# Bank Information Production Over the Business Cycle*

Cooper Howes† Gregory Weitzner‡

November 2025

**Abstract**

The information banks produce drives their lending decisions and macroeconomic outcomes, but this information is inherently difficult to analyze because it is private. We construct a novel measure of bank information quality from confidential regulatory data that include banks' private risk assessments for US corporate loans. Information quality improves as local economic conditions deteriorate, particularly for new loans, large loans, and loans with higher expected losses. Information quality also declines during periods of rapid local house price appreciation. Our results provide empirical support for theories of countercyclical information production in credit markets.

*We thank Thomas Chaney (the editor) and five anonymous referees as well as Hassan Afrouzi, Cynthia Balloch, Javier Bianchi, Olivier Coibion, Mariela Dal Borgo, Adolfo De Motta, Miguel Faria-e-Castro, Daniel Greenwald, Stefan Jacewitz, Gustavo Joaquim, Artashes Karapetyan, Anya Kleymenova, Alexandre Kohlhas, Yueran Ma, Blake Marsh, Johannes Matschke, Ralf Meisenzahl, Karel Mertens, Atanas Mihov, Lars Norden, Guillermo Ordoñez, Pablo Ottonello, Matt Pritsker, Samuel Rosen, Kasper Roszbach, Guillaume Roussellet, Jane Ryngaert, Padma Sharma, Lee Smith, Wenting Song, Andrea Vedolin, Jeff Wooldridge, Yufeng Wu and Choongryul Yang as well as seminar and conference participants at Arizona State University (Economics), the Bank of Canada, Bentley University (Economics), BSE Summer Forum, EFA, Federal Reserve Bank of Kansas City, Federal Reserve Board, IBEFA Annual Meeting, McGill University (Finance), Norges Bank, UNC Greensboro (Economics), University of Notre Dame (Economics), University of Oklahoma (Economics), Oxford Saïd – Risk Center at ETH Zürich Macro-finance Conference, George Washington University (Finance), UVA Darden (Finance), CAFRAL, CICF, FDIC Bank Research Conference, the Finance Forum, Fixed Income and Financial Institutions Conference, MFA, NFA, SEA annual meetings, NAMES summer meetings and the Procyclicality Symposium for helpful comments and discussions. We also thank Alex Zhang for excellent research assistance. These views are those of the authors and do not reflect the views of the Federal Reserve Board of Governors or the Federal Reserve System.

†Federal Reserve Board of Governors. Email: cooper.a.howes@frb.gov.

‡McGill University. Email: gregory.weitzner@mcgill.ca.

# 1 Introduction

A fundamental role of banks is to produce information about prospective borrowers.[1] Banks use this information to determine the recipients and terms of financing; hence, their information production decisions can affect real economic activity and financial stability through the supply of credit to firms. If the returns to distinguishing between different types of borrowers change with economic conditions, banks' incentives to produce information can affect and be affected by business cycles. Despite policymaker interest and an extensive theoretical literature emphasizing the importance of banks' information, there is little evidence of its empirical properties.

The key empirical challenge in testing theories of bank information production is that banks' information is intrinsically private and, therefore, unobservable to the econometrician. Because of this data limitation, the existing literature typically relies on indirect evidence; however, without access to banks' private information, researchers are severely constrained in their ability to test these theories. In this paper, we address this challenge using confidential regulatory data that contain banks' private risk assessments for corporate bank loans over one million dollars in the US. We use county-level variation in unemployment rates to show that banks' risk assessments discriminate better across borrowers in local downturns, suggesting that banks' information quality is countercyclical. We then provide evidence that this countercyclical information quality results from endogenous information production. Specifically, we find that the cyclical variation in banks' information quality is concentrated in loans that theory predicts to be more information sensitive: new loans, larger loans, and loans with higher expected losses. Finally, consistent with higher collateral values reducing information production incentives, we also show that information quality is lower during periods of rapid local house price appreciation. Overall, our results provide empirical support for theories of countercyclical information production in credit markets.

Our analysis uses the Federal Reserve's Y-14Q Schedule H.1 data, which include all corporate loans larger than one million dollars extended by large bank holding companies. In addition to detailed loan and borrower characteristics, qualified bank holding companies must report their internal estimate of the borrower's probability of default (PD) for each loan. Because the data also reveal whether loans ultimately default, these PDs—which incorporate both "hard" and "soft" information—allow us to quantify bank information quality.

We first show, using linear regressions and random forest regressions that allow for nonlinearities and interactions, that banks' probabilities of default predict realized default even after controlling for a rich set of loan- and firm-level controls. These results suggest that banks' risk assessments contain private information that is i) relevant for predicting default and ii) not captured by other observables.

[1] E.g., Leland and Pyle (1977), Diamond (1984) and Boyd and Prescott (1986).

In many models of information production in credit markets, including the simple one developed in this paper, banks have stronger incentives to produce idiosyncratic information during downturns, enabling them to better distinguish borrower quality.[2] To test this prediction, we use the area under the receiver operating characteristic curve (AUC), which measures the discriminatory ability of forecasts of binary outcomes, such as default, and is the most commonly used approach by practitioners (Engelmann and Rauhmeier (2011)), bank regulators (Basel Committee on Banking Supervision (2005)), and academics (Puri, Rocholl, and Steffen (2017), Berg and Koziol (2017) and Berg, Puri, and Rocholl (2020)). We describe the AUC in much more detail below; however, the receiver operating characteristic curve uses each possible value of the probability of default as a classification threshold (loans with a probability of default above the threshold are classified as predicted defaults, while loans below are classified as predicted non-defaults) and measures how well these classifications match actual default outcomes. The area under this curve, i.e., the AUC, measures banks' discriminatory ability, where an AUC of 1.0 indicates perfect discrimination between defaulting and non-defaulting borrowers, while an AUC of 0.5 indicates completely random prediction.

We then analyze how the AUC, our measure of information quality, evolves over the business cycle. Specifically, we split our sample into periods of high and low unemployment based on whether a county's unemployment rate was above or below its median across our sample period. We find that the AUC derived from banks' probabilities of default for newly originated loans is higher in periods of high unemployment and that this difference is statistically significant based on a DeLong test (DeLong, DeLong, and Clarke-Pearson (1988)).

While our main result is consistent with banks producing more information when economic conditions weaken, bank information quality could also vary exogenously over the business cycle. For example, if more firms become delinquent in periods of high unemployment, banks may be able to better distinguish between borrower types from the information they receive exogenously from borrowers. Next, we conduct tests that provide support for the endogenous information production channel by analyzing how banks' information quality varies across loans based on their information sensitivity (Dang, Gorton, and Holmström (2013)), i.e., the value of information for a given loan.

First, banks' information production incentives should be more sensitive to the business cycle for new loans because they require risking additional capital, in contrast to existing loans, for which banks' capital has already been sunk. To test this hypothesis, we separately compare the AUC of high- and low-unemployment periods for new and existing loans. Consistent with this prediction, we find that banks' information quality increases more during periods of high unemployment for newly originated loans. When we expand our sample to include existing loans, we also show that economic conditions at origination have persistent effects on infor-

[2] See Ruckes (2004), Dell'Ariccia and Marquez (2006), Gorton and He (2008), Dang, Gorton, and Holmström (2013), Gorton and Ordonez (2014), Gorton and Ordonez (2020), Fishman, Parker, and Straub (2020), Petriconi (2015), Farboodi and Kondor (2020) and Asriyan, Laeven, and Martin (2022).

mation quality, even more so than current economic conditions. This result is consistent with the theories that motivate our analysis, in which economic conditions drive banks' information production decisions when deciding whether to grant a loan.

Second, we test whether banks' information quality is higher for larger loans and loans with higher expected losses. According to several theories of information production in credit markets, such as Dang, Gorton, and Holmström (2012) and Gorton and Ordonez (2014), banks should produce more information about these loans, as they will have higher returns to distinguishing between borrowers. Consistent with these predictions, when we split loans by their size or expected loss and re-estimate each AUC separately, we find that information quality increases with loan size and expected losses. Moreover, for both characteristics, we show that the difference in AUCs between high- and low-unemployment periods is larger for the top quartile than for the bottom quartile, suggesting that information quality is more cyclically sensitive for large loans and loans with higher expected losses.

A key determinant of expected losses is collateral values, which recent theories suggest play an important role in how banks' information production incentives evolve over the business cycle. For example, rapid increases in collateral values can lead to reduced incentives for banks to screen borrowers as expected losses decrease (e.g., Gorton and Ordonez (2020), Asriyan, Laeven, and Martin (2022)). Motivated by these theories, we first show that local housing prices—a commonly used proxy for collateral values in the literature—are associated with lower losses given default and expected losses. We then show that the AUC is lower in areas with high house price growth, suggesting that increasing collateral values dampen banks' information production incentives.

**Literature review.** While the theory literature has long recognized the importance of bank information production (e.g., Leland and Pyle (1977), Diamond (1984) and Boyd and Prescott (1986)), testing for it is notoriously difficult given the private nature of banks' information. For this reason, the existing empirical literature focuses on proxies or indirect evidence of information production (e.g., James (1987), Cerqueiro, Ongena, and Roszbach (2016), Gustafson, Ivanov, and Meisenzahl (2021), Iyer et al. (2016) and Bedayo et al. (2020)). However, there are many different interpretations of these proxies. For example, Gustafson, Ivanov, and Meisenzahl (2021) create a measure of monitoring based on the number of visits banks take to firms. However, it is unclear whether banks visit firms to collect private information, in response to receiving information (public or private), or both. In contrast, our data allow us to observe banks' private information directly and test how well this information predicts subsequent defaults.

The closest to our empirical approach is Becker, Bos, and Roszbach (2020). They also find that banks' internal credit ratings better predict default in bad times, but they attribute this to exogenous variation in information over the business cycle. There are three key differences in our analysis and the interpretation of our results. First, their data are at the firm level rather than the loan level. This difference allows us to explore the relationship between loan characteristics and information production and how this relationship changes over the business cycle. Second,

we provide evidence that the countercyclicality of information quality is driven by endogenous bank information production by showing that the effects are stronger for new loans, large loans, and loans with higher expected losses, which is difficult to rationalize solely through exogenous variation in information quality over the business cycle. Third, their approach uses time-series variation in country-wide aggregate economic conditions for a single Swedish bank, while we exploit rich cross-sectional variation in economic conditions across US counties.

Our paper also relates to the theoretical work analyzing the cyclicality of information production in credit markets. This includes many theories in which information production is countercyclical (e.g., Ruckes (2004), Dell'Ariccia and Marquez (2006), Gorton and He (2008), Dang, Gorton, and Holmström (2013), Gorton and Ordonez (2020), Fishman, Parker, and Straub (2020), Petriconi (2015), Farboodi and Kondor (2020) and Asriyan, Laeven, and Martin (2022)). A common feature in these models is that lending standards tighten in downturns as banks produce more information. While countercyclical lending standards are widely acknowledged empirically (e.g., Asea and Blomberg (1998), Lown and Morgan (2006), Maddaloni and Peydró (2011), Dell'Ariccia, Igan, and Laeven (2012), Bassett et al. (2014) and Rodano, Serrano-Velarde, and Tarantino (2018)), the mechanism behind them remains unclear given the unobservability of banks' information. For example, countercyclical lending standards could arise solely from banks imposing stricter lending thresholds. Conversely, and consistent with our evidence, banks could lend more selectively exactly because they produce more information. Hence, our results speak to the mechanisms behind changes in lending standards over the cycle.

# 2 Data

Our main data source is Schedule H.1 of the Federal Reserve's Y-14Q filings. The Federal Reserve began collecting these data to support the Dodd-Frank mandated stress tests and the Comprehensive Capital Analysis and Review. The sample includes commercial and industrial loans from bank holding companies with $50bn or more in total assets[3], accounting for 85.9% of all assets in the banking sector (Frame, McLemore, and Mihov (2025)). Qualified institutions are required to report detailed quarterly loan-level data on corporate loans of at least $1mm. The universe of loans we analyze is large: Bidder, Krainer, and Shapiro (2020) show that the Y-14Q data cover 70% of all commercial and industrial loan volume extended by bank holding companies that file an FR Y-9C report.

The data include detailed loan characteristics (interest rates, maturity, amount, collateral, and purpose) and performance measures (defaults, past-due payments, non-accruals, and charge-offs). They also include income statement, balance sheet, and geographic information about borrowers. Crucially, banks must also report their internal estimates of the borrower's prob-

[3] In 2019, this threshold was increased to $100bn. The most recent list of participating institutions can be found in Table 3 of the 2024 Federal Reserve Stress Test Results.

ability of default (PD) and loss given default (LGD) for each loan. According to the Basel Committee on Banking Supervision, internal estimates of PD and LGD "must incorporate all relevant, material and available data, information and methods. A bank may utilize internal data and data from external sources (including pooled data)."[4]

Our primary analysis focuses on newly originated loans to study banks' information production incentives at the time financing is committed; however, we also consider several extensions that include existing loans. We exclude demand loans, which can be recalled by the lender at any time, as well as loans with government guarantees[5], tax-exempt loans, loans to foreign borrowers, and loans to firms in the finance, insurance, and real estate (FIRE) sectors. We drop loans with negative interest rates and those with missing company identifiers, PD, or loan amount at origination. We incorporate several additional filters to minimize the impact of reporting errors, including excluding loans with maturities of greater than 30 years, interest rates or probabilities of default of more than 100%, or companies with under $100k in reported assets at origination. Finally, we drop all publicly traded firms and private firms with assets above the 99th percentile, as these firms are likely to be more geographically diverse and, thus, less sensitive to changes in local economic conditions.

We define the following firm-level financial variables: profitability (EBITDA/assets), size (log assets), tangibility (tangible assets/assets), and leverage (debt/assets), which we winsorize at the $1\%$ and $99\%$ levels. Our primary measure of loan performance is default, a dummy variable that equals one if the borrower defaults within two years after origination. Focusing on a two-year default window strikes a balance between our data's limited time series and the fact that the median loan maturity is close to five years. Our sample starts in 2014Q4 when the probability of default variable first becomes well populated. To allow for consistent measurement of default rates, we include loans on banks' balance sheets until 2021Q4 and track whether they ultimately default through 2023Q4.

Table 1 includes firm, loan, and county summary statistics. Panel A shows summary statistics at the loan level for newly originated loans, where the average and median loan size are approximately $13.3mm and $3.6mm, respectively. Over our sample period, 1.13% of loans default within two years after origination, compared to an average ex-ante expected probability of default of 1.63%. We average each reported measure at the firm-quarter level across all outstanding loans to calculate the firm-level statistics in Panel B. The median firm has $21.5mm in assets and a leverage ratio of 0.26. These loan and firm sizes are small relative to other sources of loan data, such as DealScan, because our sample contains many small, private firms. Panel C shows characteristics aggregated at the county level. The median number of loans outstanding for each county-quarter is 5, while the median new loan volume is about $40mm. Finally, Table 2 includes additional loan-level summary statistics, including splits by high and

[4]The most recent instructions are available at Calculation of RWA for credit risk.

[5]We conduct a test in Online Appendix Section C.4 using loans with government guarantees, which provides further additional support for the endogenous information production mechanism.

low-unemployment periods.

The first three panels of Figure 1 show the distributions of PD and log(PD). If PD contained valuable information for predicting default, then there should be a positive correlation between PD and future realized default. In the bottom-right panel, we place loans into PD quintiles where the number below each column indicates the average PD in that quintile of loans, while the vertical axis indicates the average realized default rate. The figure shows a clear positive relationship between PDs and realized defaults, suggesting that PD contains valuable information regarding the borrower's default risk. We test this relationship more formally in the next section.

# 3 Banks' Reported Probabilities of Default Predict Realized Default

In this section, we validate banks' reported probabilities of default (PD) as a measure of their private information by showing that they are a statistically and economically significant predictor of realized default, even after controlling for various loan and firm characteristics. We start by estimating the following regression:

$$Default_i = \beta PD_i + \Omega X_i + \delta_{b,t} + \gamma_{j,t} + \sigma_{b,c} + \epsilon_i, \tag{1}$$

where $i$, $b$, $t$, $j$ and $c$, index loan, bank, quarter, industry, and county, respectively. $Default_i$ is a dummy variable that equals one if loan $i$ defaults within eight quarters following origination. $PD_i$ is the bank's estimated probability of default described in Section 2. $X_i$ is a vector of firm and loan characteristics which include firm size (log of total assets), leverage ratio (total debt to total assets), profitability ratio (EBITDA to total assets), tangibility ratio (tangible assets to total assets), log loan size, the log of the original loan maturity in months, and the bank's estimate of loss given default per dollar of exposure (LGD), as well as loan type fixed effects. We include bank-quarter fixed effects ($\delta_{b,t}$) to absorb any differences in banks' risk assessment models and cost of capital, industry-quarter fixed effects ($\gamma_{j,t}$) to absorb variation in average loan performance across industries, and bank-county fixed effects ($\sigma_{b,c}$) to absorb persistent differences in risk assessment models or credit analysts across counties. In all regressions, we cluster standard errors by county.

