

Assessing Human Information Processing in Lending Decisions: A Machine Learning Approach

Miao Liu¹

The University of Chicago, Booth School of Business

December 2019

Abstract

Effective financial reporting requires efficient information processing. This paper studies factors that determine efficient information processing. I exploit a unique small business lending setting where the entire codified demographic and accounting information set that loan officers use is observable (to the researcher). I decompose the loan officers' decisions into a part driven by codified hard information and a part driven by uncoded soft information. I show that a machine learning model substantially outperforms loan officers in processing hard information. Using the machine learning model as a benchmark, I find that limited attention and overreaction to salient accounting information largely explain the loan officers' weakness in processing hard information. However, the loan officers acquire more soft information after seeing salient accounting information, suggesting salience has a dual role: it creates bias in hard information processing, but facilitates attention allocation in new information acquisition.

¹ mliu7@chicagobooth.edu. I am grateful to my dissertation committee members Philip Berger, Christian Leuz (chair), Sendhil Mullainathan, and Valeri Nikolaev for their guidance and support. I also appreciate helpful comments from Ray Ball, Pietro Bonaldi, Jonathan Bonham, Matthias Breuer, Robert Bushman, Jung Ho Choi, Anna Costello, Hans Christensen, John Gallemore, Pingyang Gao, Joao Granja, Anya Kleymenova, Yun Lee, Rebecca Lester, Ye Li, Jinzhi Lu, Yao Lu, Mark Maffett, Charles McClure, Danqing Mei, Michael Minnis, Maximilian Muhn, Sanjog Misra, Thomas Rauter, Ethan Rouen, Haresh Sapra, Douglas Skinner, Gurpal Sran, Andrew Sutherland, Rimmy Tomy, James Traina, Felix Vetter, Xian Xu, Anastasia Zakolyukina, workshop participants at Chicago Booth, CMU emerging scholar session, and LBS Trans-Atlantic Doctoral Conference. I am indebted to the company executives who preferred to remain anonymous for helpful discussions and access to data. I do not have a financial interest in the outcomes of this research. I gratefully acknowledge financial support from the University of Chicago Booth School of Business. Any errors are my own.

1. Introduction

Research has shown that financial reporting facilitates decision-making and affects a wide range of financial and real outcomes (Beyer et al. 2010; Dechow et al. 2010; Leuz and Wysocki 2016; Roychowdhury et al. 2019). The extent to which the objective of financial reporting is met, however, depends on how decision-makers process the information. One strand of research in economics and psychology highlights human weaknesses in processing information, due to cognitive constraints (Blankespoor et al. 2019b) and behavioral biases (Kahneman 2011). Another strand in accounting and finance, however, emphasizes that humans have strengths in discovering new information (Goldstein and Yang 2017), especially soft information (Liberti and Petersen 2018). In this paper, I investigate the trade-off between human strengths and weaknesses in the context of credit information processed by loan officers in their lending decisions, using a machine learning model-based decision rule as a benchmark.

I employ detailed loan-level data from a large Chinese small business lender. Loan officers observe hard information on borrowers' demographics and accounting reports, exercise discretion to acquire additional soft information by making phone calls, and finally make lending decisions. While the loan officers might not process hard information efficiently, as predicted by theories of cognitive constraints and behavioral biases, they should play a crucial role in collecting soft information by interacting directly with borrowers, as demonstrated in other similar settings (e.g., Petersen and Rajan 1994, 1995). In addition, the officers determine when and how to acquire soft information after observing hard information, suggesting the two types of information may interact. I use this setting to study which factors impede efficient hard information processing and how these factors further affect soft information acquisition.

Three challenges emerge when assessing information processing efficiency. First, it is difficult to study how information users process information if the underlying information is unobservable, as is usually the case. Second, which information to observe is often a choice. Consequently, different people might appear to process information differently not because they have varying abilities but because they have different

information sets. Third, evaluating errors requires a benchmark. Which is the “correct” way to process a given piece of information? While a benchmark can often be established in a laboratory (such as the correct answer to a test), one is usually missing in the real world.

I combine my unique setting with a novel research design to overcome these challenges. Two key features of the setting help address the first two challenges. First, the data allows observation of the loan officers’ entire hard information set about a borrower, sidestepping the unobservability problem. Second, borrowers are randomly assigned to loan officers, meaning each officer has the same pool of borrowers on average. As a result, any systematic difference in lending decisions across officers stems from their differing abilities in processing information and, not because they are endogenously matched with different types of borrowers. This helps me to overcome the second challenge that the information set is usually an endogenous choice.

To address the third challenge, I develop an approach using a machine learning model as a benchmark to assess human decisions. I split my data randomly into a training sample and a hold-out sample. I train a machine learning model on the training sample to predict a borrower’s repayment and design a feasible lending decision rule by reallocating larger loans to borrowers who are more likely to repay, as predicted by the model. Next, using the hold-out sample, I show that this model-based decision rule can boost the lender’s profit by at least 38%, making it a valid benchmark for examining loan officers’ limitations in processing hard information.

To make the machine learning model a benchmark, I must address the fact that, while it only uses codified hard information as an input, loan officers can acquire additional soft information, including photos and phone calls. I decompose officers’ decisions into a part driven by codified hard information and a part driven by uncoded soft information. Specifically, I fit a separate machine learning model for each loan officer, this time to predict the officer’s lending decisions based on hard information. Unlike the first model, which predicts borrower repayment, the purpose of this one is to mimic how each loan officer

processes hard information. Soft information is then captured by the residual, as it represents variation in officers' decisions that cannot be explained by hard information. To validate the residual as a measure of soft information, I show that it strongly predicts loan outcomes.

These results suggest that while loan officers have weaknesses in analyzing hard information and are outperformed by machine learning, they have strengths in acquiring soft information. I next test which factors explain loan officers' underperformance in processing hard information and whether these factors also affect their ability to acquire soft information. I rely on two streams of theory to guide my search.

The first emphasizes bounded rationality (Blankespoor et al. 2019b). Agents in these models confront costs in processing information and allocate attention within their cognitive constraints. This line of theory predicts that loan officers can process only a subset of all useful variables. Regressing the fitted values of the two machine learning models on borrower characteristics using OLS, I find that, while the first model, which predicts borrower risk, identifies 147 variables with strong predictive power about repayment, the second one, which mimics how each loan officer processes hard information, suggests that officers only use between 25 to 56 variables in their decisions. Moreover, these 25 to 56 variables explain between 83% to 92% of the variation in officers' decisions in these linear regressions, suggesting officers process hard information in a linear fashion. In contrast, the 147 variables explain only 66% of the variation in the first model's prediction about borrower repayment, indicating machine learning's ability to incorporate nonlinear signals in the data that are systematically ignored by loan officers. These results are consistent with information processing being costly.

The second stream of theory emphasizes that, even in the set of variables used for decision-making, people make systematic probabilistic errors, often because they employ representativeness heuristics (Kahneman 2011). Bordalo et al. (2016) formalize this concept in economics as probability judgments based on the most distinctive differences between groups and show that representativeness can exaggerate perceived differences. In my setting, this line of theory predicts that loan officers will approve loan sizes

too small for borrower groups with distinct characteristics representative of high risk because such characteristics catch officers' eye and exaggerate their perception of the risk. One such distinct characteristic of risky borrowers is negative salient information. Indeed, among borrowers who default, 28.1% have (negative) salient characteristics, defined as large negative realizations in accounting variables. The proportion is only 15.8% among borrowers who do not default. Using the machine learning model as a benchmark, I find that loan officers overreact to salient information and approve loan sizes too small to borrowers with salient information, in line with such information being representative of a risky borrower.

Having established that both bounded rationality and representativeness help explain loan officers' underperformance in processing hard information, I next test how they affect soft information acquisition. Consider representativeness heuristics. Although theories do not directly model information acquisition, it is plausible that overreaction to salient hard information due to representativeness bias might impede soft information acquisition. Intuitively, interacting with borrowers with a biased perception can undermine officers' ability to extract unbiased soft information signals. Perhaps surprisingly, however, I find that officers acquire more soft information after seeing salient hard information. Why would salience impede hard information processing but facilitate soft information acquisition? Theories of representativeness are silent on this question, but I next examine how bounded rationality helps explain this puzzling result.

Theories of bounded rationality predict that acquiring new information is costly and thus that loan officers must allocate their effort to such activity. I build a simple model to show that, faced with information acquisition costs, salience can guide this allocation. In this model, a loan officer tries to infer a borrower's type from hard accounting signal but faces uncertainty about the precision of the signal (for example, does a jump in cash flow reflect business fundamentals or noise?).² The officer can incur a cost to acquire additional soft information on the precision of the accounting signal (for example, asking the borrower to

² This setup is built on a stream of accounting theory considering the impact of earnings disclosures when investors face uncertainty over the variance of cash flows (Beyer 2009; Heinle and Smith 2017) or the precision of the earnings disclosure (Hughes and Pae 2004; Kirschenheiter and Melumad 2002; Subramanyam 1996).

explain the jump in cash flow). I show that it is more efficient to incur the cost to acquire soft information when the accounting signal has a larger realization (i.e., more salient). To see this, assume the signal is cash flow. The officer would have more incentive to call a borrower with a large jump in cash flow. This is because learning whether the jump is a precise or noisy signal tells the officer a lot about a borrower's type, while learning whether a report of no jump in cash flow is precise reveals much less about a second borrower. Therefore salience serves a dual role: it distorts loan officers' belief when processing hard information but facilitates their attention allocation in their acquisition of soft information.

My findings suggest replacing loan officers with the machine can increase profits substantially. However, doing so would also sacrifice all the valuable soft information. Is there a way to combine the strengths of humans with machines, making them complements? I develop such a procedure in the final section. The key step is to train the machine learning algorithm on a subsample of data generated by the loan officers most skilled at acquiring soft information. The machine then puts more weight on hard variables that have higher correlations with useful soft signals and consequently captures a portion of the soft signals. I show that this procedure results in better decisions than humans or machines working alone, suggesting a means of combining humans and machines in settings where both hard and soft information matter.

