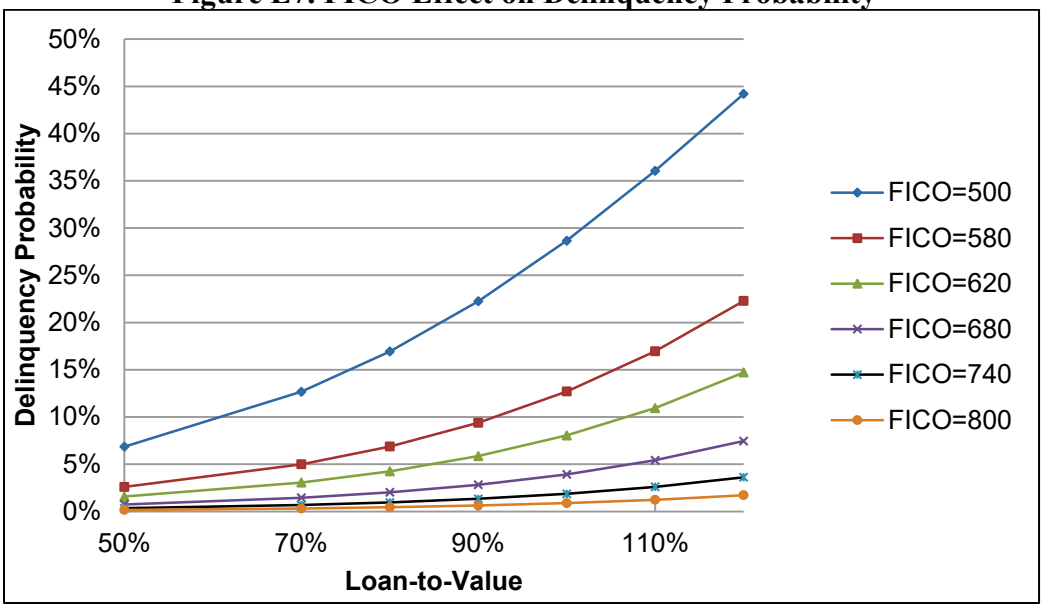
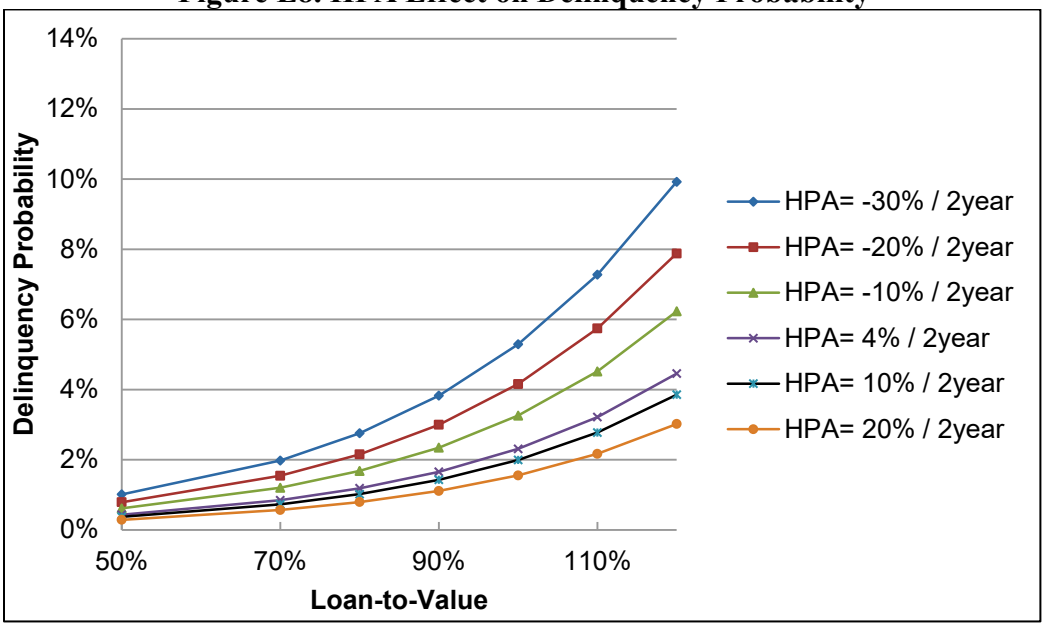
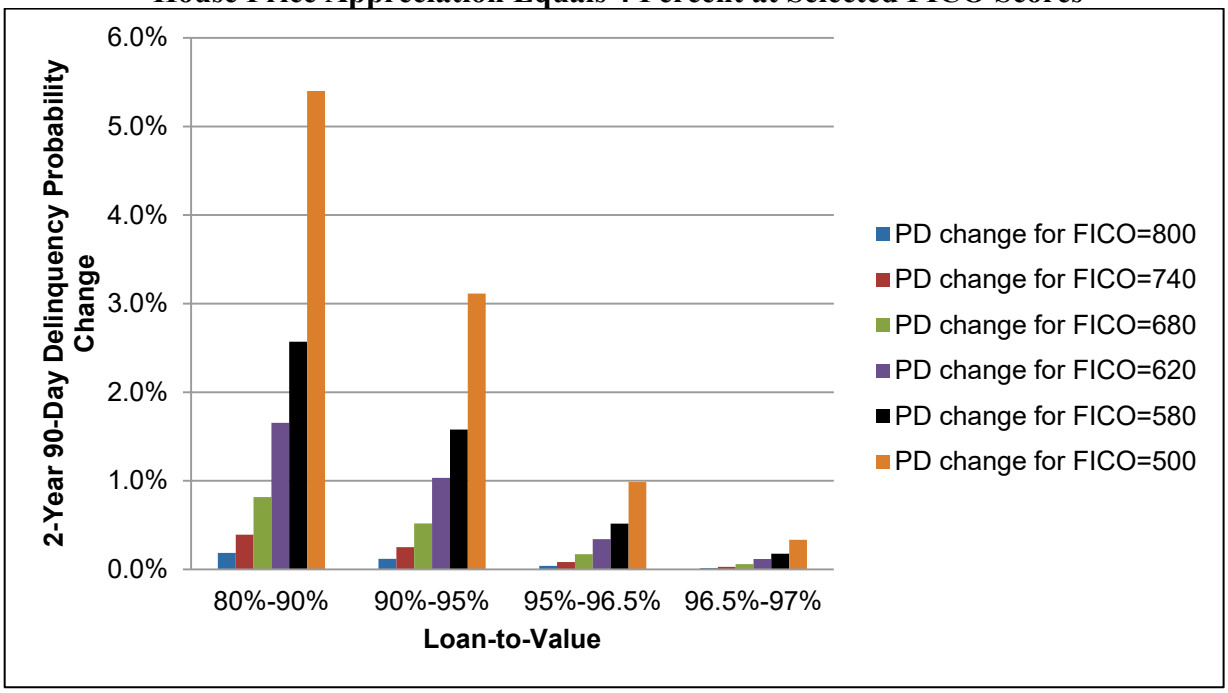