The results are shown in Table 3. The primary coefficient of interest is $\beta$, which represents the expected increase in realized default (measured in percentage points) from a one percentage point increase in a loan's PD. In Column (1), the coefficient estimate is 0.407, which means that an increase in PD of 1pp increases the probability of realized default by about 41bps.[6] In Column (2), we display the results with firm and loan characteristics and find a similar

[6]Online Appendix Table OA.4 shows that these results are robust to alternative measures of loan performance.

coefficient of 0.444. Under strict rational expectations, regressing a realized outcome on its forecast should yield a coefficient of one (Muth (1961)); however, empirically, this is often not the case (e.g., Mincer and Zarnowitz (1969)). We discuss potential reasons why the coefficient estimates may deviate from one in Online Appendix Section C.5.

For comparison, Columns (3) and (4) repeat the same exercise using interest rates as an alternative measure of borrower risk.[7] As expected, loans with higher interest rates default more frequently. However, when we include both interest rates and PD in the same specification in Column (5), the coefficient for interest rate attenuates substantially, while the coefficient for PD remains basically unchanged from Column (2).

One shortcoming of the regression approach in (1) is that it imposes a linear relationship between default and firm/loan characteristics. If firms' actual default probabilities reflect non-linearities or interactions between these characteristics, the linear specification in (1) could understate the explanatory power of observable characteristics. This would lead us to incorrectly attribute the predictive power of banks' reported PDs to private information when, in fact, they are simply capturing public information in a more sophisticated way.[8] To address this concern, we produce predicted default estimates using the random forest regression algorithm developed in Breiman (2001). This approach, described in detail in Appendix A, generates predictions by sequentially partitioning observations based on their observable characteristics and calculating the average default rate within each partition. The ability to flexibly accommodate complex nonlinearities and high-dimensional data has made random forests a popular tool across a range of fields, including econometrics (Wager and Athey (2018)), asset pricing (Gu, Kelly, and Xiu (2020)), and macroeconomics (Goulet Coulombe et al. (2022)).

We first estimate predicted default using several different specifications that combine the same firm and loan characteristics in (1) with combinations of dummy variables for industry, bank, and time. In each case, we first estimate the random forest using half of our new loans. We then estimate the following regression using the predictions $PD^{RF}$ generated for the remaining half of our sample:

$$Default_i = \beta PD_i + \theta PD_i^{RF} + \epsilon_i.$$

The results are reported in Table 4. Across all specifications, both $\beta$ and $\theta$ remain statistically and economically significant. In addition to providing out-of-sample validation that the random forest algorithm produces reasonable default estimates, this result shows that banks' PDs contain information useful for predicting default that is not fully reflected in observables and

[7] The number of observations drops when we include interest rates in the regression because we drop undrawn credit lines, for which banks are instructed to record an interest rate of zero.

[8] In principle, collinearity between PD and other observables could also lead to potential difficulties in accurately assessing the private information content of banks' PDs in this framework. However, we show in Online Appendix Table OA.2 that PD is only weakly correlated with other firm and loan characteristics, and in Online Appendix Table OA.3 that statistical tests soundly reject collinearity.

validates PD as a meaningful measure of banks' private information regarding firms' default risk.[9]

While we refer to all loan and firm characteristics as "observables", in reality, banks only observe a subset of these (i.e., from the firms they specifically lend to). Hence, these empirical tests, which control for all these characteristics across the entire sample, likely underestimate banks' private information. That we still find banks' reported PDs predict future defaults, even with these stringent controls, strengthens our conclusion that PDs capture meaningful private information.

# 4 Bank Information Quality Over the Business Cycle

In the previous section, we show that banks' reported probabilities of default (PDs) contain information regarding future default that is not reflected in observable characteristics, suggesting that PDs reflect banks' private information. In this section, we analyze how the quality of banks' information evolves over the business cycle. We present our approach to assessing information quality in Section 4.1 and then implement this approach in Section 4.2.

## 4.1 Approach to Measuring Information Quality

In the theories that motivate our analysis, banks have increased incentives to produce *idiosyncratic* information about borrowers during bad times (e.g., Ruckes (2004) Dang, Gorton, and Holmström (2012), Gorton and Ordonez (2014), Asriyan, Laeven, and Martin (2022), Fishman, Parker, and Straub (2020), Farboodi and Kondor (2020)).[10] Thus, we would like to have a statistical measure of how well PDs *discriminate across* borrowers. Henceforth, we refer to information quality and discrimination power interchangeably.

The main way that banks, practitioners, and regulators measure the discriminatory power of PDs is by estimating the area under the receiver operating characteristic (ROC) curve (AUC).[11] At a high level, the AUC measures how well banks' internal risk assessments rank borrowers based on their relative default risk. Estimating the AUC first requires constructing the ROC

[9] In Online Appendix Tables OA.5 and OA.6, we perform similar exercises using interest rates to show that PDs contain valuable information for pricing loans, which is consistent with Beyhaghi, Fracassi, and Weitzner (2025). In Appendix A, we show that PDs provide incremental *discriminatory power* over observables using the area under the receiver operating characteristic curve approach, which we describe in the next section.

[10] This can be due to several reasons. For example, in recessions, the difference in payoff between lending to a "good" and "bad" borrower may increase (e.g., Dang, Gorton, and Holmström (2012)), or there may be more "bad" borrowers (e.g., Farboodi and Kondor (2020)).

[11] For example, in referring to the receiver operating characteristics curve and the cumulative accuracy curve (whose test statistic is a simple transformation of the AUC), Engelmann and Rauhmeier (2011) say they are "...the most important and the most widely applied in practice." The Basel Committee on Banking Supervision stated, "The Group has found that the Accuracy Ratio (AR) and the ROC measure appear to be more meaningful than the other above-mentioned indices because of their statistical properties" (Basel Committee on Banking Supervision (2005)). The AUC is also highly prevalent in other contexts measuring the performance of binary classifier models, particularly in medicine (e.g., Pepe (2003)) and machine learning (e.g., Bradley (1997)).

curve. The ROC curve considers every possible value of PD in the data as a threshold for classifying loans: any loan with a PD above this threshold is classified as a predicted default, while any loan with a PD below this threshold is classified as a predicted non-default. For each threshold PD, the true positive rate is defined as the ratio of correctly predicted defaults over the total number of defaults, whereas the false positive rate is the ratio of incorrectly predicted defaults over the total number of non-defaults. Using each observed value of PD as a classification threshold, the ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis. The AUC is the area under the ROC curve.

The AUC also has a simple probabilistic interpretation: given a randomly chosen defaulting loan and a solvent loan, the AUC is the probability that the defaulting loan's probability of default is higher than the solvent one's. Hence, a higher AUC indicates that banks' PDs have higher discriminatory power. A completely random prediction model would have an AUC of 0.5, while a perfect prediction model would have an AUC of 1. As a rule of thumb, an AUC of 0.6 is generally considered desirable in environments with less information, whereas AUCs of 0.7 or greater are desirable in information-rich environments (Iyer et al. (2016) and Berg, Puri, and Rocholl (2020)). We use the Stata function `roccomp` to construct ROC curves and numerically integrate them to estimate the AUC. We test for statistical significance of differences in AUCs using the DeLong test (DeLong, DeLong, and Clarke-Pearson (1988)), which is the standard approach for testing differences in AUCs.

The AUC is particularly well-suited for measuring bank information quality for several reasons. First, because it depends only on the relative ordering of PDs, it is unaffected by any transformation of PDs that preserves the relative ordering of loans.[12] Second, the AUC primarily measures the type of information quality we are focused on, i.e., PDs' discrimination ability. This contrasts with other measures of forecast accuracy. For example, the mean-squared forecast error, also known as the Brier Score (Brier (1950)) when applied to binary outcomes, also reflects other factors unrelated to discrimination (Murphy (1973)). In Online Appendix Section C.8, we show that there is a strong mechanical relationship between the Brier Score and the underlying default risk of borrowers, particularly when PDs are close to zero. Moreover, a well-known problem with forecast errors is that they do not distinguish forecast accuracy well for rare events. We discuss these issues and provide numerical examples in Online Appendix Section C.8; however, based on our understanding, these are the main reasons forecast errors are not as commonly used in practice for credit scoring, particularly when measuring the discrimination ability of PD models. In contrast, the AUC is far less affected by the underlying level of borrowers and is better able to distinguish rare events.[13]

[12] For example, banks may have incorrect priors, incentives to misreport PDs (e.g., Behn, Haselmann, and Vig (2016) and Plosser and Santos (2018)), or behavioral biases that cause them to underreact to information. See Online Appendix Section C.5 for a further discussion of these issues.

[13] See Appendix Section B.4 and Online Appendix Section C.8 for a formal analysis of these issues. Another common approach is the accuracy ratio (AR) obtained from the cumulative accuracy profile curve (Engelmann and Rauhmeier (2011)). However, the accuracy ratio is a simple linear transformation of the AUC: $AR = 2AUC - 1$. We use the AUC because it has a simpler probabilistic interpretation.

Figure 2 displays the estimated AUC of 0.703 over our sample of new loans. While the thresholds for interpreting AUCs will vary depending on the context, the average AUC in our sample is slightly larger than recent studies analyzing consumer loans in a large German bank (Puri, Rocholl, and Steffen (2017), Berg and Koziol (2017), and Berg, Puri, and Rocholl (2020)).[14] As further validation of the AUC as a measure of information quality, in Appendix B we develop a simple model in which banks produce more information in bad times, resulting in a higher AUC.

## 4.2 Testing the Cyclicality of Bank Information Quality

In this section, we test the cyclicality of bank information quality by analyzing how the area under the receiver operating characteristic curve (AUC) varies over local economic conditions. Our measure of county-level economic conditions is the unemployment rate from the BLS.[15] Figure 3 highlights the substantial cross-sectional variation in the changes in county-level unemployment rates over this period, with roughly one-quarter of counties experiencing an increase. The top-right panel displays a histogram of defaults across county unemployment rates, suggesting that the variation in defaults for new loans is not coming solely from high-unemployment areas.

In the top-left panel of Figure 4, we split our primary sample of new loans based on whether the county's unemployment rate is above or below its county-specific median over the sample. The AUC in periods of high unemployment is 0.724 versus 0.681 in periods of low unemployment. This difference in AUCs implies that, given a randomly chosen defaulting loan and non-defaulting loan, the probability that the defaulting loan's PD is higher than the solvent loan's is 4.3pp higher in periods of high unemployment. Hence, this result suggests that PDs better discriminate across borrowers in bad times. In the other three panels, we find qualitatively similar results if we instead define a high-unemployment period as one in which 1) the county's unemployment rate is higher than the median county-level unemployment rate across counties with at least five new loans in a given quarter (top right), 2) the county's unemployment rate increased from the previous quarter (bottom left), or 3) the quarterly change in a county's unemployment rate was greater than the quarterly change in the aggregate US unemployment rate (bottom right).[16]

Local economic conditions should be more relevant for loans whose cash flows are more sensitive to local economic conditions. We test this hypothesis by comparing firms in tradeable

[14] See also Lessmanna et al. (2013) and Hayashi (2022) for surveys of AUCs in different credit scoring contexts.

[15] The Y-14Q data use ZIP codes as geographical identifiers, so we first use the ZIP-to-county crosswalk from the Department of Housing and Urban Development to assign a county to each ZIP code before merging it with the unemployment rate data.

[16] Online Appendix Figure OA.3 also shows that our main results hold when we exclude counties with few loans, and Online Appendix Table OA.8 shows that they hold when we estimate AUCs separately by bank.

and nontradeable industries.[17] Because firms in nontradeable industries will be more dependent on local markets, the same change in local economic conditions should have a more significant effect on their underlying businesses. As a result, to the extent that the business cycle drives banks' information quality, we would expect this effect to be larger for firms in nontradeable industries. In Figure 5, we test this prediction by comparing the AUC across high and low unemployment periods separately for each group of industries. Consistent with our hypothesis, the difference in AUCs across high- and low-unemployment periods is larger and only statistically significant for nontradeable firms. This result suggests that local economic conditions are an important driver of the cyclicality of banks' information quality.

One concern could be that differences in the distribution of underlying borrowers evolve over the business cycle, which could mechanically explain these results. First, as shown below in Section 5.3, loan and firm characteristics do not vary significantly over the business cycle. Second, the bottom-left panel of Figure 1 shows that the distribution of PDs appears fairly similar across high and low unemployment periods, while Table 2 shows only small differences in the mean and standard deviation of PD for newly originated loans (1.66pp versus 1.61pp and 2.88pp versus 2.58pp, respectively). In Appendix Section B.4, we show that, in the absence of differences in banks' information over the cycle, the observed difference in the distribution in PDs across high and low unemployment regimes has a quantitatively small effect on the AUC and is hence unlikely to explain our results.[18]

Another concern is that banks may have incentives to adjust their reported PDs because they are used as inputs to calculate capital requirements. Because the ROC curve is purely ordinal, any adjustments to PDs would not affect the AUC so long as they do not change their relative ordering. For example, if a bank reduced all of its PDs by 0.1pp, the AUC would be completely unaffected. Nonetheless, in Online Appendix Figure OA.9, we split our sample into loans granted by banks with high amounts of capital (as defined by having a total risk-based capital ratio above the median of all banks in our data in each quarter) versus those with low amounts of capital. We find that the AUC is actually *higher* for the low-capital banks despite the potentially stronger incentives to manipulate PDs.

Overall, the results in this section suggest that periods of elevated unemployment are associated with improvements in bank information quality. Hence, we conclude that bank information quality is countercyclical.

[17] The list of nontradeable industries includes utilities, construction, wholesale trade, retail trade, transportation, accommodation, food services, information and communication, and professional and administrative services.

[18] We also show, that for similar reasons, differences in the distribution of aggregate shocks over the business cycle are unlikely to explain our results.

# 5 Mechanisms

In Section 4, we show that banks have better information about borrowers during downturns; however, these results are entirely agnostic to the underlying mechanism. In this section, we provide evidence that banks' improved information during downturns results from endogenous information production.

As previously discussed, many models of endogenous information production, including the one developed in Appendix B, predict that banks will produce more information during downturns. However, because we cannot directly observe banks' information production decisions—only the ability of their reported probability of default (PD) to predict default—it is possible that our empirical results could instead arise purely through exogenous variation in information quality over the business cycle. For example, if more firms become delinquent during periods of high unemployment, this may exogenously provide banks with more private information, enabling them to better distinguish across borrower types. Alternatively, downturns can *publicly* reveal information relevant to default. For example, Warren Buffett famously said, "Only when the tide goes out do you learn who has been swimming naked," suggesting that downturns can publicly reveal which firms are performing well and which are performing poorly.[19] Finally, if the set of borrowers who approach banks for loans changes, banks may exogenously have better information regarding these specific borrowers. While exogenous variation in information quality is not necessarily mutually exclusive with banks endogenously producing more information as economic conditions deteriorate, this section develops several additional tests to rule out the possibility that it is the primary driver of our results.

In Section 5.1, we first show that the cyclicality of information quality is higher for new loans and that, when we expand the sample to include existing loans, the unemployment rate at origination has persistent effects on information quality. In Section 5.2, we show that the effects are concentrated in larger loans and loans with higher expected losses. Finally, in Section 5.3, we analyze how lending volume and borrower risk evolve over the business cycle and argue that both are consistent with models of endogenous information production.

## 5.1 Information Sensitivity of New Loans

We first compare the cyclicality of bank information quality for newly issued loans to those that were issued in prior quarters. Intuitively, the marginal value of information about a borrower's creditworthiness should be highest prior to the capital being sunk, and thus, new loans should be more information sensitive. If banks' incentives are driving them to produce more information in bad times, we would thus expect these effects to be stronger for new loans than for loans made in the past.

