Received: 7 August 2021 | Revised: 17 June 2022 | Accepted: 6 August 2022

DOI: 10.1111/irfi.12392

ORIGINAL ARTICLE

INTERNATIONAL REVIEW OF FINANCE WILEY

Impact of mortgage soft information in loan pricing on default prediction using machine learning

Thi Mai Luong | Harald Scheule | Nitya Wanzare

Finance Discipline Group, UTS Business School, University of Technology Sydney, Sydney, New South Wales, Australia

Correspondence
Harald Scheule, Finance Discipline Group, UTS Business School, University of Technology Sydney, PO Box 123, Broadway, Sydney, NSW 2007, Australia.
Email: harald.scheule@uts.edu.au

Funding information
Australian Prudential Regulation Authority; Brian Gray Scholarship of the Australian Prudential Regulation Authority; Hong Kong Institute for Monetary Research

Abstract

We analyze the impact of soft information on US mortgages for default prediction and provide a new measure for lender soft information that is based on the interest rates offered to borrowers and incremental to public hard information. Hard and soft information provide for a variation in annual default probabilities of approximately 3%. Soft information has a lesser impact over time and time since origination. Lenders rely more on soft information for high-risk borrowers. Our study evidences the importance of soft information collected at loan origination.

KEYWORDS
credit risk, default, hard information, lending, mortgage, prediction, pricing, soft information, yield spreads

JEL CLASSIFICATION
G01, G20, G21, C51, C55

1 | MOTIVATION

Mortgage lending plays a crucial role in financial markets and accounts for a high proportion of commercial banks' balance sheets.[1] Home ownership and finance play key roles in consumers' lives. However, mortgage lending was found to be a critical cause of the Global Financial Crisis of 2008–2009. Financial institutions rely

on credit scoring of borrowers to determine their ability to repay debt for loan pricing at origination and refinance, loan provisioning and calculating bank capital requirements.

Lenders include both "hard" and "soft" information to assess the likelihood and magnitude of future loan losses and then make lending decisions. We follow the definition by Stein (2002) in our paper and define hard information as information that is measurable, digitally stored, and publicly verifiable by all lenders. Such information is easily incorporated into scoring models and applied to a large number of borrowers. On the other hand, individual lenders collect private soft information based on personal interactions, customer visits, and trust relationships between lenders and customers. This kind of information is difficult to quantify and verify but remains significant in credit assessment processes.

To illustrate, a borrower submits her mortgage loan application to a lender and loan officers analyze hard information such as the FICO score, debt-to-income (DTI) ratio and loan-to-value (LTV) ratio. They further collect soft information from interviews or other sources, such as transactional records and determine lending terms. Soft information may result in the rejection (acceptance) of a loan application if the loan officer believes that the borrower is riskier (less risky) than a threshold. In addition, the assessment of risk will be embedded in the pricing process, whereby customers may have to pay a higher interest rate for a higher risk. It is apparent that soft information gathered by lenders adds a human touch to the approval and pricing process.[2]

This paper provides a new measure for lender soft information based on the interest rates offered to borrowers incremental to measurable hard information. We rely on the credit spreads embedded in loan rates offered to borrowers following the accept decision of lenders. Through the credit spread channel, we identify and measure the net position of soft information in which the more adverse the soft information is, the higher the credit spread on loans is, and hence the higher the default likelihood is. This approach is novel as it defines soft information as the variation after controlling for hard information. It includes soft information that may be observed through proxy variables such as geographic distance between borrowers and lenders as well as soft information that is not observed. Further, this paper also tests the additional accuracy of our soft information measure in predicting credit risk.

The econometric contributions of this approach are twofold. First, we are able to measure unobserved soft information. Prior literature considers observed soft information via proxy variables but has not considered soft information that is not observed by such variables. Second, soft information is measured in our paper on a continuous scale and on an interpretational level of the credit spread. A higher value corresponds to a higher credit spread. Prior literature has confirmed the existence of soft information but not the degree to which soft information predicts credit risk.

Relative to the existing finance literature, we make the following contributions in this paper. First, we provide evidence that soft information is predictive for mortgage default risk. Hard and soft information provide for a variation in annual default probabilities of 3%. Second, we find that soft information is less predictive over time and time since origination. This may signal a shrinkage of soft information collection and depreciation since loan origination. Third, we document that lenders rely more on soft information for high-risk borrowers as more soft information is collected and priced for borrowers when information is more binding as information has a greater sensitivity on default risk.

This paper is organized as follows, Section 2 reviews related literature. Section 3 outlines the model framework. Section 4 describes the data used, main empirical analysis and robustness checks. Section 5 provides an economic impact analysis and Section 6 discusses our findings and concludes.

2 | LITERATURE REVIEW

The existing literature supports the importance of soft information in business and consumer lending. There is a vast literature on relationship lending to firms.[3] Examples include Stein (2002), DeYoung et al. (2008), Degryse and Van Cayseele (2000), Chakraborty and Hu (2006), and Brick and Palia (2007). For example, Stein (2002) argues for the

importance of soft information on lending for small to medium firms. This is due to the relationship between the firm's executives and lenders. Berger et al. (2005) find that smaller banks collect more soft information.

The creation of soft information underpins the relationship between a lender and their borrowers. A key measure for this relationship is geographic distance. Agarwal and Hauswald (2010) as well as Berger and Udell (2002) show that borrower proximity facilitates the collection of soft information. Petersen and Rajan (2002) find that consolidation in the banking industry makes geographic distance a less precise measure as the location of the ultimate parent (i.e., bank holding company) may not be representative for the location which collects soft information (i.e., bank branch).

Literature on soft information in mortgage lending is sparser and more recent. Examples include Ergungor and Moulton (2014) who use distance to the nearest branch, bank size, and bank deposit share as proxies. The authors find that borrowers who receive a loan from a local bank are less likely to default on their loan as opposed to receiving a loan from other banks. Agarwal et al. (2011) analyze the lender decisions in a dynamic contract setting where a loan application undergoes a secondary screening (collection of soft information) and contract terms are dynamically adjusted. Saengchote (2013) analyses distance, broker competition, and regulation. Furthermore, Rajan et al. (2015) show that over the duration of a loan, the interest rate becomes a poor predictor of default.

A key contribution of this paper to the existing literature is to identify the existence and importance of soft information for mortgage loan prices relative to corporate loans. Prior literature uses indirect proxy variables. We follow a different approach as soft information is internalized by lenders and not disclosed to investors, and is thus hard to test for explicitly. Further, we define soft information as the variation after controlling for hard information, and therefore include soft information that is explained by proxy variables for relationship lending, as well as soft information, which is unobserved by such variables. Our methodology differs to the extant literature as we use two-stage model and use its residuals from a regression of the loan credit spread at origination on hard information to measure soft information in a first stage and analyze the predictive role of these residuals for credit risk in a second stage.

Literature on mortgage loan credit spreads includes Levitin et al. (2020), Justiniano et al. (2017) and Rajan et al. (2015). Levitin et al. (2020) analyze the impact of observable risk factors on mortgage pricing over time. Justiniano et al. (2017) analyze the impact of US treasury yield term structures on mortgage rates and find that the relation diminishes over time. Rajan et al. (2015) find that LTV and FICO scores bear an increasing importance on the pricing of securitized mortgages, which may be explained by borrowers window-dressing their loan applications.

Literature finds that lenders collect soft information. Prior studies in the literature rely on soft information through indirect proxy variables. Our contribution is to provide a novel metric for lender soft information that is based on the interest rates offered to borrowers incremental to measurable hard information. Further, we find that this soft information is predictive for default and establish links to the decay over time and time since origination as well as the interaction with hard information.

Using this methodology, we analyze the capability of soft information to predict mortgage default and condition our findings on vintage, time since origination and hard information. Our research null hypotheses are:

Hypothesis 1. Soft information is not predictive for mortgage default.

Hypothesis 2. The predictive power of soft information is independent of (a) vintage, (b) time since origination, and (c) hard information.

We interpret significance of test variables in Stage 2 regressions for soft information and their interaction with (a) vintage, (b) time since origination, and (c) hard information, as an indication to reject the null hypothesis and support of the alternative hypotheses. We acknowledge that we work with large data. P-values shrink to zero with the number of observations. Hence, we provide an economic impact analysis of the measures for soft information to further support our findings in Section 5.

3 | MODEL FRAMEWORK

To understand the role that soft information plays for interest rates charged and credit risk, we use a two-stage regression model. In the first stage (Stage 1) regression, we analyze the impact of hard information on the credit spread charged for a given borrower. In the second stage (Stage 2) regression, we test the residuals from the Stage 1 regression as a proxy of soft information next to hard information on default. This approach is consistent with Goetzmann et al. (2004) who propose an alternative way of thinking about soft information as noise in a quantitative model. All Stage 2 models control for loan payoffs due to early termination of contracts with lenders.

3.1 | Stage 1 – Identifying soft information

The general definition of soft information within our paper is information linked to lenders who approve the loan but cannot be linked to other observable information. Agarwal et al. (2011) argue that soft information is collected during the origination of the loan. Following this argument, the Stage 1 models are based on an origination panel in which every loan has a line entry with information known at the origination time. All variables are observed at the origination time of the loan. We have used a Linear Regression and a Random Forest in the empirical analysis. The models are the foundation to the following Stage 2 analysis.

3.1.1 | Linear regression model

We estimate the OLS Linear Regression (LR) model below:

$$\text{CreditSpread}_{i\tau} = \alpha_1 + \beta_1 X_{i\tau} + \gamma_1 \Delta_i + \delta_1 \Delta_\tau + \pi_1 \Delta_i \Delta_\tau + \varepsilon_{i\tau,\text{LR}} \quad (1)$$

with loan i ($i = 1,\ldots, I$) and origination period τ ($\tau = 1,\ldots T$). All Stage 1 parameters have the index "1". We calculate credit spreads as the difference between the mortgage rate at origination and Treasury bond yields. Then, we develop a pricing model based on hard information observed at origination. $X_{i\tau}$ is a vector of explanatory variables, including vectors of borrower features, loan features and macroeconomic factors at loan origination. We include spline effects to control for non-linear relations. Δ_i and Δ_τ are lender and time dummies.

We use variables published by periodic investor reports for mortgage securitizations (residential mortgage-backed securities) for hard information that is verifiable (compare Stein (2002)). Keys et al. (2010) argue that lenders may have weak incentives to screen due to securitization. They show that securitization changes the lenders' screening process, because investors have access to hard information but not soft information when making an investment decision and data templates are homogeneous over lenders. The authors argue that lenders practice lax screening methods for borrowers with high-FICO scores because those loans are easier to securitize. As a result, lenders are less motivated to collect soft information. In other words, securitization may introduce a bias as lenders pass on "lemon" mortgages based on soft information. However, our findings are conservative (i.e., we are likely to underestimate soft information) for loans that lenders retain on their balance sheet. Note that this is the lower proportion of total loans (see Federal Deposit Insurance Corporation, 2019).

Banks have also different securitization ratios over time compared to non-bank lenders. We aim to control for securitization bias using lender and time fixed effects as well as their interaction. We control for exogenous lending standards across lenders by adding lender effects (Δ_i), time variation by adding time effects (Δ_τ), as well as changes

across lenders and time by adding interaction terms $\Delta_i\Delta_\tau$. After accounting for these factors, we are left with the residuals $\varepsilon_{i\tau,LR}$ of the model to capture soft information.