Figure E8. HPA Effect on Delinquency Probability



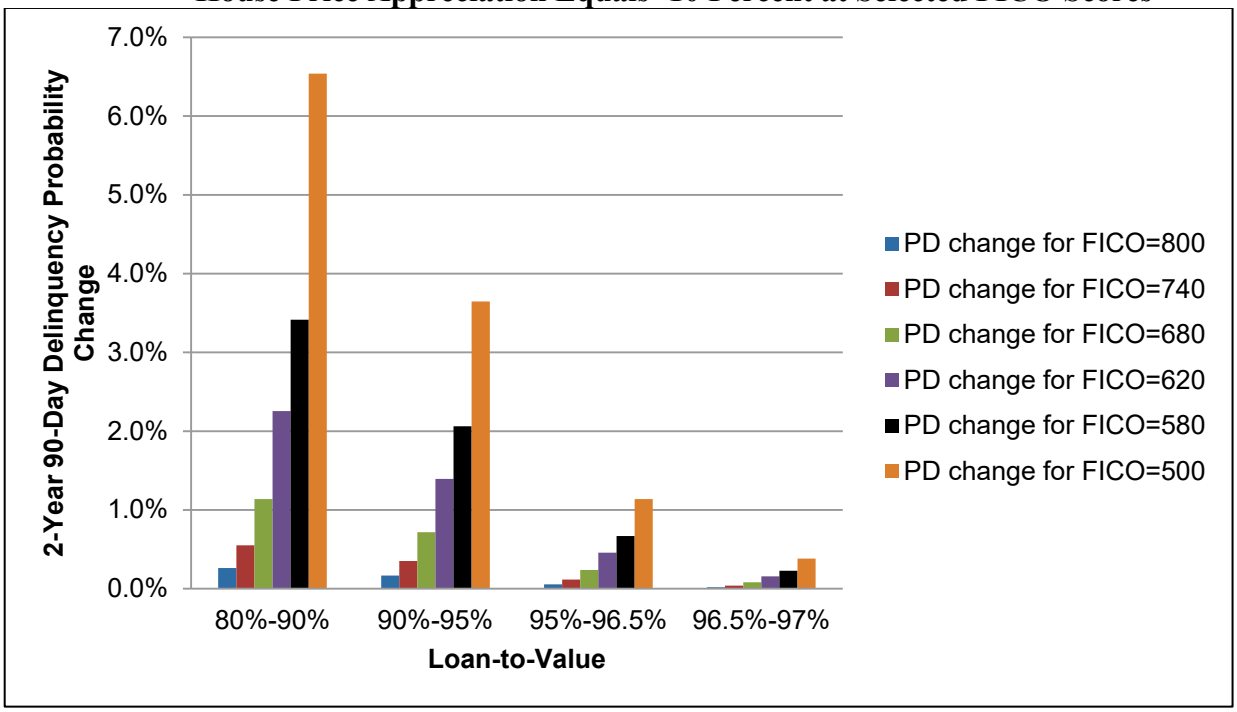
HPA = house price appreciation.