To test this hypothesis, we extend our sample to include observations for every quarter in

[19] See Berkshire Hathaway 2001 Annual Report.

which each loan was on banks' balance sheets rather than focusing exclusively on the quarter of origination.[20] In the left panel of Figure 6, we reestimate the receiver operating characteristic (ROC) curves in high- and low-unemployment periods based on this larger sample. The area under the ROC curve (AUC) is larger during periods of high unemployment, confirming our earlier findings that banks' reported probabilities of default discriminate better in bad times, even for previously issued loans. However, while this difference in AUC is statistically significant, the magnitude of the difference (0.830 versus 0.806) is smaller in both absolute and relative terms than our baseline sample of newly issued loans shown in Figure 4 (0.724 versus 0.681).

The previous test suggests that banks' information quality is most sensitive to the business cycle when loans are originated. If banks are indeed producing more information at origination, we would expect the economic conditions at origination to have a persistent effect on information quality. To test this hypothesis, we estimate AUCs using current PDs for separate groups based on each loan's unemployment rate at *origination*. We report the results of this test in the right panel of Figure 6, which shows the estimated ROC curves for high and low unemployment periods based on the origination date of the loan. The difference in AUC across periods of high and low unemployment at origination (0.806 versus 0.768) is larger than the differences based on the current unemployment rate shown in the left panel (0.830 versus 0.806).[21] This result suggests that economic conditions at origination have large and persistent effects on information quality and provides support for endogenous information production as a driving force behind the cyclicality of information quality that we observe in the data.

## 5.2 Information Sensitivity, Loan Size, and Expected Losses

In the previous section, we document that information quality is more cyclical for new loans than existing loans. In this section, we analyze how information quality varies across different types of new loans, which theory predicts to be more information sensitive. Specifically, several theories of endogenous information production, such as Manove, Padilla, and Pagano (2001), Dang, Gorton, and Holmström (2012) and Asriyan, Laeven, and Martin (2022) predict that lenders have stronger incentives to produce information about loans that are larger and have higher expected losses. Intuitively, because these are loans for which banks face more severe consequences for lending to low quality borrowers, they should produce more information, and hence, their information quality should be both better and more cyclically sensitive.

We first analyze the effects of loan size on information quality by grouping loans based on

[20] The summary statistics for this extended sample, which includes recent observations of loans originated prior to the start of our sample in 2014Q4, are shown in Online Appendix Table OA.1.

[21] The loan-level correlations between the current and origination unemployment indicators are 0.19 for the entire sample and 0.14 after excluding new loans (where the two will be the same).

their exposure at default, which is a measure of expected loan size at the time of default.[22] The top panel of Figure 7 shows receiver operating characteristic curves for loans split into quartiles of exposure at default within each bank/quarter. As the Figure shows, the AUC increases from 0.666 in the smallest quartile to 0.716 in the largest, suggesting that the discriminatory power of PDs increases with loan size.

We next examine how the sensitivity of information quality to exposure at default evolves over the business cycle. In addition to suggesting that information quality should improve with loan size, theories such as Dang, Gorton, and Holmström (2012) and Biswas (2022) also predict that information quality should be more cyclically sensitive for these loans. We test this prediction in the bottom panels of Figure 7 by comparing differences in information quality for new loans made during periods of high and low unemployment for the largest and smallest exposure at default quartiles. During periods of low unemployment (lower-left panel), the difference in AUCs between the largest and smallest loans is small (0.010) and statistically insignificant. However, when the unemployment rate is high (lower-right panel), the difference in AUCs is much larger (0.109) and statistically significant. These results suggest that information quality is both higher and more cyclically sensitive for larger loans, which provides additional support for the endogenous information production channel. They are also, to our knowledge, the first direct empirical evidence that a loan's size affects its information sensitivity and the cyclicality of that information sensitivity.

All else equal, larger loans expose banks to higher losses. However, expected losses are also affected by the loss given default and the likelihood of default. Thus, we next analyze how information quality varies across a bank's *expected losses* for each loan, which are defined as the product of exposure at default, loss given default, and PD. Beyond providing a more direct measure of banks' exposure to loans, expected losses are a key measure of credit risk used to calculate regulatory capital under Basel (Basel Committee on Banking Supervision (2015)). If banks endogenously produce information in response to their incentives, we expect information quality to be both higher and more cyclically sensitive for loans with higher expected losses.

To test this hypothesis, we repeat our previous analysis but instead use quartiles of expected losses in Figure 8. As with exposure at default, the top panel shows that information quality is lower for loans with the smallest expected losses and monotonically improves as expected losses become larger. The bottom panels compare the cyclical sensitivity of information quality for the top and bottom quartiles across periods of low and high local unemployment. The bottom-left panel shows that differences in information quality across loans with the largest and smallest EL remain significant even during periods of low unemployment (0.655 vs 0.552). However, similar to the results regarding exposure at default in Figure 7, the bottom-right panel shows that this difference is larger during periods of high local unemployment (0.761 vs 0.592).

[22] Specifically, the exposure at default incorporates the expected utilization and prepayment of principal by the borrower. Using this as a measure of loan size is particularly advantageous for credit lines in which the utilization rate changes over time. For this reason, exposure at default is a key parameter for calculating regulatory capital under Basel (e.g., Calculation of RWA for credit risk).

One potential concern is that the difference in the underlying distribution of banks' reported probabilities of default in high- and low-unemployment periods may differ across the various loan characteristics we analyze in Sections 5.1 and 5.2. As discussed in Section 4.2, differences in the underlying distribution of PDs can mechanically affect the area under the reciver operating characteristic curve. However, in Online Appendix Section C.4, we show that, in the absence of differences in information production, mechanical differences in the distribution of underlying PDs are unlikely to explain our cross-sectional results.

The results in Sections 5.1 and 5.2 help rule out the possibility that our results in Section 4.2 are driven purely by exogenous variation in information over the cycle. Specifically, if downturns exogenously reveal which firms are "naked," we would expect this to hold for all loans rather than just new loans, large loans, or loans with higher expected losses. Similarly, if banks choose to lend only to the borrowers they happen to have more information about in bad times, we do not see how this alternative mechanism can explain why banks have exogenously better information for large loans and those with higher expected losses.[23]

## 5.3 Lending Outcomes Over the Business Cycle

The evidence in the previous two sections is consistent with banks endogenously producing more information in downturns because they have stronger incentives to do so. In this section, we show that while the characteristics of new loans in a county do not meaningfully change as the local unemployment rate rises, the number and volume of new loans decline sharply. We argue that this is consistent with banks lending more selectively in downturns because of increased information production.

We first estimate the following regression across different outcome variables $y_i$:

$$y_i = \beta UR_{c,t} + \Omega X_i + \delta_{b,t} + \gamma_{j,t} + \sigma_{b,c} + \epsilon_i.$$

This regression includes the same firm-level characteristics and fixed effects that we use in (1); however, we exclude loan characteristics as controls and instead include them as dependent variables. The coefficient $\beta$ reflects how each of these characteristics changes when the dummy variable $UR_{c,t}$ takes on a value of one, indicating that the local unemployment rate is above its median. We cluster standard errors by county. Table 5 displays the results.

Loan amounts and loan maturities do not vary over the business cycle in a statistically significant way. Interest rates and banks' estimated probabilities of default (PDs) are only marginally higher in bad times—about 3bps and 6bps higher when the local unemployment rate is above its median, respectively—and neither difference is statistically significant. Similarly, Table 6, which reports the results of a similar regression using firm outcomes as the dependent

[23] As we discuss below in Section 5.3, we do believe that riskier borrowers are approaching banks for loans; however, this should not explain our main result so long as banks do not have exogenously better information about these firms, which our results in this and the previous section help rule out.

variable while controlling for loan characteristics, shows that firm characteristics do not appear to move meaningfully over the cycle. Hence, while the pool of potential borrowers is likely to be riskier in downturns, the pool of loans actually granted does not seem substantially riskier.

Despite minimal changes in loan and firm characteristics over the business cycle, we do observe large changes in lending volume over the business cycle. We aggregate the number and total volume of loans to the county level, take logs, and then regress these measures on the high-unemployment indicator and county fixed effects. The results are reported in Table 7, which shows a decline in both the number and total volume of loans in a county as its economic conditions worsen. Specifically, periods of above-median local unemployment are associated with a 6.7% decrease in the number of loans and an 8.9% decrease in total loan volume. Together, these results suggest that local downturns primarily affect the number and volume of loans issued by banks rather than the composition of loan types or borrowers.

Our model in Appendix B shows that these results, combined with our earlier results on the higher areas under the receiver operating characteristic curves (AUCs) in bad times, can be rationalized if the average default risk of *potential* borrowers is higher in downturns.[24] In our model, the value of producing information is to screen out lower-quality borrowers. Hence, increased information production in downturns results in lower aggregate lending volume, a higher AUC due to more informative PDs, and the possibility that the average risk of borrowers receiving credit remains the same.[25]

# 6 Information Quality and House Price Growth

The previous sections establish that bank information quality is countercyclical and that this effect is concentrated among loans that are more information sensitive, such as loans with higher expected losses. A key determinant of expected losses is collateral values, which recent theories suggest play an important role in how banks' information production incentives evolve over the business cycle. For example, Asriyan, Laeven, and Martin (2022) show that booms driven by high collateral values result in reduced information production as lenders face lower expected losses.[26] This section tests the implications of these theories by analyzing how information quality varies with changes in local house prices, which are commonly used in the

[24] See Proposition 3 for more details. At first glance, it may seem puzzling that more information is produced in downturns even though the pool of borrowers receiving credit is not riskier. What matters in both our model and other models in the literature is not the average risk of borrowers, but the difference in expected payoffs across unobservable borrower types, which is higher in downturns.

[25] Of course, there is no requirement that the average risk of borrowers remains the same; the net effect could be positive or negative. We only argue that it is theoretically possible for this mechanism to explain what we observe in the data. The mechanism in our model in which lending volume and information production are inextricably linked is very similar to those in the literature, such as Dang, Gorton, and Holmström (2013), Gorton and Ordonez (2014), and Asriyan, Laeven, and Martin (2022), but increased information production need not result in lower lending volume. For instance, banks may produce more information but maintain the same lending standards.

[26] See also Gorton and Ordonez (2020) for a theory in which "good booms" are driven by increases in productivity, while "bad booms" are driven by increases in collateral values.

literature as a proxy for collateral values.

We first provide evidence of the link between local house prices and loan-level measures of expected losses. Specifically, we estimate the following regression:

$$\Delta V_{i,t} = \beta R_t + \epsilon_{i,t},$$

where the dependent variable, $\Delta V_{i,t}$, is either the quarterly change in a loan's loss given default or change in log expected loss. To the extent that local house price increases are associated with rising collateral values, higher house prices should lead to higher recovery values and lower expected losses.[27] The independent variable of interest is $R_t$, which is the county-level quarterly house price return (from $t-1$ to $t$) obtained from Zillow. In some specifications, we also include county and/or loan fixed effects. We cluster our standard errors by county.

Table 8 displays the results. The heading at the top row reports the dependent variable in each regression. The first three columns under each heading report results with and without loan or county fixed effects. The second three columns in each heading repeat the same exercises while also controlling for whether the county-level unemployment rate was above or below its median. Across all specifications, stronger appreciation in house prices leads to a statistically significant decrease in loss given default and expected losses, which is consistent with the positive relationship between collateral values and real estate prices documented in Chaney, Sraer, and Thesmar (2012). Specifically, a 1% increase in housing prices leads to just over a 1% decrease in expected losses. The fact that this relationship holds even after controlling for local economic conditions suggests that collateral values—and, in turn, expected losses—do not simply move uniformly with the business cycle.

Having verified the empirical link between local house prices and measures of expected losses, we next test whether changes in house prices drive differences in information quality. The left panel of Figure 9 shows ROC curves estimated from our baseline sample of new loans split by whether county-level house price growth at origination was above or below its median for that county across our sample period. The AUC is much higher in periods of low house price growth (0.744) than in periods of high house price growth (0.655), and this difference is statistically significant and is consistent with banks having weaker incentives to produce information when house prices increase. The correlation between the loan-level indicator for "high unemployment" and "low house price growth" is 0.08 in our baseline sample of new loans, suggesting that these two measures reflect independent sources of variation in local conditions.

We next show that house price growth affects banks' information quality even when we re-

[27] While the Y-14Q data do include a field for the market value of collateral, this variable is either missing or zero for the vast majority of our sample. However, the value of collateral should be reflected in losses given default, and, in turn, expected losses (see Frye et al. (2000)). While we could compare collateralized and non-collateralized loans, roughly 90% of all new loans in our sample report having some collateral. Moreover, in most models, collateral typically matters because it affects expected losses, and we analyze expected losses directly in Figure 8.

strict the sample to periods of low unemployment in the right panel of Figure 9. This exercise is motivated by Asriyan, Laeven, and Martin (2022), who show that booms driven by high collateral values result in information depletion due to diminished information production incentives. For this analysis, we restrict the sample to periods of low unemployment (i.e., booms) and test for differences in information quality based on house price returns. The difference in information quality remains statistically and economically significant in this sample, which provides direct empirical support for theories of endogenous information production, such as Asriyan, Laeven, and Martin (2022), by showing that variation in collateral values affects information quality.

Taken together, these exercises provide additional support for the countercyclical information quality we document being driven by banks' endogenous information production decisions. However, they also highlight that not all economic expansions and downturns are alike in terms of their effect on information quality. Booms (or busts) accompanied by rapid growth in collateral values are more prone to reduced information production.

# 7 Conclusion

Information plays a crucial role in banks' lending decisions and, in turn, macroeconomic outcomes, but it is difficult to analyze empirically. In this paper, we analyze bank information quality from confidential regulatory data containing banks' private risk assessments of their borrowers. Using county-level variation in unemployment rates, we find that information quality improves as local economic conditions worsen. We provide evidence that these results are consistent with theories of endogenous information production by showing that information quality is particularly higher during downturns for newly originated loans and loans with higher expected losses, and that information quality is lower during booms accompanied by high house price growth. To our knowledge, our findings are the first in the empirical banking literature providing evidence of countercyclical information *production*.

While the focus of our analysis is on how banks' information production evolves over the business cycle, our paper demonstrates how banks' private risk assessments in Y-14Q data can be used to analyze how banks' information production decisions vary across other dimensions. This opens up promising avenues for future research examining how various factors—such as market structure, competition, technological change, organizational design, and capital requirements—shape banks' information production decisions.

**Data Availability Statement:** The primary data source is Schedule H.1 of the Federal Reserve's Y-14Q regulatory filings. Researchers employed by the Federal Reserve System may obtain access to this data for research purposes. The code and other data underlying this article are available at https://doi.org/10.5281/zenodo.17465449.

# References


Asea, Patrick K and Brock Blomberg, 1998, Lending cycles, *Journal of Econometrics* 83, 89–128.

Asriyan, Vladimir, Luc Laeven, and Alberto Martin, 2022, Collateral booms and information depletion, *The Review of Economic Studies* 89, 517–555.

Basel Committee on Banking Supervision, 2005, Studies on the validation of internal rating systems, Working Paper No. 14, Bank for International Settlements.

Basel Committee on Banking Supervision, 2015, Guidance on credit risk and accounting for expected credit losses.

Bassett, William F, Mary Beth Chosak, John C Driscoll, and Egon Zakrajšek, 2014, Changes in bank lending standards and the macroeconomy, *Journal of Monetary Economics* 62, 23–40.

Becker, Bo, Marieke Bos, and Kasper Roszbach, 2020, Bad times, good credit, *Journal of Money, Credit and Banking* 52, 107–142.

Bedayo, Mikel, Gabriel Jiménez, José-Luis Peydró, and Raquel Vegas Sánchez, 2020, Screening and loan origination time: lending standards, loan defaults and bank failures .

Behn, Markus, Rainer FH Haselmann, and Vikrant Vig, 2016, The limits of model-based regulation .

Benedetti, Riccardo, 2010, Scoring rules for forecast verification, *Monthly Weather Review* 138, 203–211.

Berg, Tobias and Philipp Koziol, 2017, An analysis of the consistency of banks' internal ratings, *Journal of Banking & Finance* 78, 27–41.

Berg, Tobias, Manju Puri, and Jörg Rocholl, 2020, Loan officer incentives, internal rating models, and default rates, *Review of Finance* 24, 529–578.

Beyhaghi, Mehdi, Cesare Fracassi, and Gregory Weitzner, 2025, Adverse selection in corporate loan markets, *Journal of Finance* forthcoming.

Bidder, Rhys M, Nicolas Crouzet, Margaret Jacobson, and Michael Siemer, 2023, Debt flexibility .

Bidder, Rhys M, John R Krainer, and Adam Hale Shapiro, 2020, De-leveraging or de-risking? how banks cope with loss, *Review of Economic Dynamics* .

Biswas, Sonny, 2022, Collateral and bank screening as complements: A spillover effect, *Working Paper* .