We summarize the hard information to the score HIS by multiplying the estimated parameters with the factors observed at loan origination:

$$HIS_{i\tau,LR} = \widehat{\alpha_1} + \widehat{\beta_1} X_{i\tau} + \widehat{\gamma_1}\Delta_i + \widehat{\delta_1}\Delta_\tau + \widehat{\pi_1}\Delta_i\Delta_\tau \quad (2)$$

For the avoidance of doubt, every loan *i* has one observation that is measured at origination τ and hence, one residual $\varepsilon_{i\tau,LR}$. As there is only a single origination time per loan, we simplify $\varepsilon_{i,LR} = \varepsilon_{i\tau,LR}$ and $HIS_{i,LR} = HIS_{i\tau,LR}$.

3.1.2 | Machine learning with random forests

The ultimate goal in our Stage 1 is to explain the credit spread by public information to infer the credit spread implied by soft information (residual). Lenders may apply a range of approaches that vary over loans, loan products, lenders, and time. Econometric techniques that rely on a complexity reduction via defined parameters may be unable to fully account for the heterogeneity in loan pricing applied in industry. Machine learning methods such as bagged trees (bootstrap aggregations like Random Forests) and boosted trees are well suited for our application as they are able to consider any variable combination (based on public information) via random sampling and non-linearity via splitting of variables. We have tested a number of such approaches and find that the best fitting model is a random forest.[4]

Random forests are based on a random selection of data and variables which are split into nodes. Multiple trees are formed and a limited number of trees (e.g., 100) is generally sufficient to minimize errors. The final model predictions $\text{CreditSpread}_{i\tau,RF}$ are based on averaging the predictions for the out-of-bag samples to avoid overfitting. We set hard information to the model-implied credit spread:

$$\text{HIS}_{i\tau,RF} = \widehat{\text{CreditSpread}}_{i\tau,RF} \quad (3)$$

The soft information is the difference between the observed and the fitted credit spread:

$$\varepsilon_{i\tau,RF} = \text{CreditSpread}_{i\tau,RF} - \widehat{\text{CreditSpread}}_{i\tau,RF} \quad (4)$$

For the avoidance of doubt, every loan *i* has one observation that is measured in period τ and hence, one residual $\varepsilon_{i\tau,RF}$ as loans are included in different leaves in different trees and thus the average spreads are different over loans. There is only a single origination time per loan and we simplify $\varepsilon_{i,RF} = \varepsilon_{i\tau,RF}$ and $HIS_{i,RF} = HIS_{i\tau,RF}$.

3.2 | Stage 2 – Base test for the impact of soft information on default

In Stage 2, we analyze the impact of soft information on the prediction of default. After loan origination, borrowers face the following three outcomes:

$$S_{it} = \begin{cases} D \text{ if loan i defaults at time t} \\ P \text{ if loan i pays off at time t} \\ 0 \text{ otherwise} \end{cases}$$

Hence, default is a competing hazard to the other two states of payoff and non-default/non-payoff (see Deng et al., 2000). Low credit risk borrowers are more likely to pay loans off prior to maturity often as a consequence of refinancing the loan with a lender at lower rates after loans are partially amortized and LTV ratios decrease. High credit risk borrowers are more likely to default. A selection bias may be introduced when low credit risk borrowers depart and the remaining population has a greater risk following payoff.

Default risk prediction in industry follows narrow regulatory requirements. One of which is the calibration of annual PDs on annual default rates. Following industry practice, we have used multinomial logit models as they are explaining default probabilities for discrete times. The competing states default, payoff, and non-default/non-payoff are mutually exclusive. We further provide a competing risk hazard rate analysis for all main results in the robustness test section.

For the main regression results, we control for payoff by estimating multinominal logit models, which are common in the literature (e.g., Ergungor and Moulton (2014) and Agarwal et al. (2011)).[5] We model the probability of default and payoff as

$$\text{Prob}(S_{it} = s, s \in \{\text{D}, \text{P}\}) = \frac{exp\left(\alpha_2^S + \beta_2^S \varepsilon_i + \gamma_2^S HIS_i + \delta_2^S * X_{it}\right)}{1 + \sum_{s \in \{1,2\}} exp\left(\alpha_2^S + \beta_2^S \varepsilon_i + \gamma_2^S HIS_i + \delta_2^S X_{it}\right)} \tag{5}$$

with loan i ($i = 1,\ldots, I$) and observation period t ($t = 1,\ldots T$). All Stage 2 parameters have the index "2". The equation is in expected value terms (note the probability operator on the left-hand side) due to the trinomial outcome variable S. Hence, there is no residual as it is common in OLS regression models. We estimate the parameters using the maximum-likelihood method.

In the equations above, ε_i and HIS_i are the proxies of soft information and hard information respectively, that we receive from the Stage 1 regression. β_2^S and γ_2^S are two parameters of interest that show the impacts of soft information and hard information on mortgage defaults. X_{it} is a vector of control variables that includes time-changes in loan-characteristics and macroeconomic variables. Note that borrowers' variables are only included in the hard information score HIS_i if they are observed at origination and do not change over time.

The probability of a non-default and non-payoff is inferred from the default probability $\text{Prob}(S_{it} = \text{D})$ and the payoff probability $\text{Prob}(S_{it} = \text{P})$:

$$\text{Prob}(S_{it} = 0) = 1 - \text{Prob}(S_{it} = \text{D}) - \text{Prob}(S_{it} = \text{P}) \tag{6}$$

3.3 | Stage 2 – Interaction test for the impact of soft information on default

We examine a number of interactions for the residuals. First, vintages indicate whether there are changes over time, due to changes in system-wide lending standards, or the type of loans that lenders securitize and we thus observe in our data set. We further analyze time since origination, as soft information is generally collected by lenders at mortgage origination when the loan terms are negotiated. Finally, we analyze hard information at origination.

We extend the base model for the probability of default (D) and payoff (P) to accommodate these interactions as follows:

$$\text{Prob}(S_{it} = s, s \in \{D, P\}) = \frac{exp\left(\alpha_2^S + \beta_{2,a}^S \varepsilon_i \text{Interaction}_{it} + \beta_{2,b}^S \text{Interaction}_{it} + \beta_{2,c}^S \varepsilon_i + \gamma_2^S HIS_i + \delta_2^S X_{it}\right)}{1 + \sum_{s \in \{1,2\}} exp\left(\alpha_2^S + \beta_{2,a}^S \varepsilon_i \text{Interaction}_{it} + \beta_{2,b}^S \text{Interaction}_{it} + \beta_{2,c}^S \varepsilon_i + \gamma_2^S HIS_i + \delta_2^S X_{it}\right)} \tag{7}$$

where $\beta_{2,a}^S$, $\beta_{2,b}^S$ and $\beta_{2,c}^S$ are the parameters of the interaction and the standalone effects.

4 | EMPIRICAL ANALYSIS

4.1 | Data

We analyze securitized subprime mortgage loans observed from 2000 to 2015 by quarterly frequency. The data was collected from the monthly loan performance reports of residential mortgage-backed securities to investors and provided by International Financial Research. The data is comparable to Rajan et al. (2015) and has been used in prior literature (e.g., Lee et al., 2021). We construct two data sets: an origination panel for the Stage 1 model and an observation panel for the Stage 2 model.

Loan securitization may reduce the role of "soft information" in the credit assessment process and we hypothesize that lender-originated loans that are not intended for securitization include an even greater degree of soft information. However, lenders in this market generally retain the equity tranche to have skin in the game and minimum equity tranche sizes have been established by the Dodd-Frank Act. One motivation for using non-agency securitized loans is that they are less subject to automated underwriting processes than the securitization markets organized by Fannie and Freddie Mac which is an alternative popular data source for mortgage risk studies.

For the origination panel, we consider a sample of approximately three million loans originated from 1997 to 2010. Figure 1 shows the distribution of loans at origination from 1997 to 2010.

The number of originated loans is comparatively small in economic downturns, and high in economic upturns. The highest origination/securitization period is from 2004 to 2007, that is, preceding the Global Financial Crisis of 2008–2009.

For the observation panel, we observe the loans for 15 years, resulting in roughly 39 million quarterly observations from 2000 to 2015. Note that information related to borrowers that is recorded only once at origination and not updated over time is included in the hard information score. We select the first lien loans only and exclude observations with a FICO score below 450 or above 900, LTV ratios greater than 130%, and more generally, missing values.

Finally, we obtain several supplemental macroeconomic data. We analyze three indexes for house prices: Case–Shiller (CS) index which is a national home price index, Zillow home value index at the zip-code level, and Federal

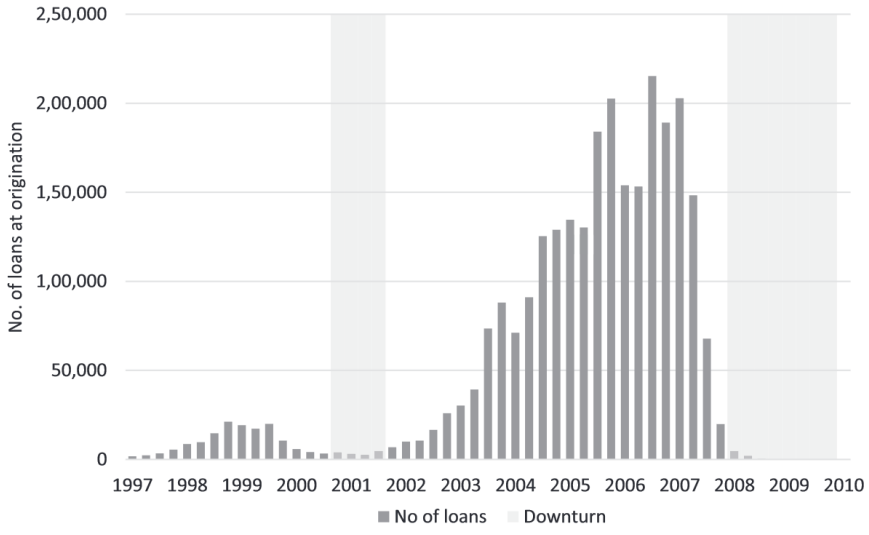

FIGURE 1 Number of originated mortgage loans, from 1997 to 2010. This figure shows the exponential growth of the observed and privately securitized mortgage loans prior to the Global Financial Crisis. The gray bars indicate economic downturns as defined by the National Bureau of Economic Research.

Housing Finance Agency (FHFA) house price index for metropolitan areas. Additionally, we collect real GDP from the US Bureau of Economic Analysis, the US Bond Yield from the Department of the Treasury.

As with all empirical studies, our results should be interpreted with care as the analyzed data relates to US mortgage loans. Individual lenders, in particular those outside the US, are likely to have different portfolios with different characteristics.

4.2 | Variable definitions

In Stage 1, we run a regression model on credit spreads, which are calculated as the difference between the original mortgage rates and the Treasury bond yields of similar maturities.[6] Figure 2 describes the credit spread of loans at origination. The gray bars indicate economic downturns as defined by the National Bureau of Economic Research (NBER). The credit spreads increase in economic downturns and reduce in economic upturns.

Our explanatory variables can be categorized into three groups: borrower characteristics, loan characteristics, and macroeconomics variables at origination time.

For borrower features, the FICO score is a popular measure to analyze the credit risk of individual borrowers. The score is provided by Fair Isaac & Company and includes the following factors (weights in brackets): payment history (35%), amounts owed (30%), length of credit history (15%), new credit (10%), and credit mix (10%). Higher FICO scores show lower credit risk and vice versa. The second factor is debt-to-income (DTI) which is calculated as a fraction of monthly debt payments to monthly incomes. The third borrower characteristic is whether borrowers are investors or owner-occupiers.