Figure E9. 2-Year 90-Day Delinquency Probability Change at House Price Appreciation Equals 4 Percent at Selected FICO Scores



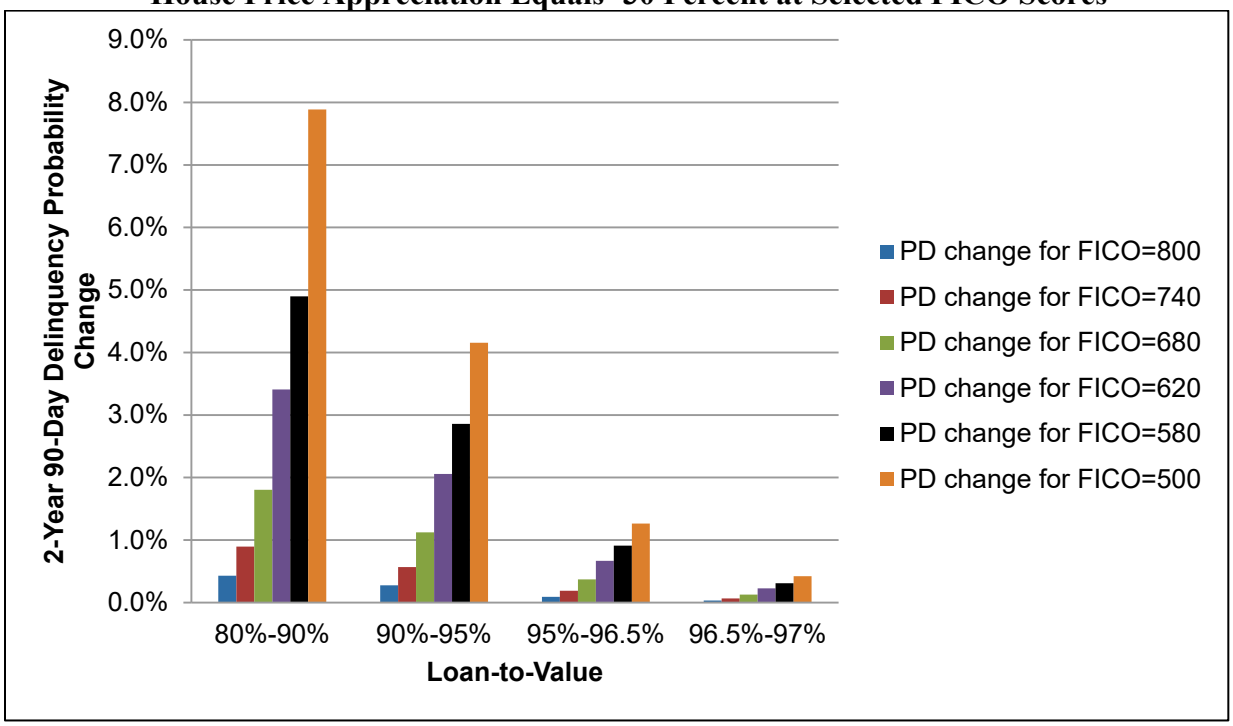
PD = probability of default.

Figure E10. 2-Year, 90-Day Delinquency Probability Change at House Price Appreciation Equals -10 Percent at Selected FICO Scores



PD = probability of default.

Figure E11. 2-Year, 90-Day Delinquency Probability Change at House Price Appreciation Equals -30 Percent at Selected FICO Scores



PD = probability of default.