Boyd, John H and Edward C Prescott, 1986, Financial intermediary-coalitions, *Journal of Economic theory* 38, 211–232.

Bradley, Andrew P, 1997, The use of the area under the roc curve in the evaluation of machine learning algorithms, *Pattern recognition* 30, 1145–1159.

Breiman, Leo, 2001, Random forests, *Machine learning* 45, 5–32.

Brier, Glenn W, 1950, Verification of forecasts expressed in terms of probability, *Monthly weather review* 78, 1–3.

Cerqueiro, Geraldo, Steven Ongena, and Kasper Roszbach, 2016, Collateralization, bank loan rates, and monitoring, *The Journal of Finance* 71, 1295–1322.

Chaney, Thomas, David Sraer, and David Thesmar, 2012, The collateral channel: How real estate shocks affect corporate investment, *American Economic Review* 102, 2381–2409.

Dang, Tri Vi, Gary Gorton, and Bengt Holmström, 2012, Ignorance, debt and financial crises, *Unpublished, Yale SOM* .

Dang, Tri Vi, Gary Gorton, and Bengt Holmström, 2013, The information sensitivity of a security, *Unpublished working paper, Yale University* 39–65.

Dell'Ariccia, Giovanni and Robert Marquez, 2006, Lending booms and lending standards, *The Journal of Finance* 61, 2511–2546.

Dell'Ariccia, Giovanni, Deniz Igan, and Luc UC Laeven, 2012, Credit booms and lending standards: Evidence from the subprime mortgage market, *Journal of Money, Credit and Banking* 44, 367–384.

DeLong, Elizabeth R, David M DeLong, and Daniel L Clarke-Pearson, 1988, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* 837–845.

Diamond, Douglas W, 1984, Financial intermediation and delegated monitoring, *The review of economic studies* 51, 393–414.

Engelmann, Bernd and Robert Rauhmeier, *The basel II risk parameters: estimation, validation, stress testing-with applications to loan risk management* (Springer Science & Business Media 2011).

Farboodi, Maryam and Peter Kondor, 2020, Rational sentiments and economic cycles, Working paper, National Bureau of Economic Research.

Fishman, Michael J, Jonathan A Parker, and Ludwig Straub, 2020, A dynamic theory of lending standards, Working paper, National Bureau of Economic Research.

Frame, W Scott, Ping McLemore, and Atanas Mihov, 2025, Haste makes waste: Banking organization growth and operational risk, *The Review of Corporate Finance Studies* cfaf003.

Frye, Jon, Lisa Ashley, Robert Bliss, Richard Cahill, Paul Calem, Matthew Foss, Michael Gordy, David Jones, Catherine Lemieux, Michael Lesiak et al., 2000, Collateral damage: A source of systematic credit risk, *Risk* 13, 91–94.

Gorton, Gary and Guillermo Ordonez, 2014, Collateral crises, *American Economic Review* 104, 343–78.

Gorton, Gary and Guillermo Ordonez, 2020, Good booms, bad booms, *Journal of the European Economic Association* 18, 618–665.

Gorton, Gary B and Ping He, 2008, Bank credit cycles, *The Review of Economic Studies* 75, 1181–1214.

Goulet Coulombe, Philippe, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant, 2022, How is machine learning useful for macroeconomic forecasting?, *Journal of Applied Econometrics* 37, 920–964.

Granger, Clive WJ, 1969, Prediction with a generalized cost of error function, *Journal of the Operational Research Society* 20, 199–207.

Grether, David M, 1980, Bayes rule as a descriptive model: The representativeness heuristic, *The Quarterly journal of economics* 95, 537–557.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Gustafson, Matthew T, Ivan T Ivanov, and Ralf R Meisenzahl, 2021, Bank monitoring: Evidence from syndicated loans, *Journal of Financial Economics* 139, 452–477.

Hayashi, Yoichi, 2022, Emerging trends in deep learning for credit scoring: A review, *Electronics* 11, 3181.

Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo FP Luttmer, and Kelly Shue, 2016, Screening peers softly: Inferring the quality of small borrowers, *Management Science* 62, 1554–1577.

James, Christopher, 1987, Some evidence on the uniqueness of bank loans, *Journal of financial economics* 19, 217–235.

Khwaja, Asim Ijaz and Atif Mian, 2008, Tracing the impact of bank liquidity shocks: Evidence from an emerging market, *American Economic Review* 98, 1413–1442.

Leland, Hayne E and David H Pyle, 1977, Informational asymmetries, financial structure, and financial intermediation, *The journal of Finance* 32, 371–387.

Lessmanna, Stefan, H Seowb, Bart Baesenscd, and Lyn C Thomasd, Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update, *Credit Research Centre, Conference Archive* (2013).

Lown, Cara and Donald P Morgan, 2006, The credit cycle and the business cycle: new findings using the loan officer opinion survey, *Journal of Money, Credit and Banking* 1575–1597.

Maddaloni, Angela and José-Luis Peydró, 2011, Bank risk-taking, securitization, supervision, and low interest rates: Evidence from the euro-area and the us lending standards, *the review of financial studies* 24, 2121–2165.

Manove, Michael, A Jorge Padilla, and Marco Pagano, 2001, Collateral versus project screening: A model of lazy banks, *Rand journal of economics* 726–744.

Mincer, Jacob A and Victor Zarnowitz, The evaluation of economic forecasts, *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, 3–46 (NBER 1969).

Murphy, Allan H, 1973, A new vector partition of the probability score, *Journal of Applied Meteorology and Climatology* 12, 595–600.

Muth, John F, 1961, Rational expectations and the theory of price movements, *Econometrica: journal of the Econometric Society* 315–335.

O'Brien, Robert M, 2007, A caution regarding rules of thumb for variance inflation factors, *Quality & quantity* 41, 673–690.

Pepe, Margaret Sullivan, *The statistical evaluation of medical tests for classification and prediction* (Oxford university press 2003).

Petriconi, Silvio, 2015, Bank competition, information choice and inefficient lending booms, *Working Paper* .

Plosser, Matthew C and Joao AC Santos, 2018, Banks' incentives and inconsistent risk models, *The Review of Financial Studies* 31, 2080–2112.

Puri, Manju, Jörg Rocholl, and Sascha Steffen, 2017, What do a million observations have to say about loan defaults? opening the black box of relationships, *Journal of Financial Intermediation* 31, 1–15.

Rodano, Giacomo, Nicolas Serrano-Velarde, and Emanuele Tarantino, 2018, Lending standards over the credit cycle, *The Review of Financial Studies* 31, 2943–2982.

Ruckes, Martin, 2004, Bank competition and credit standards, *Review of Financial Studies* 17, 1073–1102.

Schonlau, Matthias and Rosie Yuyan Zou, 2020, The random forest algorithm for statistical learning, *The Stata Journal* 20, 3–29.

Tversky, Amos and Daniel Kahneman, 1974, Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty., *science* 185, 1124–1131.

Wager, Stefan and Susan Athey, 2018, Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association* 113, 1228–1242.

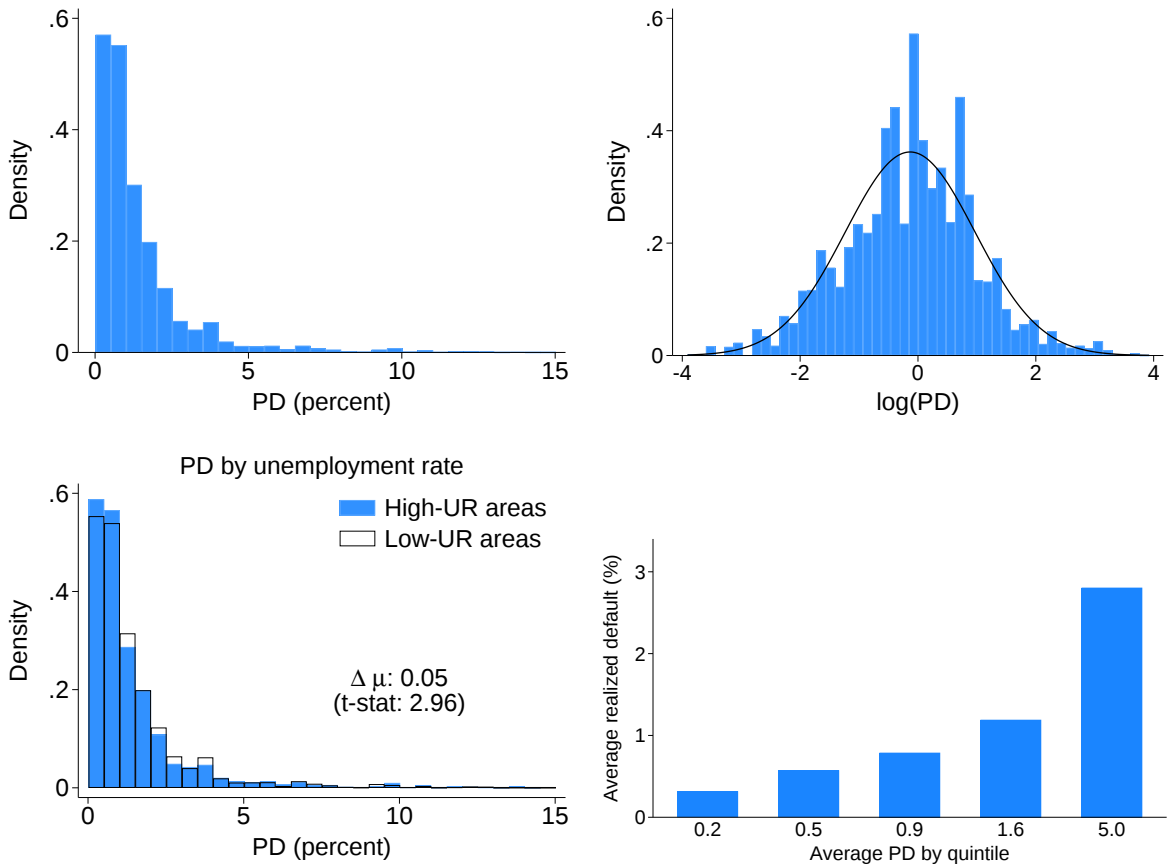


**Figure 1:** Distributions of banks' reported probabilities of default

The top-left panel plots the distribution of banks' reported probabilities of default (PD). The top-right panel plots the distribution of log(PD) with an overlaid normal distribution. The bottom-left panel plots overlaid distributions of PDs originated when a county's unemployment rate is above (solid bars) or below (hollow bars) its median from 2014Q4-2021Q4. $\Delta\mu$ reports the difference between the average PD at origination in high-UR areas minus the average PD in low-UR areas, with the corresponding t-statistic shown below in parentheses. The bottom-right panel reports average default rates within two years of origination on the y-axis by quintiles of PD at origination. The numbers beneath each bar correspond to the average PD in each quintile (rounded to the nearest 0.1pp).

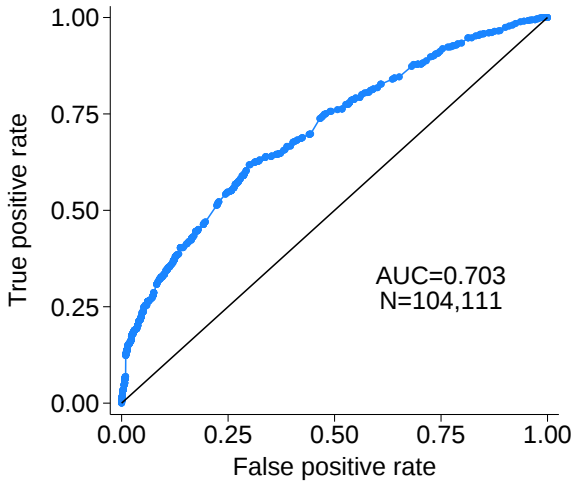


**Figure 2:** Receiver operating characteristic (ROC) curve for new loans

This figure displays the receiver operating characteristic (ROC) curve for all new loans in our sample, along with the area under the ROC curve (AUC).

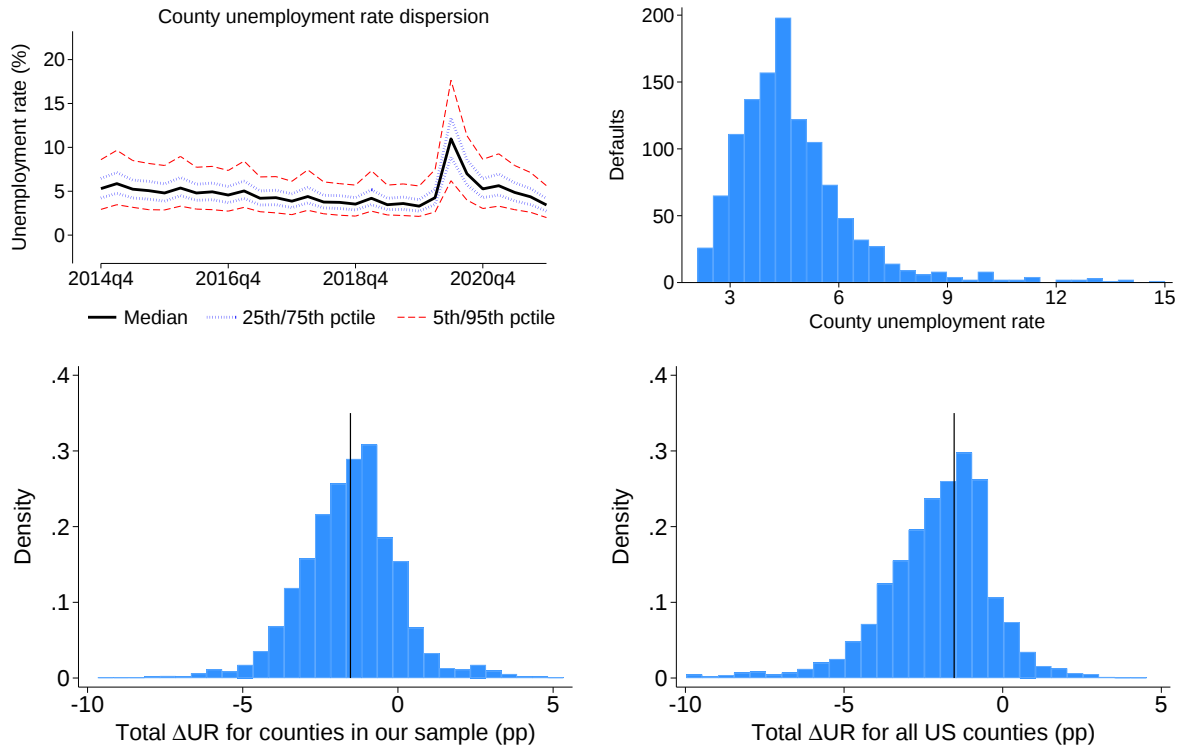


**Figure 3:** Variation in unemployment rates

The top-left panel shows the range of county-level unemployment rates for all county-quarter observations with at least one new loan in our sample. The top-right panel plots the distribution of the unemployment rate at origination for the 1,176 new loans in our sample that default within two years of origination. The bottom-left panel shows the distribution of county-level changes from the earliest observation of the unemployment rate to the latest for each county with at least two new loans in our sample. The bottom-right panel shows the distribution of county-level changes in the unemployment rate for all US counties from 2014Q4 through 2021Q4. The vertical black lines at -1.5pp show the change in the total US unemployment rate during this period. The bottom panels are truncated between -10pp and +5pp for readability. The number of counties used in our sample is shown in panel C of Table 1.