For loan characteristics, the most important variable is the loan-to-value ratio (LTV) which is the outstanding balance to appraisal value of the property. Other factors considered include maturity, original balance, property, and loan type.

Finally, we use the change in house price index and real GDP growth to control for macroeconomic conditions at origination.

Figure 3 illustrates the default and payoff rates over the observation period from 2000 to 2015. Default is defined as the foreclosure of a mortgage. We focus on loan foreclosures rather than delinquencies as these are the

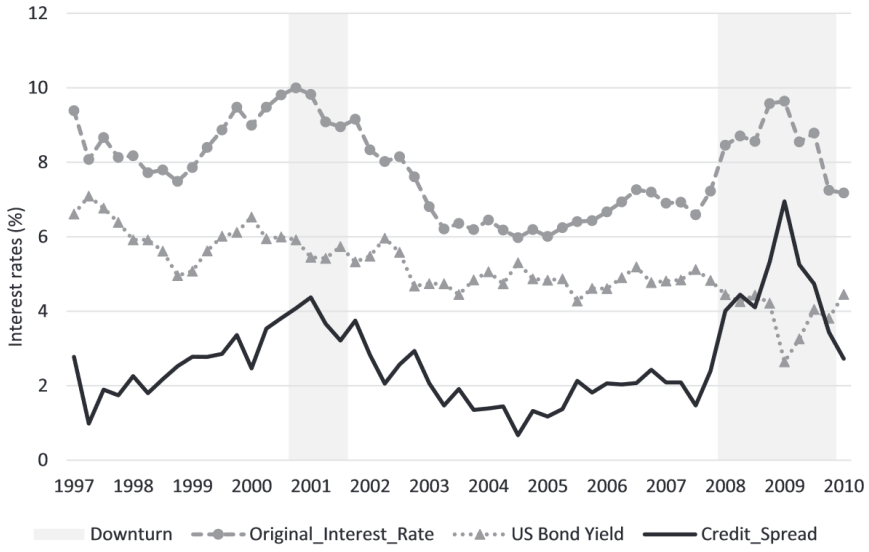

FIGURE 2 Mortgage rate, Treasury bond yield, and credit spread at origination, from 1997 to 2010. It shows the mean original mortgage rate, Treasury bond yield and credit spread over the origination period from 1997 to 2010 in quarterly frequency. Credit spread is the difference between original mortgage rate and the Treasury bond yield. The gray bars indicate economic downturns as defined by the National Bureau of Economic Research.

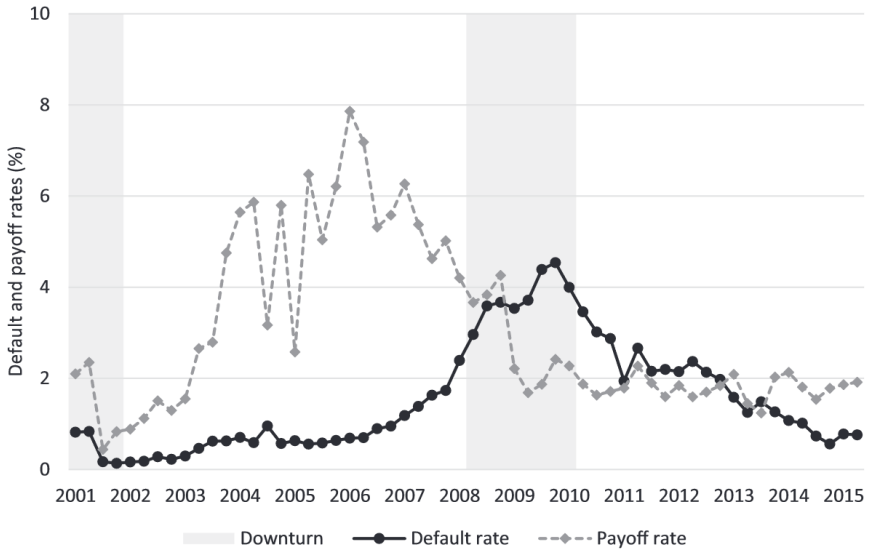

FIGURE 3 Default rate and payoff rate, from 2000 to 2015. It shows the default and payoff rates over observation periods from 2000 to 2015. The gray bars indicate economic downturns as defined by the National Bureau of Economic Research (NBER).

ultimate triggers of credit losses driven by equity and liquidity constraints. Further, strategic defaults in non-recourse states during the GFC result in immediate foreclosures and may not be considered by analyzing delinquencies. Payoff is defined as loans, which are terminated before maturity. Again, the shaded areas represent NBER economic downturn periods.

Figure 3 shows a negative relation between default and payoff rates, where default rates rise considerably in economic downturns and payoff rates are lower in economic downturns.

In Stage 2, we model the default process and use the residuals from Stage 1 as proxies for soft information and the predicted credit spreads as proxies for hard information. These are our two key test variables for default. We also control for payoff where loans are terminated prior to maturity. For all Stage 2 regressions, Panel B shows the parameters for the payoff equation. Whilst we do not discuss these in much detail, the parameters are generally opposite to the default equation, indicating that payoff is associated with low credit risk as explained above. Please refer to Pennington-Cross and Chomsisengphet (2007) and Clapp et al. (2001) for further details.

We also control for changes in loan characteristics. Change in current LTV is the current outstanding balance over the current house values and is most significant. We compute current house values based on relative changes in house price indexes over time. We further control for the remaining maturity, current outstanding, and the change in the loan interest rate, which is the difference between current interest rate and the original interest rate of a loan, as well as macroeconomic variables.

Furthermore, in interaction models, we investigate the interaction of soft information with other factors. We test the interaction of soft information with vintages, time since origination, as well as borrowers' hard information.

Table 1 summarizes the definitions of variables used in this study.

Table 2 displays summary statistics for variables for Stage 1 models.

More than half of the observed loans are adjustable rate mortgages and/or refinanced mortgages. The average maturity is nearly 30 years with an average original outstanding loan amount of $278,000. The average borrower FICO score is 675.65 and the average original LTV ratio is 76.47%.

Table 3 shows the summary statistics for variables used in Stage 2 regressions for default and payoff predictions.

In Stage 2, we observe loans quarterly and predict default probabilities using residuals and predicted values of credit spreads (HIS) from the Stage 1 models. The average residual is positive for defaulted loans and paid off loans

TABLE 1 Variables and definitions

Variable name	Definition
Panel A: Variables at origination	
Dependent variable	
Credit spreads at origination (Credit_Spread)	Difference between the mortgage rate at origination and US Treasury bond yield measured in percentage points.
Explanatory variables	
Lender	Lender ID, dummy variable that takes value of one for a given lender and takes value of zero otherwise.
Vintage	Origination time, dummy variable that takes value of one for a given vintage and takes value of zero otherwise.
Product	Product code describing mortgage amortization terms, dummy variable that takes value of one for a given product and takes value of zero otherwise.
State	State of the property location, dummy variable that takes value of one for a given state and takes value of zero otherwise.
FICO score (FICO)	Borrower credit scores provided by Fair Isaac & Company.
Debt-to-income (DTI)	Ratio of monthly mortgage payment and monthly income of the borrower at loan origination measured in percentage points.
Investor	Dummy variable that takes value of one if borrower is an investor and takes value of zero if borrower is an owner-occupier.
Loan-to-value at origination (Original_LTV)	Ratio of original loan outstanding and appraisal value of property at origination measured in percentage points.
Original term (Original_Term)	Term of loan at origination in years.
Loan size (Log_Original_Balance)	Natural logarithm of original loan balance.
Condos	Dummy variable that takes value of one if property is a condominium and takes value of zero otherwise.
Refinance	Dummy variable that takes value of one if loan purpose is for refinancing and takes value of zero otherwise.
ARM	Dummy variable that takes value of one if loan is an adjustable rate mortgage and takes value of zero otherwise.
Panel B: Variables at observation	
Dependent variable	
Outcomes	Variable that takes value of "D" for default, "P" for payoff and zero otherwise.
Default	Default is defined as foreclosure.
Payoff	Payoff is defined as loans which are repaid before maturity.
Test variables	
Residual	Error term generated from Stage 1 regressions as a measure of the soft information, in percentage points.
HIS	Predicted value generated from Stage 1 regressions as a measure of hard information, in percentage points.
Vintage	Origination year.
Time since origination (TSO)	The difference between observation time and origination time.
Control variables	
Current loan-to-value (LTV)	Ratio of current outstanding balance to current appraisal value of property. Note that current appraisal value of property is

TABLE 1 (Continued)

Variable name	Definition
	estimated from changes in the CS house price index measured in percentage points.
Remaining_Maturity	Difference between maturity time and current observation time.
Log_Current_Balance	Natural logarithm of current outstanding loan balance.
Change_Interest_Rate	Difference between current interest rate and original interest rate of a mortgage measured in percentage points.
HPI_CS	Change of CS house price index in percentage points at national level.
HPI_Zillow	Change of Zillow house price index in percentage points at zip code level.
HPI_FHFA	Change of Federal Housing Finance Agency house price index in percentage points at metropolitan area level.
GDP	Real GDP growth at current time at national level measured in percentage points.

TABLE 2 Summary statistics of loans at origination from 1997 to 2010.

Variable	Mean	*SD*	P1	P50	P99
Metric variables					
Original_Interest_Rate (%)	6.59	2.04	1.00	6.75	11.50
Yield_orig (%)	4.87	0.36	4.28	4.84	6.06
Credit_Spread (%)	1.72	2.01	−3.81	1.82	6.37
FICO	675.65	72.33	504.58	684.23	803.61
DTI (%)	34.19	24.41	5.49	27.66	124.44
Original_LTV (%)	76.47	13.88	26.80	80.00	100.00
Original_Term	29.30	4.09	15.00	30.00	40.00
Original_Balance ($ 1000)	278.41	230.53	33.30	214.00	1000.00
Dummy variables					
Investor	0.13	0.34	0.00	0.00	1.00
Condos	0.06	0.24	0.00	0.00	1.00
Refinance	0.51	0.50	0.00	1.00	1.00
ARM	0.56	0.50	0.00	1.00	1.00
No. of loans at origination	3,057,870				

Note: It shows summary statistics for variables used in Stage 1 regressions over the origination period from 1997 to 2010.

and negative for the pooled sample. The average is slightly below zero for the observation panel as the mean is weighted by the number of observation quarters per loan.

The average predicted credit spread for defaulted loans is 2.15%. This is well above the average of 1.58% for paid off loans and 1.63% for all loans. The current LTV of defaulted loans averages 98%, reflecting house price declines. The number is considerably higher than the current LTV for paid off loans with 70% and the average for all loans with 80%.

TABLE 3 Summary statistics of loans over observation years from 2000 to 2015.

