The Impact of Hard and Soft Information on Mortgage Default

Nitya Wanzare

Supervised by

Associate Professor Harald Scheule

Abstract

This paper uses U.S. loan-level dataset between 2000-2015 to identify soft information and its impact on borrower default. Unlike traditional literature encompassing this area, our dataset contains a strictly limited number of borrower characteristics and multiple lender names. Furthermore, we study the role of soft information given varying borrower, lender and economic settings. In particular, we analyse borrower FICO score, lender's size, focus on lending, profitability and safety and macroeconomic variables. From our analysis, we find soft information exists in our dataset and is a positive and significant predictor of borrower default, thus echoing its importance as displayed in existing literature.

We would like to thank APRA, RBA and The Brian Gray Scholarship in writing this report

Contents

1. Motivation

2. Literature Review

3. Hypothesis Development

    3.1. Role of Soft Information

    3.2. Role of Risk and Soft Information

    3.3. Soft Information and Economic Cycles

    3.4. Lenders and Soft Information

    3.5. Lender’s Characteristics and Soft Information

4. Framework

    4.1. Identifying Soft Information- Base Model Stage 1 Regression

    4.2. Impact of Soft Information on Default- Base Model Stage 2 Regression

    4.3. Impact of Soft Information and Change in Hard and Soft Variables on Default

    4.4. Soft Information and Interaction with Borrower and Economic Variables

    4.4.1 Borrower Characteristics and Soft Information- FICO Score

    4.4.2. Soft Information Variation with Economic Cycles

    4.5. Lender Effect and Soft Information

    4.5.1 Lender Characteristics and Soft Information

5. Empirical Analysis

    5.1. Data

    5.2. Lender Details

    5.3. Credit Spread Calculation

    5.4. Elimination within the Dataset

6. Regression Results

    6.1. Base Model Stage 1- Identifying Soft Information

    6.2. Base Model Stage 2- Residual Model and Change in Hard and Soft Information

    6.3. Residual Interaction Model

    6.3.1. FICO Model

    6.3.2. FICO Stratification Model

    6.3.3. Residual-Economic Model

    6.3.4. Lending Model

    6.3.5. Contractionary Model

6.3.6. Coincident Index Model

6.4. Estimates Model

6.4.1. Estimates Model Comparison with and without Financial Ratios

6.4.2. Estimates with Bank Size

6.4.3. Estimates with Bank Profitability

6.4.4. Estimates with Bank Capital

6.4.5. Estimates with Bank Lending

7. Critical Evaluation

7.1. Economic Impact

7.2 Conclusion

1. Motivation

Mortgage lending plays a critical role in today's financial markets. It accounts for a significant portion of commercial bank's balance sheet and has been the cause of major financial crisis in the recent decade. Owning a dwelling plays a key role in a consumer's life. Often, for the average person, it is necessary to borrow money from a financial institution in-order to complete this purchase. Financial institutions are responsible to screen borrowers and determine their ability to pay back their debt. In order to do this, the lenders have access to hard information, such as the borrowers FICO score and loan to value ratio, along with other "soft" information. Soft information is best described as; information that is hard to express numerically and is only known by the lender exposed to it (Stein 2002).

This soft information, gathered by lenders, adds a human touch to the pricing of the loan. The correct price of the loan, theoretically, should be provided by the ongoing market rate, borrowers credit score, amount of loan they wish to access along with other recordable variables. However, the human aspect amidst this is the lender's judgement, apart from the credit score of the borrower, that aids in determining their creditworthiness. This judgement is based upon factors specific to the borrower such as their character and possible income shocks as well as macroeconomic variables which indicate the current state of the economy. With the advent in technology and rise of banks with rigid heirarchial structures, this aspect of mortgage pricing is easily dissipated. As hard information is easier to transfer digitally, it has pronounced impact on the overall pricing of the loan over the soft information which could have also been a strong predictor.

This study analyses the impact of soft information on the mortgage rate and its effect on borrower default. Our data covers the recent Global Financial Crisis as it spans over 2000 to 2015. Due to this, we also analyse the impact of soft information with respect to economic cycles. Moreover, to understand how the impact of soft information changes from one borrower to another, we use a range of borrower and lender characteristics to stratify our overall results.

Given its core nature of being "*difficult to summarize in a numeric score"* as per Petersen (2004), soft information is challenging to record and transmit amongst different lenders. In contrast, the ability of hard information to be displayed numerically makes it easy store and compare. Due to its quantitative nature, transmission of hard information is significantly easier in today's banking world. FICO scores and loan to value ratios of a mortgage are examples of hard information.

The loan to value ratio (LTV) is the ratio of the loan to the value of the asset purchased. The ratio is critical for lenders as it helps them analyse their total exposure given they approve the loan. It is an indicator of the borrower's overall equity position. A number of studies show that a higher loan-to-value ratio leads to a higher probability of default thus bearing greater risk.

The FICO score is another measure of borrower risk. It is popularly used as a credit score in the United States. Originally developed by the Fair Isaac Corporation, today it is widely used by lenders to analyse the credit risk of individual borrowers. This score generally falls between 300 to 850 and takes five factors into its calculation: Payment History (35%), Amounts Owed (30%), Length of Credit History (15%), New Credit (10%) and Credit Mix (10%)[1]. The FICO score generally falls between 300 to 850 with a higher FICO score signifying lower credit risk for an individual borrower.

Contribution of this paper to the existing literature are as follows. Firstly, we analyse the significance of soft information within a loan-level dataset. Secondly, we study the significance of soft information holding as control a range of borrower characteristics, lender characteristics and economic variables. Thirdly, we study how relevance of soft information varies with economic cycles and lastly, given our diverse dataset with multiple mortgage originator's information we examine the impact of soft information given a particular lender.

2. Literature Review

The importance of soft information in business and consumer lending is supported by existing literature. Post GFC, academics began analysing soft information and its predictive power of

[1] The amounts in the brackets show the proportion of each category allocated to the FICO score.

determining loan default. Moreover, in the period 1997-2006, Rajan, Seru and Vig (2015) show that "*as the level of securitization increases, lenders have an incentive to originate loans that rate high based on characteristics reported to investors, even if other unreported variables imply a lower borrower quality*". Since soft information cannot be reported, this is the additional information Rajan, Seru and Vig are referring to. They show that over time of the loan, the interest rate becomes a poor predictor of default and the models that predict failure underestimate the importance of soft information when estimating credit risk. This poses a problem as ignoring these essential details leads to information asymmetry as the investors are unaware of the true quality of their investment product. In addition to this, Agarwal, Ambrose, Chomsisengphet and Liu (2009) [henceforth AACL] display the importance of soft information in a dynamic contract setting. They have exclusive access to home equity loans from the time of loan request by the borrower to the acceptance of the loan by the lender. AACL look at borrower's initial credit choice at origination and find that borrower who carry greater risk have a larger chance of applying for higher loan to value home equity contract. Moreover, they analyse the role soft information plays during the underwriting process and find that those borrowers who were charged a greater annual percentage rate at origination (based on soft information gathered by lender) have a greater change of default. In conclusion, they state that using soft information, a lender can reduce their credit losses.

In relation to firm lending, Chen, Huang and Tsai (2012) empirically investigate the value of soft information and find that it significantly improves the power of their default prediction models. Moreover, they explain the actual content of the soft information and find that firms with loyal customer and long-term employees have a lower chance of default.

Majority of soft information studies are conducted on small to medium firm lending. Stein (2002) argues how lending for small to medium firms is heavily based on soft information. This is due to the relationship between the firm's president and the bank lender. He also states how this kind of relationship cannot be achieved in home equity lending hence, analysing soft information becomes challenging. In his paper, Stein (2002) shows how the role of soft information is predominant given a company takes a decentralised approach to generating revenues for a project with soft information.

Along the same lines, Uchida (2011) find that banks screen loans for SMEs based on a relationship factor, financial statement factor and a collateral/guarantee factor. Upon further analysis they find that smaller banks and banks under competitive pressure place greater importance on soft information and relationship banking.

Given the difficulty in observing soft information, researchers often rely on proxies of soft information to analyse its importance. Distance is a popular proxy as the shorter the distance between a lender and a borrower, the ease of availability of soft information for the lender. Ergungor and Moulton (2012) analyse the impact of geographical distance on mortgage default for low income borrowers. Their findings conclude 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, non-local banks. Agarwal and Hauswald (2010) show that borrower proximity facilitates the collection of soft information while Petersen and Rajan (2002) analyse the impact of distance for small firm lending and conclude that small firms no longer need highest quality credit ratings to access credit hence suggesting the ease of availability of credit. Saengchote (2013) analyses the relationship between mortgage brokers and borrowers and show that borrowers with low-documentation loans[2] that are situated further away from the mortgage brokers have a higher risk of default.

The importance of soft information can also be linked to risk management. Godbillon-Camus and Godlewski (2005) in their study show that soft information allows the bankers to decrease the capital allocation for VaR coverage. In a different study, Houston and Spencer (2014) study the housing bubble of 2003-2007 with behavioural aspects of real-estate pricing. They conducted a survey asking volunteers to state if it is a good time to buy a house, why or why not. From this, they gathered the behavioural aspects of real-estate pricing and found that their regression model fits better during the 2003-2007 bubble with the behavioural factors included hence displaying significance of borrower judgement in real estate pricing.

[2] Low-documentation loans are associated with loan applications where borrowers do not submit, or are unable to provide, sufficient documentation of stable income over a period of time or proof of liquid wealth. In the most extreme cases, some loan applications have completely unverified income and asset, also known as stated-income loans.

The main contribution of this paper to the existing literature is identifying the existence and importance of soft information in a loan-level, borrower characteristic poor dataset. We call our data-set "borrower characteristic poor" as the only borrower related information we have is the FICO score and the loan to value ratio. Our methodology differs to the existing papers as we use a two stage regression model and use residuals as a measure of the unobserved soft information which we argue would have an impact on the pricing of the loan.

Once we have identified soft information and its significance in predicting default, we analyse how its role changes given varying borrower characteristics, economic cycles and lender characteristics. For borrower characteristics, we use borrower FICO score and analyse role of soft information and its impact on predicting default given a particular borrower FICO score. Generally, we should observe borrowers with a higher credit score to be less risky hence soft information impact would be lower. Moreover, over the period of our data (2000-2015), Keys, Mukherjee, Seru and Vig (2010) argue that lenders were incentivised to practice lax screening methods for borrowers with high-FICO scores given the ease with which those loans could be securitized. This is as borrowers with a high FICO score are deemed to have a lower likelihood of default as shown by Bajari, Chu and Park (2008). The loan related characteristics we look at modelling is the loan to value ratio[3]. Amromin and Paulson (2009) study prime and sub-prime mortgages and conclude that LTV, borrower FICO score and interest rate at origination are all significant factors predicting loan default. In particular, the higher the LTV ratio, the greater the chances of default.

In addition, we analyse the change in relevance of soft information given varying economic cycles. Bank lending standards are a good proxy for understanding the bank's views on the current economy and subsequently the credit market. Asea and Blomberg (1998) show that banks switch their lending standards systemically over the economic cycle. Analysing lending standards and the impact it has soft information gives us an insight on the lender's side of the story. Moreover, we also analyse the

[3] However, upon analysis of the LTV in our dataset we provide reasons why this is not a good measure for this study.

coincident economic activity index for the United States and contractionary economic phases defined by NBER.

Lastly, we observe the lender effect in relation to soft information gathered. The contribution we can make to literature in this area is that we have access to multiple lenders and also have information regarding firm characteristics for a handful. However, we do not have access to the loans that were rejected by each of the lenders similar to what papers like AACL (2009) have obtained. This information would have been extremely valuable as it would've provided us with grounds of comparison between lenders loans that have and have not been rejected.

Overall, this paper is organised as follows, section 3 outlines introduces the hypothesis development for each study we undertake, in section 4 we introduce the model framework, section 5 describes the data used and specifications regarding our dataset, section 6 shows the regression results derived and finally section 7 is a critical review where we touch upon the use of soft information in an economic sense and the conclusion.

3. Hypothesis Development

3.1. Role of Soft Information

When borrowers need credit for the purpose of buying a house or an investment property, they generally approach lenders or mortgage brokers to get the best possible rate on a loan. The lender whom they interact with, analyses the borrower's current credit situation, income streams and desired loan value. Along with this data, they collect other information such as possibility of the borrower to experience an income shock in the future or their soft-skills such as trustworthiness and character and use this extra information to price the loan. This extra information gathered by lenders through interacting with the borrower in any manner is defined as soft information and due to complications in quantifying this information, it is not stored and is only known by the lender. We hypothesize the soft information gathered by lenders would be detected in our loan-level dataset moreover, it will play an important role in predicting default and thus contain significant predictive power.

3.2. Role of Risk and Soft Information

Once the unknown soft information has been verified within our data, the next question we ask is how does soft information's impact change on a borrower and lender level? A basic example is if a lender were to be approached by two different borrowers – borrower A has a high FICO score while borrower B has an average FICO score – what is the impact of soft information collected on both types of borrowers?

To answer this query, we take into account borrower FICO scores and lending standards prevalent at the time of origination. We expect to find that borrowers with a high FICO score would be screened less. Keys, Mukherjee, Seru and Vig (2009) state that over the period of 2001-2006, lenders had weak incentives to screen borrowers with high FICO scores due to securitization. Following this argument, we should expect to see a lower impact of soft information for borrowers with high FICO scores compared to those with lower FICO scores.

3.3. Soft Information and Economic Cycles

An economic cycle is generally defined in four phases, expansionary phase, contractionary phase, peak and trough. The peak is the highest point in an economic cycle and is usually followed by the contractionary phase until it reaches the trough which is the lowest point in the economic cycle.

During periods of poor economic growth, lenders would exercise extreme caution thus screening the borrowers thoroughly as compared to expanding economic periods. Hence, we would expect to see a stronger emphasis being placed on soft information when the economy is undergoing a downturn. To test this, we use a categorical variable which indicates whether an economy is currently in a contractionary period, the Coincident Economic Activity Index for the United States (Coincident Index) and Lending Standards.