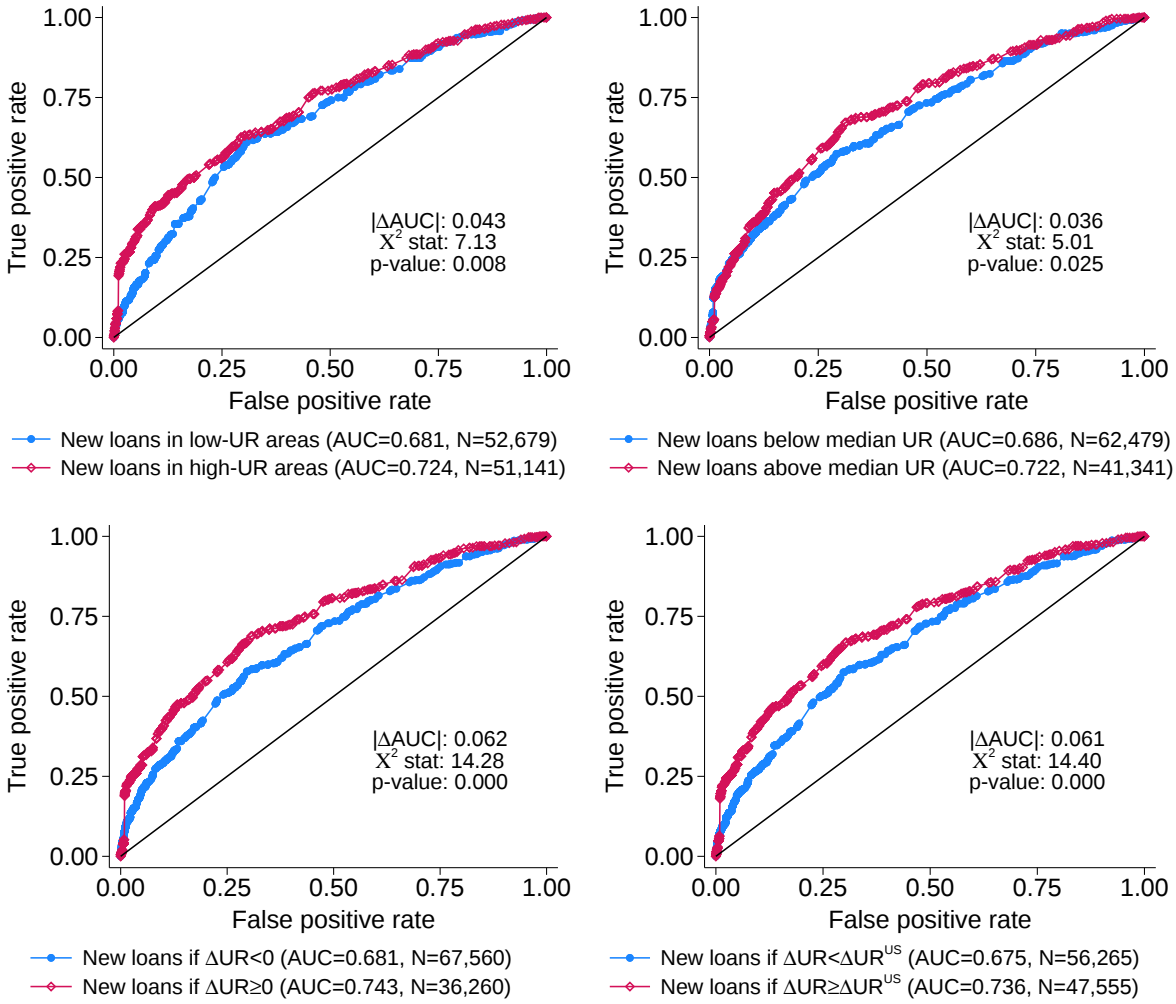


**Figure 4:** Receiver operating characteristic (ROC) curve for new loans over the business cycle

The top-left panel shows the receiver operating characteristic (ROC) curve for all new loans in our sample split by whether the unemployment rate in each period was above or below its county-level median from 2014Q4-2021Q4. The top-right panel shows ROC curves split by whether the unemployment rate at origination was above or below the median unemployment rate of all counties with at least 5 new loans in that quarter. The bottom-left panel shows ROC curves split by whether the county unemployment rate increased from the prior quarter. The bottom-right panel shows ROC curves split by whether the change in the county unemployment rate was greater than the change in the total US unemployment rate. The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

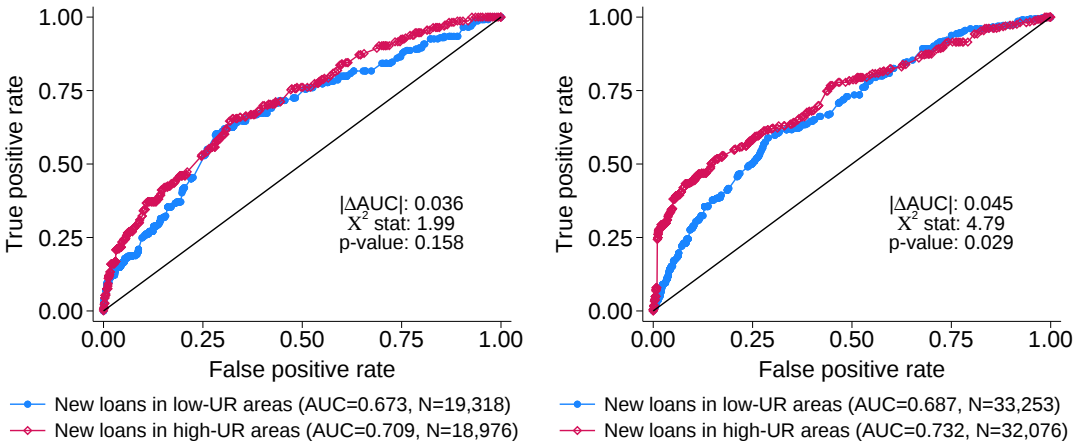


**Figure 5:** Receiver operating characteristic (ROC) curve for new loans to tradeable (left) vs nontradeable (right) industries

This figure shows receiver operating characteristic (ROC) curves for new loans split by the local unemployment rate at origination for tradeable (left) and nontradeable (right) industries. Nontradeable industries include utilities, construction, wholesale trade, retail trade, transportation, accommodation, food services, information and communication, and professional and administrative services; all other loans in our sample with non-missing industry codes are considered tradeable. The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

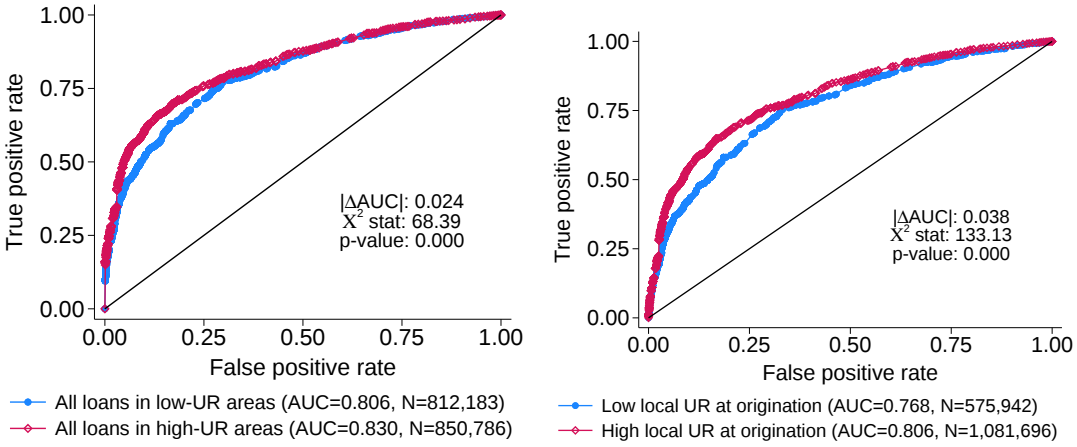


**Figure 6:** Receiver operating characteristic (ROC) curve for current (left) versus origination (right) unemployment rate among all loans

The left panel shows receiver operating characteristic (ROC) curves for both new and existing loans split by whether the contemporaneous unemployment rate in each period was above or below its county-level median from 2014Q4-2021Q4. The right panel shows ROC curves split by whether the unemployment rate at origination was above or below its county-level median from 2014Q4-2021Q4. The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

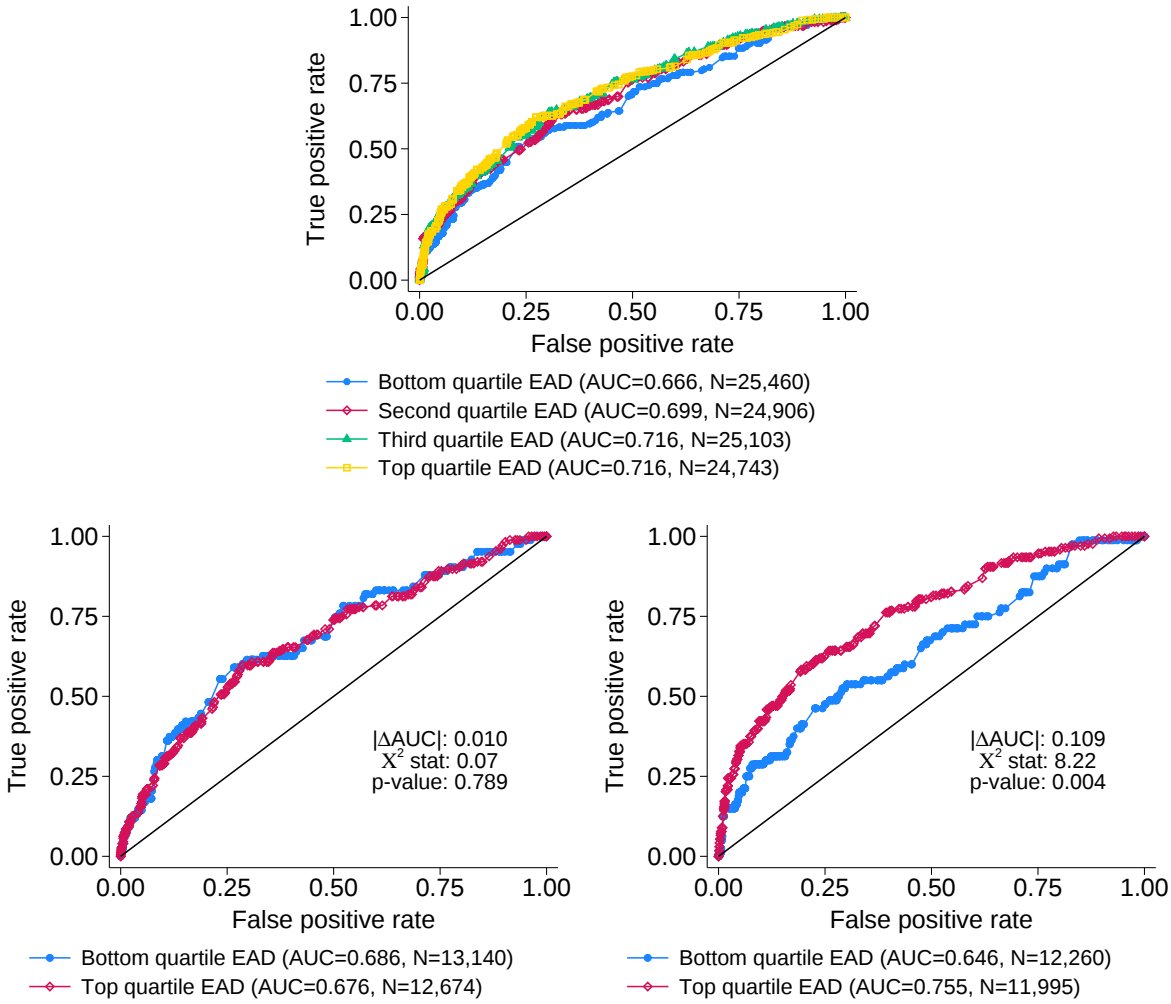


**Figure 7:** Receiver operating characteristic (ROC) curve by quartiles of exposure at default

The top panel shows ROC curves for new loans based on their quartile of exposure at default calculated within each bank-quarter. The bottom-left panel compares the ROC curves for the highest and lowest quartiles in low-UR areas. The bottom-right panel compares the ROC curves for the highest and lowest quartiles in high-UR areas. The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

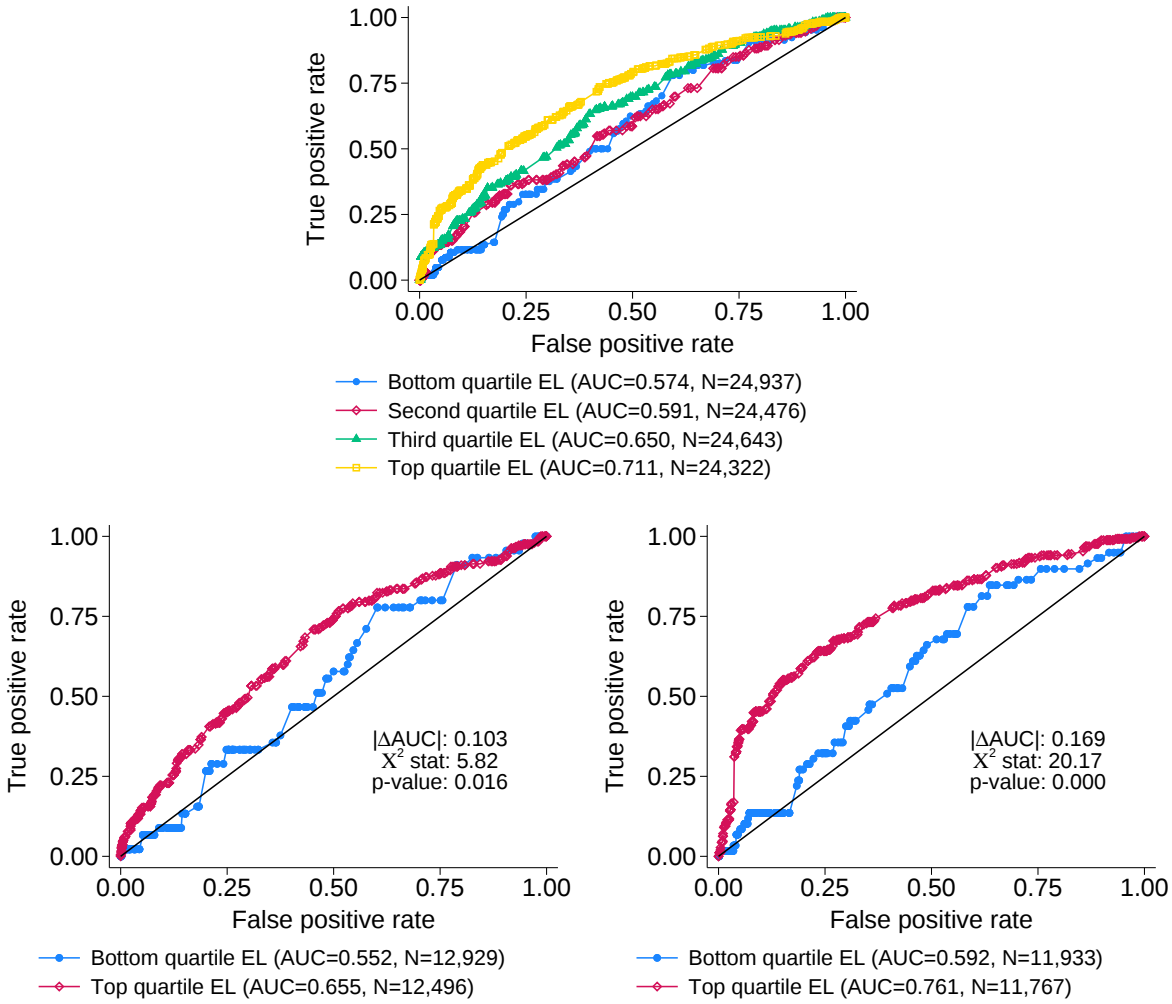


**Figure 8:** Receiver operating characteristic (ROC) curve by quartiles of expected loss

This figure shows receiver operating characteristic (ROC) curves for new loans based on their quartile of expected loss, the product of exposure at default, loss given default, and probability of default, calculated within each bank-quarter. The bottom-left panel compares the ROC curves for the highest and lowest quartiles in low-UR areas. The bottom-right panel compares the ROC curves for the highest and lowest quartiles in high-UR areas. The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

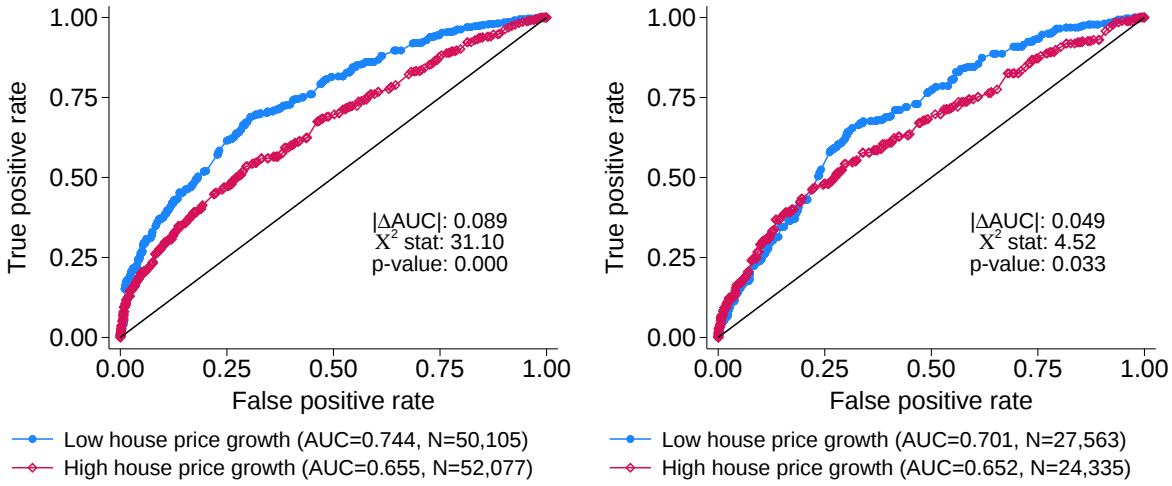


**Figure 9:** Receiver operating characteristic (ROC) curve and local house prices

This figure shows receiver operating characteristic (ROC) curves for new loans split by county-level house price growth. The left panel includes our entire sample of new loans. The right panel includes only new loans in counties where the local unemployment rate was below its county-level median during our sample period. High (low) house price growth is determined by whether the quarterly change in house prices was above (below) that county's median during our sample period. The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