This paper contributes to three strands of literature. First, it provides new insights into the literature on investors' information processing. I differentiate between bounded rationality theory and representativeness bias theory. Due to difficulty in observing the decision-maker's information set, empirical studies under the bounded rationality framework have focused on various market outcomes as indirect evidence of information processing constraints. These outcomes include investors' trading (Blankespoor et al. 2019a), consumer and manager's use of tax rates (Chetty et al. 2009; Graham et al. 2017), investors' perceptions of different reporting formats (Hirshleifer et al. 2003), stock price responsiveness to disclosure (Hirshleifer et al. 2009, 2011; Dellavigna et al. 2009; Lawrence et al. 2018), and firms' disclosure choices in response to shocks to investors' information processing costs (Dehaan et

al. 2015; Blankespoor 2019; Abramova et al. 2019). Empirical studies under the representativeness bias framework provide sparse evidence outside of experimental settings, with a few recent exceptions (e.g., Mullainathan and Obermeyer 2019).

My paper adds to both categories of research in three ways. First, by combining unique data with a novel approach, I provide direct evidence of bounded rationality and representativeness in human hard information processing in a non-experimental setting. Second, using machine learning as a benchmark, my results emphasize a trade-off between humans' weakness in hard information processing and their strength in acquiring soft information, two aspects of information processing usually studied separately. Finally, my finding on the dual role of salience bridges these two lines of research. While salience impedes hard information processing, as predicted by representativeness, it facilitates the allocation of attention, a costly resource as emphasized by bounded rationality. This new result highlights the value of combining both kinds of models to understand human decision-making.

Second, my paper contributes to the literature on soft and hard information and their differing roles in contracts (Liberti and Petersen 2018). Since soft information is hard to quantify, researchers have relied on indirect measures at bank-branch or loan-officer level, such as the geographical distance between lenders and borrowers (e.g., Petersen and Rajan, 1994, 2002; Granja et al. 2019), cultural distance between loan officers and borrowers (Fisman et al. 2017), or loan officer fixed effects (Bushman et al. 2019). A notable exception is the work of Campbell et al. (2019), who use keywords in loan officers' internal reports to construct soft information. Contrary to these studies, I design an approach to identify soft information embedded in officers' individual lending decisions at the loan level. This approach allows me to investigate the interaction between hard and soft information and examine which factors impede or facilitate soft information acquisition both across and within loan officers.

Finally, my paper relates to the nascent literature studying how machine learning tools can improve decision-making (e.g., Hoffman et al. 2018; Kleinberg et al. 2018; Einav et al. 2018; Bartlett et al. 2019;

Erel et al. 2019). I add to this literature in two ways. First, while most researchers focus on demonstrating human underperformance, I go a step further by searching for factors explaining underperformance.³ Second, I document that humans have strengths, relative to machines, in acquiring soft information. A closely related paper is by Costello et al. (2019). They find that, in a randomized experiment, allowing humans to incorporate private information into a machine-based credit score improves loan outcomes. Using a different approach to recover private soft information, my paper also shows that soft information has value, consistent with theirs. My paper differs in that I focus on examining how cognitive constraints and behavioral factors affect hard information processing and soft information acquisition.

The rest of the paper proceeds as follows. Section 2 describes the institutional setting and data. Section 3 connects the setting to conceptual underpinnings and then lays out the research design. Section 4 introduces the machine learning procedure used in the research design. Section 5 reports the main findings and interprets them in a theoretical framework. Section 6 builds on the research design and develops a technique to combine humans and machines. Section 7 concludes.

2. Setting

My data comes from a large Chinese lender with sales offices in 23 major cities spread across the country. The lender offers unsecured short-term cash loans to small businesses as well as personal loans. My sample contains all small business loans and runs through the lender's entire operating history from 2011 to 2015. As I will describe in detail in this section, the setting offers a unique opportunity to investigate human information processing because 1) I observe the entire set of codified information that loan officers observe, and 2) random assignment ensures that all loan officers with sufficient amount of observations face the same pool of borrowers. I restrict the sample to 28 officers who have approved at least 500 loan applicants. This restriction eliminates 8% of the sample. Due to random assignment, any loan officer level variation in contract terms and loan outcomes come from differences in loan officers'

³ Mullainathan and Obermeyer (2019) also explore this direction. They show that bounded rationality and representative bias explain physicians' inefficient use of test for heart attack using predictions from machine learning as a benchmark.

information processing, rather than differences in borrower characteristics.

2.1. The Lender

The lender on average receives 32,000 loan applications and approves 14,000 loans per year, with an average loan size of around 55,000 Chinese yuan (\$8,000). Most loans have a maturity term between 12 to 24 months. The lender diversifies risk by making small loans across regions and industries. None of the originated loans is sold or securitized. All borrowers face the same 24% to 25% effective annual interest rate. Fixed interest rates are common in China's unsecured short-term small business credit market, and credit demand being much more responsive to the quantity margin than the price margin is a common feature of high-risk short-term lending markets in other economies as well.⁴ As the price of loans is fixed, loan officers' decisions are solely on the quantity margin, namely loan size, and maturity. Since maturity is typically determined by the management team, based on the lender's funding liquidity condition, leaving little discretion to loan officers, I focus on loan size as the main decision variable.

The whole lending process, from the borrower's application to the final credit decision, typically takes one to two weeks. During the sample period, the lender has no algorithm to assign internal ratings to borrowers. As a result, lending decisions depend on how loan officers assess hard information and acquire soft information.⁵ I describe loan officers' duties and incentives in the following section.

2.2. Loan Officers

There are on average 40 loan officers throughout the sample period. All loan officers work independently in the headquarter and do not have any regional or industry specialization. Loan officers do not solicit loans. Borrowers' demographic information enters the lender's system once an application file is filled out.

⁴ Loan contracts being sensitive on the quantity margin (loan size) but much less on the price margin (interest rate) is consistent with the credit rationing theory a la Stiglitz and Weiss (1981) and is confirmed by empirical findings in the US small business credit market (e.g. Petersen and Rajan 1994) and other high risk credit markets (e.g. Karlan and Zinman 2008).

⁵ There are, however, certain rules that officers are required to follow. For example, the lender has a maximum lending amount of 500,000 yuan (\$80,000). The lender only lends to borrowers whose age is between 20 and 60. Starting December 2012, the lender cannot lend to borrowers who cannot provide verified bank statement.

Field employees then visit borrowers' business sites to collect additional accounting information, such as account receivables/payables and bank statements. These employees are required to take a few pictures of the borrower, the business site, and inventories (if available). After information collection is completed, the system randomly assigns the borrower to a loan officer in the headquarter.⁶

Loan officers first make an accept/reject decision and, conditional on accepting, approve a loan size. The decisions are then randomly assigned to 12 credit managers for review. Credit managers can reject loans approved by loan officers, as happened in 20% of the sample. Credit managers also sometimes revise the approved loan size. These rejections and adjustments made by credit managers are generally based on the lender's funding constraint, not on concerns about individual loan officer's ability. The correlation between loan officer approved loan size and the final loan size is higher than 0.9. Loan officers have no further interaction with the borrower once a decision is made. In particular, loan officers do not monitor loans or participate in collecting overdue debts.

In addition to a fixed salary, loan officers' bonuses and promotion prospects are linked to the number of loans they approved as well as revenues generated from these loans. Based on this compensation scheme, I assume that the objective of loan officers is to maximize the repayment of each loan. Due to random assignment, each loan officer receives a diversified pool of borrowers across regions and industries. As a result, officers do not need to manage their own portfolios to avoid large exposure to a certain region or industry. In addition, since less than 5% of borrowers return for a second loan, loan officers do not have the incentive to lend to low-quality borrowers to form relationships so that they can later hold up these borrowers as in Rajan (1992). I drop all repeated borrowers.⁷ Other than a loan size cap, the lender does not impose other constraints on loan officers that might conflict with profit maximization.⁸ For these

⁶ In rare occasions, approximately 2% of the sample, a loan officer returns a randomly assigned borrower back to the pool. This happens when loan officers take days off before a decision is made or when they feel it is extremely difficult to make a decision. All my results remain after dropping these cases.

⁷ Since my sample period covers the entire operating history, it is clear that all borrowers in sample are first-time borrowers.

⁸ For example, since all the borrowers are privately owned small enterprises without government connections, there is no political concern for doing business in this market.

reasons, maximizing individual loan repayment reasonably describes a loan officers' objective.

2.3. Hard Information

Loan officers observe codified demographic and accounting information through the internal system. The former is mostly self-reported by borrowers. While some information, such as age and gender, is easy to verify, other information, such as months living at the current address, is not. The lender obtains borrowers' credit records from the central bank's credit system. These reports contain information on borrowers' credit history, such as number and types of loans, number of credit cards, and delinquencies. Accounting information is collected during field officers' visits to the borrowers' business sites. Overall, there are 70 hard information variables, all either submitted by borrowers or collected by field employees and made available to loan officers exogenously.⁹ Table 1 reports a complete list.

2.4. Soft Information

In addition to hard information, loan officers can acquire soft information from two sources. First, they have access to pictures of the borrower, borrower's business site, and borrower's inventory. These pictures are taken by field employees and made available to loan officers through the lender's internal system. Since interpretations of pictures are ambiguous and vary across officers, I define them as soft information. Second, officers can call the borrower and the borrower's family members or coworkers.¹⁰

Figure 1 summarizes the information structure and decision process. First, borrower i fills out a loan application with demographic information. The lender then sends field employees to collect accounting information and take pictures. Next, the system randomly assigns borrower i to one of the J officers. Finally, the officer being assigned (call her officer j) processes codified information X_i , determines how much soft information s_i she would like to produce and makes a lending decision based on X_i and s_i .