	Pooled sample		Default loans		Payoff loans	
Variable	Mean	*SD*	Mean	*SD*	Mean	*SD*
Residual (linear regression) (%)	−0.10	1.26	0.05	1.21	0.05	1.00
Hard information score, HIS (linear regression) (%)	1.64	1.60	2.18	1.64	1.64	1.65
Residual (random forest) (%)	−0.12	1.02	0.10	0.99	0.02	0.78
Hard information score, HIS (random forest) (%)	1.67	1.56	2.12	1.70	1.68	1.62
Original_LTV	75.78	14.34	80.23	10.03	74.84	15.42
LTV (%)	80.10	26.77	98.10	21.97	70.58	24.82
Remaining_Maturity (in years)	24.90	5.55	26.71	3.81	25.50	5.58
Actual_Current_Balance ($ 1000)	259.47	224.64	265.63	206.34	263.51	228.98
Change_Interest_Rate (%)	0.03	1.51	0.31	1.61	0.39	1.53
TSO (in years)	4.25	2.94	3.46	2.18	3.42	2.65
HPI_CS (%)	−0.51	3.50	−1.65	3.63	−0.40	3.39
HPI_Zillow (%)	−0.27	6.23	−1.86	7.43	0.11	3.04
HPI_FHFA (%), smaller matched sample	0.04	1.98	−0.43	2.10	0.26	1.97
GDP (%)	0.97	1.60	0.29	1.95	1.23	1.41
No. of Obs.	39,044,923		840,998		1,248,713	

Note: It shows the summary statistics for metric variables used in the Stage 2 regressions over the observation period from 2000 to 2015 divided into all loans (pooled sample), default loans, and payoff loans.

4.3 | Stage 1 – Identifying soft information

To ensure a high model fit, we have tested all information that is publicly available for loan investors and considered variable interactions as well as non-linear relations between variables and outcome variables. We have tested linear regressions and machine learning techniques. We find that random forests provide the best fitting models.[7] In the following, we present and analyze two models.

The first model is a linear regression of all observed variables including lender, vintage, product[8] and state fixed effects as well as the interaction between vintage and lender effects.[9] The *SE*s are clustered by lenders. We run a piecewise linear regression for the continuous variables. We include continuous variables x_{ijt} (borrower i, explanatory variable j and time t) and their spline expansions using a truncated power function (TPF) basis with a degree of unity $T(.)$.[10] That is, we include the differences between these variables and the knot (threshold) k_j if the variable is greater than the threshold and zero otherwise:

$$T(x_{ijt}) = \begin{cases} x_{ijt} - k_j \text{ if } x_{ijt} > k_j \\ 0 \text{ if } x_{ijt} \leq k_j \end{cases} \tag{8}$$

The variables FICO score, DTI ratio and log loan balance are separated into 10 bins based on percentiles. Hence, the bins have an equal number of observations. The variable LTV ratio has limited variation at origination of around 80%

with most observations at or below 80% and five bins were manually defined.[11] The knots are defined by the upper interval boundary of the bins bar the last one:

- FICO: 574.5, 614.5, 634.4, 664.3, 684.2, 694.2, 724.1, 744.1, 774.1;
- DTI: 12.0, 16.0, 19.6, 23.4, 27.7, 32.7, 39.2, 48.3, 64.1;
- LTV: 70, 75, 80, 85;
- Log_Original_Balance: 11.2, 11.6, 11.8, 12.1, 12.3, 12.5, 12.7, 13.0, 13.2.

The second model is the best performing random forest that includes the empirically observed non-linearities and variable interactions.[12] Note that the variable set includes (like in the Linear Regression) observable public information. However, the variable set does not include the additional terms of the piecewise regressions as random forests split variables at appropriate thresholds. In order to approximate and compare univariate variable sensitivities for non-parametric random forests, we run a second Linear Regression of the predicted credit spreads on the components of the above Linear Regression. The *SE*s are clustered by lenders.

Table 4 shows the parameter estimates of the pricing model from Equation (1) for the Linear Regression and implied regression of the Random Forest where we replace $\text{CreditSpread}_{i\tau}$ by the predicted credit spread from the Random Forest and run regression Equation (1) for comparison:

We plot the mean observed and predicted credit spread for the percentile-based bins of FICO score, DTI, LTV ratio and loan balance at origination using a Linear Regression and Random Forest model. The figures allow us to assess the directional impact of continuous variables on credit spread. They also demonstrate both models are capable of including the non-linear relation to the observed credit spread for these important variables. The linear model may show a slightly better fit here as the figures focus on the univariate relation between explanatory variables and credit spreads, whilst the Random Forest includes variable interactions next to non-linearities (Figure 4).

The parameters are aligned with our prior expectations based on the above-mentioned credit spread literature (e.g., Levitin et al. (2020), Justiniano et al. (2017) and Rajan et al. (2015)) and the mortgage risk literature (e.g., Amromin & Paulson, 2009). Interest rates are higher for lower FICO scores, loan terms, loan balances, and refinances. They are higher for higher DTI ratios, LTV ratios, adjustable rate loans, condominiums, and investor loans. The refinance parameter is negative, as refinanced loans generally relate to lower risk borrowers. R-square measures the explanatory power of models and shows how well hard information in the model can explain credit spreads. The R-square is 65.66% for the piecewise linear model and 77.75% for the Random Forest. We include the residual and predicted values of credit spreads from both the Linear Regression and the Random Forest in the following Stage 2 regressions as proxies for soft information and hard information.

We rely on the credit spreads embedded in loan rates offered to borrowers following the accept decision of lenders. We do not have information on rejected loans. Soft information and hard information are metric measures where a low value corresponds to low credit spreads and hence, low credit risk. Vice versa, a high value corresponds to high credit spreads and hence, high credit risk. As a result, we expect a positive relation between default, residuals, and hard information in all models throughout this paper. This interpretation is an important distinction to prior literature that generally does not measure the degree of soft information metrically.

4.4 | Stage 2 – Baseline test of soft information in the default process

Table 5 shows the parameter estimates of Equation (5) for the baseline test of soft information in the default process using a Multinomial Logit model. We have three competing outcomes: default, payoff and non-default non-payoff and our Multinomial Logit models have a default equation and a payoff equation. Multinomial Logit models are popular in the mortgage literature as mortgage borrowers have a prepayment option (i.e., a competing hazard).[13]

TABLE 4 Stage 1 regression results – Identifying soft information

	Linear regression		Random forest	
	Estimate	*SE*	Estimate	*SE*
Panel A: Parameter estimates				
FICO	−0.0074***	(0.0022)	−0.0033***	(0.0011)
FICO Spline 1	−0.0045**	(0.0022)	−0.0056***	(0.0010)
FICO Spline 2	0.0103***	(0.0011)	0.0091***	(0.0010)
FICO Spline 3	−0.0032***	(0.0008)	−0.0017***	(0.0006)
FICO Spline 4	0.0021***	(0.0005)	0.0011***	(0.0003)
FICO Spline 5	0.0008	(0.0006)	−0.0003	(0.0003)
FICO Spline 6	0.0002	(0.0005)	0.0004	(0.0003)
FICO Spline 7	0.0006*	(0.0003)	0.0000	(0.0002)
FICO Spline 8	0.0010***	(0.0003)	0.0005**	(0.0002)
FICO Spline 9	0.0001	(0.0005)	−0.0003	(0.0003)
DTI	0.1299***	(0.0137)	0.0581***	(0.0058)
DTI Spline 1	−0.0381***	(0.0090)	−0.0036	(0.0047)
DTI Spline 2	−0.0213***	(0.0042)	−0.0143***	(0.0028)
DTI Spline3	−0.0129***	(0.0026)	−0.0093***	(0.0015)
DTI Spline 4	−0.0137***	(0.0023)	−0.0091***	(0.0015)
DTI Spline 5	−0.0104***	(0.0018)	−0.0061***	(0.0009)
DTI Spline 6	−0.0066***	(0.0016)	−0.0032***	(0.0009)
DTI Spline 7	−0.0076***	(0.0012)	−0.0028***	(0.0007)
DTI Spline 8	−0.0009	(0.0009)	−0.0017***	(0.0005)
DTI Spline 9	−0.0117***	(0.0019)	−0.0060***	(0.0013)
Original_LTV	−0.0092***	(0.0020)	−0.0035**	(0.0016)
Original_LTV Spline 1	0.0110*	(0.0063)	0.0067	(0.0049)
Original_LTV Spline 2	−0.0004	(0.0060)	−0.0046	(0.0055)
Original_LTV Spline 3	0.0191**	(0.0087)	0.0374***	(0.0067)
Original_LTV Spline 4	0.0094	(0.0163)	−0.0130	(0.0118)
Log_Original_Balance	−1.7889***	(0.2385)	−1.1981***	(0.2024)
Log_Original_Balance Spline 1	0.3825	(0.3096)	−0.0707	(0.2845)
Log_Original_Balance Spline 2	0.1852	(0.1564)	0.5160***	(0.1490)
Log_Original_Balance Spline 3	0.0198	(0.0725)	−0.1422***	(0.0605)
Log_Original_Balance Spline 4	0.1928***	(0.0731)	0.2740***	(0.0666)
Log_Original_Balance Spline 5	0.0813	(0.0521)	0.0614	(0.0374)
Log_Original_Balance Spline 6	0.2159***	(0.0717)	0.1509***	(0.0528)
Log_Original_Balance Spline 7	0.3567***	(0.1166)	0.2909***	(0.0995)
Log_Original_Balance Spline 8	0.2055***	(0.0744)	0.1397**	(0.0576)
Log_Original_Balance Spline 9	−0.1182	(0.1091)	−0.1502	(0.0976)
Original_Term	−0.0214***	(0.0075)	−0.0270***	(0.0067)
ARM_Dummy	0.5333**	(0.2083)	0.3946**	(0.1830)
CO	0.2546***	(0.0299)	0.0559***	(0.0122)

TABLE 4 (Continued)

	Linear regression		Random forest	
	Estimate	***SE***	**Estimate**	***SE***
Investor	0.3106***	(0.0439)	0.2029***	(0.0357)
Refinance	−0.1551***	(0.0200)	−0.0827***	(0.0191)
Panel B: Fixed effects				
State	Yes		Yes	
Product	Yes		Yes	
Lender	Yes		Yes	
Vintage	Yes		Yes	
Lender*Vintage	Yes		Yes	
Panel C: Goodness-of-fit				
R-square	65.7%		77.8%	
RMSE	1.179		0.961	
No. of Obs.	3,057,870.00		3,057,870.00	

Note: It shows the regression results for Stage 1. The first model is a Linear Regression model where the dependent variable is the credit spread as the difference between the mortgage rate and the Treasury bond yield at origination. The second model relates to a linear regression model where the dependent variable is the estimated credit spread from a random forest. Panel A shows the parameter estimates. Panel B shows the goodness of fit and number of observations of the models. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