**Table 1:** Summary statistics

Panel A reports loan characteristics calculated as simple averages across all new loans. Panel B reports firm characteristics, which are calculated after collapsing all loan observations (including both new and existing loans) to the firm-quarter level using simple averages. Panel C reports county characteristics, which are calculated after collapsing all loan observations (including both new and existing loans) to the county-quarter level. Section 2 describes our sample.

| | Mean | Median | 5% | 95% | SD | N |
|---|---|---|---|---|---|---|
| **Panel A: Loan characteristics** | | | | | | |
| Interest rate (pp) | 3.70 | 3.56 | 1.64 | 6.16 | 1.46 | 87,193 |
| Probability of default (pp) | 1.63 | 0.91 | 0.14 | 5.07 | 2.73 | 104,111 |
| Loss given default (ratio) | 0.35 | 0.35 | 0.08 | 0.61 | 0.16 | 101,884 |
| Realized default (pp) | 1.13 | 0.00 | 0.00 | 0.00 | 10.57 | 104,111 |
| Maturity (months) | 47.49 | 58.00 | 7.00 | 96.00 | 31.39 | 104,111 |
| Loan size ($ mil) | 13.32 | 3.58 | 1.00 | 50.00 | 41.54 | 104,111 |
| Revolver (indicator) | 0.38 | 0.00 | 0.00 | 1.00 | 0.49 | 104,111 |
| Term loan (indicator) | 0.41 | 0.00 | 0.00 | 1.00 | 0.49 | 104,111 |
| Floating rate (indicator) | 0.56 | 1.00 | 0.00 | 1.00 | 0.50 | 104,111 |
| **Panel B: Firm characteristics** | | | | | | |
| Sales | 783.80 | 44.07 | 3.34 | 1,412.12 | 36,404.71 | 948,138 |
| Assets | 1,328.81 | 21.52 | 1.63 | 1,484.10 | 253,734.53 | 947,916 |
| Leverage | 0.31 | 0.26 | 0.00 | 0.81 | 0.27 | 919,616 |
| Profitability | 0.18 | 0.13 | -0.04 | 0.56 | 0.24 | 937,776 |
| Tangibility | 0.89 | 0.99 | 0.39 | 1.00 | 0.20 | 936,201 |
| Nontradeable | 0.65 | 1.00 | 0.00 | 1.00 | 0.48 | 1,069,329 |
| Probability of default (PD) | 2.49 | 0.91 | 0.13 | 9.82 | 6.87 | 1,070,295 |
| Total number of loans | 1.56 | 1.00 | 1.00 | 4.00 | 4.11 | 1,070,295 |
| Number of new loans | 0.10 | 0.00 | 0.00 | 1.00 | 0.47 | 1,070,295 |
| Number of banks | 1.15 | 1.00 | 1.00 | 2.00 | 0.65 | 1,070,295 |
| Total loan volume ($ mil) | 20.77 | 4.16 | 1.00 | 85.00 | 119.29 | 1,070,295 |
| **Panel C: County characteristics** | | | | | | |
| Unemployment rate | 5.01 | 4.53 | 2.53 | 9.03 | 2.26 | 52,832 |
| Number of new loans | 1.97 | 0.00 | 0.00 | 9.00 | 7.14 | 52,910 |
| Number of total loans | 31.58 | 5.00 | 1.00 | 138.00 | 102.11 | 52,910 |
| Total new loan volume ($ mil) | 420.14 | 39.87 | 1.41 | 1,892.21 | 1,788.92 | 52,910 |

**Table 2:** Additional loan-level summary statistics

This table contains additional loan-level summary statistics (observations are at the same level as in panel A of Table 1). Panel A includes our entire sample, Panel B includes new loans in low-UR areas, and Panel C includes new loans in high-UR areas. Because a small number of loans are made in counties with missing unemployment rates, the sample sizes in panel A are greater than the sum of panels B and C. Section 2 describes our sample.

| | Mean | Median | 5% | 95% | SD | N |
|---|---|---|---|---|---|---|
| **Panel A: All loans** | | | | | | |
| Sales | 2,584.48 | 81.50 | 3.63 | 4,252.16 | 47,106.60 | 84,208 |
| Assets | 3,816.42 | 50.51 | 2.04 | 4,872.10 | 87,873.09 | 84,240 |
| Leverage | 0.34 | 0.31 | 0.00 | 0.81 | 0.26 | 82,291 |
| Profitability | 0.20 | 0.15 | -0.00 | 0.61 | 0.24 | 84,240 |
| Tangibility | 0.85 | 0.97 | 0.31 | 1.00 | 0.23 | 84,040 |
| Nontradeable | 0.63 | 1.00 | 0.00 | 1.00 | 0.48 | 103,914 |
| Probability of default (PD) | 1.63 | 0.91 | 0.14 | 5.07 | 2.73 | 104,111 |
| Total loan volume ($ mil) | 13.32 | 3.58 | 1.00 | 50.00 | 41.54 | 104,111 |
| **Panel B: Loans in low-UR areas** | | | | | | |
| Sales | 2,410.35 | 85.49 | 5.30 | 4,252.16 | 28,254.39 | 42,667 |
| Assets | 3,245.06 | 53.20 | 2.42 | 4,829.70 | 35,858.18 | 42,721 |
| Leverage | 0.34 | 0.31 | 0.00 | 0.80 | 0.26 | 41,761 |
| Profitability | 0.20 | 0.15 | 0.00 | 0.60 | 0.23 | 42,721 |
| Tangibility | 0.85 | 0.97 | 0.31 | 1.00 | 0.23 | 42,618 |
| Nontradeable | 0.63 | 1.00 | 0.00 | 1.00 | 0.48 | 52,571 |
| Probability of default (PD) | 1.61 | 0.93 | 0.14 | 4.66 | 2.58 | 52,679 |
| Total loan volume ($ mil) | 13.11 | 3.50 | 1.00 | 50.00 | 35.90 | 52,679 |
| **Panel C: Loans in high-UR areas** | | | | | | |
| Sales | 2,740.06 | 78.04 | 2.15 | 4,175.67 | 60,795.20 | 41,306 |
| Assets | 4,371.10 | 47.59 | 1.65 | 4,739.00 | 120,080.33 | 41,284 |
| Leverage | 0.35 | 0.31 | 0.00 | 0.81 | 0.26 | 40,299 |
| Profitability | 0.20 | 0.15 | -0.01 | 0.62 | 0.24 | 41,284 |
| Tangibility | 0.85 | 0.97 | 0.31 | 1.00 | 0.23 | 41,187 |
| Nontradeable | 0.63 | 1.00 | 0.00 | 1.00 | 0.48 | 51,052 |
| Probability of default (PD) | 1.66 | 0.90 | 0.14 | 5.37 | 2.88 | 51,141 |
| Total loan volume ($ mil) | 13.49 | 3.70 | 1.00 | 50.00 | 46.59 | 51,141 |

**Table 3:** Banks' probabilities of default predict default

This table shows the results of estimating (1), which tests whether banks' reported probabilities of default (PDs) predict default after controlling for observables. The dependent variable in each regression is a dummy variable indicating whether each loan defaults within eight quarters after origination, multiplied by 100. Interest rates and PDs are measured in percentage points. Standard errors are clustered at the county level and are shown below the parameter estimates in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

| | Default | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| PD | 0.407*** | 0.444*** | | | 0.445*** |
| | (0.049) | (0.064) | | | (0.073) |
| Interest rate | | | 0.508*** | 0.452*** | 0.227*** |
| | | | (0.068) | (0.078) | (0.073) |
| Leverage | | 1.027*** | | 1.458*** | 0.981*** |
| | | (0.310) | | (0.355) | (0.360) |
| Profitability | | 0.009 | | -0.698*** | 0.029 |
| | | (0.248) | | (0.233) | (0.249) |
| Tangibility | | -0.359 | | -0.274 | -0.289 |
| | | (0.345) | | (0.417) | (0.410) |
| Log firm size | | -0.139*** | | -0.151*** | -0.118** |
| | | (0.041) | | (0.049) | (0.047) |
| Log loan amount | | 0.251*** | | 0.267*** | 0.252*** |
| | | (0.063) | | (0.071) | (0.070) |
| Loss given default | | 0.709* | | 0.130 | 0.673 |
| | | (0.396) | | (0.492) | (0.487) |
| Log maturity | | -0.058 | | -0.203** | -0.100 |
| | | (0.064) | | (0.080) | (0.074) |
| Controls | N | Y | N | Y | Y |
| Bank-quarter FE | Y | Y | Y | Y | Y |
| Industry-quarter FE | Y | Y | Y | Y | Y |
| Bank-county FE | Y | Y | Y | Y | Y |
| Observations | 100,368 | 77,226 | 83,602 | 64,566 | 64,566 |
| $R^2$ | 0.174 | 0.196 | 0.185 | 0.206 | 0.214 |

**Table 4:** Banks' probabilities of default predict default (random forest)

This table tests whether banks' reported probabilities of default (PDs) contain additional information regarding future realized default beyond a predicted default variable estimated using a random forest regression. The dependent variable in each regression is a dummy variable indicating whether each loan defaults within eight quarters after origination, multiplied by 100. "RF predicted PD" is the random forest estimate of the loan's default probability measured in percentage points. The details of these estimates are described in Appendix A. Each column corresponds to a different set of controls used to estimate "RF predicted PD" in the random forest. All specifications include our default set of firm and loan controls; other specifications include indicator variables for industry, bank, and time and are indicated in the rows below the table. "PD" is the probability of default reported by the bank and is measured in percentage points. All regressions exclude the half of the baseline sample used to train the random forest. Standard errors are clustered at the county level and are shown below the parameter estimates in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

| | Default | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| RF predicted PD | 1.087*** | 1.279*** | 1.038*** | 1.153*** | 1.207*** |
| | (0.075) | (0.078) | (0.058) | (0.070) | (0.067) |
| PD | 0.344*** | 0.295*** | 0.315*** | 0.330*** | 0.270*** |
| | (0.045) | (0.041) | (0.040) | (0.042) | (0.039) |
| Industry controls | N | Y | N | N | Y |
| Bank controls | N | N | Y | N | Y |
| Time controls | N | N | N | Y | Y |
| Observations | 51,913 | 51,913 | 51,913 | 51,913 | 51,913 |
| $R^2$ | 0.071 | 0.104 | 0.093 | 0.096 | 0.150 |

**Table 5:** Loan characteristics over the business cycle

This table analyzes the relationship between the local unemployment rate and loan characteristics. "High UR" is a dummy variable equal to one if that county's unemployment rate was above its median during our sample period. The dependent variable in each regression is shown at the top of each column. All regressions include our default firm controls. The unemployment rate (UR), probability of default (PD), default, and interest rate are measured in percentage points. Maturity is measured in log months, and loan size is measured in log dollars. Standard errors are clustered at the county level and are shown below the parameter estimates in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

| | Loan size | Interest rate | Maturity | Default | PD |
|---|---|---|---|---|---|
| High UR | 1.437 | 0.027 | 1.911 | -0.076 | 0.060 |
| | (1.369) | (0.017) | (1.351) | (0.143) | (0.046) |
| Firm controls | Y | Y | Y | Y | Y |
| Bank-quarter FE | Y | Y | Y | Y | Y |
| Industry-quarter FE | Y | Y | Y | Y | Y |
| Bank-county FE | Y | Y | Y | Y | Y |
| Observations | 79,078 | 66,149 | 79,050 | 79,078 | 79,078 |
| $R^2$ | 0.541 | 0.552 | 0.419 | 0.187 | 0.293 |

**Table 6:** Firm characteristics over the business cycle

This table analyzes the relationship between the local unemployment rate and firm characteristics. "High UR" is a dummy variable equal to one if that county's unemployment rate was above its median during our sample period. The dependent variable in these regressions is shown at the top of each column. Firm size is the log of total assets, while profitability, leverage, and tangibility are ratios. All regressions include our default loan controls. Standard errors are clustered at the county level and are shown below the parameter estimates in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

| | Firm size | Profitability | Leverage | Tangibility |
|---|---|---|---|---|
| High UR | -1.780 | -0.004 | -0.003 | -0.004 |
| | (2.913) | (0.003) | (0.004) | (0.004) |
| Loan controls | Y | Y | Y | Y |
| Bank-quarter FE | Y | Y | Y | Y |
| Industry-quarter FE | Y | Y | Y | Y |
| Bank-county FE | Y | Y | Y | Y |
| Observations | 79,128 | 79,128 | 77,268 | 78,937 |
| $R^2$ | 0.646 | 0.295 | 0.354 | 0.455 |

**Table 7:** County-level lending over the business cycle

This table analyzes the relationship between the unemployment rate and county-level loan volume. "High UR" is a dummy variable equal to 1 if that county's unemployment rate was above its median during our sample period. Data are aggregated at the county level. The dependent variable in each regression is shown at the top of each column, and both are in logs. Standard errors are clustered at the county level and are shown below the parameter estimates in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) |
|---|---|---|
| | Loan count | Loan volume |
| High UR | -0.067*** | -0.089*** |
| | (0.009) | (0.018) |
| County FE | Y | Y |
| Observations | 18,396 | 18,396 |
| $R^2$ | 0.749 | 0.587 |

**Table 8:** Collateral values and house prices

This table shows the effects of changes in county-level house prices on loans' loss given default (Columns (1)-(6)) and log expected losses (Columns (7)-(12)). "House price growth" is the quarterly change in house prices from Zillow. "High UR" is a dummy variable equal to 1 if that county's unemployment rate was above its median during our sample period. Standard errors are clustered at the county level and are shown below the parameter estimates in parentheses. *, **, and *** indicate statistical significance at the $10\%$, $5\%$, and $1\%$ levels, respectively.

| | $\Delta$ Loss given default | | | | | | $\Delta$ log(Expected loss) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| House price growth | -0.021*** | -0.027*** | -0.020*** | -0.027*** | -0.034*** | -0.026*** | -1.143*** | -1.257*** | -1.196*** | -1.102*** | -1.207*** | -1.130*** |
| | (0.006) | (0.006) | (0.006) | (0.006) | (0.007) | (0.007) | (0.099) | (0.117) | (0.131) | (0.096) | (0.115) | (0.126) |
| High UR | | | | 0.001*** | 0.001*** | 0.001*** | | | | -0.008*** | -0.008*** | -0.011*** |
| | | | | (0.000) | (0.000) | (0.000) | | | | (0.002) | (0.002) | (0.002) |
| County FE | N | Y | N | N | Y | N | N | Y | N | N | Y | N |
| Loan FE | N | N | Y | N | N | Y | N | N | Y | N | N | Y |
| Observations | 1,386,280 | 1,386,247 | 1,359,216 | 1,386,280 | 1,386,247 | 1,359,216 | 1,355,384 | 1,355,350 | 1,328,971 | 1,355,384 | 1,355,350 | 1,328,971 |
| $R^2$ | 0.000 | 0.002 | 0.087 | 0.000 | 0.002 | 0.087 | 0.001 | 0.003 | 0.094 | 0.001 | 0.003 | 0.094 |

## Appendix A. Random Forest Regression Methodology and Additional Analysis

This section describes the details behind the random forest regression estimates we use as a nonlinear benchmark for predicting default from observables in Sections 3 and 4. The fundamental building block of a random forest is a *regression tree*. A regression tree generates predictions for an outcome variable by sequentially partitioning observations into $K$ regions (also called "leaves" or terminal nodes) based on similarly valued explanatory variables and then calculating average outcomes within each region. The number of partitions $L$ is called the "depth" of the tree. A simple illustrative example based on a similar figure shown in Gu, Kelly, and Xiu (2020) with $K = 3$ and $L = 2$ is shown below in the left panel of Figure A.1.