⁹ Many variables are categorical. There are overall 205 variables if all categorical variables are converted to dummies.

¹⁰ Personal conversations with loan officers suggest that this happens in half of the sample on average, but varies significantly across loan officers.

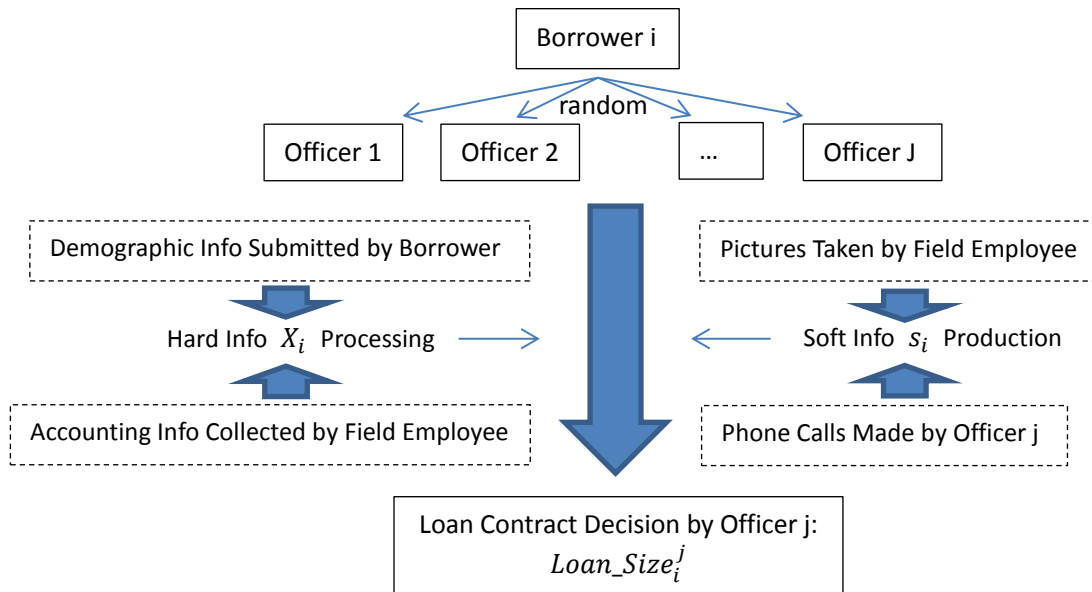


Figure 1: Information Structure and Decision Process

3. Conceptual Underpinnings and Research Design

The small business lending setting provides a suitable laboratory to study information processing by loan officers. In this section, I first connect this setting to the conceptual underpinnings in my research question. I then lay out a research design to address my research question.

3.1. Conceptual Underpinning

3.1.1. Hard Information Processing

I differentiate between two categories of theories that explain the inefficiencies in loan officers' hard information processing. The first emphasizes bounded rationality, going back at least to the work of Herbert Simon (Simon 1955). These models focus on the limits of cognitive resources, such as attention, memory, or computation (Mullainathan 2002; Sims 2003; Gabaix 2014; Bordalo et al. 2019). Blankespoor et al. (2019b) survey applications of this class of models in accounting and finance. Bounded rationality is very natural in my setting: loan officers may not be able to attend to, process, or mentally represent the

rich set of data available on borrowers and so may instead resort to a simpler model of risk. I test two predictions from this class of models. First, if attention is costly and cognitive constraints bind, officers can process only a subset of useful signals. Second, research has shown that humans have particular difficulty perceiving nonlinear relationships between variables. These relationships are usually simplified and represented mentally in a linear fashion (Stango and Zinman 2009).¹¹ For example, officers might recognize a borrower's education background and industry as individually important but have difficulty seeing how they might interact in determining risk if education is more relevant in certain industries than others. I thus test whether officers systematically fail to incorporate nonlinear signals in their decisions.

The second category of theory emphasizes that, even in the set of variables used for decision making, people make systematic probabilistic errors (Kahneman 2011; Benjamin 2019). An important class of model under this category studies representativeness heuristics and was first proposed in psychology by Kahneman and Tversky (1972) and Tversky and Kahneman (1974). Fischer and Verrecchia (1999; 2004) make early theoretical applications of representativeness heuristics in the accounting literature, focusing on trading and disclosure. Libby et al. (2002) and Bloomfield (2002) survey early experimental tests on such biases in accounting. Recently, Bordalo et al. (2016) formalize this concept in economics as probability judgments based on the most distinctive differences between groups and show that representativeness heuristics can exaggerate perceived differences between groups.¹²

In my setting, the theory of representativeness predicts that loan officers approve loan sizes too small for borrower groups with characteristics representative of high risk, because such characteristics catch officers' eye and exaggerate their perception of the risks. One such distinctive characteristic that is both relevant and have large sample in my setting is salient information, defined as large negative realizations

¹¹ Wagenaar and Sagaria (1975) and Wagenaar and Timmers (1978, 1979) provide initial experimental evidence of this phenomenon.

¹² Bordalo et al. (2016) use this formalization to generate gender stereotypes. In recent years, this theoretical framework has been shown to explain a wide range of financial and economic outcomes, including consumer behavior (Bordalo et al. 2013b), corporate investment (Gennaioli et al. 2015), aggregate stock returns (Greenwood and Shleifer, 2014), cross-sectional stock returns (Bordalo et al. 2019), bank lending standards (Baron and Xiong 2017, Fahlenbrach et al. 2017), corporate bond returns (Greenwood and Hansen 2013), and credit cycles (Lopez-Salido et al. 2017; Bordalo et al. 2018). It also has been a leading framework to explain the various episodes of the global financial crisis and its aftermath (Gennaioli and Shleifer 2018).

in accounting variables (e.g., a large negative jump in cash flows). If loan officers consider such information as being representative of a “bad type” of borrower, they might overreact. Indeed, among borrowers who default, 28.1% are salient, defined as having at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. This proportion is only 15.8% among borrowers who do not default, making negative salient information a distinctive difference between good and bad borrowers. Based on this prediction, I test whether loan officers on average approve loan sizes that are too small to borrowers with negative salient information.¹³

A remark is in order for my definition of salience. The literature does not provide a uniform definition. Empirical papers often adopt a functional definition by assuming that something is more salient if it is more visible (e.g., Chetty et al. 2009). Theoretical research has established more rigorous definitions by stating that what is salient depends on what the decision-maker compares it to. Bordalo et al. (2013) define the salience of an attribute of a particular good to a consumer (e.g., the price of a particular bottle of wine) as the distance between its value and the average value of that attribute across all goods available to the consumer (e.g., the average price of all wine on the menu). Similarly, Bordalo et al. (2012) define the salience of a particular payoff of a lottery as the difference between its value and the average payoff yielded by all other available lotteries in that state. My definition that a borrower characteristic is more salient if it has a large (negative) realization fits well with both the functional definition in empirical research, because large realizations are rare and thus more noticeable, and the more rigorous theoretical definition, because large realizations of a characteristic have a greater distance from the average of that characteristic across all borrowers.

3.1.2. Soft Information Acquisition

Despite any inefficiency in hard information processing, loan officers should play a crucial role in

¹³ Another borrower group potentially subject to representativeness bias consists of female borrowers as in the model of Bordalo et al. (2016). In appendix B, I test whether gender stereotypes explain loan officers’ underperformance. It is important to note that, as discussed in Dobbie et al. (2019), my setting is not powerful enough to distinguish between gender bias due to representativeness and that is taste-based.

collecting and processing qualitative and costly-to-verify soft information, as documented in other similar settings (e.g., Petersen and Rajan 1994, 1995; Agarwal and Hauswald 2010; Michels 2012; Cassar et al. 2015; Iyer et al. 2016; Campbell et al. 2019). For example, loan officers in my setting extract valuable signals by initiating and sustaining conversations with borrowers and it is hard to train a computer to mimic humans' adaptability in conversations.¹⁴ In addition, loan officers observe hard information before determining whether and how to acquire soft information, and certain features of hard information can trigger soft information acquisition. For example, when observing unusual patterns in cash flows, loan officers typically make calls and ask the borrower for an explanation. Consequently, factors that impede hard information processing might also affect soft information acquisition.

In settings with no friction, information acquisition is usually modeled under the rational expectation equilibrium framework in which investors learn about a fundamental (e.g., a borrower's type) by acquiring a signal of it (e.g., Verrecchia 1982; Diamond 1985). This framework imposes that investors' belief about the fundamental coincides with its true underlying distribution. If representativeness distorts loan officers' beliefs in my setting, such that their perception of a borrower's type differs from the borrower's true type, the acquired soft information might be polluted.¹⁵ Intuitively, talking to a borrower but starting with an incorrect belief might make the conversation less effective. I thus test whether soft information acquisition is less efficient when the borrower has salient information.

3.2. Research Design

My research design has three components. First, assessing information processing by loan officers requires a benchmark. In section 3.2.1, I discuss the rationale for using machine learning as such a

¹⁴ Hardening soft information plays a key role in the history of the credit scoring (Kraft 2015; Liberti and Petersen 2018). Although it is possible to ask loan officers to assign scores to borrowers based on conversations, and then feed scores to a model, we still need officers to acquire soft information before scores can be assigned. Moreover, such a process inevitably lose information as different people might interpret the same conversation differently, making scores hard to compare across officers.

¹⁵ In mathematical terms, denote the true borrower type as $\theta \sim N(\bar{\theta}, \sigma)$. Under a rational expectation equilibrium, acquired soft information x is modeled as a signal that equals θ plus a noise term ε , $x = \theta + \varepsilon$, where ε has mean 0. Consequently, the signal x is an unbiased measure of θ . In contrast, if loan officers have a distorted belief about the borrower's type $\theta' \sim N(\bar{\theta}', \sigma)$ such that $\bar{\theta}' \neq \bar{\theta}$, the acquired soft information x' might become $x' = \theta' + \varepsilon$, which is a biased measure of the true type θ .

benchmark. Second, machine is a useful benchmark for hard information processing, but not soft information acquisition. In section 3.2.2, I describe an approach to decompose officers' decisions into a part driven by hard information and a part driven by soft information. Third, an ideal comparison between humans and machines requires randomly assigning borrowers to each. Since all my data is generated by human decisions, I do not have a counterfactual about what loan outcome would be if a different loan size were to be assigned by a machine. In section 3.2.3, I address this challenge by estimating a causal parameter between loan outcome and size and using this parameter to recover the counterfactual outcome.