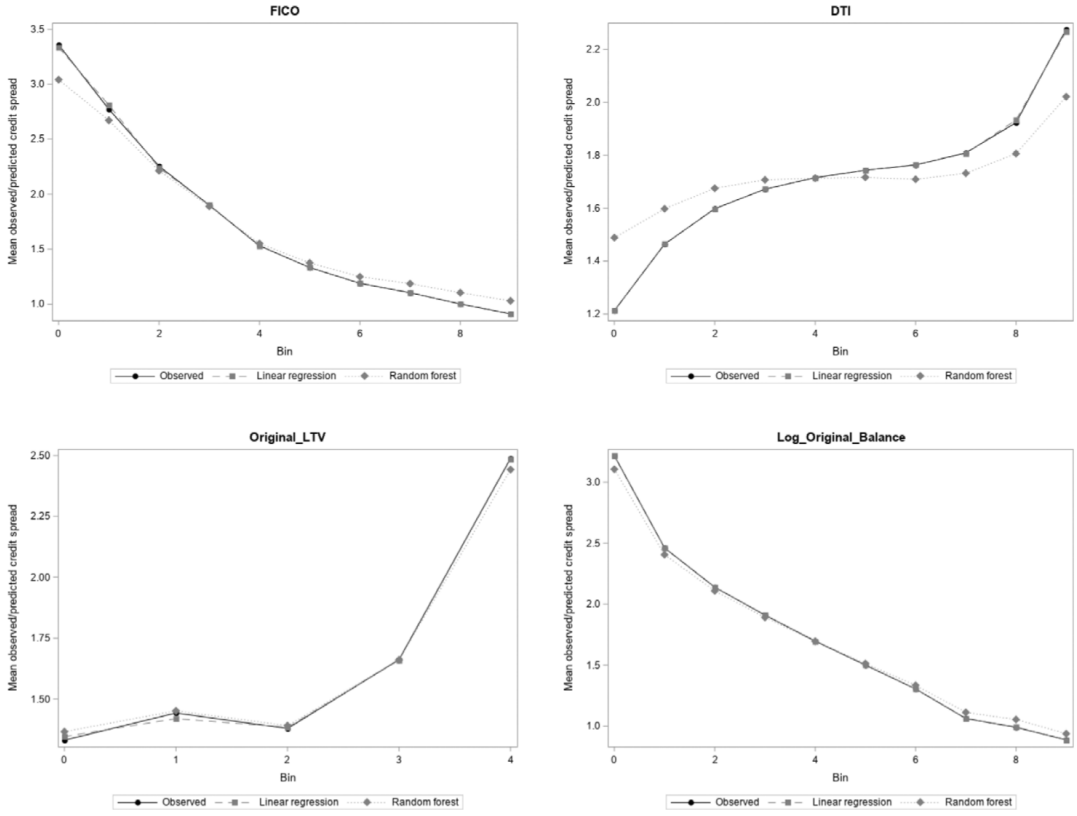

FIGURE 4 Mean observed and predicted spreads by binned continuous variables. It shows the mean observed and predicted credit spread using a Linear Regression and a Random Forest for bins of FICO, DTI, original LTV, and natural logarithm of the original loan balance.

Equation (6) shows that the probability for the third outcome can be inferred from the probabilities of the first two modeled outcomes.

Prepayment is a borrower choice. Key factors are mortgage market rates falling below current loan contract rates as borrowers can reduce mortgage payments by refinancing their loans with a more favorable rate and paying off existing loans. Adverse soft information results in higher mortgage rates with current lenders and may motivate borrowers to refinance their loans with lenders collecting adverse soft information. This is supported by the positive and significant estimate of residuals in Panel B in our base test. We are unable to further analyze the impact of soft information on payoff and refinance as borrowers are anonymous and measure the soft information of new lenders.

The estimates for key metric time-varying features generally show opposite signs for prepayment and default as borrowers with high credit quality are more likely to refinance and payoff loans. The LTV estimates are negative for

TABLE 5 Stage 2 regression results – Base test for the impact of soft information on default.

	Linear regression		Random forest	
	Estimate	***SE***	**Estimate**	***SE***
Panel A: Estimates for default				
Residual	**0.1565*****	(0.0191)	**0.2034*****	(0.0227)
HIS	**0.3707*****	(0.0448)	**0.3150*****	(0.0368)
LTV	0.0232***	(0.0022)	0.0235***	(0.0022)
Remaining_Maturity	0.0369***	(0.0047)	0.0339***	(0.0052)
Log_Current_Balance	0.3099***	(0.0638)	0.2682***	(0.0616)
Change_Interest_Rate	0.2478***	(0.0255)	0.2407***	(0.0257)
HPI_CS	−0.0173***	(0.0059)	−0.0192***	(0.0057)
GDP	−0.0579***	(0.0077)	−0.0571***	(0.0078)
Intercept	−11.3155***	(0.7345)	−10.6519***	(0.7137)
Panel B: Estimates for payoff				
Residual	0.1582***	(0.0161)	0.1630***	(0.0161)
HIS	0.1544***	(0.0325)	0.1517***	(0.0393)
LTV	−0.0162***	(0.0014)	−0.0162***	(0.0014)
Remaining_Maturity	0.0347***	(0.0065)	0.0347***	(0.0066)
Log_Current_Balance	0.1379***	(0.0416)	0.1363***	(0.0418)
Change_Interest_Rate	0.2091***	(0.0214)	0.2077***	(0.0233)
HPI_CS	−0.0001	(0.0122)	−0.0002	(0.0122)
GDP	0.0660***	(0.0203)	0.0660***	(0.0202)
Intercept	−5.0803***	(0.5005)	−5.0561***	(0.5263)
Panel C: Goodness-of-fit				
AUROC (Default)	75.09%		74.93%	
AUROC (Payoff)	65.50%		65.52%	
No. of obs.	39,044,923		39,044,923	

Note: It shows the regression results for the base test for Stage 2 using a multinominal logit model with the possible outcomes of non-default/non-payoff (reference category), default, and payoff. Panel A shows the parameter estimates for the default process. Panel B shows the parameter estimates for the payoff process and Panel C shows the goodness of fit and number of observations of the models. Residual and HIS are based on outcomes of Stage 1 using a linear regression and a random forest. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively. Bold values describe the test variables and non-bold values control variables in the models.

the prepayment equation and positive for the default equation. The GDP and HPI estimates are positive for the prepayment equation and negative for the default equation.

Some features have the same (positive) signs for both equations: the remaining maturity, current balance, and change in interest rate. This confirms that borrowers are more likely to default and payoff in earlier years when outstanding loan amounts and times to maturity are high. Further, high interest rates imply a greater default rate.

In Panel A, the estimate of residual is significant, suggesting that soft information predicts the default likelihood. Note that residuals include the amount of soft information embedded in credit spreads that is considered in addition to hard information. Higher residuals are associated with higher credit spreads. The positive signs of residuals imply that information, which results in an incremental credit spread, increases the default probabilities. Also, the hard information score from Equation (2), HIS, is positive, meaning that higher predicted credit spreads result in higher probabilities of default. We analyze the relative predictive power between soft and hard information in the next section.

In Stage 2, we link the soft information with default outcomes. We find that this relation is positive, which suggests lenders obtain soft information outside the hard information score. We see this as an indication of private information. In other words, had the lenders not included the soft information, the residuals and hence implied loan prices would be less reflective of default outcomes.

Information asymmetries (see Keys et al. (2010)) may exist. For example, borrowers may hide information or window-dress their loan applications to achieve more favorable lending decision outcomes. However, if lenders do not observe this information privately then the residuals, which consider what the lenders can observe (as they are included in the loan prices), would be independent or even negatively related to the default outcome in our Stage 2 regressions.

We do not detail the interaction between soft information and payoff risk in Tables 6–8 as we are unable to further analyze the impact of soft information on payoff and refinance as borrowers are anonymous and we are unable to measure the soft information of new lenders.

4.5 | Stage 2 – Interaction models

4.5.1 | Soft information over time

Table 6 and 7 show the parameter estimates of Equation (7) where residuals interact with vintages and time since origination.

The interaction terms $\text{Residual}_i * \text{VintageYear}_t$ are positive and significant as soft information predicts default. The results are stronger for the Random Forest residuals. The estimates of $\text{Residual}_i * \text{VintageYear}_t$ reduce by vintage year and hence time. This suggests that the role of adverse soft information diminishes over vintages. In other words, lenders rely less on soft information over time. The *SE*s are clustered by lender.

This finding is consistent with Rajan et al. (2015) who focus on hard information, with the distinction being that we document the deterioration of soft information. We find that soft information decays with time since loan origination. Possible explanations include bank mergers (see Petersen & Rajan, 2002) and the increased application of standardized digital processing of hard information in the industry.

We categorize time since origination (TSO) by the number of years, from one to ten and more years. It can be seen that the estimate of Residual*1Y_TSO is insignificant, meaning that in the first year after origination, soft information does not alter the default risk. However, after the first year, the interaction parameters are more negative. For example, the estimate of Residual*3Y_TSO is −0.15, which is well below the estimate of Residual*2Y_TSO at −0.07.

The economic interpretation is that soft information, like most hard information, becomes less informative with the passage of time. This observation is particularly relevant for the mortgage industry where most information is collected at loan origination when the lender has bargaining power, as it can approve or reject loan applications and request hard and soft information. Consumer protection laws limit lenders from including covenants in loan contracts and lenders generally do not collect soft information after loan origination. Some hard information is collected as lenders update LTV ratios by considering the loan amortizations and changes of house prices due to property

TABLE 6 Stage 2 regression results – Interaction with vintage, default equation

	Linear regression		Random forest	
	Estimate	***SE***	**Estimate**	***SE***
Panel A: Estimates for default				
Residual*Vintage2002	**0.1384*****	(0.0444)	**0.1263***	(0.0444)
Residual*Vintage2003	**0.1328****	(0.0577)	**0.2519*****	(0.0577)
Residual*Vintage2004	**0.1200**	(0.0738)	**0.1486***	(0.0738)
Residual*Vintage2005	**0.1571***	(0.0822)	**0.1240**	(0.0822)
Residual*Vintage2006	**0.1208**	(0.0865)	**0.0420**	(0.0865)
Residual*Vintage2007	**0.1410**	(0.0868)	**0.0485**	(0.0868)
Vintage2002	0.1544	(0.2936)	0.1122	(0.2936)
Vintage2003	−0.2092	(0.3126)	−0.3084	(0.3126)
Vintage2004	−0.0569	(0.3149)	−0.1716	(0.3149)
Vintage2005	−0.1106	(0.3188)	−0.1731	(0.3188)
Vintage2006	−0.0141	(0.3126)	−0.0637	(0.3126)
Vintage2007	−0.1319	(0.3181)	−0.1703	(0.3181)
Residual	0.0294	(0.0807)	0.1369*	(0.0807)
HIS	0.3686***	(0.0448)	0.3090***	(0.0448)
LTV	0.0228***	(0.0020)	0.0229***	(0.0020)
Remaining_Maturity	0.0358***	(0.0049)	0.0325***	(0.0049)
Log_Current_Balance	0.3139***	(0.0620)	0.2692***	(0.0620)
Change_Interest_Rate	0.2476***	(0.0266)	0.2373***	(0.0266)
HPI_CS	−0.0172***	(0.0055)	−0.0193***	(0.0055)
GDP	−0.0591***	(0.0078)	−0.0588***	(0.0078)
Intercept	−11.2292***	(0.7546)	−10.4287***	(0.7546)
Panel B: Goodness-of-fit				
AUROC (Default)	75.11%		74.95%	
AUROC (Payoff)	67.15%		67.44%	
No. of Obs.	39,044,923		39,044,923	

Note: It shows the regression results for a Stage 2 test with the interaction between vintage and soft information using a multinominal logit model with the possible outcomes non-default/non-payoff (reference category), default, and payoff. Panel A shows the estimates for the default process. Panel B shows the goodness of fit and number of observations of the models. Residual and HIS are based on outcomes of Stage 1 using a linear regression and a random forest. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively. Bold values describe the test variables and non-bold values control variables in the models.

revaluations or house price index changes. We include updated LTV ratios and other time-varying hard information next to the time-invariant hard information score in all Stage 2 models.

4.5.2 | Interaction of soft information with hard information

Once we have measured soft information and its significance for predicting default, we analyze the impact of hard information. We generate a dummy variable that is one if HIS is above the median value and zero otherwise. Table 8 shows the parameter estimates of Equation (7), showing the relation between the hard information dummy and soft information for the impact on default.