Lending standards of the banks on a quarterly basis were gathered from FRB and assimilated into the panel dataset. An originator's lending standards represent their judgement on the current economy and the relevant lending practices ongoing at the time. Lax lending standards usually correlate with the current economic cycle as shown by Asea and Blomberg (1998). When the lending standards are tighter, lenders are more cautious when approving new credit. Given this economic significance, we expect to see stronger influence of soft information during tighter lending cycles than otherwise.

The Coincident Economic Activity Index collates four economic indicators, the unemployment rate, non-farms payroll, number of hours worked in manufacturing and wages and salaries. Using this indicator eliminates short term noise and indicates the prevalent economic environment over each quarter. We expect to see soft information have a stronger impact on loans originated during economic downturns.

3.4. Lenders and Soft Information

In addition to the above, we control for the lender effect within our model. Since we have access to different originators, we can use this information about the different lenders and analyse loans they have originated. From this we expect to control for the lender side of the story and capture the soft information to analyse against the probability of default. Soft information with the additional lender control should still be significant and a positive determinant of probability of default.

3.5. Lender's Characteristics and Soft Information

For a few lenders within the dataset, we have information regarding their financial ratios in particular, the capital adequacy ratio, loans to total assets and profitability ratio. In addition to this, we also have access to the total size of the lender. The capital adequacy ratio is a determinant of the bank safety and we hypothesize, the safer banks should place greater importance on screening borrower hence soft information should play a larger role. With the same logic, we expect to see soft information be very important for banks with a high loan to total assets ratio as a high loan to assets ratio signifies the focus a particular bank places on lending and if this focus is high, the caution with which such a bank would lend would be higher and thus greater the impact of soft information.

It is safe to say, the larger a given lender, the greater the hierarchical structure within the bank. Given this, if a borrower were to ask for a loan from such a lender, there is a potential their application would be handled by different agents. This would lead to a dissipation of soft information along the process of the loan approval hence the role soft information plays for larger banks should be expected to be less than for smaller banks. Lastly, following the same logic, banks with higher profitability ratios would be expected to place lesser importance on soft information given that these banks are likely to be larger in size. The contents of these ratios will be explained in later sections.

Summary of Hypothesis: -

Hypothesis 1: - Soft information exists within our dataset and has significant predictive power of default.

Hypothesis 2: - Soft information is a strong predictor of default for borrower with low FICO scores compared to those with higher FICO scores.

Hypothesis 3: - Soft information should have a larger impact during economic downturns. That is, in a contractionary phase, during tight lending standards and periods with low coincident index.

Hypothesis 4: - Soft information remains significant after controlling for the lender effect

Hypothesis 5: - Soft information plays an important role for banks with high capital adequacy ratio and high loans to asset ratio while, the opposite should hold for banks larger in size and with a high profitability ratio.

4. Framework

To understand the role soft information plays in interest rate charged and credit risk, we use a two stage regression model. In the first stage regression, we analyse the impact of hard facts and soft information on the credit spread charged for a given borrower. The credit spread is the premium allocated for a particular borrower upon loan approval. The calculation and definition of the credit spread is outlined in section 5.1.2. The second stage regression involves using the soft information variable to predict the probability of default. The second stage regression is divided up into four stages and this will be discussed in detail in the following sections.

4.1. Identifying Soft Information- Base Model Stage 1 Regression

The general definition of soft information within our dataset is information which cannot be quantified and is only known by the lender who approves the loan. AACL (2009) argue that soft information is mainly gathered during the origination of the loan. Following this argument, the first stage regression is estimated at origination of the loan only. All the different dependent variables are recorded as per their value during the origination of the loan.

Given that soft information is not recorded and hence unobserved, our first regression equation involves analysing the unknown aspect of our model, namely the residuals. For the construction of this model we are stating that:

$$Credit\ Spread = Hard\ information + Soft\ information \qquad (1)$$

That is, the market is correctly pricing each of the loans originated over our sample period. The hard information within our dataset contains factors such as LTV at origination and FICO score at origination and macroeconomic variables. FICO score is a nonlinear credit score hence we add $FICO^2$ to our model. Hard facts also include categorical variables related to the real estate type; condominium, single family home or planned urban development and whether the property was purchased for investment purposes. The macroeconomic variables we use are unemployment rate, personal consumer spending, lending standards and the house price index[4]. We chose these economic variables to explain the economic environment prevalent at the time of origination. After accounting for these factors with regards to credit spread, we are left with the residuals of the model which should capture soft information that is unrecorded within our dataset. If soft information was a significant predictor of credit

[4] As macroeconomic variables are calculated either as an index or percentage value, we use percent change across all to ensure consistency in results.

risk, we should observe residual as a positive and significant variable in determining the probability of default. Hence, the Stage 1 regression model involves estimating the equation:

$$CreditSpread_{i,orig\_t} = \alpha_1 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \beta Lending_{orig\_t} + \varepsilon_i \qquad (2)$$

Credit spread is calculated as the difference in yield between the mortgage rate at origination and U.S. Treasury Bond[5]. Hence credit spread would be the premium a lender will charge for a particular borrower. It is vital we analyse this variable as it signifies the pricing of the loan from the viewpoint of the lender. Since we will analyse the first stage regression at origination, the origination data set currently includes approximately 35,000 observations. In this model, $H_{l,i,orig\_time}$ is hard information 'l' of borrower 'i' at origination this also includes the macroeconomic variables as at origination time 'orig_t'. Here, $Lending_{orig\_t}$ are the lending standards as at origination. They can be interpreted as both macroeconomic variables and a proxy for soft information as they depict lender's judgement regarding the current credit market. From this model, we derive ε_i which are our residuals.

4.2. Impact of Soft Information on Default- Base Model Stage 2 Regression

In the Stage 2 regression we wish to analyse the impact of soft information in predicting default and this is carried out by estimating the following logit function:

$$Prob(Default_{it} = 1) = F(\alpha_2 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \beta Lending_{orig\_t} + \delta\varepsilon_i) \qquad (3)$$

$$\text{Where } F(x) = \frac{e^x}{(1+e^x)}$$

We will call this model, Model A. To estimate Model A, we use the panel dataset which includes quarterly observations for an individual borrower until the loan's maturity or default. In this model, ε_i are the residuals derived from the Stage 1 regression. We regress this with respect to hard information gathered at origination. Due to its judgement based characteristics, we categorise lending standards as a proxy of soft information from the borrower's perspective and include this within our model.

[5] A proper definition and calculation of the credit spread are discussed in section 5.1.2.

Following AACL's argument regarding majority of soft information being collected during the origination of the loan and taking into consideration soft information is affected by current market conditions, we analyse equation 3 only given the hard facts at origination of the loan. However, the current economic conditions would play a significant role in determining whether a borrower might default. This might be due to the income shocks that may affect a particular borrower or their decision to strategically default. To take these additional factors into consideration we estimate the Models B, C and D as explained in the next section.

4.3. Impact of Soft Information and Change in Hard and Soft Variables on Default

To make up for the time varying macroeconomic, loan related and soft information proxy variables, we devise Model B (4), Model C (5) and Model D (6). These models take into consideration the change in every macroeconomic and loan characteristic (hard facts) as well as the change in lending standards (soft information from lender's perspective). The equation for these models are shown below.

$$Prob(Default_{it} = 1) = F(\alpha_2 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \Delta H_{l,i,t} + \delta \varepsilon_i) \tag{4}$$

$$Prob(Default_{it} = 1) = F(\alpha_2 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \Delta Lending_t + \delta \varepsilon_i) \tag{5}$$

$$Prob(Default_{it} = 1) = F(\alpha_2 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \Delta H_{l,i,t} + \Delta Lending_t + \delta \varepsilon_i) \tag{6}$$

Where $\Delta H_{l,i,t} = H_{l,i,orig\_t} - H_{l,i,t}$ and it represents the change in hard information within our dataset and $\Delta Lending_t = Lending_{orig\_t} - Lending_t$ representing the change in soft information within our dataset.

4.4. Soft Information and Interaction with Borrower and Economic Variables

The Base Model helps us identify the effect of soft information in our overall dataset. Yet we are unaware as to how soft information varies for different borrowers. The next question we ask is, what might affect the significance of soft information for a particular borrower? To analyse the impact of soft information with respect to a specific borrower characteristic, we plot absolute value of residuals against the borrower characteristic. We plot against the absolute value of residuals because the residuals can take positive or negative values. We argue that positive residuals[6] would increase the credit spread charged for a given borrower (from equation 2) hence we will simply term this as "bad" soft information. On the other hand, a negative residual would decrease the credit spread charged and we will call this "good" soft information. Since we wish to analyse the total variation in soft information given particular characteristics we plot absolute value of residuals against the variable of interest. In addition, we wish to observe the impact of soft information in predicting default due to which we introduce an interaction term in our second stage regression Models A, B, C and D. The interaction term involves a categorical variable and the residuals. The value of this interaction terms helps us understand the "quality" of the soft information in predicting default for a given loan observation. In order to ensure consistency within the soft information gathered during the origination of the loan, the first stage regression estimated as in equation 2 remains consistent throughout this study. Thus the Stage 2 regression for this part of the study is shown in equation 7.

$$P(Default_{it} = 1) = F(\alpha_2 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \beta Lending_{orig\_t} + \beta_d Dummy + \delta_1 (Dummy * \varepsilon_i) + \delta_2 \varepsilon_i \quad (7)$$

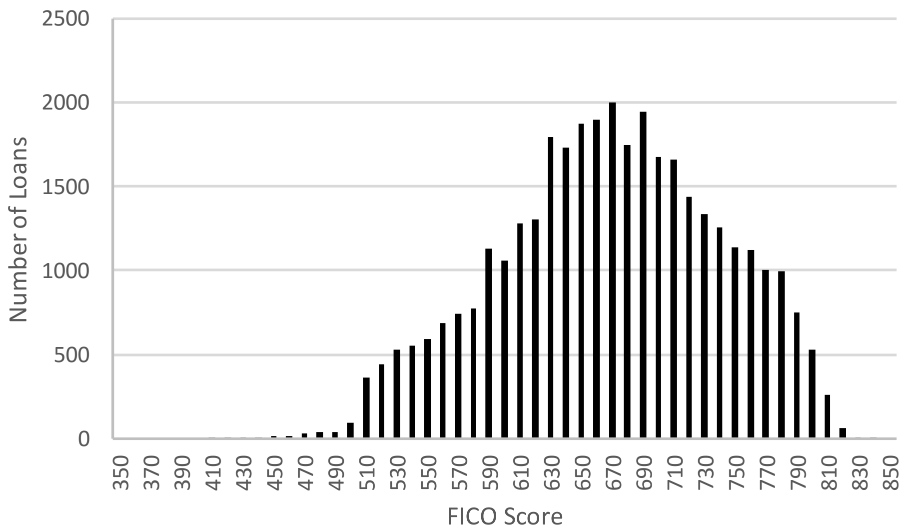
The categorical variable (*Dummy*) helps us comprehend the impact of the borrower or lender characteristic on the overall probability of default of a loan originated within that period. Moreover, the interaction term between the dummy variable and residual signifies the total impact the dummy variable has on the soft information's predictive power of probability of default. Model B, C, and D follow the same structure as the Base Model B, C and D with the additional variables $\beta_d Dummy + \delta_1 (Dummy * \varepsilon_i)$.

[6] Residuals = Observed – Predicted

4.4.1. Borrower Characteristics and Soft Information- FICO Score

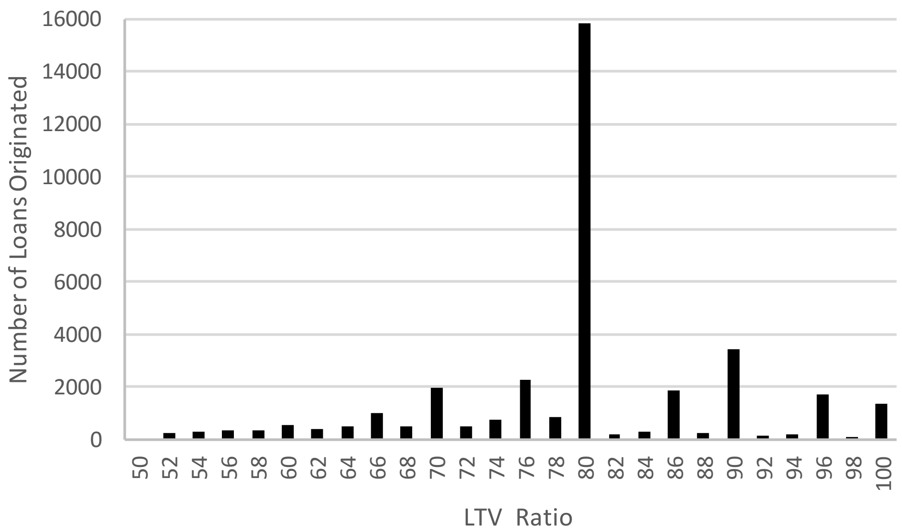
We generate categorical variables to analyse the impact of soft information for borrowers with FICO scores higher than the median score. Extending the FICO interaction model, we stratify our sample into different FICO score percentiles to analyse the effect of soft information on borrower falling in each of the percentile brackets. The FICO score for an individual borrower is one of the key characteristics in our dataset. This score represents the total capacity of a borrower to pay back his/her loan based on their individual credit history. The score ranges from 300 to 850 and the higher the score a borrower has, the greater their creditworthiness. To assess how different FICO score affects the impact of soft information for particular borrowers, we divide our dataset into four sections based on the percentile bracket a borrower falls into. The brackets are divided up into four sections: the first section consists of borrowers in the 25th percentile (FICO25), the second- borrowers between the 25th and 50th percentile (FICO50), the third- borrowers between the 50th and 75th percentile (FICO75) and the fourth between 75th and 100th percentile (FICO100). The distribution of the FICO scores within our dataset is as shown in Figure 1.