3.2.1. Machine Learning as a Benchmark

Suppose we observe the outcome of a decision made in the previous period:

$$\text{Human Decision}(\text{Info}_t) \rightarrow \text{Outcome}_{t+1}$$

How should we assess strengths and weaknesses of Human Decision(Info_t)? Defining decision error requires a benchmark. Such a benchmark is generally absent in the real world because the ideal, error-free decision is unobservable. One potential solution is to use the outcome (*Outcome*_{t+1}) as the benchmark. This approach has two limitations. First, it is usually impossible to know what *Outcome*_{t+1} would be if no error were to exist in *Human Decision* (Info_t). As a result, we can only cross-sectionally estimate errors of worse decision-makers versus better ones. Second, any unpredictable shock between t and t+1 that change *Outcome*_{t+1} should not be used in assessing decision at t. In my setting, this means that any unpredictable shock to a borrower's willingness and ability to repay should not be used to evaluate loan officers' decisions. But since shocks are often not observable and cannot be separated from *Outcome*_{t+1}, they would inevitably induce bias if I were to follow this approach.¹⁶

To avoid these difficulties, I instead use a machine learning model to predict a borrower's risk (e.g., expected repayment), using only information available at t and treat this machine prediction as the

¹⁶ This point has been raised by Einav et al. (2018).

benchmark. If the model makes considerably fewer decision errors than loan officers, this helps me avoid the challenge that human error-free decision is unobservable. Since machine predictions are based only on information available when human decisions are made, it also avoids the look-ahead problem. Denote X_t as the set of machine-readable information. I train a machine learning model $M()$ that maps X_t into a machine predicted borrower riskiness and then further map this machine predicted riskiness into a loan contract. Call this loan contract *Machine Decision* _{t} :

$$\textit{Machine Decision}_t = M(X_t) \tag{1}$$

It is important to ensure that the machine-based contract is feasible. In particular, does the lender have sufficient funding to implement this contract? Moreover, would the borrower accept this contract if it differed from the original one? To achieve feasibility, I follow a conservative approach with four steps.

Step 1: For each month, sort borrowers by their machine predicted riskiness.¹⁷

Step 2: For the same month, sort the original human assigned contracts by loan size.

Step 3: Reallocate human-assigned loan sizes to borrowers according to the machine-predicted riskiness, where bigger loans are reallocated to borrowers with lower machine-predicted riskiness.

Step 4: If machine-based loan size exceeds the borrower's requested loan amount, I reset the machine-based loan size equal to the borrower's requested loan amount.

Figure 2 illustrates a hypothetical example with 3 borrowers ($i=1,2,3$). Suppose the model predicts that borrowers 1, 2, and 3 have small, medium, and large risk, respectively. And suppose loan officers give medium, small, and large loans to borrowers 1, 2, and 3, respectively. My machine-based decision rule would first reallocate the large loan to borrower 1, the medium loan to borrower 2, the small loan to borrower 3 and then reset the reallocated loan size to the requested loan size if the latter is smaller.

¹⁷ I defer the details of how I train the machine learning model to Section 4.

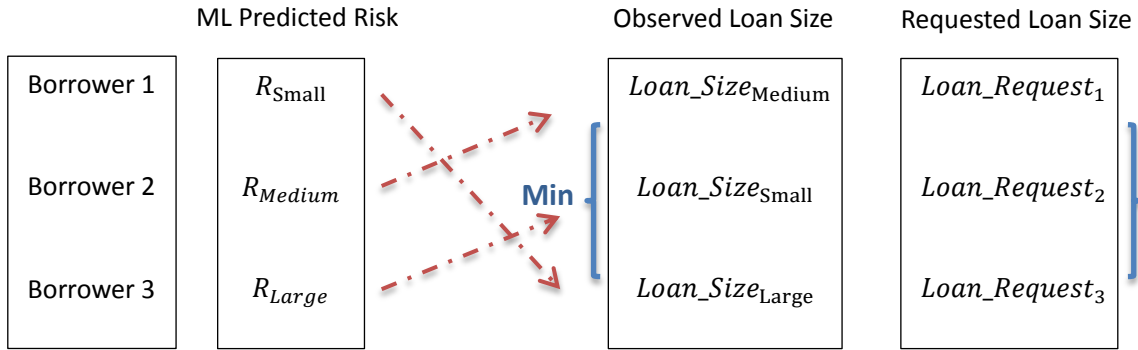


Figure 2: Generating machine-based Contracts

With Steps 1–3, I do not allow the machine to transfer funds across months as the lender might face time-varying funding liquidity shocks. Neither do I allow the model to optimize loan size distribution within a month. This step ensures that the strategy is feasible under the lender’s credit supply constraint.

With respect to credit demand, there are two scenarios to consider. First, would a borrower accept the machine-based loan size $M(X_{i,t})$ if it were bigger than the original loan size? Second, would a borrower accept the machine-based loan size if it were smaller than the original loan size? Step 4 deals directly with the first scenario by not allowing the machine-based contract to surpass borrowers’ credit requests. This is not an important constraint. Most borrowers request an amount far larger than the approved loan size, and only 0.17% request loans smaller than the approved size. Figure A1 in the appendix reports a histogram of excess demand, defined as the difference between requested and approved loan size, together with a histogram of actual approved loan size. The median excess demand is greater than the 93rd percentile of the approved loan size. This universal under-funding is consistent with credit rationing theory, such as by Stiglitz and Weiss (1981), and is a common feature in other highly risky credit markets (Adams et al. 2009).¹⁸ Indeed, I only need to take Step 4 for less than 3% of the sample to implement $M(X_i)$.

Next, I consider the second scenario. 14% of approved loans are turned down by borrowers. I, therefore,

¹⁸ While such gaps could be the result of a strategic game where borrowers request larger loans and anticipate underfunding, it is unlikely to explain all of the variation. Early repayment fees and immediate first month payment make the costs of excessive borrowing non-trivial. In addition, the average borrower may not be able to predict the loan officer’s own lending preference.

test whether underfunding is a primary reason that some borrowers turn down an approved offer. The results are reported in Table A1. Controlling for credit demand (i.e., requested loan size), column (3) suggests that reducing approved loan size from the 75th percentile (60,000 yuan or \$9,000) to the 25th percentile (27,000 yuan or \$4,000) raises the probability of turning down an offered loan by 11.7%.¹⁹ Since this effect is modest, I do not adjust for the likelihood of borrowers turning down a loan if the machine-based loan is smaller than the original loan. I revisit this assumption in the robustness checks.

Having established the rationale for using the machine learning model as the benchmark and a feasible implementation strategy, to make an apples-to-apples comparison between human decisions and this benchmark, I must take into account that loan officers have access to private, soft information not available to the machine. In the next section, I discuss a method that decomposes human decisions into variation driven by hard information X_i and variation driven by soft information s_i :

$$\text{Human Decision}_i = H(X_i) + s_i \quad (2)$$

3.2.2. Hard and Soft Information: a Decomposition

To perform decomposition (2), I must observe the entire hard information set X_i and know what $H()$ looks like. My data satisfies the first requirement. To obtain $H()$, I search for the combination of X_t that best explains each loan officer's usage of hard information. This is not an inference problem, but rather a prediction problem suitable for machine learning (Mullainathan and Spiess 2017). Specifically, I allow different officers to have different $H()$ and train a machine learning model for each officer j to predict her decision, call it $H^j(X_{i,t})$. Unlike $M(X)$, whose purpose is to predict borrower risk, the purpose of $H^j(X_{i,t})$ is to mimic officer j 's behavior in processing hard information $X_{i,t}$. Importantly, $H^j(X_{i,t})$ captures any limitation or bias that officer j might have in processing $X_{i,t}$. Finally, I recover soft information $s_{i,j,t}$ as the difference between officer j 's actual decision and $H^j(X_{i,t})$:

¹⁹ 11.7% = (60,000-27,000)*0.355/100,000

$$s_{i,j,t} = Loan_Size_{it}^j - H^j(X_{i,t}) \quad (3)$$

Note that (3) identifies variations in $s_{i,j,t}$ across loans but not its average effect in (2) as it is absorbed in the intercept of $H(X_i)$. One concern is that some variation in $Loan_Size_{it}^j$ might be driven by human noise, such as mood and sentiment. Research suggests that mood generates noise in human information processing (Hirshleifer and Shumway 2003; Bushee and Friedman 2016; Dehaan et al. 2016; Cortes et al. 2016). Such noise is pooled with soft information in $s_{i,j,t}$. To test whether there is a valid signal about borrower's risk in $s_{i,j,t}$, I show that soft information identified by (3) predicts loan performance.

In Figure 7, I compute the standard deviation of $s_{i,j,t}$ for each officer and plot it against two measures of officer performance—the average default rate and average profit rate, defined as (total repayment – loan size)/loan size. If $s_{i,j,t}$ captures a valid signal, officers better at acquiring soft information should have larger dispersion in $s_{i,j,t}$. Indeed, Figure 7 shows that officers with larger dispersion in $s_{i,j,t}$ have lower default rates and generate higher profits. Switching from officer-level to loan-level evidence, Table 2 shows that $s_{i,j,t}$ strongly predicts loan profit at the loan level. For example, column (3) indicates that moving $s_{i,j,t}$ from 25% percentile (-0.67) to 75% percentile (0.56) is associated with 1.5% higher profit rate. These tests confirm that soft information identified by (3) captures valid signal about borrower risk.