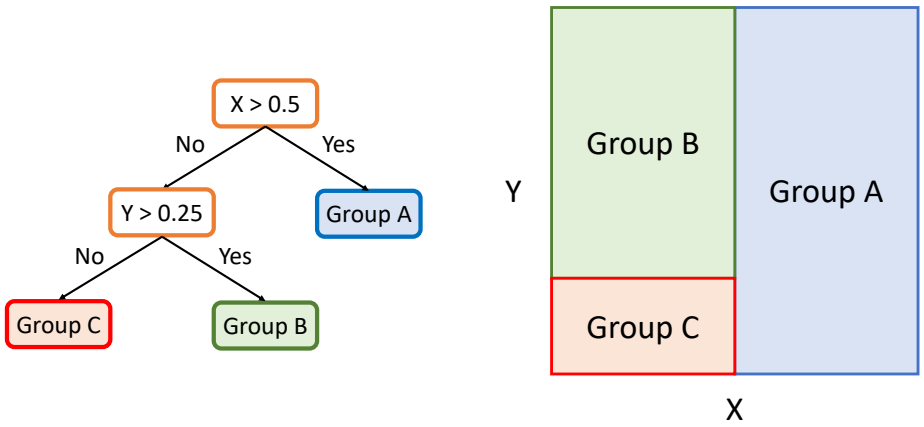


**Figure A.1:** Example Regression Tree

There are two predictors, $X$ and $Y$, each of which takes on values between 0 and 1. To obtain a prediction for a given observation $i$, the node at the top of the tree first partitions the data based on $X_i$. If $X_i > 0.5$, the observation is assigned to Group A. If $X_i \leq 0.5$, the remaining observations are further partitioned by $Y_i$; observations with $Y_i > 0.25$ are assigned to Group B, while observations with $Y_i \leq 0.25$ are assigned to Group C.[28] An equivalent graphical representation of this partition is shown in the right panel. A prediction is obtained by taking the sample average of the outcome variable for all observations in a group.

Because regression trees can easily accommodate outliers and idiosyncrasies in the training data, they tend to perform very well in sample, but are prone to over-fitting. Breiman (2001) showed that the random forest model, which averages predictions across a large number of regression trees estimated from bootstrapped data, substantially improves out-of-sample forecast accuracy. We generate random forest predictions of loan defaults using the `rforest` Stata command developed in Schonlau and Zou (2020). We use the default settings for the number of trees (100) and the number of explanatory variables used in each tree (3), and we do not specify a maximum depth or minimum number of observations per node.

[28] The threshold values shown here at each node are chosen arbitrarily, but in practice, they are usually determined as the solution to an optimization problem such as minimizing the mean squared forecast error.

Following their recommendation, 50% of the data are used for training, while the other 50% are used for validation. We assign observations to each set by first sorting our baseline sample of newly originated loans by loan ID number, and then alternatively placing each loan in the training set or the validation set. The training set is used to estimate the parameters of the random forest, and with these parameters, we generate predictions for the validation set. For consistency, we maintain the same training and validation sets across the range of specifications we estimate. All specifications shown throughout the paper include the observable characteristics described in Section 3: for loans, these are log size, LGD, and log maturity, and for firms, these are leverage, profitability, tangibility, and log assets. Several of the specifications shown in Table 4 also include separate dummy variables for each industry, bank, or time period.

Figure A.2 uses this random forest approach to evaluate the marginal discriminatory power of PD using ROC curves. The blue curve uses default estimates from a random forest regression that uses only the baseline set firm and loan controls, while the red curve includes these same controls plus PD. Neither specification includes any firm, bank, or time dummies. As before, we train each random forest on 50% of the sample and compare the out-of-sample predictive power using the remaining 50%. The difference between the two is statistically significant and larger than the difference between periods of high and low unemployment reported in our baseline results in Figure 4. This provides further evidence that PD contains information useful for predicting default that is not fully captured by nonlinear combinations of other observable characteristics.

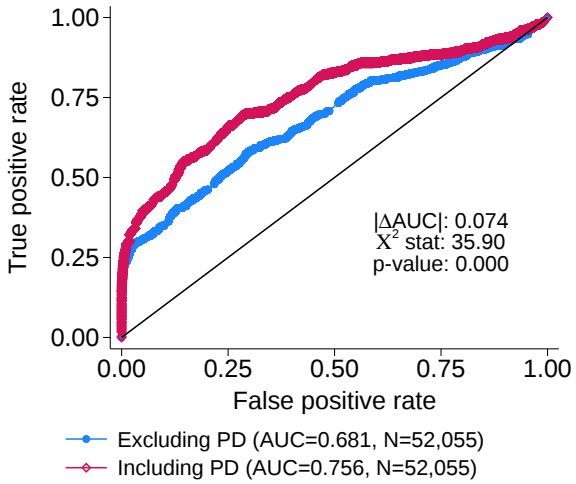


**Figure A.2:** Comparing random forest predictions with and without banks' probabilities of default

This figure compares default predictions from two random forest regressions. The blue curve shows estimates using the following controls: log loan size, LGD, log maturity, firm leverage, firm profitability, firm tangibility, and firm size. The red curve uses the same controls plus banks' reported probability of default (PD). The area under each ROC curve (AUC) is reported along with the number of observations in the legend. $|\Delta\text{AUC}|$ reports the difference between the two AUCs. Below $|\Delta\text{AUC}|$, the DeLong, DeLong, and Clarke-Pearson (1988) statistics are reported: the $\chi^2$ test statistic and its corresponding p-value, which tests the null hypothesis that the difference between the two AUCs equals zero.

## Appendix B. Simple Theoretical Framework

In this section, we develop a simple model, from first principles, which shows that increased information production by banks in bad times leads to a higher area under the receiver operating characteristic curve (AUC) in bad times, consistent with our empirical results.

Before proceeding, we emphasize two important points. First, we keep the model as simple as possible to make our main argument—information production in bad times leads to increased forecast accuracy of PDs—as clearly as possible. Hence, the model will have little to say about ancillary issues we analyze in the paper, such as collateral and real estate prices, though the model can easily be adapted to incorporate these features. The main goal of the model is not to offer a new explanation for why banks produce more information in bad times but rather to show how information production affects the discrimination ability of PDs. Nonetheless, the way in which we generate higher information production incentives in bad times—lower expected cash flows for low-quality borrowers—seems fairly reasonable. Second, there are some ingredients in the model that we include to fit it into our empirical framework that would not be necessary in a model purely meant to generate intuitions (e.g., different classes of borrowers and three types of borrowers rather than two).

### *B.1. Setup*

There is a single borrower seeking funds from a bank at $t = 0$ for a project that yields a random payoff at $t = 1$. The borrower and bank are risk-neutral, and there is no discounting. At $t = 0$, there is a publicly observable aggregate state $\omega \in \{H, L\}$ (High or Low) which represents current economic conditions. The borrower belongs to a publicly observable class $\alpha \in [\underline{\alpha_\omega}, \bar{\alpha_\omega}]$ where $\underline{\alpha_\omega} > 0$ and $\bar{\alpha_\omega} < 1$, which is distributed according to the density function $f_\omega(\alpha)$, where the distribution potentially depends on the aggregate state. The borrower's class can be thought of as the borrower's public credit score based on observable characteristics.[29] Within each class of borrowers, there are three types of borrowers $\theta \in \{G, M, B\}$ (Good, Medium or Bad) where $\theta$ is initially unknown to all, and each type is equally likely.[30]

The borrower has an investment opportunity that requires an initial investment of $I$, which we normalize to 1, at $t = 0$ and yields a cash flow at $t = 1$ of $R^\theta_\omega(\alpha) > 0$ if it succeeds and 0 if it fails, where the cash flow in the case of success can depend on the state, class of the borrower, and its unobservable type. We assume that the cash flows in the case of success are increasing in the quality of the borrower, i.e., $R^G_\omega(\alpha) \geq R^M_\omega(\alpha) \geq R^B_\omega(\alpha)$, for $\omega \in \{H, L\}$ and $\alpha \in [\underline{\alpha}, \bar{\alpha}]$. The average probability of success within class $\alpha$ of borrowers is $\alpha$. If the borrower

[29] As will become clear below, the purpose of having a borrower class is such that we obtain multiple probabilities of default to derive a receiver operating characteristics curve based on the model. We assume the class is continuously distributed for analytic tractability; however, we have also analyzed a discrete version which is available upon request.

[30] As shown below, by having three borrower types, the bank produces information and screens out the bad borrower, but still has improved information among those that it ultimately lends to.

is good ($\theta = G$), the probability of success is $\alpha + \epsilon$; if the borrower is medium ($\theta = M$), the probability of success is $\alpha$, and if the borrower is bad ($\theta = B$), the probability of success is $\alpha - \epsilon$.

The borrower offers the bank a loan contract that raises 1 at $t = 0$ to finance the investment and promises to repay $F$, which is endogenously determined, at $t = 1$. The firm has limited liability, so if the project's realized cash flow is lower than $F$, the firm defaults, and the bank collects the realized cash flow. Although the borrower's type $\theta$ is initially unknown, the bank can pay a cost $c > 0$ to learn $\theta$. The potential value of information for the bank is to screen out lower-quality borrowers.[31]

### *B.2. Information Production*

There are several ways we could model higher information production incentives in the low state. However, one that is natural and convenient for modeling purposes is simply to assume that the project is always ex-ante NPV positive in the high state and information production is unprofitable for the bank, while the project is ex-ante NPV negative in the low state absent information production. Formally,

**Assumption 1.** *The project is ex-ante NPV positive in the high state and NPV negative in the low state regardless of its class. It is unprofitable for the bank to produce information in the high state.*

1. $\frac{1}{3}\left[(\alpha+\epsilon)R_H^G(\alpha) + \alpha R_H^M(\alpha) + (\alpha-\epsilon)R_H^B(\alpha)\right] \geq 1 \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$,
2. $\frac{1}{3}\left[(\alpha+\epsilon)R_L^G(\alpha) + \alpha R_L^M(\alpha) + (\alpha-\epsilon)R_L^B(\alpha)\right] < 1 \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$,
3. $\frac{\epsilon}{3\alpha} < c \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$.

We also make the following assumptions:

**Assumption 2.**

1. $\frac{1}{3}\left((\alpha+\epsilon)R_L^G(\alpha) + \alpha R_L^M(\alpha)\right) - \frac{2}{3} \geq c \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$,
2. $c \geq \frac{1}{3}\left(\frac{\alpha+\epsilon}{\alpha} - \alpha R_L^M(\alpha)\right) \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$,
3. $R_L^M(\alpha) \geq \frac{1}{\alpha} \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$,
4. $R_H^B(\alpha) \geq \frac{1}{\alpha} \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$.

The purpose of this second set of assumptions will become clearer in the proof below. However, Assumption 2.1 ensures that producing information and lending to the good and medium borrower is NPV positive in the low state, Assumption 2.2 ensures that if the bank

[31] In a more complicated bargaining game, there could be value from being able to adjust interest rates. This force is not present here because the firm simply makes a take-it-or-leave-it offer.

produces information in the low state, it lends to both the good type and the medium type, not just the good type, and Assumption 2.3 ensures that this is feasible.[32] Finally, Assumption 2.4 is not strictly necessary but simplifies the expressions by guaranteeing the borrower never defaults in the high state when the project succeeds.[33]

We next characterize the equilibrium.

**Proposition 1.** *The bank does not produce information in the high state and lends to the borrower regardless of the type and class of the borrower. The bank always produces information in the low state and lends to the good and medium borrowers regardless of their class. Expected lending volume is lower in the low state than in the high state.*

*Proof.* Throughout the proof, we can consider a fixed $\alpha$; however, the proof applies to all $\alpha \in [\underline{\alpha}, \bar{\alpha}]$. First, consider the case in which the state is high. Assuming the bank does not produce information, the participation constraint $\alpha F \geq 1$ must hold. Because the firm has all of the bargaining power, it can offer the bank a zero profits contract with face-value $F_H = \frac{1}{\alpha}$. We can then check whether the bank would have incentives to produce information and only lend to the medium and good borrower if offered this contract:

$$\frac{2}{3}\left(\left(\alpha + \frac{\epsilon}{2}\right) F_H\right) - c > 0, \quad \implies \frac{\epsilon}{3\alpha} > c. \tag{2}$$

However, Assumption 1.3 implies that (2) is violated. Hence, the bank would not produce information and only lend to the medium and good borrower if offered the zero profits contract. We can also rule out the possibility that the bank produces information and only lends to the good borrower:

$$\frac{1}{3}\left(\left(\alpha + \epsilon\right) F_H - 1\right) > c, \quad \implies \frac{\epsilon}{3\alpha} > c.$$

Hence, the firm offers the bank a contract with face value $F_H$ and the bank does not produce information and lends to the firm regardless of its type. Notice also that Assumption 2.4 implies that this contract is always feasible because the firm will never default if the project succeeds.

Now consider the low state. Because of Assumption 1.2, the bank will not lend without producing information. There are two potential alternatives. First, the bank could produce information and only lend to the good type. If this were the case, the firm would offer the following zero-profits face-value of debt $F'_L = \frac{1+3c}{\alpha+\epsilon}$, and the firm's profits would be:

$$\frac{\alpha + \epsilon}{3}\left(R_L^G(\alpha) - \frac{1+3c}{\alpha+\epsilon}\right). \tag{3}$$

[32] While this assumption is not strictly necessary from an economic intuition perspective, we need the bank to lend to at least two types of borrowers within each class to generate an improvement in discrimination from information production.

[33] Note that this assumption directly implies the project is NPV positive in the high state (i.e., Assumption 1.1).

Notice that from Assumption 2.3 this contract is feasible since $R_L^G(\alpha) > R_L^M(\alpha) \quad \forall \alpha \in [\underline{\alpha}, \bar{\alpha}]$.

Alternatively, the firm could offer a contract that induces the bank to lend to both the medium and good borrower. This contract cannot have zero profits, as this would mean the bank earns negative profits from the medium type and would not be willing to lend. Here, it must be incentive-compatible for the bank to lend to the medium type, i.e, $\alpha F \geq 1$. Again, since the firm has all of the bargaining power, we can consider a face-value of debt such that this constraint binds, i.e., $F_L'' = \frac{1}{\alpha}$. Notice again that Assumption 2.3 ensures that this loan contract is feasible. Under this contract, the firm's profits would be:

$$\frac{1}{3}\left((\alpha+\epsilon)(R_L^G(\alpha) - F_L'') + \alpha(R_L^M(\alpha) - F_L'')\right). \tag{4}$$

To check which contract the firm offers the bank, we need to compare (4) to (3). Subtracting (3) from (4) we have:

$$c - \frac{1}{3}\left(\frac{\alpha+\epsilon}{\alpha} - \alpha R_L^M(\alpha)\right),$$

which is positive from Assumption 2.2. Hence, the firm earns higher profits from offering a loan with face value $F_L''$ than $F_L'$, and so the bank produces information and lends to both the good and medium borrowers. Finally, based on these lending decisions, the expected lending volume is 1 in the high state and $\frac{2}{3}$ in the low state.

□

### *B.3. Area Under the Curve (AUC) Derivation*

In this section, we derive the AUC from the model and analyze its properties. First, it will also be useful to convert the class of borrower $\alpha$ into a corresponding probability of default: $p \equiv 1-\alpha$. The corresponding density function is then $g_\omega(p) \equiv f_\omega(1-\alpha)$ with support $[\underline{p_\omega}, \bar{p_\omega}]$ where $\underline{p_\omega} \equiv 1 - \bar{\alpha_\omega}$ and $\bar{p_\omega} \equiv 1 - \underline{\alpha_\omega}$.

The bank's perceived probability of default is:

$$\hat{PD} = \begin{cases} p & \text{if } \omega = H, \\ p-\epsilon & \text{if } \omega = L, \theta = G, \\ p & \text{if } \omega = L, \theta = M, \\ p+\epsilon & \text{if } \omega = L, \theta = B. \end{cases}$$

Notice that because the bank produces information in the low state, its perceived probability of default is more precise than in the high state.