Figure 1: Number of Loans Originated at FICO Score



Another choice of variable to stratify was the loan to value ratio. However, it was soon apparent that LTV is a poor measure of borrower quality as most banks lend out at LTV between 70-80 (as can be seen from the graph below). For this reason, it is safe to say that a particular borrower might go to a bank in hopes of getting a loan. The bank will likely offer them something around the 70-80% of asset value mark hence they have 80% additional along with the current deposit they are willing to place. Within this scenario, the true characteristic of the borrower has no effect on the amount the bank is willing to lend out to the borrower hence LTV becomes a poor measure to stratify borrower quality to observe the impact of soft information. In addition to borrower FICO score and LTV ratio, we had access to a limited number of lender's zip codes which, along with property zip code, helped us calculate the total distance between the lender and the property. As previously mentioned, distance is a popular proxy for ease of availability of soft information however given we had access to only a limited number of lender zip codes, our sample size was drastically reduced and the results from the regression estimation were insignificant. For this reason, we have excluded this variable interaction from our study. In conclusion, the variables we interact with residuals to measure change in soft information given a particular borrower characteristics are: High FICO, FICO100, FICO75, FICO50 and FICO25.

Figure 2: Number of Loans Originated and LTV



4.4.2. Soft Information Variation with Economic Cycles

Do economic cycles have a major impact on the role soft information plays during pricing and the credit risk associated with a home loan? To answer this question, we look at different economic indicators and indices. There is a plethora of economic variables that indicate the past, present and future economic environment. Since we are interested in the economic situation during the origination of the loan, we use coincident indicators[7].

To analyse business cycles, we first define what a business cycle is. Data on peaks and troughs can be found on NBER dataset and using these indicators, we developed the contractionary dummy variable which would be 1 from a peak to a trough and 0 for all other periods. Along with this categorical variable, we also use the coincident economic activity index and lending standards. These were also represented as categorical variables and their descriptions are outlined in their relevant sections.

We mainly work with percentage change for each macroeconomic variable as macroeconomic variables we use are a mixture of indexes and percentage (like unemployment rate). To standardise them, we use the percent change over our sample period and using the mean we develop categorical variables for each economic indicator. Since we have 4 Stage 2 models, we will mainly focus on interpreting model A for regressions concerning economic interactions. This is as Models B, C and D contain change in macroeconomic variables and soft information as described by lending standards. Since these are included in the second stage regression, we would expect the inclusion of these effects would negate the effect of the dummy variable making it harder to interpret and analyse. We have reported the models B, C and D in our final results regardless.

4.5. Lender Effect and Soft Information

Our dataset contains approximately 1000 different originators. As a robustness check, the last study we undertake includes analysing the role of soft information after accounting for the lender effect. The lender effect is described as the extra information known by a particular lender, (this is part of the

[7] Lagging indicators, such as GDP and leading indicators such as stock futures.

definition of soft information) and is compared against the hard facts. For this model we categorise the loans originated by each lender and using this categorical separation we derive parameter for every lender in relation to one arbitrary reference group. This parameter estimate is calculated in the first stage regression as the model in equation 2 changes to:

$$CreditSpread_{i,orig\_t} = \alpha_1 + \sum_{l=1}^{L} \beta_l \, H_{l,i,orig\_t} + \beta Lending_{orig\_t} + \gamma Lender\_Effect + \varepsilon_i \qquad (8)$$

Due to the change in the first stage regression, we will observe a subsequent change in the residuals when the lender effect is taken into consideration. Hereon, we estimate the second stage Models A, B, C and D as done previously to analyse the change in significance of soft information once the lender effect is taken into consideration.

It would have been possible for us to include the parameter estimates derived for each lender in the second stage regression and analyse its impact on probability of default however, given that lender parameter estimates are calculated holding a certain lender[8] as the reference, interpretation of this coefficient in the Stage 2 regression becomes arbitrary[9].

4.5.1. Lender Characteristics and Soft information

For a limited number of lenders, we derived different information regarding their total assets, liabilities, risk weighted assets, net income and total Tier 1 and 2 Capital. From these values we generated different ratios which helped us explain the profitability, size, safety and proportion of business in lending. These ratios were treated similarly to the interaction variables in discussed in second 4.2 with the goal of analysing how changing lender characteristics affects the role soft information plays in predicting default.

After including these variables in our regression model, the total number of available observations for this model reduces from approximately 35,000 in Stage 1 to 3,000 and from 480,000

[8] Lender chosen by program generally the last within the list of lenders.
[9] We have estimated the second stage regression with parameter estimates and the values and significance of the coefficients in models A, B, C and D remain the same throughout.

in Stage 2 to 60,000. For this reason, we analyse the change in Stage 1 model by comparing an estimation with the financial ratios and without. Following the methodology in section 4.2, we estimate the Stage 2 regression using an interaction variable along with a categorical variable.

5. Empirical Analysis

5.1. Data

For this research, we are using United States loan level data collected on a quarterly basis between 2000 and 2015. Since we are interested in analysing origination level data for part of the research, it should be noted that some of the loans originated around 1990. The total dataset contains 480,914 observations after data cleaning. The process undertaken to clean data is described in the subsequent sections.

The data includes loan specific information such as the property type bought- single family, condominium or planned urban development and whether the property was meant for investment purposes. Other mortgage specific information such as the balance of mortgage at the time of recording, maturity, the interest rate at origination, interest rate during the period and balance time are all recorded. Borrower specific information includes FICO score at origination, LTV ratio at origination and time of observation. We also have data regarding the zip code of the property and the name of the lender who originated the loan and the issuer of the mortgage backed security of the loan.

Descriptive statistics of key factors of the complete dataset are presented in Table 1 and Table 2 below.

Table 1: Descriptive Statistics

Panel A: Loan-level data at origination

Variable	N	Mean	Std Dev	1st Pctl	Q1	Median	Q3	99th Pctl
LTV at Origination	35,936	79.479	9.841	52.7	75	80	84.6	100
FICO at Origination	35,936	663.472	72.258	504	615	665	717	800
Interest Rate at Origination	35,936	7.349	1.527	4.875	6.25	7.05	8.125	11.99
Credit Spread	35,936	2.465	1.449	0.115	1.43	2.227	3.25	6.835
Unemployment Rate	35,936	-1.532	2.661	-6	-3.571	-1.852	0	4.444
Personal Consumer Spending	35,936	1.414	0.334	0.65	1.2	1.307	1.742	2.094
Lending Standards	35,936	10.445	5.964	-5.7	5.6	11.5	14.5	21.2
Coincident Index	35,936	0.657	0.187	0	0.606	0.684	0.806	0.921

Panel B: Loan-level data

				No Default				
Variable	N	Mean	Std Dev	1st Pctl	Q1	Median	Q3	99th Pctl
LTV at Origination	465,919	78.765	10.195	52.2	75	80	80	100
FICO at Origination	465,919	675.439	71.316	506	628	680	730	801
Interest Rate at Origination	465,919	6.984	1.608	2.116	5.99	6.75	7.75	11.84
Credit Spread	465,919	2.243	1.411	0.105	1.24	1.95	2.975	6.66
Unemployment Rate	465,919	-1.51	2.654	-6	-3.571	-1.852	0	4.444
Personal Consumer Spending	465,919	1.424	0.332	0.736	1.2	1.307	1.742	2.094
Lending Standards	465,919	10.503	5.84	-5.6	5.6	11.5	14.5	21.2
Coincident Index	465,919	0.658	0.188	0	0.606	0.684	0.806	0.921

				Default				
Variable	N	Mean	Std Dev	1st Pctl	Q1	Median	Q3	99th Pctl
LTV at Origination	11,312	80.85	8.441	56.8	80	80	85	100
FICO at Origination	11,312	650.688	67.005	501	607	653	697	789
Interest Rate at Origination	11,312	7.574	1.515	4	6.625	7.45	8.375	11.875
Credit Spread	11,312	2.82	1.35	0.33	1.865	2.635	3.54	6.81
Unemployment Rate	11,312	-1.644	2.617	-6	-3.774	-1.961	0	4.444
Personal Consumer Spending	11,312	1.36	0.32	0.846	1.161	1.307	1.59	2.094
Lending Standards	11,312	10.311	5.883	-3.7	5.6	11.5	14.5	21.2
Coincident Index	11,312	0.681	0.148	0.155	0.606	0.684	0.806	0.881

Table 2: Descriptive Statistics of Categorical Variables

Variable	Value	Frequency	Percentage
Default	No	465,919	97.63
	Yes	11,312	2.37
Condominium	No	446,222	93.5
	Yes	31,009	6.5
Planned Urban Development	No	417,372	87.46
	Yes	59,859	12.54
Single Family Home	No	189,291	39.66
	Yes	287,940	60.34
Investment Property	No	404,115	84.68
	Yes	73,116	15.32
High FICO	No	237,327	49.73
	Yes	239,904	50.27
FICO100	No	332,884	69.75
	Yes	144,347	30.25
FICO75	No	350,211	73.38
	Yes	127,020	26.62
FICO50	No	367,277	76.96
	Yes	109,954	23.04
FICO25	No	381,321	79.9
	Yes	95,910	20.1
Coincident Index	No	420,039	88.02
	Yes	57,192	11.98
Loans to Assets	No	1,419	43.58
	Yes	1,837	56.42
Netincome to Assets	No	1,557	47.82
	Yes	1,699	52.18
Capital Adequacy Ratio	No	1,260	38.7
	Yes	1,996	61.3
Log(Total Assets)	No	1,899	58.32
	Yes	1,357	41.68

The graph below shows the frequency of loan origination during each quarterly cycle. After the 2007-2008 Global financial crisis, the origination levels dropped drastically and are far from their peak in early and mid-2000's.

Figure 3: Frequency of Total Loans Originated

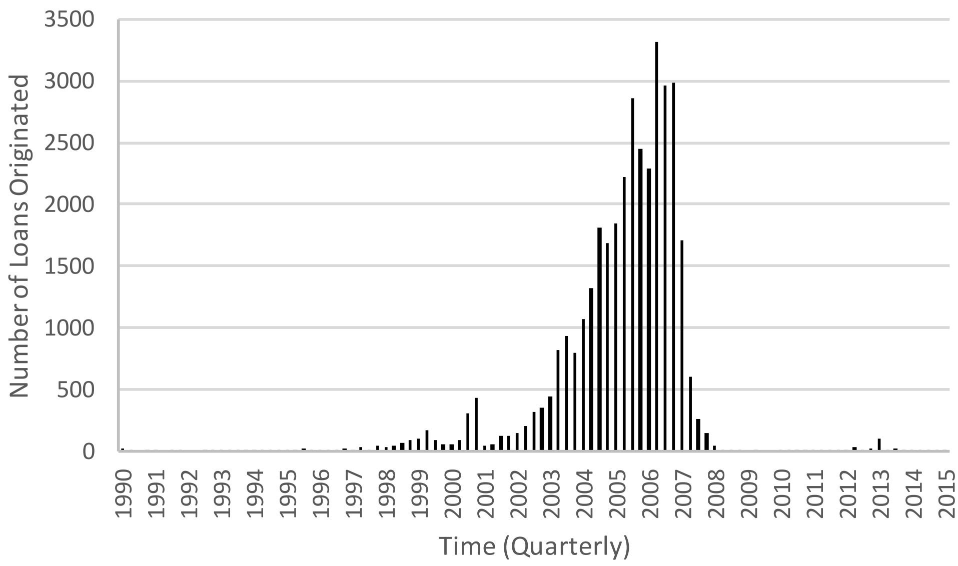
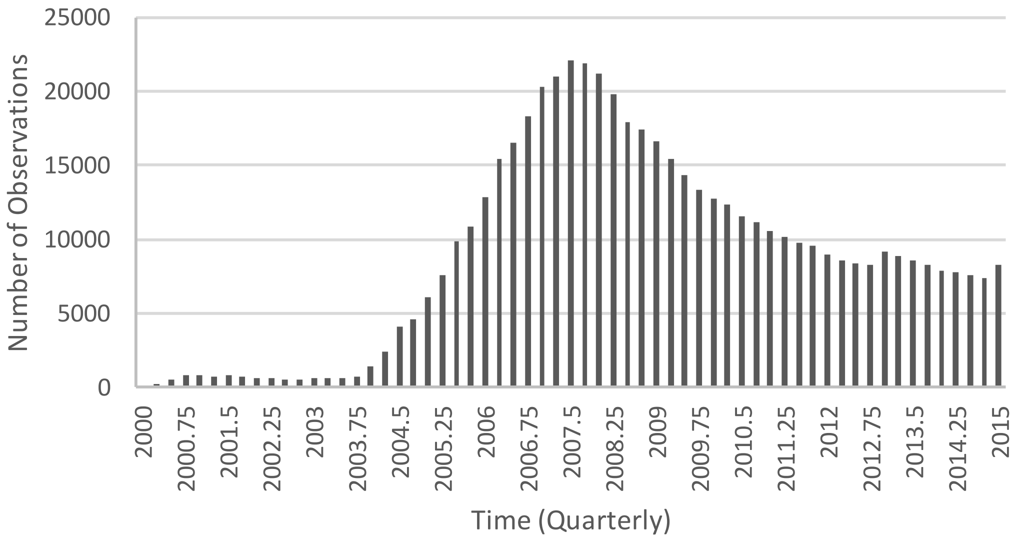


Figure 4 shows the total loan observations recorded within the dataset we use over each quarter. The distribution of the total number of loan observations is low towards the beginning of our sample and increases as time passes by.

Figure 4: Total Number of Loan Observations- Quarterly



5.2. Lender Details

The data also includes the name of the lender originating the loan and the issuer for the mortgage backed securities offered to investors. In its rawest form, the data is missing lender names for some observations. In order to maximise the number of observations to be used, two steps were taken:

1) Origination dataset was created where there is only the observation at origination for each borrower id. The total number of observations in the origination dataset were 50,000.
2) For a specific borrower [identified using borrower id] the lender name was available for some of the quarterly collected data but not for others. These lender names were backfilled using the borrower id.
3) 15,398 loan observations did not have a value for lender name after carrying out the above step. These missing values are named "other" within the dataset. The main motivation is to maximise the number of observations useable[10].

[10] To analyse if there was significant impact if the missing lender names were called "other", we generated two regression models with and without the filled lender names and did not observe any significant different amongst the two.

4) The new lender names were then merged into the original dataset with no missing values for originator name.

The following table displays the top 10 lenders and the composition of their originated loans within the dataset.