To summarize what I have gained so far, Figure 3 provides an overview of my methodological framework. I observe hard information set X_t and actual human-determined loan contracts $Loan_Size_{it}^j$ (second box in the left column) directly from data. In section 3.2.1, I train a machine learning model $M(X_t)$ (first box in the right column) and obtain machine-based loan contracts (second box in the right column). In section 3.2.2, I decompose $Loan_Size_{it}^H$ into $H(X_{i,t})$ and $s_{i,t}$ (first box in the left column). The final step is to compare the performance of $M(X_{i,t})$ and $H(X_{i,t})$. If $M(X_{i,t})$ considerably outperforms $H(X_{i,t})$, it can be used as a valid benchmark to study errors in $H(X_{i,t})$. Here, I face the lack of a

counterfactual problem. While the performance of $H(X_{i,t}) + s_{i,t}$ is readily observable (third box in the left column), neither the performance of $H(X_{i,t})$ nor the performance of $M(X_{i,t})$ is observable.

That is, to the extent that a borrower's repayment behavior is a function of loan size, this behavior would certainly differ if no soft information is used (i.e., $H^j(X_{i,t}) \neq H^j(X_{i,t}) + s_{i,j,t}$) or a different loan contract is assigned by the machine (i.e., $M(X_{i,t}) \neq H^j(X_{i,t}) + s_{i,j,t}$). In the next section, I describe a method to generate these unobservable counterfactuals.

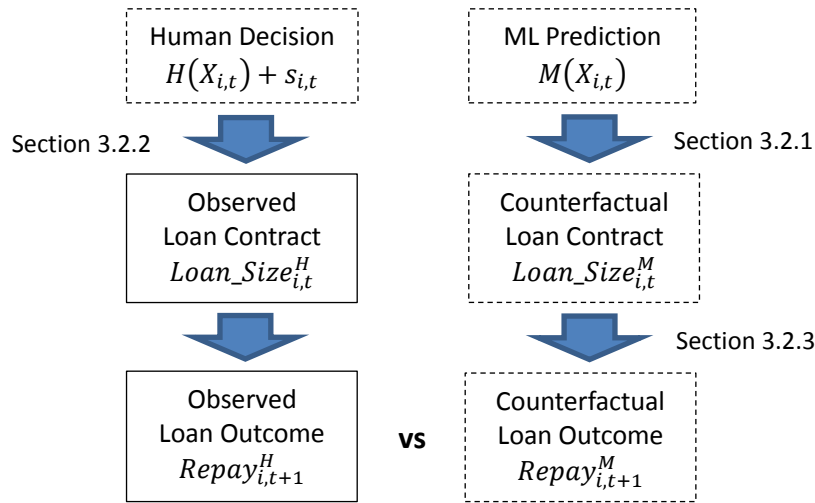


Figure 3: Overview of Methodological Framework

3.2.3. Generating Counterfactual

The goal of this section is to generate the following two unobserved counterfactuals.

1. The loan outcome if no soft information were used in loan officers' decisions: The counterfactual decision rule here is $H^j(X_{i,t})$. Call this counterfactual $Repay_{i,t+1}^{Hard}$.
2. The loan outcome if the loan size were determined by the machine: The counterfactual decision rule here is $M(X_{i,t})$. Call this counterfactual $Repay_{i,t+1}^M$.

Both counterfactuals require estimating changes in loan outcomes induced by changes in loan size while

keeping everything else fixed. The first counterfactual requires estimating changes in loan outcome when loan size changes by $s_{i,j,t}$. The second counterfactual requires estimating changes in loan outcome when loan size changes by $M(X_{i,t}) - Loan_Size_{it}^H$. Denote the parameter governing the (causal) relation between changes in loan size and changes in loan outcome by β :

$$\Delta Repay = \beta \Delta Loan_Size \quad (4)$$

Once β is estimated, the two counterfactuals can be generated by

$$Repay_{i,t+1}^{Hard} = Repay_{i,j,t+1}^H + \beta s_{i,j,t} \quad (5)$$

$$Repay_{i,t+1}^M = Repay_{i,j,t+1}^H + \beta (M(X_{i,t}) - Loan_Size_{i,j,t}^H) \quad (6)$$

I examine two measures of loan outcomes. The first is *Repay_Ratio*, defined as the total repayment amount over the loan amount. Total repayment includes interest payments, management fees, early repayment fees, and principal repayment. My second measure of loan outcome *Repay_Dollar* is simply total dollar repayment (i.e., nominator of *Repay_Ratio*). I estimate (4) for both measures and obtain β_{ratio} for *Repay_Ratio* and β_{dollar} for *Repay_Dollar*.

Both measures are useful because their strengths lie in different aspects. Since *Repay_Ratio* is unit free, it better captures a borrower's type in that a higher quality borrower repays a higher proportion of obligations, irrespective of loan size. On the contrary, *Repay_Dollar* has a dollar unit. β_{dollar} thus confounds the effect of β_{ratio} with a mechanical effect that larger loans on average have more money paid back. Therefore the economic meaning of β_{dollar} is hard to interpret. β_{dollar} , however, has practical value in that it maps dollar changes in loan size directly to dollar changes in profit, facilitating model performance comparison. For this reason, I focus on β_{ratio} when discussing the underlying economics and use β_{dollar} when computing profits for each counterfactual.

One point is worth discussing before diving into details about estimating β . The four-step rule to implement $M(X_{i,t})$ specified in section 3.2.1 does not depend on how precisely β is estimated. This is because I merely reallocate credit across borrowers and any change in loan outcome induced by some borrower receiving a larger loan should be offset by some other borrower receiving a loan smaller by the same amount.²⁰ In equation (6), it implies that the term $\beta(M(X_{i,t}) - Loan_Size_{i,j,t}^H)$ disappears after aggregating across borrowers.²¹ Consequently, the entire profit gain by implementing $M(X_{i,t})$ comes from reallocating larger (smaller) loans to borrowers who are more (less) likely to repay, as predicted by $M(X_{i,t})$. Nonetheless, it is useful to estimate β for other strategies to implement $M(X_{i,t})$.

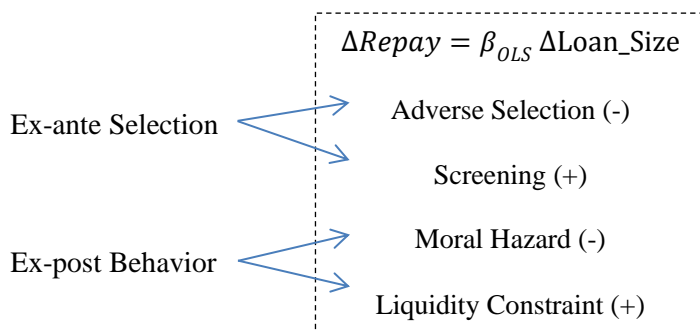


Figure 4: Economic Channels Captured by β_{OLS}

It is helpful to first consider an OLS estimator for (4). In principle, there are four economic channels that connect loan size and repayment, summarized in Figure 4. First, adverse selection predicts a negative relationship as the low-quality type has the incentive to self-select larger loans (Jaffee and Russell 1976). Second, screening serves as a counterforce where loan officers try to assign larger loans to the high-quality type. Third, moral hazard predicts a negative relation since borrowers who receive larger loans have an incentive to take too much risk ex-post (Stiglitz and Weiss, 1981) because they have less skin-in-the-game. Finally, the theory of liquidity constrained entrepreneur suggests a positive relationship because more funding allows more positive NPV projects to be undertaken and improves repayment

²⁰ This is not true for cases that require Step 4. But since such cases represent less than 3% of the sample, and the adjustments made in Step 4 for these cases are typically small, these cases only have a marginal effect.

²¹ The only assumption required for this logic is that the relation between loan size and profit is linear as specified in (4). Figure A2 in the appendix provides reduced-form supporting evidence that this relation is linear throughout the range of loan size.

ability (Evans and Jovanovic, 1989; Holtz-Eakin et al. 1994).²²

The OLS estimator pools all four effects. But for the purpose of generating the counterfactuals, my goal is to capture the two ex-post effects of changing loan size and eliminate the two ex-ante selection effects. This is because the counterfactuals should capture what happens to a borrower’s repayment behavior if a different loan size is assigned to the same borrower. If loan officers’ screening effort cannot eliminate adverse selection (i.e., the ex-ante selection effect is negative), β_{OLS} underestimates β .

I exploit random assignment and use officers’ average loan size as an instrument for actual loan size to estimate the causal relationship between loan size and repayment. Since all officers face the same pool of borrowers due to random assignment, their average loan size (i.e., “leniency”) should be orthogonal to borrower characteristics and affect repayment only through officer-specific characteristics, such as their style or level of risk aversion. In this way, the IV estimator is not polluted by borrowers with different characteristics, observable or unobservable, systematically self-selecting into different loan contracts. Specifically, I run the following 2SLS regression.

$$\text{First Stage: } Loan_Size_{i,j,t}^H = \gamma Leniency_j + controls_{i,j,t} \quad (7)$$

$$\text{Second Stage: } Repay_{i,j,t+1}^H = \beta_{IV} \widehat{Loan_Size}_{i,j,t}^H + controls_{i,j,t}$$

$Leniency_j$ is defined as the average loan size by officer j . $Repay_{i,j,t+1}^H$ is the observed repayment ratio or dollar repayment. Table 3 reports the results of (7). In both Panel A (*Repay_Ratio*) and Panel B (*Repay_Dollar*), the IV estimates are larger than the OLS estimates, consistent with the previous conjecture that OLS estimator is biased downward due to adverse selection. Note that this result is unlikely to be driven by weak IV, as the first stage is very strong.

²² It is well understood in the empirical contract literature that β_{OLS} pools adverse selection and moral hazard effects in the consumer credit market and insurance market (Adam et al. 2009; Chiappori and Salanie 2000). I add a liquidity constraint effect since the borrowers are entrepreneurs and the amount of credit should affect repayment by changing future cash flows.