TABLE 7 Stage 2 regression results – Interaction with time since origination, default equation.

	Linear regression		Random forest	
	Estimate	***SE***	**Estimate**	***SE***
Panel A: Estimates for default				
Residual*1Y_TSO	**−0.0157**	(0.0224)	**−0.0251**	(0.0251)
Residual*2Y_TSO	**−0.0687**	(0.0647)	**−0.0976**	(0.0810)
Residual*3Y_TSO	**−0.1527***	(0.0839)	**−0.2033***	(0.1064)
Residual*4Y_TSO	**−0.1527***	(0.0839)	**−0.2013***	(0.1131)
Residual*5Y_TSO	**−0.1567***	(0.0885)	**−0.2051***	(0.1104)
Residual*6Y_TSO	**−0.1854****	(0.0916)	**−0.2594****	(0.1132)
Residual*7Y_TSO	**−0.2429****	(0.0944)	**−0.3332*****	(0.1182)
Residual*8Y_TSO	**−0.2328****	(0.0929)	**−0.3315*****	(0.1172)
Residual*9Y_TSO	**−0.2628*****	(0.0928)	**−0.3572*****	(0.1171)
Residual*10Y_TSO	**−0.2770*****	(0.0950)	**−0.3741*****	(0.1176)
Residual*10 + Y_TSO	**−0.2881*****	(0.0944)	**−0.3949*****	(0.1179)
1Y_TSO	0.5192***	(0.0286)	0.5265***	(0.0300)
2Y_TSO	0.7856***	(0.0501)	0.7917***	(0.0533)
3Y_TSO	0.9223***	(0.0659)	0.9216***	(0.0689)
4Y_TSO	0.8457***	(0.0731)	0.8255***	(0.0770)
5Y_TSO	0.7232***	(0.0831)	0.6865***	(0.0863)
6Y_TSO	0.7730***	(0.0821)	0.7283***	(0.0849)
7Y_TSO	0.6939***	(0.0841)	0.6380***	(0.0877)
8Y_TSO	0.6561***	(0.0955)	0.5886***	(0.0976)
9Y_TSO	0.7352***	(0.1053)	0.6580***	(0.1096)
10Y_TSO	0.8380***	(0.1031)	0.7457***	(0.1093)
10 + Y_TSO	1.1446***	(0.1254)	1.0772***	(0.1231)
Residual	0.3031***	(0.0922)	0.4003***	(0.1157)
HIS	0.3829***	(0.0456)	0.3200***	(0.0378)
LTV	0.0220***	(0.0023)	0.0226***	(0.0023)
Remaining_Maturity	0.0472***	(0.0053)	0.0403***	(0.0062)
Log_Current_Balance	0.3229***	(0.0642)	0.2745***	(0.0620)
Change_Interest_Rate	0.2576***	(0.0252)	0.2435***	(0.0250)
HPI_CS	−0.0128**	(0.0058)	−0.0129**	(0.0057)
GDP	−0.0298***	(0.0073)	−0.0275***	(0.0071)
Intercept	−12.4171***	(0.7671)	−11.5618***	(0.7453)
Panel B: Goodness-of-fit				
AUROC (Default)	75.49%		75.30%	
AUROC (Payoff)	68.30%		68.31%	
No. of Obs.	39,044,923		39,044,923	

Note: It shows the regression results for a Stage 2 test with the interaction between time since origination and soft information using a multinominal logit model with the possible outcomes non-default/non-payoff (reference category), default, and payoff. Panel A shows the parameter estimates for the default process. Panel B shows the goodness of fit and number of observations of the models. Residual and HIS are based on outcomes of Stage 1 using a linear regression and a random forest. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively. Bold values describe the test variables and non-bold values control variables in the models.

TABLE 8 Stage 2 regression results – Interaction with hard information, default equation

	Linear regression		Random forest	
	Estimate	***SE***	**Estimate**	***SE***
Panel A: Estimates for default				
Residual*HIS_D	**0.0646****	**(0.0284)**	**0.1019*****	**(0.0329)**
Residual	0.0857***	(0.0298)	0.1351***	(0.0360)
HIS_D	**0.5870*****	**(0.0891)**	**0.7210*****	**(0.0927)**
HIS	0.2675***	(0.0419)	0.1726***	(0.0328)
LTV	0.0215***	(0.0019)	0.0217***	(0.0019)
Remaining_maturity	0.0317***	(0.0038)	0.0270***	(0.0042)
Log_Current_Balance	0.3512***	(0.0544)	0.3145***	(0.0493)
Changes_in_interest_rate	0.2360***	(0.0254)	0.2167***	(0.0251)
Hpi_CS	−0.0159***	(0.0057)	−0.0184***	(0.0056)
GDP	−0.0615***	(0.0077)	−0.0617***	(0.0078)
Intercept	−11.7047***	(0.6425)	−11.0533***	(0.6017)
Panel B: Goodness-of-fit				
AUROC (Default)	75.30%		75.15%	
AUROC (Payoff)	65.53%		65.56%	
No. of Obs.	39,044,923		39,044,923	

Note: It shows the regression results for a Stage 2 test with the interaction between soft information and a dummy variable for low risk and high risk loans based on hard information (HIS_D is one if HIS is above the median and zero otherwise). The regression is using a Multinominal Logit model with the possible outcomes of: non-default/non-payoff (reference category) default, and payoff. Panel A shows the parameter estimates for the default process. Panel B shows the goodness of fit and number of observations of the models. Residual and HIS are based on outcomes of Stage 1 using a Linear Regression and a Random Forest. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively. Bold values describe the test variables and non-bold values control variables in the models.

The estimate of Residual*HIS_D is positive and significant, suggesting that soft information from borrowers with higher HIS has a greater impact on default risk than lower HIS borrowers. Lenders rely more on soft information for high-risk borrowers as more soft information is collected and priced for borrowers where information is more binding as information has a greater sensitivity on default risk. Vice versa, lenders rely less on soft information for low risk where information is less binding as it has a lower impact on default risk.

We have included HIS_D as a standalone effect as this is common for econometric interaction models. The standalone effects of HIS and HIS_D are positive and significant, meaning that HIS is positively correlated with default risk.

4.6 | Robustness checks

We perform a number of robustness checks to confirm our findings of the impact of soft information on default risk. We run the base test of default equation using Residual and HIS outcomes from Stage 1 based on the random forest method on different subsamples (robustness check 1–3). We also analyze logit models (robustness check 4) and different HPI proxies (robustness check 5 and 6):

- Robustness check 1: Subsample for negative residuals;
- Robustness check 2: Subsample for positive (and zero) residuals;

- Robustness check 3: Subsample for conventional 30-year fixed rate mortgage only;
- Robustness check 4: Full sample using Logit model (instead of Multinomial Logit);
- Robustness check 5: Full sample using HPI proxy sourced from Zillow at zip code level; and
- Robustness check 6: Subsample using HPI proxy sourced from FHFA for metropolitan areas.

In robustness check 1 and 2, we test negative and positive residuals separately to investigate whether lenders differentiate between negative and positive soft information on loan pricing. We find that the results between the two subsamples are comparable.

In robustness check 3, we run a robustness check for our Stage 2 regressions for 30-year fixed rate loans. These loans are the default mortgage loan in the United States. Borrowers who choose other types (e.g., 30-year adjustable rate loans) may actively select a different type of mortgage and this may reflect a riskier types of borrowers (see e.g., Liu and Sing (2013)). We have controlled for a large range of mortgage features including LTV and mortgage products in our main regressions. We find that 30-year fixed rate loans result in a greater significance of the residual effect, implying that selection of non-standard mortgages reduces the collection of soft information.

In robustness check 4, we use a Logit regression for Stage 2 to test whether the results uphold if we do not control for competing risks using multinomial logit models. This is relevant as the industry mainly uses logistic regressions in credit scoring applications. We find for logistic regressions that the impact of soft information on default outcomes is consistent.

In previous analyses, the Case-Shiller house price index is used as the national index. We use the Zillow house price index at the zip-code level and the Federal Housing Finance Agency (FHFA) house price index for metropolitan areas, to run the robustness check 5 and 6. The advantage of the Zillow and FHFA house price indexes, compared to Case-Shiller, is that we can more precisely capture changes in the housing market at particular locations. Note that the FHFA House Price Index is collected for all 50 states and over 400 American cities. However, it is not available for all zip codes. We find that the impact of soft information on default outcomes is consistent.

Table 9 summarizes all robustness checks and presents the regression results.

In Stage 2, the estimates of Residual and HIS are positive significant, suggesting that both soft and hard information predict default. Again, this robustness confirms our main hypothesis that soft information impacts mortgage default. In summary, all findings are robust for different subsamples and HPI indices. Furthermore, the signs and magnitudes of all parameter estimates are very comparable.

Further, we have added a competing risk hazard model for our base model following the extended Cox proportional hazard model by Fine and Gray (1999). The model explains the baseline hazard over TSO whilst our base model does not condition on age.

The model is defined as:

$$\lambda_j(TSO, x) = \lambda_{j0}(TSO)\, exp\left(x'\beta_j\right) \tag{9}$$

λ_j is the proportional hazard rate of event j ($j = 1$ – Default, 2 – Payoff, 0 - Performing) and λ_{j0} is the baseline sub-hazard of event j. Table 10 shows the results for Random Forest residuals for the main regressions. The results are consistent with the main regression results for the Random Forest residuals in Tables 5–8.

5 | ECONOMIC IMPACT OF SOFT INFORMATION

We work with a large data set and there is a debate on the appropriateness of p-values (see Demidenko, 2016), which shrink and hence increase significance with observation counts. There is related discussion on the appropriateness of the areas under the receiver operating characteristics curve (AUROC, see Blochwitz et al. (2005) and Basel Committee on Banking Supervision (2005)) which depend on the composition of credit portfolios and the numbers

TABLE 9 Robustness checks – Stage 2 base test regression results for different subsamples, logit model and HPI proxies, default equation.