Table 3: Top 10 Originators

Originator Name	Frequency	Percent
Other	310,579	57.91
Wells Fargo	15,923	2.97
Countrywide	15,406	2.87
New Century	9,336	1.74
Fremont	8,894	1.66
American Home Mortgage	8,265	1.54
Greenpoint	7,097	1.32
WMC	6,839	1.28
Chase	6,812	1.27
Option One	6,677	1.25
Remainder	140,464	26.19

5.3. Credit Spread Calculation

The mortgage rate charged on a given loan could be argued to be the single most important number in displaying a lender's judgement with regards to a particular borrower. It not only displays borrower characteristics as determined by the lender but also displays the prevalent economic conditions. Credit spread is defined as the difference in yield between a U.S. Treasury bond and a debt security with the same maturity but of a lesser quality. To understand the impact hard and soft information has on the interest rate charged for a particular borrower we decided to account for the risk free market rate ongoing at the time of origination. Given its risk-free status, we use U.S. Treasury bond rates as the relevant risk-free rate.

Once we subtract the risk-free rate form the mortgage rate for a given loan at origination, we have the premium a borrower has charged a lender. This premium is of great importance as it excludes the expected rate lenders hope to earn on a mortgage and thus represents the riskiness as judged by the lender.

We first calculated the total maturity of the loan using origination time and maturity time we use maturity matching to analyse the treasury bond rate that should be used for each loan observation. Maturities of the loans varied from 1 month to 30 years and they were matched with the Treasury Bond maturity closest to the loan maturities given the time of observation.

Once each loan observation had its relevant treasury bond maturity, from a downloaded dataset of treasury bond rates from years 1990-2015, the treasury bond rate was the relevant risk-free rate for a given loan. From here, credit spread was calculated as the difference between mortgage rate at origination of the loan and calculated risk free rate. Figure 5 shown below shows the series progression of the mean risk-free rate, mortgage rate and credit spread during quarterly periods of origination of various loans in our dataset.

Figure 5: Credit Spread Relative to U.S. T-Bond and Mortgage Rate

Note: This figure shows the results from the credit spread calculation against the U.S Treasury Bond Rate the Mortgage Rate at origination. The credit spread is calculated as the Mortgage Rate less the U.S. Treasury Bond Rate

5.4. Elimination within the dataset

Since we are interested in observing soft information as captured during origination, we use the credit spread calculated at origination. Within the dataset, we eliminated loans where credit spread was negative and interest rate at origination was zero. This is because, origination of these low interest rate loans would not have required a lot of screening and therefore do not add much value towards identifying soft information within our dataset. From this, the useable dataset contained approximately 480,000 observations in the panel dataset and 36,000 loan observations at origination.

6. Regression Results

6.1. Base Model Stage 1- Identifying Soft Information

As we wish to analyse the existence of soft information within our dataset, we state that the soft information must be captured in the residuals of the Stage 1 regression. After estimating the first model, we arrive with the following results. Within Table 4 we have estimated 4 different models to analyse the impact of the independent variables on the dependent variable (credit spread for borrowers). Out of all the dependent variables used in the model, the most significant are LTV ratio and FICO score as their pronounced importance echoes in current literature. Hence, Model II displays the singular effect of FICO score and LTV on the credit spread[11]. Furthermore, to add to this model, we take into consideration different categorical variables indicating the type of property bought. Lastly, the fourth model analyses the impact of the main borrower characteristics combined with macroeconomic factors. The adjusted R-square for Model I is the greatest depicting approximately 40% of variation is explained by this model. Model I is the one we will be mainly using for the rest of this study.

[11] FICO2 is $FICO^2$, we use $FICO^2$ to account for the non-linearity in the variable FICO.

Table 4: Base Residual Model

	Model I	**Model II**	**Model III**	**Model IV**
Intercept	16.503***	15.304***	15.68***	16.29***
	(0.42)	(0.445)	(0.439)	(0.422)
FICO at Origination	-0.036***	-0.035***	-0.036***	-0.035***
	(0.001)	(0.001)	(0.001)	(0.001)
FICO2	2x10-5***	2x10-5 ***	2x10-5 ***	2x10-5 ***
	(0)	(0)	(0)	(0)
LTV at Origination	0.023***	0.022***	0.023***	0.022***
	(0.001)	(0.001)	(0.001)	(0.001)
Unemployment Rate	-0.002***			-0.03***
	(0.002)			(0.003)
Personal Consumer Spending	-0.03***			0.201***
	(0.022)			(0.024)
GDP	-0.198***			-0.532***
	(0.005)			(0.02)
HPI at Origination	-0.016***			-0.161***
	(0.001)			(0.005)
Lending Standards	-0.248***			-0.021***
	(0.028)			(0.001)
Condominum	-0.351***		-0.187***	
	(0.022)		(0.029)	
Planned Urban Development	-0.137***		-0.306***	
	(0.016)		(0.023)	
Single Family Home	0.541***		-0.075***	
	(0.018)		(0.016)	
Investor at Origination	0.54***		0.556***	
	(0.018)		(0.019)	
Number of Observations	35,936	35,936	35,936	35,936
Adjusted R-Square	0.372	0.2861	0.3077	0.362

From the results, we can see that high FICO score at origination reduces the predicted credit spread and a higher LTV ratio at origination increases the credit spread. These relationships between FICO score and LTV ratio are expected as a borrower with a high LTV ratio will have a greater risks associated which could lead to him or her paying a higher interest rate on their loan. The greater the FICO score, the better credit worthiness of the borrower which leads to a lower credit spread. If a borrower intends to use the property as an investment, he is charged a higher credit spread than those who wish to use the property as their primary dwelling.

In terms of the macroeconomic variables, a high unemployment rate decreases the credit spread charged at origination. This may be as during periods of high unemployment, the economy would be slowing down and as a result the banks would make up for this loss in business by charging attractive rates to worthy customers. It could also be attributed to the characteristics of the borrower as if he or she is employed during periods of high unemployment, they might communicate promising prospects of paying back their loans and not having drastic income shocks. Due to the lack of data available regarding borrower income and employment, we unfortunately cannot test for this. If the housing price index during origination is high, the overall credit spread is lower suggesting that during periods of growth (positive HPI) the interest rate charged for borrowers is lower thus promoting further growth in the housing market. This could also be attributed to optimism in the current economy in which case this result would be expected.

Tighter lending standards are depicted by a negative percentage change in the number of lenders willing to originate loans. Thus, given our parameter estimate, the tighter the lending standards, the larger the credit spread charged on an originating loan. To understand how effective these practices are, we consider Table 5 which shows the frequency of default given a loan originated during tighter vs lax lending practices. We see that the percent of loans which default during tighter lending standards is 0.8% less than defaults during lax lending standards. This shows that loans originated to borrowers during tighter lending periods were of better quality as they amount that default was lower than loans originated during lax lending standards. All variables are significant at the 1% level.

Table 5: Lending Standards and Default

Tight Lending Standards

Default (Y/N)	N	Percent	Cumulative Frequency	Cumulative Percent
0	12,982	98.41%	12,982	98.41%
1	210	1.59%	13,192	100%

Lax Lending Standards

Default (Y/N)	N	Percent	Cumulative Frequency	Cumulative Percent
0	452,937	97.61%	452,937	97.61%
1	11,102	2.39%	464,039	100%

6.2. Base Model Stage 2- Residual Model and Change in Hard and Soft Information

Next we evaluate the second stage regression model. This model takes the residuals derived from the Stage 1 regression and includes them as an additional independent variable to examine its relevance in determining credit risk. In the Stage 2 Model A regression, we collate all the hard facts at origination and display them under "Hard Facts"[12] to help analyse the changing effects of these across the different models.

From the results, we see that the residuals derived from the first stage regression are significant and positive displaying that they have predictive power of borrower default. In other words, there is additional information gathered by lenders during the origination of the loan which is unrecorded within the dataset but has significant predictive power.

Model B includes the additional impact which change in hard facts would have on the probability of default. As a quick reminder, the change in hard facts is calculated as:

$$Change\ in\ Hard\ Fact_{i,t} = Hard\ Facts_{i,orig} - Hard\ Facts_{i,t} \tag{9}$$

Even though residuals remain significant throughout the models, the magnitude of it decreases when the change in hard information is included. Since additional information is being added to our

[12] Note: these hard facts do not include lending standards as this is categorised as soft information for these models.

model (change in hard facts) the reliance on residuals to predict probability of default decreases as a new dependent variable is significant for predicting borrower default. Alongside this, the parameter estimate of hard facts increases suggesting that they are predictive of explaining default.

The change in HPI implies that given a loan originated during a downturn, the probability of its default is lower as the HPI at origination would be less than HPI at observation time which would lead to the change in HPI being an overall negative value and since the parameter estimate of change in HPI is positive, the overall probability of default is reduced.

As a borrower pays off their debt, the loan to value ratio reduces over time thus, the change in LTV should be positive (as LTV at origination is greater than LTV at observation) hence given that the parameter estimate of change in LTV is negative, the overall probability of default is reduced for a responsible borrower. With regards to personal consumer spending, as the percentage change in personal consumer spending increases during observation time, the probability of default also increases. In an economic context this result makes sense as if consumer spending increases, those consumers with poor self-control could face the problem of not being able to repay their debt. Likewise, given a loan originates during a high unemployment rate, the change in unemployment rate should be higher unemployment rate at observation time would be lower thus resulting in an overall positive change in unemployment rate. This when multiplied by the parameter estimate, leads to a lower probability of default.

Lending standards are categorised as soft information from the lender's point of view within our model. Model C includes the change in lending standards, that is determining whether they are lax or tighter compared to lending standards at origination and using this data as the change observed in soft information within our dataset. The change in lending standards is calculated similar to the change in hard facts. Given that loans originated during tighter lending periods, the value of lending standards at origination would be negative and thus the overall change in lending would have a negative sign. Given that the parameter estimate of change in lending is positive, this would lead to an overall lower probability of default for a given borrower.

The pseudo R-Square is the R-Square equivalent measure for logistic regression however, they cannot be interpreted in the same way. The pseudo R-Square is a goodness-of-fit measure which can only be evaluated within the same dataset. Hence, the pseudo R-Square values we derive in this study cannot be compared with the pseudo R-Squares in other studies due to the varying dataset. The higher the pseudo R-Square, the higher the goodness-of-fit measure. For the Stage 2 regression, the pseudo R-Square is greatest for Model D which includes change in hard and soft information along with the hard facts as gathered at origination.

Table 6: Base Model Stage 2 Regression Model A, B, C and D

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-4.949***	-5.377***	-5.173***	-5.355***
	(0.03)	(0.035)	(0.031)	(0.035)
Residual	0.123***	0.124***	0.118***	0.122***
	(0.007)	(0.008)	(0.007)	(0.008)
Hard Facts at Origination	0.485***	0.514***	0.503***	0.513***
	(0.01)	(0.011)	(0.01)	(0.011)
Change in HPI		0.049***		0.041***
		(0.004)		(0.004)
Change in LTV		-0.017*** (0)		-0.017***
				(0)
Change in Personal Spending		-0.118***		-0.139***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.037***		-0.025***
		(0.002)		(0.002)
Change in Lending			0.019***	0.007***
			(0)	(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.0255	0.0598	0.0404	0.0604

6.3. Residual Interaction Model

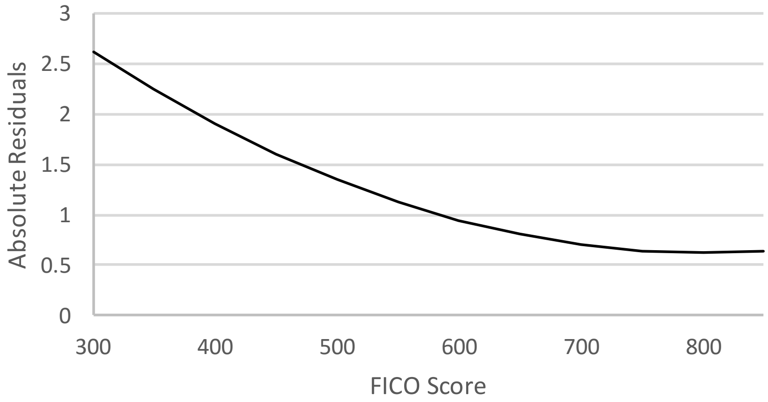
Since we have shown the existence and importance of soft information within our dataset, the next step is to analyse the change in impact of soft information given various borrower and lender characteristics. The first stage residual of our models will remain constant so as to ensure variability in the residuals is minimal. The second stage regressions will include an additional dummy variable and an interaction term between the dummy and the residuals.

6.3.1. FICO Model

A key borrower characteristic we have access to is the borrower's FICO score. For the interaction model, we create a FICO binary variable which takes the value of 1 given a borrower's FICO score is above the median and 0 otherwise. Given residuals from the Stage 1 model, we wish to analyse the impact of soft information for varying FICO scores. Residuals from Stage 1 model can be either positive or negative[13]. We are interested in the total soft information gathered for particular borrowers, hence we want to measure the total variability in soft information regardless of the sign of the residuals. To do this, we generate absolute value of the residuals and derive their mean for different FICO score buckets. From Figure 6 we can see that the borrowers with a low FICO score experience a greater impact of soft information than those borrowers with a higher FICO score. The graph can be interpreted as; borrowers with low FICO Score have on average, 2.5% of their credit spread explained by the residuals captured by our model. The proportion of credit spread defined by soft information reduces as a borrower's FICO score increases. One thing to keep in mind is this relation is not just true for residuals but is also true for credit spreads. That is borrowers with low FICO scores have higher credit spreads and borrowers with high FICO scores have lower credit spreads. This follows logic as borrowers with a high FICO score are less risky.

[13] Reminder: A positive residual would depict a negative impression cast on the agent as this would lead to a higher credit spread, on the other hand negative residual would indicate positive borrower characteristics hence affecting the pricing of the loan positively.