One might worry that $Leniency_j$ captures an officer's ability to acquire soft information because officers good at this might have the confidence to approve larger loans. This is unlikely to violate the exclusion restriction condition if officers only use soft information to make decisions at the intensive margin (i.e., loan size) but not at the extensive margin (i.e., whether to grant the loan). The identification assumption is that leniency can only affect repayment through its impact on loan size. Since officers do not interact with borrowers after the contract is signed, they can only affect repayment through loan size. There is no other channel at the intensive margin that officers with different soft information acquisition abilities can affect repayment differently other than loan size. However, if officers use soft information to make grant/reject decisions as well, β_{IV} might overestimate β . This is because if more lenient officers are better at using soft information to reject bad borrowers, they will face a better pool of borrowers at the intensive margin, leading to inflated β_{IV} . I therefore interpret β_{IV} as the upper bound and β_{OLS} as the lower bound of β . In robustness checks, I show that my results are robust to any value of β within this range.

All the pieces in Figure 3 are now complete and ready to take to data. Since I use machine learning tools extensively, I take a short detour in the next section to introduce the procedure to train these models.

4. Machine Learning Models

Machine learning is a powerful tool for prediction problems²³. In my setting, it involves fitting a model $M(X)$ on borrower characteristics X to predict an outcome, such as loan repayment. In principle, one can fit $M(X)$ with OLS. OLS usually does not perform well in prediction problems, however, because it imposes a linear functional form on $M(X)$. For example, if the effect of borrower's education on repayment depends on the borrower's industry, OLS does not capture this interactive feature, unless one puts in an interaction between education and industry. Without strong theoretical reasons to guide which

²³ Hastie et al. (2009) provide comprehensive treatment of machine learning models. Varian (2014), Mullainathan and Spiess (2017), and Athey and Imbens (2019) provide introductions to machine learning for economic research. Though my focus is on prediction, an emerging strand of literature studies how machine learning can aid causal inference (e.g. Belloni, et al. 2014; Athey and Imbens 2017). In the accounting literature, machine learning models have been used to predict misstatements and fraud (Bertomeu et al. 2018; Bao et al. 2020) and in textual analysis (Li 2010; Loughran and McDonald 2016; Brown et al. 2020).

interactions to include in $M(X)$, it is computationally infeasible to put in all pairwise interactions in OLS because the number of variables would far exceed the number of observations.

As alternatives, many machine learning models do not impose a priori what to include but let the data “speak” to identify variables and interactions that have out-of-sample predictive power. This approach is problematic for inference but suitable for prediction problems (Mullainathan and Spiess 2017). Following prior research (Bertomeu et al. 2018; Kleinberg et al. 2018), I use gradient boosted decision trees (GBM) as my main model (Friedman 2001). I test whether my results are robust to other models in the robustness checks. To emphasize the importance of the ability of machine learning models to capture nonlinear and interactive signals in the data, I also report results that use OLS to fit $M(X)$.

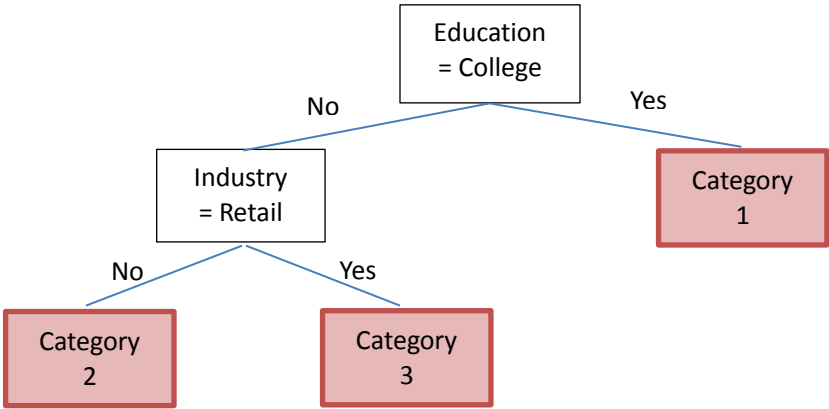


Figure 5: Regression Tree Example

The basic building block of GBM is a decision tree, in which the data are divided through a sequence of binary splits. Figure 5 illustrates an example. To predict loan repayment, the first split might be a borrower’s education (whether the borrower has a college degree). In the next step, we can split each of the two nodes created by that first split by different variables, such as industry (whether the borrower is in retail). Based on these two characteristics, the sample of borrowers is divided into three categories. The predicted repayment of a new borrower in the hold-out sample who falls into category i is set to equal to the average repayment of borrowers in the training sample in category i . This sequential procedure allows

for a high degree of interactivity in the data, a key advantage of machine learning models over OLS.

I use machine learning for two purposes. Figure 6 demonstrates the analysis strategy for both purposes. First, I train a model $M(X)$ to predict a borrower's repayment. An accurate evaluation of $M(X)$ requires being performed out-of-sample. I, therefore, divide the whole sample randomly into a training sample that $M(X)$ is fitted on and then use the remaining data, a hold-out sample, to evaluate $M(X)$ against $H(X)$. This is a standard procedure to prevent $M(X)$ from appearing to do well simply because it is being evaluated on data it has already seen. A concern with this strategy is that, at each period t , it allows $M(X)$ to see future data not available to loan officers. To rule out this as the main reason for $M(X)$'s outperformance, I use a second strategy that trains $M(X)$ month-by-month by feeding it with data up to month t and compare its performance with loan officers out-of-sample in month $t+1$.

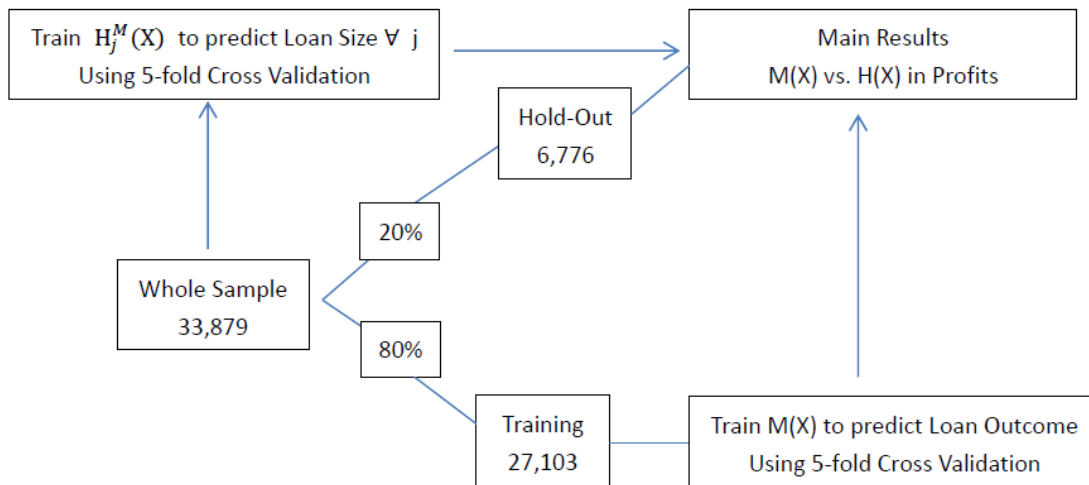


Figure 6: Data Partition, Model Training, and Model Evaluation

The second purpose of using machine learning is to predict loan officer decisions. I train a separate model $H_j^M(X)$ for each officer j 's decision, allowing different officers to have different models to process X . After obtaining $H_j^M(X)$, I compare the performance of $H_j^M(X)$ with that of $M(X)$ on the hold-out sample by evaluating their ability to generate profits (or equivalently, to rank borrowers correctly).

The performance of a GBM model depends on a few parameters.²⁴ Following standard practice (e.g., Kleinberg et al. 2018), I search for the optimal values of these parameters using fivefold cross-validation. I then estimate the final model using the full training sample. In robustness checks, I test whether my main results are robust to other machine learning models, including Random Forest, LASSO, and Neural Nets. I describe the details of the GBM and other models in the online appendix.

5. Results

In this section, I first test whether $M(X)$ outperforms loan officers. If so, $M(X)$ can be used as a valid benchmark to assess loan officers' weaknesses in processing hard information. Then I explore which factors best explain the different decisions and performance between $M(X)$ and loan officers' hard information processing. I next examine how these factors affect loan officers' soft information acquisition. Finally, I interpret these findings using a unifying theoretical framework.

5.1. Human Performance and Machine Performance

Table 4 compares human and machine performance using the loan profit aggregated over the entire hold-out sample as a performance measure.²⁵ I compare the performance of three types of decisions: A) the observed loan officer decision based on both hard and soft information $H(X_t) + S_t$, B) the officer decision based on hard information only $H(X_t)$, and C) the machine decision $M(X_t)$.

The actually observed profit is 15.6%, as shown in column 2. The difference between columns 1 and 2 is due to the effect of soft information. This should be interpreted as a lower bound for the contribution of soft information for two reasons. First, some soft signal is captured by $H(X)$ if hard and soft information is correlated. This happens if, for example, officers ask specific questions when they observe certain patterns in hard information. Second, as discussed in Section 3.2.2, since soft information is identified as

²⁴ These parameters include the depth of each tree, the total number of trees averaged together, and the weighting scheme for each subsequent tree.

²⁵ Profit is defined as the repayment ratio (Total Repayment – Total Lending/Total Lending).

the residual term, its average effect is absorbed in the intercept and not identified. Moving to column 4, the GBM model generates a profit rate of 21.5%, a 38% increase from observed profit. To demonstrate the importance of the machine learning model's ability to capture nonlinear and interactive features of the data, I report in column 3 the performance of $M(X)$ trained with OLS. The 13.5% profit generated is worse than human performance and far worse than $M(X)$ trained with GBM. Table A2 in the appendix shows similar results with other machine learning models.