Henceforth, we will assume that the bank interacts with an infinite number of borrowers independently drawn from both the class distribution and the ex-ante unobservable type distribution within each class. The receiver operating characteristic curve (ROC) plots the true pos-

itive rate against the false positive rate. Specifically, for a given threshold $t$, the ROC considers any probability of default larger than $t$, i.e., $\hat{PD} > t$ a predicted positive and any probability of default less than $t$, i.e., $\hat{PD} \leq t$, a predicted negative. A predicted positive is a true positive if the borrower actually defaults and a false positive if it does not. A predicted negative is a true negative if the borrower does not default and a false negative if it does default. The true positive rate is equal to the ratio of true positives to total positives and in the high state is:

$$\begin{aligned} TPR_H(t) &= \frac{\int_t^{\bar{p_H}} \left(\frac{1}{3}(p-\epsilon) + \frac{1}{3}p + \frac{1}{3}(p+\epsilon)\right) f_H(p)dp}{\int_{\underline{p_H}}^{\bar{p_H}} \left(\frac{1}{3}(p-\epsilon) + \frac{1}{3}p + \frac{1}{3}(p+\epsilon)\right) f_H(p)dp} \\ &= \frac{\int_t^{\bar{p_H}} p f_H(p)dp}{\int_{\underline{p_H}}^{\bar{p_H}} p f_H(p)dp}. \end{aligned}$$

The false positive rate is equal to the ratio of false positives over actual negatives, which in the high state is:

$$FPR_H(t) = \frac{\int_t^{\bar{p_H}} (1-p) f_H(p)dp}{\int_{\underline{p_H}}^{\bar{p_H}} (1-p) f_H(p)dp}.$$

Because the probability of a true positive or false positive only depends on the average realized default rate for a given PD, the $\epsilon$ term drops out in both expressions, given that it has a mean of zero.

The receiver operating characteristic curve plots the true positive rate against the false positive rate, and the AUC is the area under this curve. Given the continuous distribution, we can write the AUC in the high state as:

$$AUC_H = \int_{\underline{p_H}}^{\bar{p_H}} TPR_H(t) |FPR'_H(t)| dt, \tag{5}$$

where $|FPR'_H(t)|$ denotes the absolute value of the derivative of $FPR_H$ with respect to $t$.

In the low state, the true positive rate is as follows:

$$TPR_L(t) = \begin{cases} \frac{\frac{1}{2}\int_{t+\epsilon}^{\bar{p_L}} (p-\epsilon) f_L(p)dp + \frac{1}{2}\int_{\underline{p_L}}^{\bar{p_L}} p f_L(p)dp}{\int_{\underline{p_L}}^{\bar{p_L}} \left(p - \frac{\epsilon}{2}\right) f_L(p)dp} & \text{if } t < \underline{p_L}, \\ \frac{\frac{1}{2}\int_{t+\epsilon}^{\bar{p_L}} (p-\epsilon) f_L(p)dp + \frac{1}{2}\int_{t}^{\bar{p_L}} p f_L(p)dp}{\int_{\underline{p_L}}^{\bar{p_L}} \left(p - \frac{\epsilon}{2}\right) f_L(p)dp} & \text{if } t \in [\underline{p_L}, \bar{p_L} - \epsilon], \\ \frac{\frac{1}{2}\int_{t}^{\bar{p_L}} p f_L(p)dp}{\int_{\underline{p_L}}^{\bar{p_L}} \left(p - \frac{\epsilon}{2}\right) f_L(p)dp}, & \text{if } t > \bar{p_L} - \epsilon, \end{cases}$$

and the false positive rate is as follows:

$$FPR_L(t) = \begin{cases} \frac{\frac{1}{2}\int_{t+\epsilon}^{\bar{p_L}}(1-p+\epsilon)f_L(p)dp + \frac{1}{2}\int_{\underline{p_L}}^{\bar{p_L}}(1-p)f_L(p)dp}{\int_{\underline{p_L}}^{\bar{p_L}}\left(1-p+\frac{\epsilon}{2}\right)f_L(p)dp} & \text{if } t < \underline{p_L}, \\ \frac{\frac{1}{2}\int_{t+\epsilon}^{\bar{p_L}}(1-p+\epsilon)f_L(p)dp + \frac{1}{2}\int_{t}^{\bar{p_L}}(1-p)f_L(p)dp}{\int_{\underline{p_L}}^{\bar{p_L}}\left(1-p+\frac{\epsilon}{2}\right)f_L(p)dp} & \text{if } t \in [\underline{p_L}, \bar{p_L} - \epsilon], \\ \frac{\frac{1}{2}\int_{t}^{\bar{p_L}}(1-p)f_L(p)dp}{\int_{\underline{p_L}}^{\bar{p_L}}\left(1-p+\frac{\epsilon}{2}\right)f_L(p)dp} & \text{if } t > \bar{p_L} - \epsilon. \end{cases}$$

In contrast to the high state, the $\epsilon$ term is present in the expressions for the TPR and FPR because the bank is only lending to good and medium borrowers. Finally, the area under the curve in the low state is then:

$$AUC_L = \int_{\underline{p_L}-\epsilon}^{\bar{p_L}} TPR_L(t)|FPR_L'(t)|dt.$$

### *B.4. Analysis of AUC*

In general, the expressions for AUC are not easily expressed analytically. However, we can solve for the AUC in closed-form for both the low and high states, assuming a uniform distribution of the class of borrowers. Specifically, suppose that $p \sim U[\underline{p_\omega}, \bar{p_\omega}]$, where $\underline{p_\omega} - \epsilon \geq 0$, $\bar{p_\omega} + \epsilon \leq 1$ and $\bar{p_\omega} - \epsilon \geq \underline{p_\omega}$ for $\omega \in \{H, L\}$.[34]

First, we analyze the case in which the distribution of borrower class is the same in the high and low state, i.e., $f_H(\alpha) = f_L(\alpha)$. Under the uniform distribution, this amounts to $\bar{p_H} = \bar{p_L} = \bar{p}$ and $\underline{p_H} = \underline{p_L} = \underline{p}$. In this case, so long as the average probability of default is below $50\%$, the AUC is always higher in the low state. Formally,

**Proposition 2.** *Assume the class of borrowers is the same in the high and low state, i.e., $f_H(\alpha) = f_L(\alpha)$ and follows a uniform distribution as described above. If the average probability of default is less than 50%, the AUC is always larger in the low state.*

*Proof.* If we subtract $AUC_H$ from $AUC_L$ and remove the $\epsilon$ term we have

$$\frac{4(\overline{p}-\underline{p})^3(1-\underline{p}-\overline{p}) + (\underline{p}-\overline{p})\left(\underline{p}^2 - (6-\overline{p})\overline{p} + 2\underline{p}(5\overline{p}-3)\right)\epsilon - (2-\underline{p}-\overline{p})(\underline{p}+\overline{p})\epsilon^2}{6(\underline{p}-\bar{p})^2(2-\underline{p}-\bar{p})(\underline{p}+\bar{p})(2-\underline{p}-\bar{p}+\epsilon)(\underline{p}+\bar{p}-\epsilon)}$$

Since the denominator is clearly positive, it suffices to show the numerator is positive. The first term of the numerator is positive because $\overline{p} - \underline{p} > 0$ and $\underline{p} + \overline{p} < 1$. Thus, it is sufficient to show that the remaining terms are positive. Hence, we need:

$$-(\overline{p}-\underline{p})(\underline{p}^2 + \overline{p}^2 - 6(\underline{p}+\overline{p}) + 10\underline{p}\overline{p}) > (2-(\underline{p}+\overline{p}))(\underline{p}+\overline{p})\epsilon$$

[34] Note this is equivalent to $\alpha \sim U[1-\bar{p_\omega}, 1-\underline{p_\omega}]$.

Since $\overline{p} - \epsilon > \underline{p}$, then $\overline{p} - \underline{p} > \epsilon$ and since $(2 - (\underline{p} + \overline{p}))(\underline{p} + \overline{p}) > 0$, it suffices to show

$$\begin{aligned} &- (\overline{p} - \underline{p})(\underline{p}^2 + \overline{p}^2 - 6(\underline{p} + \overline{p}) + 10\underline{p}\overline{p}) > (2 - (\underline{p} + \overline{p}))(\underline{p} + \overline{p})(\overline{p} - \underline{p}) \\ &\iff (8\underline{p}\overline{p} - 4(\underline{p} + \overline{p}))(\overline{p} - \underline{p}) < 0 \end{aligned}$$

The first term is negative because both $\overline{p}$ and $\underline{p}$ are less than one and the second term is positive because $\overline{p} - \underline{p} > 0$. Hence, the entire term is negative. □

Proposition 2 says that so long as the average perceived PDs are below 50%, the AUC is always higher in the low state. Why does an average of 50% matter? The reason is that while the improved discrimination across borrowers from information production raises the AUC, there is also a second effect due to the change in the average probability of default. To best understand this, notice that even in the absence of information production, the AUC in the high state depends on the average risk of borrowers:

$$AUC_H = \frac{1}{6}\left(3 + \frac{2(1 - \underline{p})}{2 - \underline{p} - \bar{p}} - \frac{2\underline{p}}{\underline{p} + \bar{p}}\right). \tag{6}$$

Suppose we fix the range between the upper and lower bound of the uniform distribution, i.e., $\delta \equiv \bar{p} - \underline{p}$, and substitute $\underline{p}$ with $\bar{p} - \delta$ into the expression for $AUC_H$, then we have:

$$\frac{1}{6}\left(3 - \frac{2\bar{p}}{\delta - 2\bar{p}} - \frac{2(1 - \bar{p})}{2 - 2\bar{p} + \delta}\right). \tag{7}$$

Differentiating (7) with respect to $\bar{p}$ and substituting back in $\underline{p}$ we have:

$$-\frac{4(\bar{p} - \underline{p})(1 - \bar{p} - \underline{p})}{3(2 - \bar{p} - \underline{p})^2(\bar{p} + \underline{p})^2},$$

which is positive if $\bar{p} + \underline{p} > 1$ and negative otherwise. Hence, shifting the distribution of PDs upwards while fixing the distance between the upper and lower bound of the distribution increases the AUC if the average probability of default is above one-half. Because the bank produces information and screens out bad borrowers in the low state, the distribution of PDs shifts downwards, causing a reduction in the average PD. When $\bar{p} + \underline{p} > 1$, it is possible for information production to decrease the AUC. The intuition for this is as follows: when the average PD is close to 50%, uncertainty is highest.[35] Probabilities close to 0 and 1 are more certain and hence provide clearer separation. For example, a difference in PDs of 1% versus 2% is much more discriminative than a difference of 49% versus 50%.

While this effect is theoretically possible, it is extremely unlikely to be driving our results for several reasons. First, this is a necessary, but not sufficient, condition for the AUC to decrease from information production. There are still parameters (particularly when $\epsilon$ is large

[35] As discussed below, this is also true, although to a much larger extent, when assessing forecast errors.

enough) such that even when the average PD is above 50%, the AUC is higher in the low state. Second, PDs in the data average less than 2% unconditionally, and hence, are very far from 50%. Lastly, in the data, we do not actually observe lower PDs in bad times. Rather, they are quite similar and, if anything, slightly larger (1.66% versus 1.61%), which is likely due to the distribution of underlying borrowers also changing in bad times. Indeed, we next show that if the distribution of potential borrowers has higher PDs in the low state, but the average PD among borrowers who actually receive loans is the same in the low and high state, which is approximately what we see in the data, the AUC is always higher.

**Proposition 3.** *Suppose that $p \sim U[\underline{p}, \bar{p}]$ in the low state and $p \sim U[\underline{p} - \frac{\epsilon}{2}, \bar{p} - \frac{\epsilon}{2}]$ in the high state. While the average PD of potential borrowers is higher in the low state, the average PD of firms that actually receive financing is the same, and the AUC is always higher in the low state.*

*Proof.* First, note that the bank lends to all borrowers in the high state from Proposition 1. Hence, the mean of both the true PD and perceived PD is $\frac{1}{2}\left(\underline{p} + \bar{p} - \epsilon\right)$, which is the same as in the low state because the borrower only lends to the medium and good borrowers. If we subtract $AUC_H$ from $AUC_L$ we have:

$$\frac{\epsilon^2(3\bar{p} - 3\underline{p} - \epsilon)}{6(\bar{p} - \underline{p})^2(2 - \bar{p} - \underline{p} + \epsilon)(\bar{p} + \underline{p} - \epsilon)},$$

which is strictly positive. Hence, the AUC is always higher in the low state than the high state when the average PD of granted loans is the same in both states. □

This result suggests that it is the improved discrimination in bad times that is driving the increase in the AUC in our results, not the fact that information production changes the average quality of borrowers.[36] Finally, since PDs are slightly higher in bad times in our data, if anything, this should go against our main result.

A higher variance of underlying PDs generally raises the AUC, even in the absence of differences in information production. To see this, suppose we fix the average PD in the high state for the uniform case, i.e., $\sigma \equiv \bar{p} + \underline{p}$ and replace $\underline{p}$ with $\sigma - \bar{p}$ into (6), then we have: $\frac{4\bar{p}+4\sigma-3\sigma^2}{12\sigma-6\sigma^2}$, which is increasing in $\bar{p}$. This implies that a higher variance of PDs increases the AUC in the uniform case. In our data, the standard deviation of PD for new loans is slightly higher in periods of high unemployment (2.88pp versus 2.58pp). However, it is highly unlikely that this effect can quantitatively explain our results. For instance, if we assume PDs follow a uniform distribution, we can solve for $\bar{p}$ and $\underline{p}$ given the mean and variance that we observe in the data in low and high unemployment periods. We can then plug these values into the AUC in the high state without information production (5). Doing so gives us an AUC of 0.542 and

[36] Relatedly, under the uniform case it can easily be shown that if the class distribution is the same in both states, but the bank exogenously receives information in the low state and lends to all borrowers, the AUC is always higher in the low state.

0.551 in the high and low states (a difference of 0.009). Hence, it is unlikely that this small difference in variance mechanically explains all of our results. For example, in Figure 4, we find a difference in AUCs of over four times as large (0.043) among newly issued loans.

Although we cannot solve the AUC analytically for other distributions, we can do a similar exercise as above and simulate the data assuming the class of borrowers follows a lognormal distribution[37], matching the mean and variance of PDs to what we observe in the data in periods of high and low unemployment. Figure B.1 displays the distribution of the difference between the low and high state AUC over 100 simulations. The average difference is 0.011, and the maximum difference is 0.042, which is less than the 0.043 difference we find in the data.[38] Hence, mechanical differences in the mean and variance of PD in periods of high unemployment are unlikely to explain our main results. In Online Appendix Section C.4, we also show how mechanical differences in the distribution of PD are unlikely to explain our cross-sectional results comparing information quality across different types of loans.

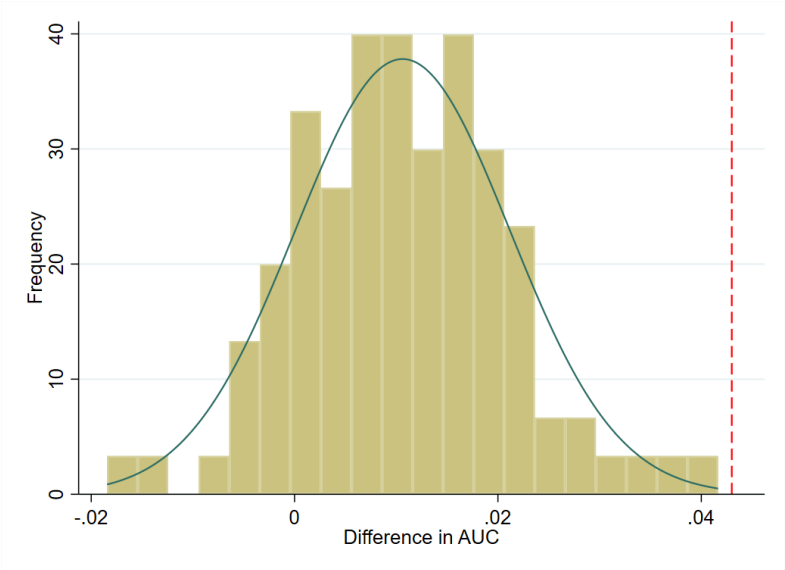


**Figure B.1:** Difference in AUCs for simulated lognormally distributed PDs

This figure plots the simulations of the difference in AUCs across periods of high and low unemployment, assuming no information production and PDs are lognormally distributed. The number of observations in each simulation is equal to that in our main sample (104,111). The parameters of each distribution are estimated to match the means and variances we observe in the data. The sample average of the difference in AUCs is 0.011. The dashed red line at 0.043 marks the difference in AUC between high and low unemployment periods that we estimate in the data.

[37] In Figure 1, log(PD) approximates a normal distribution.

[38] It is also comforting that the average difference under a lognormal distribution (0.011) is fairly similar to the difference when the class of borrowers follows a uniform distribution (0.009).