Figure 6: Variability in Absolute Residuals and FICO Score



Next we analyse the predictive power of this gathered soft information in relation to borrower default. From Table 7 *HIGH Fico Dummy* is significant in Model B and D and insignificant in Model A and C. Given the high Pseudo R-square value for Model D, this is the model we will be interpreting. Residuals and the interaction term *Residual*High FICO Dummy* are positive and significant throughout the table. To get this result into perspective, residuals, as obtained at origination, for a particular borrower were negative (that is positive soft information) given this is the case, and the borrower has a high FICO score, their overall probability of default is reduced as the product of the parameter estimate (which is positive) and the interaction term (which is negative) will be negative. This shows that given the borrower has a high FICO score and their soft information component was positive, the overall probability of their default is reduced.

Thus, from our analysis, borrowers with a high FICO score have a lower probability of default. Moreover, the soft information obtained for these borrowers was a significant and strong predictor of default.

Table 7: Stage 2 Regression for High FICO Score[14]

	Model A	Model B	Model C	Model D
Intercept	-4.947***	-5.267***	-5.157***	-5.237***
	(0.046)	(0.051)	(0.047)	(0.051)
Residual	0.08***	0.073***	0.076***	0.072***
	(0.008)	(0.009)	(0.008)	(0.009)
High FICO Dummy	-0.008	-0.085***	-0.019	-0.09***
	(0.025)	(0.026)	(0.025)	(0.026)
Residual*High FICO Dummy	0.174***	0.192***	0.173***	0.191***
	(0.015)	(0.017)	(0.016)	(0.017)
Hard Facts at Origination	0.485***	0.488***	0.499***	0.484***
	(0.013)	(0.014)	(0.013)	(0.014)
Change in HPI		0.046***		0.039***
		(0.004)		(0.004)
Change in LTV		-0.017***		-0.017***
		(0)		(0)
Change in Personal Spending		-0.119***		-0.14***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.038***		-0.026***
		(0.002)		(0.002)
Change in Lending			0.019***	0.007***
			(0)	(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.027	0.061	0.042	0.062

6.3.2. FICO Stratification Model

We extend the FICO Interaction Model to analyse the impact FICO score has given different quantiles. In order to carry this out, we have created four new dummy variables divided up into quantiles. They are assigned the value of 1 given a borrower falls in the specified quantile and 0 otherwise. The distribution of default amongst the different quantiles is as given in Table 8.

[14] Stage 1 Regression was Model I from Base Model

Table 8: Default Frequency and FICO quartiles

	Default (1=Yes, 0=No)	Frequency	Percent	Cumulative Frequency	Cumulative Percent	TOTAL
	0	93,043	96.61	93,043	96.61	
FICO25	1	3,264	3.39	96,307	100	96,307
	0	106,992	96.55	106,992	96.55	
FICO50	1	3,828	3.45	110,820	100	110,820
	0	124,623	97.39	124,623	97.39	
FICO75	1	3,344	2.61	127,967	100	127,967
	0	143,759	98.59	143,759	98.59	
FICO100	1	2,061	1.41	145,820	100	145,820

FICO Category	FICO Score
FICO25	300-614
FICO50	615-654
FICO75	655-716
FICO100	717-850

The second stage results for borrower lying in the first quartile are displayed below in Table 9. This is the category of borrowers with the lowest FICO scores. The parameter estimate of the dummy is negative indicating loans originated with lower FICO scores had a lower probability of default. Moreover, the results are significant at the 1% level. This is a surprising result as borrowers with lower FICO scores are expected to be riskier hence have a higher chance of repeating history and defaulting on their loans. This might raise the question of the true strength of the FICO score as the main determinant of creditworthiness of the borrower. One might also argue, given that a borrower lies in the 25th percentile they would have a poor credit history as a result of being unsuccessful in meeting their debt obligations. Perhaps these borrowers have learnt from their mistakes and are concerned with making more timely payments to not loose on of the greatest assets they may have ever acquired—their home. However, this variable should not be analysed standalone as the interaction variable between the residuals and the dummy are also included within this model.

The residuals for quartile 1 models are significant and positive and their parameter estimates are greater than that for quartile 2, 3 and 5 models. This result is in line with Figure 6. However, the interaction variable *Residual*FICO25* is negative and significant and hence can be interpreted as given a borrower lies in the first quartile, the predictive power of the residuals would be lower although, not completely irrelevant. We can say that the quality of the soft information gathered for these borrowers is not very high in terms of predicting default. *Hard Facts at Origination* for borrowers in the 25th percentile have the largest parameter estimate compared to the other quartiles and is also significant at the 1% level. This might suggest that borrowers on the lower end of the spectrum are more affected by macroeconomic variables faced at origination.

Similar to quartile 1, Table 10 displays the results for quartile 2. The interaction variable for quartile 2 is negative. These results might suggest that there is noise within our model and this is consistent with recent events as subprime loans were a major catalyst for the GFC in 2008. We tested for this noise by attempting to remove the period in our sample when the house prices were accelerating rapidly. Demyanyk and Hemert (2009) define the period between 2003-2005 as the period of highest appreciation in prices. We tested for this excess noise by first disregarding the loans in our dataset between 2003-2005 and then between 2005-2015[15]. After running the models, the results derived were consistent with Table 10, hence eliminating our theory of the housing bubble caused majority of the noise within our model.

Quartiles 2 and 3 represent borrowers collectively falling between the 25th and 75th quartile. Due to similarity in their signs, we will study them together. The second stage results for quartile 2 and 3 are displayed in Tables 10 and 11. The parameter estimate for both dummy variables are positive and significant suggesting that loans originated for borrowers falling in this category have a greater chance of default. This is supported by Table 8 wherein we see the number of defaults for borrowers in quartiles 2 and 3 is greatest amongst the four categories. Moreover, from the FICO frequency in Figure 1 we observe that majority of the loans were originated for borrowers with FICO scores between 615-665

[15] This was to completely remove the impact of GFC from our model.

which is the interval of quartile 2 and 3. Recalling that Griffin and Maturana (2016) state that lenders originated loans to borrowers with good credit scores but other characteristics that display poor creditworthiness can be linked to this result.

Overall, the value of residuals remains positive and significant showing the soft information gathered by lenders contains significant predictive power. The interaction variable for quartile 2 is negative and this could be due to similar noise related reasons as per the first quartile. However, for quartile 3, the interaction variable is positive and consistently significant suggesting that soft information gathered for these borrowers at origination has a positive impact in predicting default. Similar results are observed for borrowers lying in the top quartile. This result could be linked to the Keys, Mukherjee, Seru and Vig (2010) in the sense that lenders do not have strong incentives to screen borrowers with high FICO scores hence, whatever little interaction that might have taken place between them is significant in predicting default.

Lastly, we look at quartile 4. This quartile consists of the borrowers with the highest FICO scores between 717 to 850. The parameter estimate of their dummy is negative and significant indicating that borrowers with a FICO lying in the top quarter have a lower probability of default this is in line with our hypothesis. From the means graph below, we can see that for total loans originated over each quarter, the mean is around 717 before 1995 and after 2010. Thus we could conclude that there is a possibility of loans in quartile 4 mainly originating during these time periods. This time period falls beyond the bounds of the GFC. Hence, from all the models we have derived so far, Quartile 4 model is least affected by the noise caused by the GFC. The interaction variable is positive and consistently significant across the four models keeping in line with the logic that lenders began screening borrowers thoroughly outside of the housing bubble period thus validating our use of residuals as soft information for this category.

Figure 7: Mean FICO Score in Each Quarter

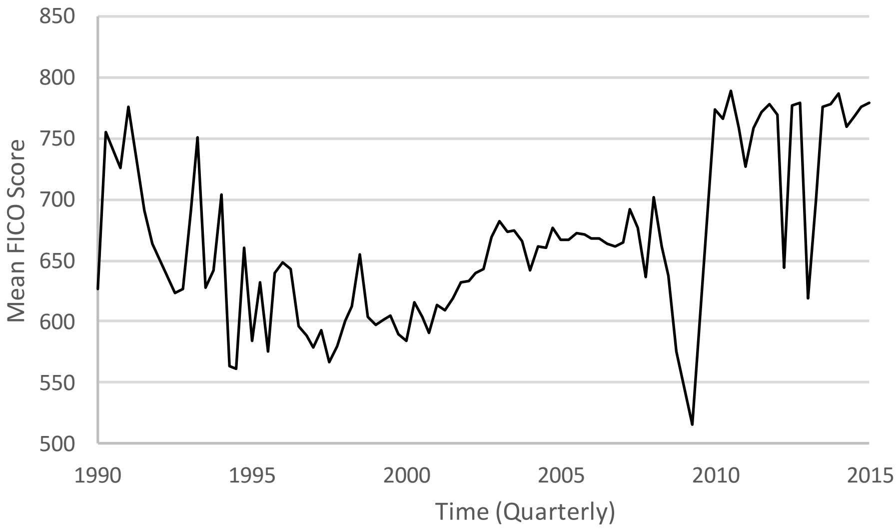


Table 9: Stage 2 Regression for Quartile 1

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-5.486***	-5.812***	-5.681***	-5.787***
	(0.038)	(0.043)	(0.039)	(0.043)
Residual	0.151***	0.162***	0.146***	0.16***
	(0.009)	(0.009)	(0.009)	(0.009)
FICO25	-0.723***	-0.539***	-0.691***	-0.531***
	(0.029)	(0.03)	(0.03)	(0.03)
Residual*FICO25	-0.046***	-0.084***	-0.047***	-0.083***
	(0.014)	(0.016)	(0.014)	(0.016)
Hard Facts at Origination	0.742***	0.716***	0.747***	0.712***
	(0.014)	(0.016)	(0.015)	(0.016)
Change in HPI		0.06***		0.053***
		(0.004)		(0.004)
Change in LTV		-0.016***		-0.016***
		(0)		(0)
Change in Personal Spending		-0.114***		-0.133***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.035***		-0.024***
		(0.002)		(0.002)
Change in Lending			0.018*** (0)	0.006***
				(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.032	0.064	0.046	0.064

Table 10: Stage 2 Regression for Quartile 2

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-5.036***	-5.438***	-5.255***	-5.415***
	(0.032)	(0.036)	(0.033)	(0.036)
Residual	0.138***	0.138***	0.133***	0.136***
	(0.008)	(0.009)	(0.008)	(0.009)
FICO50	0.315***	0.269***	0.305***	0.268***
	(0.02)	(0.02)	(0.02)	(0.02)
Residual*FICO50	-0.056***	-0.044***	-0.055***	-0.044***
	(0.015)	(0.017)	(0.016)	(0.017)
Hard Facts at Origination	0.485***	0.51***	0.502***	0.508***
	(0.01)	(0.011)	(0.01)	(0.011)
Change in HPI		0.049***		0.041***
		(0.004)		(0.004)
Change in LTV		-0.017***		-0.017***
		(0)		(0)
Change in Personal Spending		-0.116***		-0.137***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.037***		-0.025***
		(0.002)		(0.002)
Change in Lending			0.019***	0.007***
			(0)	(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.028	0.062	0.043	0.062

Table 11: Stage 2 Regression for Quartile 3

	Model A	Model B	Model C	Model D
Intercept	-5.142***	-5.545***	-5.358***	-5.521***
	(0.034)	(0.039)	(0.035)	(0.039)
Residual	0.118***	0.117***	0.114***	0.116***
	(0.008)	(0.009)	(0.008)	(0.009)
FICO75	0.294***	0.247***	0.283***	0.244***
	(0.022)	(0.022)	(0.022)	(0.022)
Residual*FICO75	0.052***	0.053***	0.048***	0.052***
	(0.017)	(0.019)	(0.018)	(0.019)
Hard Facts at Origination	0.527***	0.551***	0.543***	0.549***
	(0.011)	(0.011)	(0.011)	(0.011)
Change in HPI		0.05***		0.043***
		(0.004)		(0.004)
Change in LTV		-0.017***		-0.017***
		(0)		(0)
Change in Personal Spending		-0.117***		-0.138***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.037***		-0.025***
		(0.002)		(0.002)
Change in Lending			0.019*** (0)	0.007***
				(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.027	0.061	0.042	0.062

Table 12: Stage 2 Regression for Quartile 4

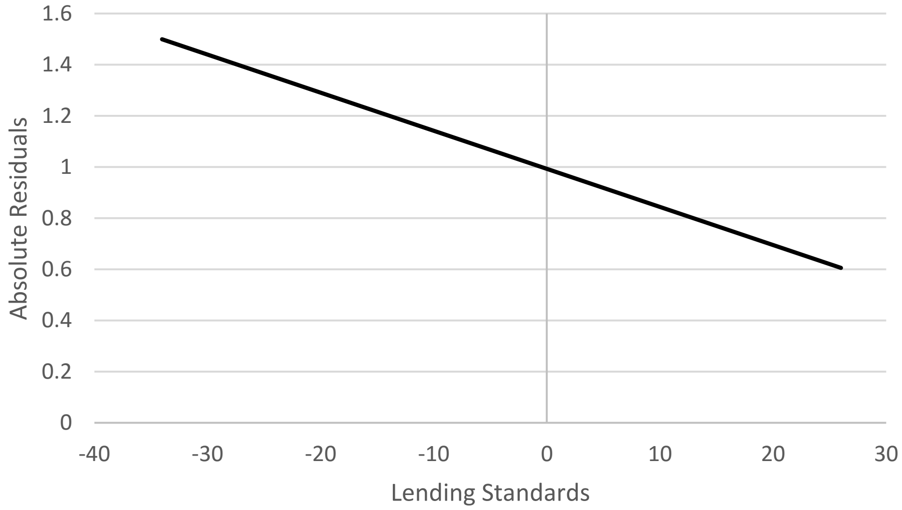
	Model A	Model B	Model C	Model D
Intercept	-4.662***	-5.026***	-4.884***	-4.998***
	(0.038)	(0.043)	(0.039)	(0.043)
Residual	0.094***	0.093***	0.09***	0.092***
	(0.007)	(0.008)	(0.007)	(0.008)
FICO100	-0.373***	-0.43***	-0.376***	-0.434***
	(0.029)	(0.029)	(0.029)	(0.029)
Residual*FICO100	0.246***	0.281***	0.249***	0.279***
	(0.02)	(0.022)	(0.02)	(0.022)
Hard Facts at Origination	0.409***	0.421***	0.426***	0.418***
	(0.012)	(0.013)	(0.012)	(0.013)
Change in HPI		0.043***		0.035***
		(0.004)		(0.004)
Change in LTV		-0.017***		-0.017***
		(0)		(0)
Change in Personal Spending		-0.117***		-0.139***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.038***		-0.026***
		(0.002)		(0.002)
Change in Lending			0.019***	0.007***
			(0)	(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.028	0.063	0.043	0.064