Another way to see the different performances between loan officers and the machine is to compare how they differ in ranking borrowers. While machine learning models (OLS and GBM) produce predicted profit as output, which can be used to infer their rankings of borrowers, it is not immediately clear how loan officers rank borrowers. Since officers have an incentive to assign larger loans to higher-quality borrowers, conditional on credit demand, I regress observed loan size on requested loan size and treat the residual as a basis for officers' ranking through revealed preference. Indeed, this residual is small for a borrower who receives a small loan, despite having requested a large loan, indicating officers perceive this borrower as less profitable. I obtain the officers' ranking of borrowers by sorting this residual.

For each of the three models (human, OLS, and GBM), I first sort borrowers into deciles by their rankings. I then compute the average observed profit for each decile of borrowers. A model performs better if loans it predicts to be more profitable (and thus ranks higher) are indeed more profitable. Figure 8 plots average observed profits across predicted profit deciles as predicted by each model. A better performing model should have a more positive slope in the graph. Two messages emerge. First, while GBM has a clear positive slope throughout all deciles, OLS and human models are rather flat, suggesting their predicted profit does not match well with observed profit.²⁶ Second, compared to GBM, officers show similar

²⁶ Flat curve does not mean loan officers (or OLS) give out loans randomly (or have zero predication ability). If loan officers (or OLS) have zero screening ability and give out loans randomly, their curves should be downward sloping, due to adverse selection. An easy way to see this is through the standard asset substitution argument of Jensen and Meckling (1976). Suppose there are two borrowers with the same expected future cash flows but different levels of risk. The first borrower's future cash flow is 0.5 for sure. The second borrower has a future cash flow of 0 or 1, each with a probability of 0.5. Borrower 2 is riskier and thus a worse type from a lender's perspective. Borrower 1 will not accept any loan obligation larger than 0.5. In contrast, borrower 2 will accept loan obligation up to 1 when there is limited liability to protect the downside. Therefore adverse selection suggests that

ability to differentiate borrowers in middle deciles but have severe trouble toward the two tails. Table 5 reports the performance of the three models in the tails. It is clear that borrowers who GBM identifies as bad (good) indeed have bad (good) observed profits. On the contrary, borrowers who officers and OLS identify as bad generate the same level of profit as borrowers they identify as good. Figure A3 and Table A3 in the appendix show similar results with other machine learning models.

The machine learning model might outperform because it does not have human limitations or because it has access to more data. To rule out the latter as the main explanation, I train a model month-by-month by feeding it with data up to month t and compare its performance with loan officers in month $t+1$. Figure 9 reports the results. The model starts to show its superiority in the sixth month, and its advantage increases over time. Since most loan officers work much more than six months, this result suggests that the amount of data is not the most important factor explaining the performance gap. Figure A4 in the appendix shows that this result generalizes to other machine learning models. As additional supporting evidence, I will show in the next section that officers' bias does not disappear with experience.

To sum up, I have shown strong evidence that the machine learning model substantially outperforms loan officers, making it a valid benchmark to assess weaknesses in human decisions. The substantial gap in performance suggests that loan officers mis-rank a considerable number of borrowers when compared to the machine. Figure 10 plots the distribution of the mis-ranking. Bar X represents the percentage of borrowers that officers rank X deciles away from the machine's ranking. Officers rank 74% of borrowers more than one decile away and 22% of borrowers more than five deciles away from machine's ranking.

5.2. Assessing Hard Information Processing

I first study whether cognitive constraint helps explain the human underperformance, as predicted by bounded rationality theories. I examine whether loan officers can only process a subset of useful variables.

worse type borrowers select bigger loans, leading to a downward sloping curve if loans are handed out randomly.

To show this, I regress $M(X)$ and each $H_j^M(X)$ on the entire set of X using an OLS forward stepwise selection procedure. This procedure begins with an empty model and adds in variables in X one by one. In each step, a variable that has the lowest p-value is added. The procedure stops if no variable left has a p-value smaller than 5% if added. Figure 11 reports the results. Among a total of 205 codified variables, $M(X)$ (performer 0 on the vertical axis) finds 147 useful to predict loan outcomes. In contrast, officers use only 25 to 56 of these variables in their decisions. I next test whether officers systematically fail to incorporate nonlinear signals in their lending decisions. I compare the R-squared of the stepwise OLS regressions in Figure 11 and report the results in Figure 12. The 147 variables that $M(X)$ uses explain 66% of the variation in its predictions, the much smaller sets of variables that loan officers use explain a much larger portion of the variation in their decisions, ranging from 83% to 92%. This finding suggests that, while loan officers process information in a linear fashion, the machine learning model captures a remarkable nonlinear and interactive feature of the data that contains useful signals of borrowers' risk.

Having established that cognitive constraint impedes hard information processing, I move on to test whether loan officers systematically make probabilistic errors, as predicted by theories of representativeness heuristics. In particular, I test whether salient information explain loan officers' mis-ranking of borrowers in the following regression.

$$Mistranking_{ijt}^K = \beta Salienc_{ijt} + \varepsilon_{it}$$

$Mistranking_{ijt}^K$ is a 0-1 indicator variable turning on if officer j ranks borrower i more than K deciles away from machine's ranking. I consider $K=1$ and $K=5$. *Salienc* is an indicator equals 1 if the borrower has at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. Table 6 (for $K=1$) and Table 7 (for $K=5$) summarize the results.

Column (2) in Table 6 suggests that loan officers are 28% more likely to mis-rank when they observe salient information in borrowers' hard information. Table 7 changes the outcome variable from $K=1$ to

$K=5$ (i.e., a borrower is ranked by loan officers at least five deciles away from her machine ranking). All results are similar but more pronounced. Table A4 and A5 in the appendix provide further supporting evidence by changing the dependent variable from absolute mis-ranking to directional mis-ranking. Consistent with theories of representativeness, loan offices tend to rank borrowers with salient information too low comparing to the machine.²⁷

Do biases disappear with experience? Loan officers in my sample have a large dispersion of working experience at the lender, ranging from having processed 1,089 to 7,531 loan applications, with a median of 3,793 applications.²⁸ In Table 8, I split the sample into loans that are processed by officers with above and below median experience (i.e., 3,793 applications) and test whether experience affects biases. If anything, the results suggest that experience worsens biases. Why this is the case is outside the scope of this paper and is left for future research. But at the minimum, it provides strong evidence that having access to less data does not explain loan officers' underperformance.

5.3. Assessing Soft Information Acquisition

The results in the previous section suggest that salience impede hard information processing, as predicted by the theory of representativeness heuristics. Next, I study how salience affects their ability to acquire new soft information, a task where human strength may lie.

In Table 9, I split the sample into a subsample with borrowers who have salient information and a subsample without. A different pattern emerges. Soft information produced on the salient subsample has stronger predictive power on loan repayment, as evidenced by both the magnitude of coefficient and the R-squared. To provide further supporting evidence, the following table shows that the standard deviation of soft information, $s_{i,j,t}$, is 40% larger for the salient subsample. Therefore, while Table 9 indicates that

²⁷ Table A6-A8 in the appendix report that the findings in Table 6 are robust to using other machine learning models as benchmarks, including Random Forest (Table A6), LASSO (Table A7), and Neural Nets (Table A8). Results in Table 7 also generalize to these other machine learning models (untabulated).

²⁸ These numbers include rejected loan applications.

per unit of $s_{i,j,t}$ contains a greater signal of a borrower's risk for the salient subsample, loan officers also produce more amount (i.e., dispersion) of $s_{i,j,t}$ for the salient subsample. In addition, officers spend 13% more time processing borrowers with salient information.

| Std of Soft Info | | Loan Processing Time (Min) | |
|------------------------|------|----------------------------|----|
| salient (n= 7,766) | 0.17 | salient (n= 7,766) | 77 |
| non-salient (n=26,113) | 0.12 | non-salient (n=26,113) | 68 |

Why would salience impede hard information processing but facilitate soft information acquisition? Existing theories are silent on this issue. In the next section, I present a simple model to explain this result.

5.4. Interpretation

In this model, a loan officer tries to infer a borrower's type $\theta \sim N(0,1)$ from a hard accounting signal X . X measures θ up to a random noise, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$:

$$X = \theta + \varepsilon$$

The loan officer faces uncertainty in the accounting signal's precision σ_ε^2 . For example, cash flow is an accounting signal for θ . But without knowing the borrower's payment policy with her customers, the officer does not know for sure how noisy cash flow is as a signal. Therefore σ_ε^2 is a random variable, rather than a known constant, to the officer. I model this in a simple way by assuming σ_ε^2 to be binomial:

$$\sigma_\varepsilon^2 = \begin{cases} H & \text{with probability } 1/2 \\ L & \text{with probability } 1/2 \end{cases}$$

This setup is similar to accounting theories considering the impact of earnings disclosures when investors face uncertainty over the variance of cash flows (Beyer 2009; Heinle and Smith 2017) or the precision of earnings disclosures (Hughes and Pae 2004; Kirschenheiter and Melumad 2002; Subramanyam 1996). I depart

from this literature by studying how uncertainty in signal precision σ_ε^2 affects information user's incentive to acquire another signal on σ_ε^2 . In particular, I assume that the officer has an option to acquire additional soft information by incurring a cost C . I model this soft information as a signal on σ_ε^2 . For example, the loan officer can call the borrower and ask directly about her customer payment policy. To make the model simple and similar to that of Grossman and Stiglitz (1980), I assume the officer can completely resolve the uncertainty if she acquires additional soft information. That is, by incurring a cost C , the loan officer learns whether σ_ε^2 is H or L.

I assume that the loan officer has a standard quadratic loss function and thus has the goal to learn the mean of θ , conditional on her information set. Denote the lowercase x as the realization of the accounting signal X . I prove in the following proposition that the loan officer chooses to acquire additional soft information if and only if the magnitude of x is large enough.