	1: Residuals < 0	2: Residuals >= 0	3: 30-year FRM	4: Logit model	5: HPI Zillow	6: HPI: FHFA
Panel A: Estimates for default						
Residual	**0.1999*** **	**0.1912*** **	**0.3360*** **	**0.1983*** **	**0.2035*** **	**0.2017*** **
	(0.0357)	**(0.0341)**	**(0.0249)**	**(0.0226)**	**(0.0228)**	**(0.0214)**
HIS	**0.3162*** **	**0.3351*** **	**0.2176** **	**0.3105*** **	**0.3102*** **	**0.3311*** **
	(0.0407)	**(0.0414)**	**(0.0935)**	**(0.0361)**	**(0.0368)**	**(0.0377)**
LTV	0.0282***	0.0192***	0.0229***	0.0240***	0.0224***	0.0183***
	(0.0022)	(0.0021)	(0.0024)	(0.0022)	(0.0023)	(0.0019)
Remaining_Maturity	0.0161***	0.0471***	0.0123	0.0328***	0.0312***	0.0423***
	(0.0055)	(0.0052)	(0.0095)	(0.0051)	(0.0056)	(0.0046)
Log_Current_Balance	0.2188***	0.3141***	0.1558*	0.2645***	0.2345***	0.2202***
	(0.0591)	(0.0665)	(0.0813)	(0.0609)	(0.0589)	(0.0579)
Change_Interest_Rate	0.2982***	0.1721***	−0.1352***	0.2332***	0.2252***	0.2390***
	(0.0310)	(0.0238)	(0.0388)	(0.0247)	(0.0253)	(0.0233)
GDP	−0.0572***	−0.0600***	−0.0549***	−0.0589***	−0.0214***	−0.0597***
	(0.0086)	(0.0084)	(0.0102)	(0.0076)	(0.0064)	(0.0071)
HPI_CS	−0.0042	−0.0310***	−0.0168**	−0.0194***		
	(0.0065)	(0.0055)	(0.0082)	(0.0056)		
HPI_Zillow					−0.0868***	
					(0.0040)	
HPI_FHFA						−0.0457***
						(0.0035)
Intercept	−10.0652***	−11.1687***	−8.8268***	−10.6429***	−10.1389***	−9.8058***
	(0.7262)	(0.7415)	(1.2211)	(0.7076)	(0.6874)	(0.6607)
Panel B: Goodness-of-fit						
AUROC (Default)	75.41%	73.71%	72.40%	74.92%	75.41%	73.40%
No. of default/delinquency	381,129	459,869	141,321	840,998	840,998	356,203
No. of Obs.	21,745,529	17,299,394	9,356,842	39,044,923	39,044,923	16,165,334

Note: It shows the different robust regression results for the Stage 2 base test of the default process. Robustness check 1 uses the subsample for negative residuals. Robustness check 2 uses the subsample for zero and positive residuals. Robustness check 3 uses the subsample for conventional 30-year fixed rate mortgages. Robustness check 4 uses the full sample and a Logit model (instead of multinomial logit). Robustness check 5 uses the full sample using HPIs from Zillow at the zip code level. Robustness check 6 uses the subsample using HPIs from the FHFA for metropolitan areas. Residual and HIS are based on outcomes of Stage 1 using a Random Forest. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively. Bold values describe the test variables and non-bold values control variables in the models.

TABLE 10 Robustness checks – Stage 2 regression results using a competing risk hazard model, default equation.

	1: Base test	2: Interaction with vintage	3: Interaction with TSO	4:Interaction with HIS_D
Residual	0.1945***	0.1487***	0.3893***	0.1078***
HIS	0.3225***	0.3150***	0.2875***	0.1723***
Residual*VintageYear2002		0.1382***		
Residual*VintageYear2003		0.2427***		
Residual*VintageYear2004		0.2006***		
Residual*VintageYear2005		0.1006***		
Residual*VintageYear2006		0.0192***		
Residual*VintageYear2007		0.0162**		
Residual*1Y_TSO			−0.0592***	
Residual*2Y_TSO			−0.1165***	
Residual*3Y_TSO			−0.2114***	
Residual*4Y_TSO			−0.2063***	
Residual*5Y_TSO			−0.2085***	
Residual*6Y_TSO			−0.2448***	
Residual*7Y_TSO			−0.3213***	
Residual*8Y_TSO			−0.3339***	
Residual*9Y_TSO			−0.3595***	
Residual*10Y_TSO			−0.3711***	
Residual*10 + Y_TSO			−0.3753***	
Residual*HIS_D				0.1328***
HIS_D				0.8014***
LTV	0.0172***	0.0119***	0.0195***	0.0152***
Remaining_Maturity	0.0980***	0.0952***	0.0706***	0.0925***
Log_Current_Balance	0.2575***	0.2318***	0.2351***	0.3246***
Change_Interest_Rate	0.1288***	0.1353***	0.1222***	0.1233***
HPI_CS	−0.0328***	−0.0329***	−0.0325***	−0.0337***

(Continues)

TABLE 10 (Continued)

	1: Base test	2: Interaction with vintage	3: Interaction with TSO	4:Interaction with HIS_D
GDP	−0.0149***	−0.0214***	−0.0169***	−0.0199***
Vintage	No	Yes	No	No
TSO	No	No	Yes	No
AUROC	71.43%	72.50%	68.40%	72.51%
No. of default	840,998	840,998	840,998	840,998
No. of Obs.	39,044,923	39,044,923	39,044,923	39,044,923

Note: It shows the Stage 2 regression results using a competing risk hazard model for the main test results. Robustness check 1 includes the baseline results, Robustness check 2 tests the interaction between vintage year and soft information. Robustness check 3 tests the interaction between time since origination and soft information. Robustness checks 4 to 6 test the interaction between hard information and soft information. Hard information includes FICO score, current LTV and original LTV. Panel A shows the parameter estimates for the default process. Panel B shows the goodness of fit and number of observations of the models. Residual and HIS are based on outcomes of Stage 1 using a Random Forest. *, **, and *** represent statistical significance at the 10%, 5%, and 1% level, respectively.

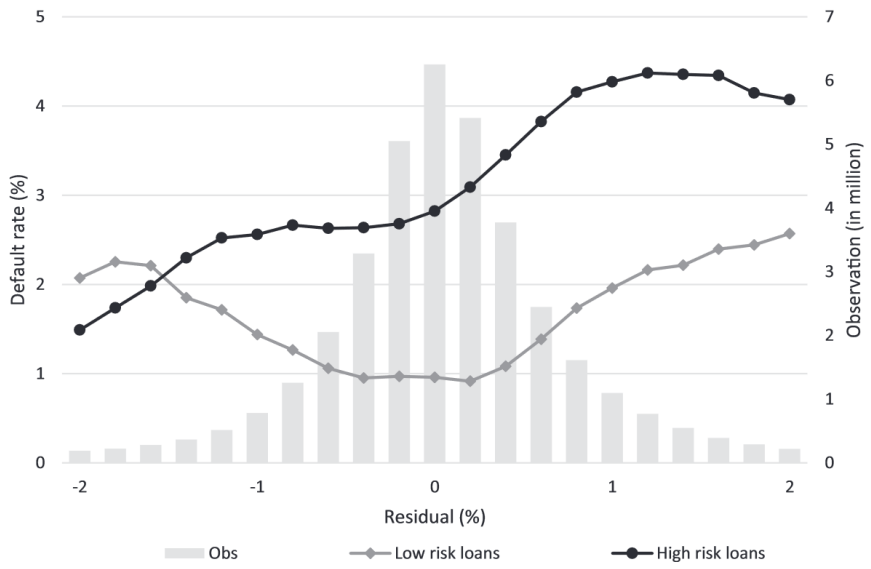

FIGURE 5 Sensitivity of default risk between low risk and high risk loans in relation to residuals (soft information). It shows the relation between observed default rates and residuals for low risk loans (HIS below the median) and high risk loans (HIS below the median). The residuals are from the Stage 1 Random Forest model and binned in 0.5 intervals. The gray histogram shows the distribution of total residual observations for the pooled data sample, measured in millions.

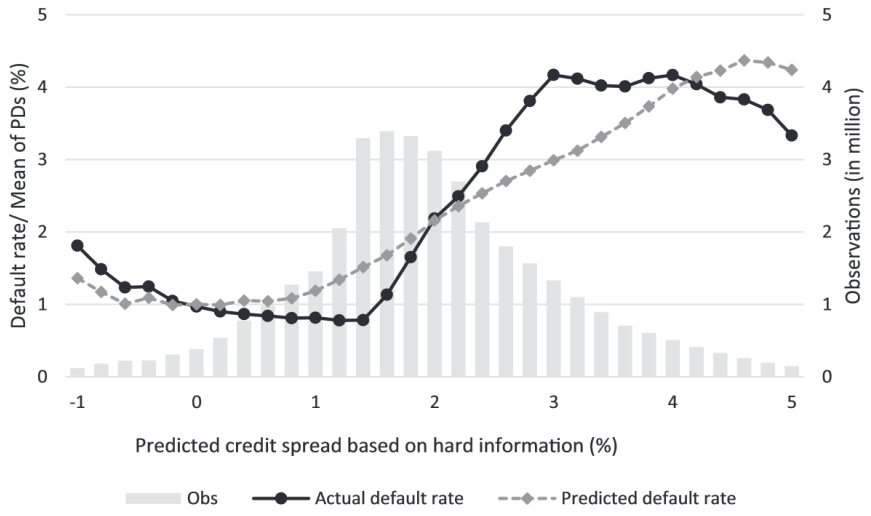

FIGURE 6 Observed default rate and mean predicted default probability in relation to hard information. It shows the relation between observed default rates, mean predicted default probabilities and the predicted credit spreads. Predicted credit spreads from Stage 1 using the Random Forest method are binned in 0.5 intervals. The gray histogram shows the distribution of predicted credit spread observations, measured in millions.

of defaulters in the estimation sample. As a result, we follow the guidance of Kellner and Rösch (2019) and analyze the economic impact of the residuals on probabilities of default.

The Stage 1 models decompose the observed credit spread into a credit spread based on observed information (hard information score, HIS) and residuals. Residuals may be negative or positive and measure the deviation from credit spread based on hard information for subprime loans.

Figure 5 shows the relation between soft information and default rate observations and predictions. Default rates are predicted by the mean estimated default probabilities, which are based on soft information and the hard

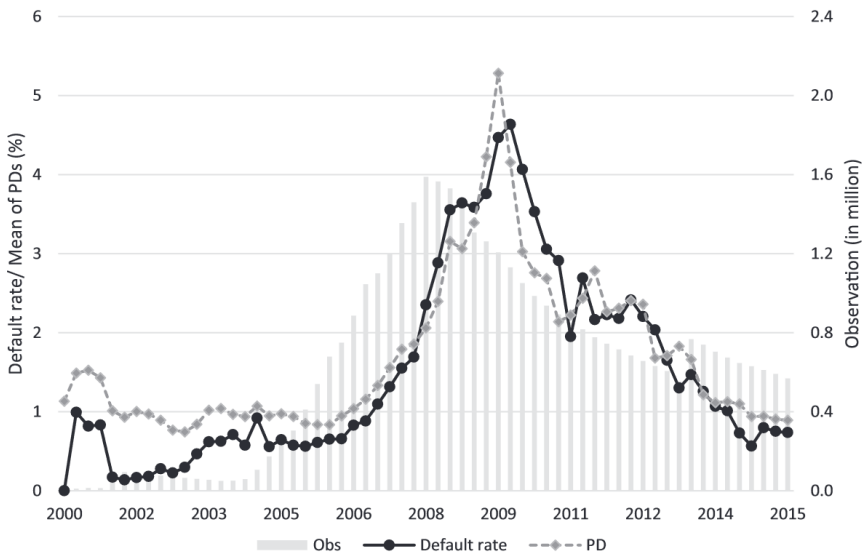

FIGURE 7 Observed default rate and mean predicted default probability over observed period of 2000–2015. It shows the observed default rate and mean predicted default probability over time. The gray histogram shows the distribution of residual observations, measured in millions.

information score. Following an increasing trend, higher residuals tend to result in higher probabilities of default as well as default rates. The residuals are from the Stage 1 random forest model and are binned in intervals of length 0.5.

We show the implied PD variation (Stage 2) derived from residuals for loans with an HIS below median (gray solid line) and above the median (black solid line). The figure confirms that high-risk loans are more sensitive to soft information than low risk loans as the gap of default rates between low risk and high risk increases with the residual. The variation of default rates for low risk loans is approximately 1.5% and for high risk loans 3%.

Figure 6 shows the relation between predicted credit spreads based on hard information and default rate observations and predictions.

The variation in default probabilities from low to high hard information scores is approximately 3%. Hard and soft information may have a similar impact on default rates.

Finally, Figure 7 shows the actual default rate and predicted default probability over time.