6.3.3. Residual-Economic Model

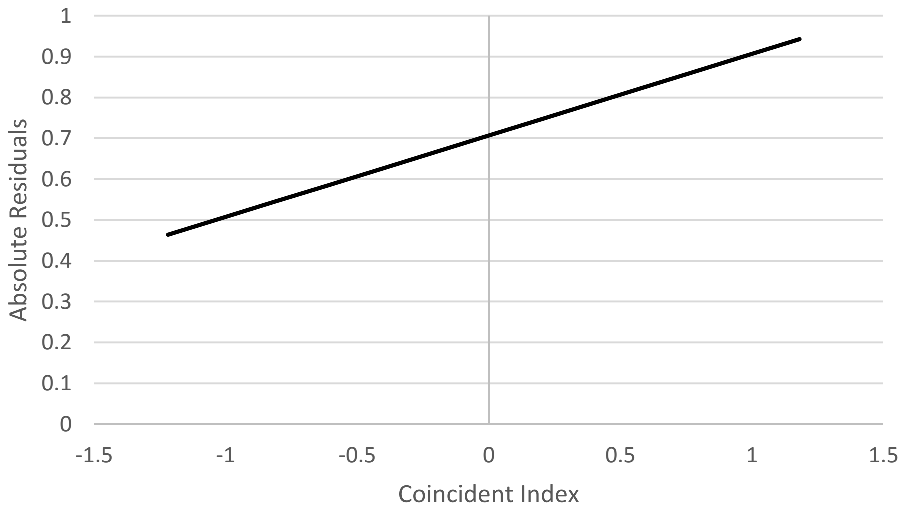
Next we analyse the economic cycles and its relevant impact on soft information gathered by borrowers. From the recent financial crisis, we know that as the economy goes through different cycles, banker's readiness to lend gets influenced as a result. To observe the impact varying economic cycles have on soft information gathered, we analyse interaction between soft information and three economic variables: the prevalent lending standards, contractionary period indicator and the coincident index.

Figure 8: Variation in Residuals and Lending Standards and Coincident Index

a) Variation in Residuals and Lending Standards



b) Variation in Residuals and Coincident Index



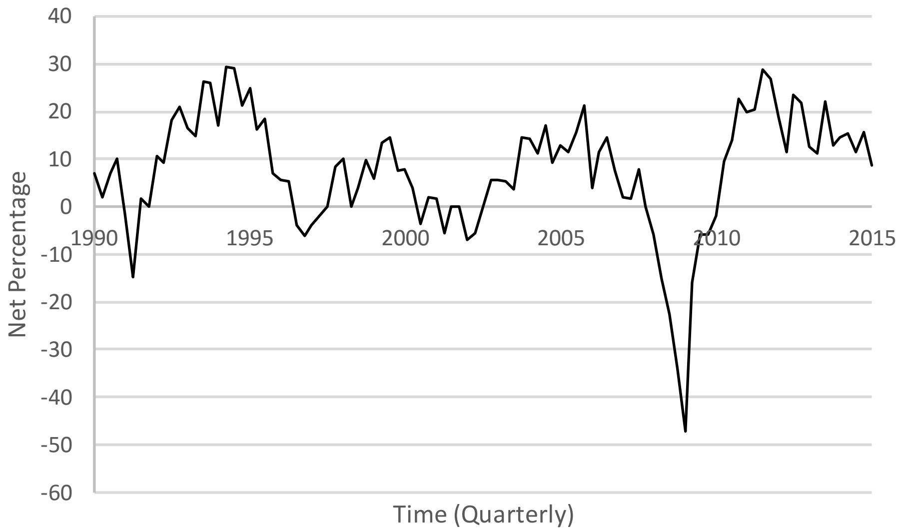
In Figure 8, we show the variation in residuals against the lending standards and the coincident index. From this we see that lending standards follows our original hypothesis as the impact of soft information during periods when domestic banks are willing to lend is low as opposed to tighter lending standards. However, the variation amongst residuals and coincident index does not match our hypothesis as coincident index increases, the variation in residuals increases with it.

6.3.4. Lending Model

The Lending Model aims to analyse the change in soft information given a domestic bank's willingness to originate consumer instalment loans. A consumer instalment loan is a loan that has scheduled payments. The maturities of these loans can last from a number of months up to 30 years. Hence, mortgages fall under this category. The Board of Governors of the Federal Reserve System (FRB) conduct an electronic survey to be completed by senior loan officers. This survey asks a number of questions about the existing credit market conditions as the FRB has an overarching goal to gauge the qualitative and quantitative information regarding credit availability, demand, evolving developments and lending practices in the U.S. market. This survey is of utter importance as it helps regulators understand the views of banks regarding the credit market and other banking developments.

For this study, we gathered data on the willingness of domestic banks to originate consumer instalment loans. This data is a net percentage of banks which reported an increase in willingness to make consumer loans over the period of 1990 to 2016. We calculate a lending dummy called *Low_Lending* which is 1 if the net percentage increase is less than or equal to 0. *Low_Lending* periods depict reluctance to originate loans and as a result, actioning with extreme caution to ensure least loss possible. The following figure graphs the net percentage increase in willingness of domestic banks to make loans.

Figure 9: Willingness of Domestic Banks to Originate Consumer Instalment Loans



The result derived from the second stage regression is displayed in Table 13. Here we can see that after introducing *Low Lending*, overall residual is positive and significant. The Stage 2 regressions help us analyse the importance of *Low Lending* in determining default. *Low Lending* is negative and significant showing that loans originated during periods when domestic banks were unwilling to originate instalment loans have a lower probability of default than loans originated in other periods. The interaction variable *Residual*Low Lending* depicts the additional impact of soft information gathered during origination. This parameter estimate is positive and significant at the 1% level in model A and positive and significant at the 10% level in model B suggesting that the additional soft information gathered during the origination of a loan in a poor credit market is a significant predictor of default.

Table 13: Stage 2 Regression for Low Lending Standards[16]

	Model A	Model B	Model C	Model D
Intercept	-4.965***	-5.383***	-5.18***	-5.361***
	(0.03)	(0.035)	(0.031)	(0.035)
Residual	0.118***	0.119***	0.113***	0.118***
	(0.007)	(0.008)	(0.007)	(0.008)
Low Lending	-0.8***	-0.081	-0.511***	-0.017
	(0.074)	(0.076)	(0.074)	(0.077)
Residual*Low Lending	0.141***	0.109***	0.144***	0.106***
	(0.039)	(0.039)	(0.039)	(0.039)
Hard Facts at Origination	0.497***	0.517***	0.51***	0.515***
	(0.01)	(0.011)	(0.01)	(0.011)
Change in HPI		0.049***		0.042***
		(0.004)		(0.004)
Change in LTV		-0.017***		-0.017***
		(0)		(0)
Change in Personal Spending		-0.118***		-0.14***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.037***		-0.025***
		(0.002)		(0.002)
Change in Lending			0.018*** (0)	0.007***
				(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.027	0.06	0.041	0.061

6.3.5. Contractionary Model

For this model, we derive data regarding business cycles from NBER and use an indicator variable to represent 1 if the economy was undergoing a contractionary phase during the quarter and 0 otherwise. The results from the second regression model are mixed. The parameter estimate is negative for Model A and C and positive for Model B and D. The key change between the models is Model B and D include changes in economic variables while A and C do not. The inclusion of change in economic variables in Models B and D colours the impact of the Contractionary variable.

[16] Low Lending standards refers to quarters where domestic banks were unwilling to originate loans.

Given a loan has originated during a contractionary period of the lending cycle, the credit spread charged on the loan would be expected to be higher (since the economy is in a downturn) this would mean, the value of the residuals would be positive (signifying negative soft information). The main lender judgement which would have a play in this area would be related to the economic factors in the current environment. Hence for safety, the credit spread charged would be expected to be higher. Given this, the likelihood of a residual being positive is greater thus given the parameter estimate is negative, the total impact on the probability of default is reduced.

These results could be coloured due the few number of observations during the contractionary period that we have access to. Another reason behind this could be related to the nature of the contractionary indicator. Between 1990-2015, we detected 12 quarters where the economy was undergoing a contractionary phase, on the other hand, in the same time period, there were 23 quarters when lenders were not willing to originate consumer instalment loans. From this we could gather that contractionary indicator acts somewhat like a lagging indicator for extreme situations in an economic cycle. However, lending standards is a lot more contingent as it is based on lender's judgement and are a better representation of the current economic environment. This could explain the differing results between the two.

Table 14: Stage 2 Regression for Contractionary Periods

	Model A	Model B	Model C	Model D
Intercept	-4.951***	-5.375***	-5.172***	-5.353***
	(0.03)	(0.035)	(0.031)	(0.035)
Residual	0.13***	0.127***	0.123***	0.125***
	(0.007)	(0.008)	(0.007)	(0.008)
Contractionary	-0.665***	0.353***	-0.418***	0.296**
	(0.116)	(0.119)	(0.117)	(0.119)
Residual*Contractionary	-0.101**	-0.125***	-0.097**	-0.124***
	(0.043)	(0.046)	(0.042)	(0.046)
Hard Facts at Origination	0.488***	0.513***	0.504***	0.511***
	(0.01)	(0.011)	(0.01)	(0.011)
Change in HPI		0.049***		0.041***
		(0.004)		(0.004)
Change in LTV		-0.017*** (0)		-0.017*** (0)
Change in Personal Spending		-0.119***		-0.14***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.037***		-0.026***
		(0.002)		(0.002)
Change in Lending			0.018*** (0)	0.007***
				(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.026	0.06	0.041	0.061

6.3.6. Coincident Index Model

The Coincident Economic Activity Index for the United States consist of four indicators: the unemployment rate, nonfarm payroll employment, average hours worked in manufacturing and wages and salaries. Due to this reason, part of the noise associated with short-term indicators is eliminated hence the index proves to be a better measure. In our study, we use this index as a prime indicator to gauge economic strength given a particular period. We derive the data from The Federal Reserve Bank of St. Louis over the period 1990 to 2015 and it is shown in Figure 10.

Figure 10: Coincident Economic Activity Index for the United States

Table 15 displays the Stage 2 models for the Coincident Index. The Coincident Index Dummy is defined to be 1 if the percent change of the index is less than the median percent change and 0 otherwise. In the stage 2 model, soft information has a standalone significant impact on the probability of default as it is a positive and significant predictor of default. Similar to previous models, loans originated during downturns have a lower probability of default and the interaction variable has a significant parameter estimate however this value is insignificant.

In conclusion, given a loan originates during an economic downturn, the probability of default is lower across all models. The soft information gathered is most significant during tight lending standards and least prominent in predicating default during contractionary periods. We state this could be due to the nature of contractionary and lending variable as lending is more contingent than contractionary.

Table 15: Stage 2 Regression for Low Coincident Index

	Model A	Model B	Model C	Model D
Intercept	-4.911***	-5.362***	-5.131***	-5.34***
	(0.031)	(0.035)	(0.031)	(0.035)
Residual	0.12***	0.123***	0.114***	0.122***
	(0.008)	(0.008)	(0.008)	(0.008)
CoincidentIndex	-0.793***	-0.193***	-0.608***	-0.191***
	(0.038)	(0.04)	(0.038)	(0.04)
Residual*CoincidentIndex	0.021	0.008	0.026	0.007
	(0.021)	(0.023)	(0.021)	(0.023)
Hard Facts at Origination	0.497***	0.518***	0.51***	0.516***
	(0.01)	(0.011)	(0.01)	(0.011)
Change in HPI		0.047***		0.04***
		(0.004)		(0.004)
Change in LTV		-0.016***		-0.016***
		(0)		(0)
Change in Personal Spending		-0.112***		-0.134***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.037***		-0.025***
		(0.002)		(0.002)
Change in Lending			0.017***	0.007***
			(0)	(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.031	0..06	0.043	0.061

6.4. Estimates Model

In this section, we study if the significance of the residuals remains consistent after controlling for a lender effect. For this, we include the lender effect in our Stage 1 regression and this would influence the residuals derived. Since soft information is only known by the lender, accounting for a lender effect in Stage 1 regression would capture some of this information as we are regressing each lender against the credit spread of the loans they have originated. The lender estimates are derived holding one lender as a reference category.

From the Stage 1 regressions, the signs of the variables are very similar to the signs of the base model. Due to the large number of originators available, we do not present each lender’s parameter estimate instead, we indicate that this effect has been accounted for under “*Lender Effect”*.

Following Stage 1 regression, our main objective is to study the change in soft information after accounting for a lender effect. To do this, we estimate Stage 2 regression Model A, B, C and D. From the results shown in Table 17, the residuals remain positive and significant, hence from our analysis, this role soft information plays in determining default does not change greatly after including a lender effect within our model.

Table 16: Stage 1 Regression with inclusion of Lender Estimate

	Model I	**Model II**	**Model III**	**Model IV**
Intercept	19.511***	19.078***	19.414***	19.076***
	(0.887)	(0.932)	(0.918)	(0.901)
Lender Effect	INCLUDED	INCLUDED	INCLUDED	INCLUDED
FICO at Origination	-0.039***	-0.039***	-0.04***	-0.038***
	(0.001)	(0.001)	(0.001)	(0.001)
FICO*FICO	2.2x10-5***	2.2x10-5***		2.18x10-5***
	(0)	(0)		(0)
LTV at Origination	0.023***	0.022***	0.023***	0.022***
	(0.001)	(0.001)	(0.001)	(0.001)
Unemployment Rate	-0.001***			0*** (0.002)
	(0.002)			
Personal Consumer Spending	-0.035***			-0.035***
	(0.022)			(0.022)
HPI	-0.184***			-0.183***
	(0.005)			(0.005)
Lending Standards	-0.019***			-0.019***
	(0.001)			(0.001)
Condominium	-0.184***		-0.136***	
	(0.027)		(0.028)	
Planned Urban Development	-0.291***		-0.249***	
	(0.022)		(0.023)	
Single Family	-0.067***		-0.02***	
	(0.016)		(0.016)	
Investor at Origination Time	0.563***		0.579***	
	(0.018)		(0.019)	
Number of Observations	35,937	35,937	35,937	35,937
R-Square	0.423	0.36	0.381	0.403

Table 17: Stage 2 Regression with Lender Effect Included

	Model A	Model B	Model C	Model D
Intercept	-4.755***	-5.116***	-4.967***	-5.091***
	(0.026)	(0.03)	(0.027)	(0.031)
Residual	0.111***	0.119***	0.106***	0.118***
	(0.007)	(0.008)	(0.008)	(0.008)
Hard Facts at Origination	0.418***	0.426***	0.432***	0.424***
	(0.008)	(0.009)	(0.008)	(0.009)
Change in HPI		0.044***		0.036***
		(0.004)		(0.004)
Change in LTV		-0.017***		-0.017***
		(0)		(0)
Change in Personal Spending		-0.123***		-0.144***
		(0.013)		(0.013)
Change in Unemployment Rate		-0.039***		-0.027***
		(0.002)		(0.002)
Change in Lending			0.019***	0.007***
			(0)	(0.001)
Number of Observations	480,914	480,914	480,914	480,914
Pseudo R-Square	0.0241	0.0575	0.0389	0.058

6.4.1. Estimates Model Comparison with and without Financial Ratios

Within our dataset, we had access to a range of different lender characteristics for a handful of lenders. This limited availability of data reduced our Stage 1 sample size from 35,000 to 3,256. We calculated the loans to total bank assets ratio, capital adequacy ratio, net income to assets ratio and the size of the total bank as logarithm function of its net income. The loans to total bank assets ratio is an indication of what proportion of the bank's total assets is related to the lending business a bank participates in. The capital adequacy ratio is the ratio of Tier 1 and 2 Capital over total risk weighted assets. This ratio signifies the safety of a bank as it is used to protect the bank's depositors and to promote economic stability of the bank. It is also used by regulators around the globe to monitor safety of the financial system. The net income to assets ratio defines the profitability of the bank. It signifies the banks total return on earnings. Lastly, to measure the size of a bank, we consider its total assets. However, the total assets vary from one lender to another in magnitude and often comparison between a large lender and a smaller lender is made difficult. For this reason, we normalise the total bank assets by calculating their log.

First we compare the estimates model with and without financial ratios and in order to do this, we ensure our dataset contains the same number of observations. Column 2 shows results without inclusion of financial ratios while column 3 shows the change once financial ratios are included in the model. The results show that banks which focus more on lending as their business function, are larger in size and profits have a positive parameter estimates suggesting they might charge their lenders a higher premium however these values are insignificant at the 10% level. A bank with greater capital, defined by its CAR, tends to not charge a large premium for its customers, this parameter estimate is also insignificant at the 10% level. The insignificance of the parameter estimates could be attributed to the small sample size we are working with for these models. Given similar information on more borrowers would result in a more significant estimation model. The sign change in Table 18 only takes place for housing price index. Interpretation would suggest as the percent change in housing price index increases, the credit spread decreases. This is the correct sign for the relationship between housing price index and credit spread hence we could say inclusion of financial ratios improves the model slightly. The R-square is slightly greater for models with financial ratios than the model without.

Table 18: Stage 1 Comparison between Lender Financial Ratios Included and Excluded

	Without Financial Ratios	**With Financial Ratios**
Intercept	15.335***	13.256***
	(1.261)	(2.756)
Lender Effect	Included	Included
FICO at Origination	-0.04***	-0.04***
	(0.004)	(0.004)
FICO*FICO	2.61x10-5***	2.61x10-5*** (0)
	(0)	
LTV at Origination	0.017***	0.019***
	(0.002)	(0.002)
Loans to Assets		0.474***
		(0.967)
Net Income to Assets		1.659***
		(4.347)
Capital Adequacy Ratio		-0.225***
		(1.994)
Log(Assets)		0.124***
		(0.152)
Unemployment Rate	0.014***	-0.013***
	(0.006)	(0.006)
Personal Consumer Spending	-0.413***	-0.048***
	(0.053)	(0.062)
HPI	0.009***	-0.187***
	(0.001)	(0.015)
Lending Standards	-0.024***	-0.004***
	(0.003)	(0.003)
Condominium	-0.1***	-0.085***
	(0.069)	(0.069)
Planned Urban Development	-0.14***	-0.139***
	(0.061)	(0.061)
Single Family	-0.089***	-0.097***
	(0.043)	(0.043)
Investor at Origination Time	0.515***	0.518***
	(0.046)	(0.046)
Number of Observations	3,256	3,256
R-Square	0.552	0.553

Table 19: Stage 2 Regression with Lender Estimate and Lender Financial Ratios

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-5.542***	-5.739***	-5.568***	-5.737***
	(0.075)	(0.082)	(0.075)	(0.083)
Residual	0.138***	0.125***	0.133***	0.125***
	(0.025)	(0.025)	(0.025)	(0.025)
Hard Facts	0.82***	0.826***	0.816***	0.826***
	(0.024)	(0.025)	(0.024)	(0.025)
Change in HPI		0.069***		0.069***
		(0.01)		(0.01)
Change in LTV		-0.005***		-0.005***
		(0.001)		(0.001)
Change in Personal Spending		-0.001947		-0.061*
				(0.034)
Change in Unemployment Rate		-0.003		-0.002
		(0.004)		(0.006)
Lending Change			0.005***	0.001
			(0.001)	(0.002)
Number of Observations	59,767	59,767	59,767	59,767
Pseudo R-Square	0.0801	0.0889	0.0812	0.0889

Table 19 shows the Stage 2 regression model with financial ratios included in Stage 1 regression. The value of residuals is positive and significant as is Hard Facts as gathered at origination. The parameter estimate of hard facts gathered at origination increases drastically from Table 6 however, this direct comparison can be questioned due to the drastically different dataset size. The change in hard facts is insignificant even though the sign of the parameter estimates remains constant. From this analysis, we can conclude soft information has a positive and significant effect in predicting default after controlling for a lender effect as well as lender characteristics. Table 20 shows the percent of borrowers who defaulted and the financial ratios of the lenders originating those loans.

Table 20: Frequency Table of Default given Lender Financial Ratio

	Large Bank		**High CAR**		**High Net Income to Assets**		**High Loan to Assets**	
Default Time	0	1	0	1	0	1	0	1
0	42.80%	54%	68.20%	28.60%	38.79%	58.01%	33.97%	62.82%
1	2.17%	1.03%	2.08%	1.12%	2.09%	1.11%	1.10%	2.10%

Next, we are interested in analysing the variation in residuals given the different lender characteristics, as a quick recap, the variation is analysed by obtaining the absolute value of residuals and plotting their marginal effect[17]. We develop figures for lender size, profitability, focus on lending and safety to study the relationship.

Generally, larger banks have a multitude of managerial layers and are organised in a heirarchial structure. Berger and Udell (2002) model inner workings of relationship banking and argue that restructuring banks to be smaller hence reducing the layers of mangers eliminates the agency problems created as it would promote the flow of soft information much easily within the banks. From this we can state that for the larger banks, soft information might play a smaller role as its communication across the banking structure would be limited due to the magnitude of managerial levels they contain. This is depicted in Figure 11 (a) where we see as the size of the bank increases, the variation in residuals as they explain the credit spread charged reduces.

Bank profitability may also be linked to bank size as the larger the bank, the greater the number of opportunities present to make good business. From the correlation matrix, high bank profitability and high size are positively and strongly correlated. From Figure 11 (b) we observe a similar relationship as for bank size and absolute residuals.

Since the GFC, a number of regulatory reforms have taken place around the globe as means to curb the probability and impact of the failure of a significant and important bank. Under the Basell III regulation for banks, the minimum capital adequacy ratio required to be held by banks is around 8%. From the distribution below, we see that greater the capital held by banks, the less importance they place on the soft information gathered when pricing default. Moreover, given they have to be safer compared to other banks, the total number of defaults for banks with a high capital adequacy ratio is lower than for those with a lower ratio.

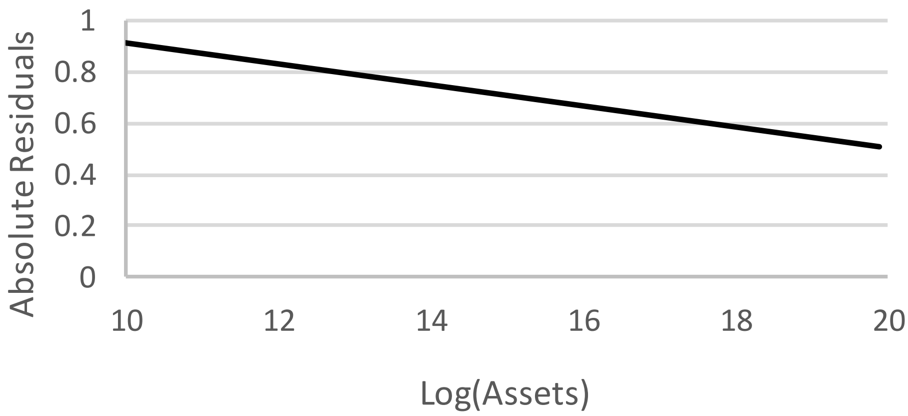
[17] The marginal effect is calculated as the mean of the residuals in different buckets of the lender and borrower qualitative features which is then plotted against the absolute residuals.

A bank may have multiple business functions in which they operate. The last ratio analyses the number of banks whose focus is on lending solely. The higher the ratio depicts a stronger focus on lending for a given bank. We would expect the smaller banks to be more focused on lending as their main business function while the larger banks would be diversified with regards to participation in various business functions. Given this, and from the graph below, we find a positive linear relationship between the focus on lending for a given bank against the absolute residuals. This states that banks with stronger focus on lending place greater emphasis on soft information as a determinant of the total credit spread charged for a given borrower. This could be linked to technological advancements as larger lenders would have easier access to better modelling tools which could be substituted by soft information for the smaller lenders.

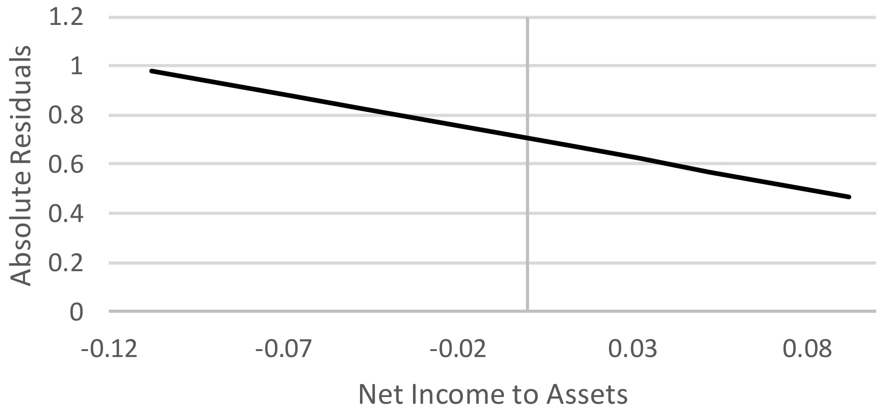
The graphs only show the linear relationship between the various lender quality characteristics and soft information. We tested the impact of non-linearity on each of the lender characteristics and they were an insignificant in predicting the credit spread unlike the FICO model. For this reason, we have excluded non-linear terms from our model.

Figure 11: Variation in Residuals and different lender characteristics

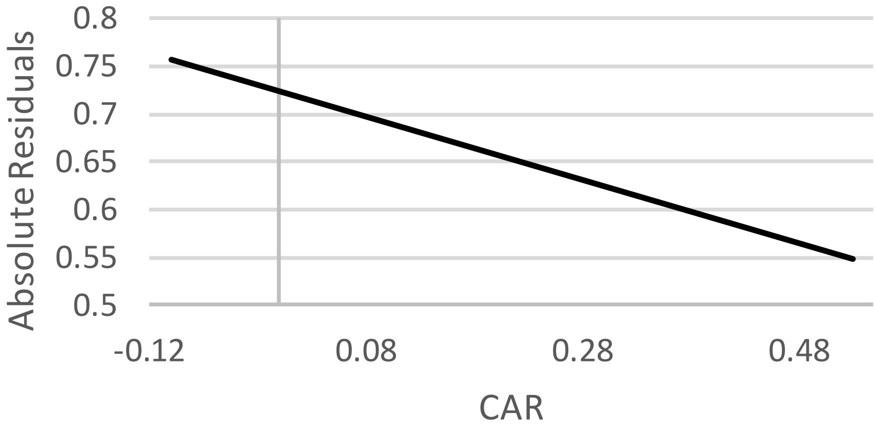
a) Variation in Residuals and Size of Bank



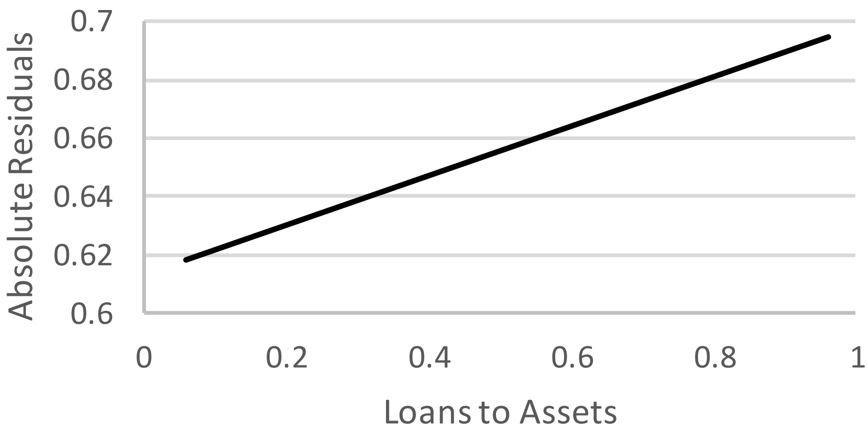
b) Variation in Residuals and Profitability



c) Variation in Residuals and Bank Safety



d) Variation in Residuals and Focus on Lending



6.4.2. Estimates with Bank Size

Bank size is determined by taking the log of the total assets of a given lender. The categorical variable is 1 given the size of the bank is less than the mean and 0 otherwise. From the results in Table 21, residuals are positive and significant consistently across the four models. The categorical variable indicating 1 given a bank is above the median is higher signifying the larger banks originate loans which do not default as much as smaller banks. This result is consistent from Table 20. Moreover, the interaction term between the residuals and bank size is negative as is expected because if a bank is large in size, it would generally have a heirarchial structure which would lead to the dissipation of soft information through the origination process of the loan. Hence, the reliance on soft information to determine probability of default is lower overall.