Actual and predicted values of default rates are close to each other, suggesting a high degree of calibration of mean model implied default probabilities to observed default rates over time of our model. The fit over time is an important model performance aspect for lenders, as the default rate and hence, loss rate for a given time period, needs to be offset by loan loss provisions and bank capital allocations.[14]

6 | CONCLUSION

We study the impact of soft information and its sensitivity on mortgage default. The econometric technique enables us to measure observed and unobserved soft information. Prior literature has not considered unobserved soft information. Soft information is measured on the interpretational level of credit spreads and higher values correspond to higher credit spreads. The literature has measured whether there is soft information but not the degree to which it impacts credit risk measures such as default probabilities.

First, we find strong evidence that soft information exists and predicts default probabilities. Second, although lenders may have decreased their attention on collecting soft information, its effect on default remains. Soft information is sensitive to the time and survival time (i.e., time since origination) of borrowers. The importance of soft

information collected at loan origination diminishes over time and survival times. Third, we provide evidence for the importance of soft information by showing that hard information is positively aligned with soft information, suggesting that the impact of soft information on default risk is stronger for high-risk borrowers. This may indicate that lenders rely more on soft information for high-risk borrowers as more soft information is collected and priced for borrowers when information is more binding as information has a greater sensitivity on default risk.

Soft information is captured and internalized by lenders and should not be ignored because it is significant for predicting default risk. This is a particular current concern, as lenders may shift to automatic, computer driven processes that include machine learning and artificial intelligence to save costs and reduce processing times to compete in a digital economy. Our study is important to policy makers to assist in reducing information asymmetries between lenders and investors in the securitization process. Soft information from borrowers is captured and internalized by lenders and should not be ignored because it is significant for predicting default risk.

Future research may explore other ways in which soft information may be digitally collected and made available to investors. This may include the analysis of digital footprints using machine learning algorithms and artificial intelligence.

ACKNOWLEDGMENTS

The authors would like to thank the participants of the seminars at the Australian Prudential Regulation Authority, Hong Kong Monetary Authority, University of Regensburg, University of Technology Sydney and the participants and discussants of the annual event of Finance Research Letters, Frontiers in Credit Risk, 2021. The support of the Hong Kong Institute for Monetary Research and the Brian Gray Scholarship of the Australian Prudential Regulation Authority are gratefully acknowledged. Open access publishing facilitated by University of Technology Sydney, as part of the Wiley - University of Technology Sydney agreement via the Council of Australian University Librarians.

ORCID

Harald Scheule https://orcid.org/0000-0001-7283-5654

ENDNOTES

[1] According to the International Monetary Fund's financial soundness indicators, real estate loan ratios of banks are 63.5% for Australia, 18.5% for Germany, 36.6% for Canada and 31.0% for the US.

[2] Some lenders do not offer risk-based interest rates but set accept thresholds and price the group of accepted mortgage loans with the same loan rate. However, the accept thresholds and loan prices vary across lenders. For the avoidance of doubt, we do not consider bank effects (e.g., risk appetite) as soft information as we control for lender fixed effects in our Stage 1 regressions.

[3] There are related contributions in the broader finance and accounting literature. For example, Bertomeu and Marinovic (2016) examine unverifiable information disclosure such as news, forecasts, unaudited statements, and document cases of misreporting.

[4] See Footnote 8 for further details.

[5] Further examples include Heinen et al. (2019), Pennington-Cross and Chomsisengphet (2007) and Clapp et al. (2001).

[6] Treasury bond yields are available for 1 month, 2 months, 3 months, 6 months, 1 year, 2 years, 3 years, 5 years, 10 years, 15 years, 20 years and 30 years. We use the bond yield which is closest to the original maturity.

[7] We find that other linear regressions with more granular non-linear spline extensions have similar model accuracies. We considered quadratic and cubic TPF splines as well as cubic penalized and non-penalized B splines (R-square of 65.7%). We find that other machine learning techniques have slightly lower but comparable model accuracy. We have considered boosted trees with an R-square of 77.1%. We did not use these extensions considering the trade-off between modest model accuracy improvements and greater model complexity.

[8] Liu and Sing (2013) find that mortgage choices have an impact on ex-post default risk. We control for mortgage products in our models.

[9] Vintage is a strong predictor for credit spreads (see e.g., Levitin et al., 2020). The vintage fixed effects are proxies for the general macro economy at loan origination. The interactions between vintage and lender effects explain how

time-varying lender specific effects are proxies for time-varying bank underwriting criteria – commonly known as the lending standard. The interaction effects use annual rather than quarterly time effects to limit the number of estimated parameters.

[10] We exclude the loan terms as the variation is limited.

[11] Bins 1 to 5 include 20.11%, 8.06%, 11.88%, 38.92%, and 21.02% of observations.

[12] Hyperparameters of the selected model include number of trees: 100, in-bag fraction: 60% and maximum depth 20. For more details, please refer to Breiman (2001) or Rösch and Scheule (2020).

[13] See Heinen et al. (2019), Ergungor and Moulton (2014), Agarwal et al. (2011)), Pennington-Cross and Chomsisengphet (2007) and Clapp et al. (2001). We provide results for competing risk hazard models in the robustness check section. We have also tested a number of alternative models including Logistic Regression, Probit Regression, sample selection and regularization (L1 and L2) techniques. The results for the different methods are comparable and implied default probabilities are highly correlated across the various econometric techniques. Rösch and Scheule (2020) analyze the predictive accuracy of a broader range of classification models for mortgage loans including bagged and boosted trees and find that the predictive performance measured by AUROC and Brier score slightly improves. We do not apply these models in our Stage 2 regressions as the accuracy improvements are less pronounced than in our Stage 1 regressions due to the binary nature of defaults. Further, some techniques do not allow for a model-based estimation of the impact of soft and hard information and the various interactions due to their non-parametric (black box) nature.

[14] Calibration measures such as the Hosmer-Lemeshow test statistic or the Brier score (see Rösch and Scheule (2020) for more details) confirm our visual analysis

REFERENCES

Agarwal, S., Ambrose, B. W., Chomsisengphet, S., & Liu, C. (2011). The role of soft information in a dynamic contract setting: Evidence from the home equity credit market. *Journal of Money, Credit and Banking*, *43*(4), 633–655.

Agarwal, S., & Hauswald, R. (2010). Distance and private information in lending. *The Review of Financial Studies*, *23*(7), 2757–2788.

Amromin, G., & Paulson, A. L. (2009). Comparing patterns of default among prime and subprime mortgages. *Economic Perspectives*, *33*(2), 18-37.

Basel Committee on Banking Supervision (2005), Studies on the validation of internal rating systems, Working Paper No. 14.

Berger, A. N., Miller, N. H., Petersen, M. A., Rajan, R. G., & Stein, J. C. (2005). Does function follow organizational form? Evidence from the lending practices of large and small banks. *Journal of Financial Economics*, *76*(2), 237–269.

Berger, A. N., & Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *The Economic Journal*, *112*(477), F32–F53.

Bertomeu, J., & Marinovic, I. (2016). A theory of hard and soft information. *The Accounting Review*, *91*(1), 1–20.

Blochwitz, S., Hamerle, A., Hohl, S., Rauhmeier, R., & Rösch, D. (2005). Myth and reality of discriminatory power for rating systems. Wilmott Magazine, 2–6, *2005*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brick, I. E., & Palia, D. (2007). Evidence of jointness in the terms of relationship lending. *Journal of Financial Intermediation*, *16*(3), 452–476.

Chakraborty, A., & Hu, C. X. (2006). Lending relationships in line-of-credit and nonline-of-credit loans: Evidence from collateral use in small business. *Journal of Financial Intermediation*, *15*(1), 86–107.

Clapp, J. M., Goldberg, G. M., Harding, J. P., & LaCour-Little, M. (2001). Movers and shuckers: Interdependent prepayment decisions. *Real Estate Economics*, *29*(3), 411–450.

Degryse, H., & Van Cayseele, P. (2000). Relationship lending within a bank-based system: Evidence from European small business data. *Journal of Financial Intermediation*, *9*(1), 90–109.

Demidenko, E. (2016). The p-value you can't buy. *The American Statistician*, *70*(1), 33–38.

Deng, Y., Quigley, J., & Van Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, *68*, 275–307.

DeYoung, R., Glennon, D., & Nigro, P. (2008). Borrower–lender distance, credit scoring, and loan performance: Evidence from informational-opaque small business borrowers. *Journal of Financial Intermediation*, *17*(1), 113–143.

Ergungor, O. E., & Moulton, S. (2014). Beyond the transaction: Banks and mortgage default of low-income homebuyers. *Journal of Money, Credit and Banking*, *46*(8), 1721–1752.

Federal Deposit Insurance Corporation (2019). *Quarterly banking profile*, 13(4). Federal Deposit Insurance Corporation.

Goetzmann, W. N., Pons-Sanz, V., & Ravid, S. A. (2004). Soft information, hard sell: The role of soft information in the pricing of intellectual property. (No. w10468). National Bureau of Economic Research.

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, *94*(446), 496–509.

Heinen, A., Kau, J. B., Keenan, D. C., & Kim, M. L. (2019). Spatial dependence in subprime mortgage defaults. *The Journal of Real Estate Finance and Economics*, *62*, 1–24.

Justiniano, A., Primiceri, G. E., & Tambalotti, A. (2017). The mortgage rate conundrum (No. w23784). National Bureau of Economic Research.

Kellner, R., & Rösch, D. (2019). A Bayesian re-interpretation of "significant" empirical financial research. *Finance Research Letters*, *38*, 101402.

Keys, B. J., Mukherjee, T., Seru, A., & Vig, V. (2010). Did securitization lead to lax screening? Evidence from subprime loans. *The Quarterly Journal of Economics*, *125*(1), 307–362.

Lee, Y., Rösch, D., & Scheule, H. (2021). Systematic credit risk in securitised mortgage portfolios. *Journal of Banking & Finance*, *122*, 105996.

Levitin, A. J., Lin, D., & Wachter, S. M. (2020). Mortgage risk premiums during the housing bubble. *The Journal of Real Estate Finance and Economics*, *60*(4), 421–468.

Liu, B., & Sing, T. F. (2013). Unobserved risks in mortgage contract choice. *Real Estate Economics*, *41*(4), 958–985.

Pennington-Cross, A., & Chomsisengphet, S. (2007). Subprime refinancing: Equity extraction and mortgage termination. *Real Estate Economics*, *35*(2), 233–263.

Petersen, M. A., & Rajan, R. G. (2002). Does distance still matter? The information revolution in small business lending. *The Journal of Finance*, *57*(6), 2533–2570.

Rajan, U., Seru, A., & Vig, V. (2015). The failure of models that predict failure: Distance, incentives, and defaults. *Journal of Financial Economics*, *115*(2), 237–260.

Rösch, D., & Scheule, H. (2020). Deep credit risk, machine learning with Python. KDP.

Saengchote, K. (2013). Soft information in the subprime mortgage market. Proceedings, 26th Australasian Finance and Banking Conference.

Stein, J. C. (2002). Information production and capital allocation: Decentralized versus hierarchical firms. *The Journal of Finance*, *57*(5), 1891–1921.

How to cite this article: Luong, T. M., Scheule, H., & Wanzare, N. (2022). Impact of mortgage soft information in loan pricing on default prediction using machine learning. *International Review of Finance*, 1–29. https://doi.org/10.1111/irfi.12392