Table 21: Stage 2 Regression with Lender Financial Ratios Included

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-5.299***	-5.444***	-5.321***	-5.434***
	(0.09)	(0.097)	(0.09)	(0.098)
Residual	0.059**	0.051*	0.054*	0.05**
	(0.028)	(0.028)	(0.028)	(0.028)
Low Log(Assets)	-0.325***	-0.375***	-0.331***	-0.377***
	(0.061)	(0.062)	(0.061)	(0.062)
Residual*Low Log(Assets)	0.402***	0.385***	0.403***	0.388***
	(0.064)	(0.066)	(0.065)	(0.066)
Hard Facts	0.767***	0.759***	0.762***	0.758***
	(0.027)	(0.028)	(0.027)	(0.028)
Change in HPI		0.067***		0.066***
		(0.01)		(0.01)
Change in LTV		-0.006***		-0.006***
		(0.001)		(0.001)
Change in Personal Spending		-0.058*		-0.00221
		(0.033)		
Change in Unemployment Rate		-0.003		0.001
		(0.004)		(0.006)
Lending Change			0.005***	0.002
			(0.001)	(0.002)
Number of Observations	58,890	58,890	58,890	58,890
Pseudo R-Square	0.0846	0.0936	0.0857	0.0936

6.4.3. Estimate with Bank Profitability

Bank profitability is measured as the ratio between net income of the bank and its total assets. The categorical variable takes the value 1 if the bank has a high profitability ratio and 0 otherwise. The residuals are positive however their significance is mixed across the four models. Given a lender with a strong profitability ratio originated a loan, the probability of that loan defaulting is less than otherwise. The interaction term between soft information and residual is positive and significant across the four models. Upon analysis, we could conclude given that a particular lender is profitable compared to its counterparts, the soft information gathered is of good quality and thus is predicative of future default. Furthermore, this can also be interpreted as given a loan is originated by a highly profitable bank, the overall probability of a borrower with positive soft information defaulting is very low.

Table 22: Stage 2 Regression for High Lender Profitability

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-5.308***	-5.443***	-5.329***	-5.436***
	(0.102)	(0.111)	(0.102)	(0.111)
Residual	0.062**	0.055*	0.058**	0.054*
	(0.028)	(0.028)	(0.028)	(0.028)
High Net Income to Assets	-0.221***	-0.261***	-0.227***	-0.262***
	(0.062)	(0.063)	(0.062)	(0.063)
Residual*Net Income to Asset	0.382***	0.36***	0.382***	0.363***
	(0.062)	(0.063)	(0.062)	(0.063)
Hard Facts	0.765***	0.756***	0.76***	0.755***
	(0.029)	(0.031)	(0.029)	(0.031)
Change in HPI		0.064***		0.063***
		(0.01)		(0.01)
Change in LTV		-0.006***		-0.006***
		(0.001)		(0.001)
Change in Personal Spending		-0.002013		-0.002278
Change in Unemployment Rate		-0.004		-0.001
		(0.004)		(0.006)
Lending Change			0.005***	0.001
			(0.001)	(0.002)
Number of Observations	58,890	58,890	58,890	58,890
Pseudo R-Square	0.0833	0.0921	0.0845	0.0921

6.4.4. Estimates with Bank Capital

In the next model we analyse the change of the role of soft information with regards to different lender ratios. In this section we look at the capital adequacy ratio which defines the bank's capital and safety. Similar to Residual interaction models, we create a categorical variable which takes the value of 1 given the ratio is above the median and 0 for others.

After stating this definition, from the results derived for the stage 2 model in Table 23 we can see that the parameter estimate of a company with a high CAR is negative and significant, this would suggest companies that are safer tend to have a lower probability of default which follows logic as banks with greater capital on hold would be safer hence be cautious of the borrower they lend to. The value of the residuals standalone is positive and significant suggesting it has an impact in predicting overall default of a given loan however, the interaction term between residuals and high CAR is negative and significant across all the models. This suggests a lower quality of the soft information derived from the borrowers which is not highly predictive of default.

Table 23: Stage 2 Regression for High Lender Safety

	Model A	Model B	Model C	Model D
Intercept	-5.529***	-5.718***	-5.556***	-5.712***
	(0.075)	(0.082)	(0.075)	(0.083)
Residual	0.205***	0.19***	0.205***	0.189***
	(0.049)	(0.05)	(0.049)	(0.05)
High CAR	-0.415***	-0.49***	-0.421***	-0.491***
	(0.056)	(0.057)	(0.056)	(0.057)
Residual*High CAR	-0.092	-0.089	-0.097***	-0.089
	(0.057)	(0.058)	(0.057)	(0.058)
Hard Facts	0.911***	0.929***	0.908*	0.928***
	(0.027)	(0.028)	(0.027)	(0.028)
Change in HPI		0.068***		0.067***
		(0.01)		(0.01)
Change in LTV		-0.006***		-0.006***
		(0)		(0.001)
Change in Personal Spending		-0.058*		-0.063*
		(0.033)		(0.034)
Change in Unemployment Rate		-0.004		-0.001
		(0.004)		(0.006)
Lending Change			0.005***	0.001
			(0.001)	(0.002)
Number of Observations	58,890	58,890	58,890	58,890
Pseudo R-Square	0.0838	0.0939	0.085	0.0939

6.4.5. Estimates with Bank Lending

The bank loans to assets model is defined as the proportion of the total bank's business in lending. In the Stage 2 regression in Table 24, the value of the residuals remains constant and significant throughout the different model variations. Given a lender had majority of their business in lending, the loans originated by them generally leads to lower default in the future. This is consistent with logic as we would expect lenders with a stronger focus on lending activities to screen borrowers more as a loss experienced would negatively affect their business.

The interaction term between soft information and high loans to assets is negative and significant across the four models. Figure 11 (d) shows as loans to asset ratio increases, so does the variation in soft information as an explanatory variable of credit spread charged. This combined with results in Table 24 would suggest that companies with a stronger focus on lending activities gather soft information however, this soft information does not always add to the total predictive power of soft information. The interaction terms are significant at the 1% level across the four models.

Table 24: Stage 2 Regression for Strong Focus on Lending

	Model A	**Model B**	**Model C**	**Model D**
Intercept	-5.447***	-5.606***	-5.478***	-5.599***
	(0.075)	(0.083)	(0.075)	(0.084)
Residual	0.251***	0.253***	0.25***	0.252***
	(0.038)	(0.04)	(0.038)	(0.04)
High Loan to Assets	-0.479***	-0.572***	-0.493***	-0.573***
	(0.048)	(0.05)	(0.049)	(0.05)
Residual*High Loan to Assets	-0.176***	-0.192***	-0.183***	-0.192***
	(0.051)	(0.052)	(0.051)	(0.052)
Hard Facts	0.877***	0.884***	0.875***	0.884***
	(0.025)	(0.025)	(0.025)	(0.025)
Change in HPI		0.062***		0.06***
		(0.01)		(0.01)
Change in LTV		-0.006***		-0.006***
		(0)		(0)
Change in Personal Spending		-0.072**		-0.078**
		(0.033)		(0.034)
Change in Unemployment Rate		-0.007*		-0.005
		(0.004)		(0.006)
Lending Change			0.006***	0.002
			(0.001)	(0.002)
Number of Observations	59767	59767	59767	59767
Pseudo R-Square	0.0874	0.0986	0.0889	0.0984

7. Critical Evaluation

7.1. Economic Impact

In this paper we have attempted to showcase the impact of soft information on default and how it changes given different borrower lenders. In this section, we wish to show the influence this soft information would have on the lender as well as its overall economic impact. In order to put things into perspective, we created a categorical variable which took the value of 1 given a particular borrower default at some point in the future, this value is forward looking. Next we merged this new variable with our origination dataset as we are interested in analysing the impact soft information has on the pricing of the loan which is best observed during the loans origination. In essence, this categorical variable is the crystal ball looking into the future hence taking the value of 1 given a borrower defaults. Next we estimate our base model Stage 1 regression and find the predicted credit spread. The predicted credit spread is the credit spread charged for a particular borrower given the hard facts we have used in Stage 1 regression, that is, the predicted credit spreads do not include soft information which is captured within the residuals for the purposes of our study. The actual credit spread includes the residuals and thus the soft information in this dataset.

We calculate the mean of the predicted credit spread and compare this with the mean of the true credit spread by the default status of a borrower at a point in the future. From this we find that given a borrower does not default, that is- have a stronger propensity to payback their debt, on average, they are charged less with the addition of soft information compared to without. This is in line with the effect we would expect soft information to have given a lender judges a person positively in terms of their propensity to pay back their loan.

On similar lines, when observing the borrowers who default at a point in the future, the credit spread with soft information charged for such borrowers is higher than the predicted credit spread which is solely based on the hard facts. Given a borrower has a choice to make between two lenders, one who uses credit spread and one who does not, the borrower with positive soft information will likely go to the lender who uses soft information in their screening process and thus charges a lower credit spread. On the other hand, the borrower who has a poor propensity to pay back a loan would be offered a lower credit spread by the lender who does not use soft information in screening as their average credit spread is lower overall. This shows the impact soft information has on an economic scale as the "human touch" when pricing loans is significant.

Table 25: Credit Spread Charged with and without Soft Information

Variable	Number of observations	Mean	Std Dev	Minimum	Maximum
No Default					
Credit Spread	24,628	2.30%	1.463	0.01%	14.47%
Predicted Credit Spread	24,628	2.35%	0.884	0.07%	6.50%
Default					
Credit Spread	11,312	2.82%	1.35	0.01%	10.68%
Predicted Credit Spread	11,312	2.72%	0.829	0.33%	6.52%

7.2. Conclusion

We use a loan-level dataset in this study to analyse the impact of soft information on mortgage default. Soft information is information that is difficult to summarize in a numeric score and is only known by the lender who is exposed to it. Academic literature consistently shows its significance in improving default predicting models.

We use the credit spread at origination to capture the soft information unrecorded by the lenders in the residuals of our base model. Furthermore, provided the nature of soft information being vague, we explain how soft information varies with different borrower, lender and economic characteristics. From this we find that given a borrower has a strong FICO score, the probability of his or her default is significantly lower and the soft information gathered for these borrowers is of a high quality and is significant in predicting default. Similarly, in analysing changing economic periods, we find that lenders who originate loans during tight lending standards prevalent in the market, place a stronger emphasis on soft information when pricing loans and this information is a significant predictor of default at a given point in the future.

One of the strengths in our dataset is the access to different originator names, with this additional information, we control for it to analyse whether there is a considerable change in the significance and impact of soft information and find the significance of soft information remains the same. We extend this model by including a range of different financial ratios from the lender's perspective which depicts their size, profitability, safety and focus on lending. All in all, we show that soft information is important and useful in predicting default for a given borrower and using this soft information has a positive overall economic impact.

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 & Banking (Wiley-Blackwell),* vol. 43, no. 4, pp. 633-655.

Agarwal, S. & Hauswald, R. 2010, "Distance and Private Information in Lending", *Review of Financial Studies,* vol. 23, no. 7, pp. 2757-2788.

Amromin, G. & Paulson, A.L. 2009, "Comparing patterns of default among prime and subprime mortgages", *Economic Perspectives,* vol. 33, no. 2, pp. 18-37.

Asea, P.K. & Blomberg, B. 1998, "Lending cycles", *Journal of Econometrics,* vol. 83, no. 1–2, pp. 89-128.

Bajari, P, Chu, C. S., Park, M., 2008, "An Empirical Model of Subprime Mortgage default from 2000 to 2007", *NBER Working Paper #14625.*

Berger, A.N. & Udell, G.F. 2002, "Small Business Credit Availability and Relationship Lending: The Importance of Bank Organisational Structure", *Economic Journal,* vol. 112, no. 477, pp. F32.

Demyanyk, Y. & Van Hemert, O. 2011, "Understanding the Subprime Mortgage Crisis", *Review of Financial Studies,* vol. 24, no. 6, pp. 1848-1880.

Ergungor, O.E. & Moulton, S. 2014, "Beyond the Transaction: Banks and Mortgage Default of Low-Income Homebuyers", *Journal of Money, Credit & Banking (Wiley-Blackwell),* vol. 46, no. 8, pp. 1721-1752.

Godbillion-Camus, B. & C. J. Godlewski 2005, "Credit Risk Management in Banks: Hard Information, Soft Information and Manipulation", *Strasbourg, University of Strasbourg*

Keys, B.J., Mukherjee, T., Seru, A. & Vig, V. 2010, "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans", *Quarterly Journal of Economics,* vol. 125, no. 1, pp. 307-362.

Petersen, M.A. & Rajan, R.G. 2002, "Does Distance Still Matter? The Information Revolution in Small Business Lending", *Journal of Finance,* vol. 57, no. 6, pp. 2533-2570.

Rajan, U., Seru, A. & Vig, V. 2015, "The failure of models that predict failure: Distance, incentives, and defaults", *Journal of Financial Economics,* vol. 115, no. 2, pp. 237-260.

Stein, J.C. 2002, "Information Production and Capital Allocation: Decentralized versus Hierarchical Firms", *The Journal of Finance,* vol. 57, no. 5, pp. 1891-1921.

Uchida, H. 2011, "What Do Banks Evaluate When They Screen Borrowers? Soft Information, Hard Information and Collateral", *Journal of Financial Services Research,* vol. 40, no. 1, pp. 29-48.