Proposition 1: The loan officer chooses to incur cost C to acquire additional soft information if and only if

$$x^2 > C / \text{Var}\left(\frac{1}{1+\sigma_\varepsilon^2}\right). \quad (8)$$

Proof: In the appendix

$\frac{1}{1+\sigma_\varepsilon^2}$ is the precision of the accounting signal, relative to the variance of fundamental (which equals 1).

The right-hand side of condition (8) states that the officer has more incentive to acquire soft information when the cost of doing so is low and when the variance of the precision of the accounting signal is high.

The reason for the latter is that resolving uncertainty is more valuable when the uncertainty is high. The left-hand side of (8) highlights the key takeaway. It says that the officer has more incentive to acquire soft information if the accounting signal is more salient, as measured by the magnitude of realization x . The intuition, building on the previous example, is that learning a borrower's customer payment policy has more value when the borrower has bigger jumps in her cash flows. Therefore salience serves as an attention allocation device that helps the officer decide when and where to pay costly attention.

This simple model shows that salience has a dual role: It distorts belief but at the same time helps attention allocation. Two important implications follow. First, the dual role of salience bridges the theory of bounded rationality and the theory of representativeness heuristics. While salience impedes hard information processing, as predicted by representativeness, it facilitates attention allocation, a costly resource, as emphasized by bounded rationality. This result highlights the value of combining the two categories of theories to understand human decision-making. Second, hiding salient information might improve humans' ability to process hard information but with the cost of losing an attention allocation device. This trade-off echoes a fundamental principle in economics. When there are multiple frictions (e.g., cognitive constraint and behavioral bias), it cannot be presumed that removing one will necessarily improve overall efficiency because other frictions might become more severe.

6. Combining Human and Machine Strength to Make Better Decisions

Although the machine substantially outperforms humans in my setting, simply replacing humans with the machine would sacrifice valuable soft information in $s_{i,t}$. Is there a way to combine human strength and machine strength so that we can make better decisions than machines or humans working alone? In this section, I propose an approach to achieve this.

To see the intuition, first note that officers acquire soft information after observing hard information, making soft and hard information correlated. For example, a key finding in section 5 is that much of the soft information is acquired after officers observe some salient feature of $X_{i,t}$. The key to this approach is to exploit such correlation and train $M(X_t)$ on a subsample generated by officers with better soft information acquisition skills. Intuitively, $M(X_t)$ trained in this way can better capture useful soft signals by putting more weights on variables in $X_{i,t}$ that have a higher correlation with useful signals in $s_{i,t}$, which is otherwise not observable by the machine. This approach also filters out human noise in $s_{i,t}$, since noise does not predict loan outcome and thus will not be captured by $M(X_t)$.

This procedure is particularly useful when there is a large dispersion in soft information acquisition skills across loan officers. Figure 13 confirms this. I regress repayment rate on $s_{i,t}$ for each officer j 's sample with a full set of controls and time fixed effects. I then plot the coefficient of $s_{i,t}$ for each officer j as a measure of soft information acquisition skills. The graph shows that, whereas $s_{i,t}$ produced by officers with low soft information acquisition skills has no predictive power on loan repayment, that produced by high skill officers strongly predicts repayment.

In Table 10, I divide the whole training sample into three subsamples by loan officers' soft information acquisition ability as identified in Figure 13. I then show that $M(X_t)$ trained on the highest ability subsample outperforms that trained on the other two subsamples and, more importantly, outperforms that trained on the whole sample. Comparing column (1) (whole sample) and column (3) (soft = medium) reveals a trade-off. While training on a subsample with high ability enhances the machine's power to capture more valuable soft information, it also significantly shrinks the sample size and reduces the machine's power to capture signals in hard information. A promising future research direction is to develop an optimal weighting procedure that keeps the whole sample but gives observations from humans with higher soft information acquisition skills larger weights.

This technique is generic and should be generalizable to other settings and industries. An important implication is that it is productive to keep and hire humans with high soft information acquisition ability, even if they have low overall performance due to poor hard information processing skills, because their strengths complement those of machines.

7. Discussion and Conclusion

Accounting information facilitates decision-making to the extent that it is properly processed by investors. Which factors determine information processing efficiency is not well understood (Blankespoor et al. 2019b). In this paper, I investigate such factors by leveraging a unique lending setting where the entire

hard information set used by loan officers is observable. I show that a machine learning model substantially outperforms loan officers in processing hard information. Using the machine learning model as a benchmark, I find that limited attention and overreaction to salient information largely explain loan officers' weakness in processing hard information. However, officers acquire more soft information after seeing salient hard information, suggesting salience has a dual role: it creates bias in hard information processing, but facilitates attention allocation in soft information acquisition.

My finding on the attention allocation role of salience may generalize from soft information acquisition by loan officers in my setting to acquisition of any type of information by investors in other markets. Research highlights the key role of information acquisition to mitigate frictions in the capital market (Goldstein and Yang 2017; Blankespoor et al. 2019b) and the credit market (Minnis 2011; Berger et al. 2017; Breuer et al. 2017; Carrizosa and Ryan 2017; Ertan et al. 2017; Lisowsky et al. 2017; Minnis and Sutherland 2017; Sutherland 2018; Darmouni and Sutherland 2019). In such settings, although salience can induce overreaction and inefficiency in information processing, it might still be a desirable feature of disclosure as it facilitates attention allocation and increases new information discovery. I leave such a trade-off between the quality of information processing and the total amount of information discovered in other markets for future research.

My methods to 1) decompose human decisions into hard information processing and soft information discovery and 2) evaluate human ranking using machine ranking as a benchmark are applicable to settings without randomization. Generating unobservable counterfactuals (such as loan profit under a machine decision), however, requires random assignment. Randomization is not uncommon in other high-risk markets for small short-term loans (for example, the UK lender in Dobbie et al. 2019 and Liberman et al. 2019, and the call-center loans of the US credit union in Campbell et al. 2019). In addition, this approach can be applied to two other types of settings. The first is laboratory and field experiments, where humans (e.g., auditors) are randomly matched with objects (Duflo et al. 2013, 2018; Cole et al. 2015; see Floyd

and List 2016 for a survey). The second is settings with an exogenously determined rotation policy, (for example, routinely reassigning loan officers to different borrowers as in Hertzberg et al. 2010 and Fisman et al. 2017). In these settings, my approach for generating counterfactual outcomes is readily applicable.

Advances in AI never fail to stir the public's fear.²⁹ There is no doubt that technological progress will continuously reshape the relation between humans and machines, creating more research questions and challenges. What seems clear today, as this paper tries to demonstrate, is that AI can be a tool, rather than a threat, to help clarify what makes us human.

²⁹ Recently there has been a heated debate in academia on how AI might create unemployment (Acemoglu et al. 2018, 2019), polarization (Autor et al. 2015), inequality, and social imbalance (Rajan 2019).

Reference

- Abramova, I., Core, J. and Sutherland, A., (2019). “Institutional Investor Attention and Firm Disclosure.” Working Paper.
- Acemoglu, D. and P. Restrepo (2019), “Automation and New Tasks: How Technology Displaces and Reinstates Labor,” *Journal of Economic Perspectives*, 33, 3–30.
- Acemoglu, D. and P. Restrepo (2018), “The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment,” *American Economic Review*, 108 (6), 1488–1542.
- Adams, William, Liran Einav, and Jonathan Levin (2009), “Liquidity Constraints and Imperfect Information in Subprime Lending,” *American Economic Review*, Vol. 99, pp. 49–84.
- Agarwal, S., and R. Hauswald. (2010), “Distance and private information in lending,” *Review of Financial Studies* 23:27,57–88
- Athey, S. (2017), “Beyond Predictions: Using Big Data for Policy Problems.” *Science* 355: 483-385.
- Athey, S. and G.W. Imbens. (2017), “The State of Applied Econometrics: Causality and Policy Evaluation,” *Journal of Economic Perspectives*, 31, 3-32.
- Athey, Susan, and Guido W Imbens. (2019), “Machine learning methods that economists should know about.” *Annual Review of Economics*, vol. 11.
- Autor, D. H. (2015), “Why Are There Still So Many Jobs? The History and Future of Workplace Automation,” *Journal of Economic Perspectives*, 29, 3–30.
- Bao, Y., Ke, B., Li, B., Yu, J., and Zhang, J., (2020), “Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach,” *Journal of Accounting Research*, Forthcoming
- Baron, Matthew, and Wei Xiong, (2017), “Credit expansion and neglected crash risk,” *Quarterly Journal of Economics* 132, 765–809.
- Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2019), “Consumer Lending Discrimination in the FinTech Era,” Working paper, UC Berkeley.
- Belloni, Alexandre, Chernozhukov Victor, and Hansen Christian, (2014), “Highdimensional Methods and Inference on Structural and Treatment Effects,” *Journal of Economic Perspectives*, 28, 29–50.
- Benjamin, Daniel J. (2019), “Errors in Probabilistic Reasoning and Judgment Biases.” Chap. 2 in *Handbook of Behavioral Economics*, vol. 2, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson. Amsterdam: Elsevier.
- Berger, P. G.; M. Minnis; and A. Sutherland. (2017), “Commercial Lending Concentration and Bank Expertise: Evidence from Borrower Financial Statements.” *Journal of Accounting and Economics* 64: 253–77.
- Bertomeu, J., E. Cheynel, E. Floyd, and W. Pan. (2018), “Ghost in the Machine: Using Machine Learning to Uncover Hidden Misstatements”, Working Paper

- Beyer, A. (2009), "Capital market prices, management forecasts, and earnings management," *The Accounting Review* 84: 1713-1747.
- Beyer, A., Cohen, D., Lys, T., Walther, B., (2010). "The financial reporting environment: review of the recent literature." *Journal of Accounting and Economics* 50, 296–343.
- Blankespoor, E. (2019), "The Impact of Information Processing Costs on Firm Disclosure Choice: Evidence from the XBRL Mandate." *Journal of Accounting Research*, 57: 919-967.
- Blankespoor, E., E. Dehaan, J. Wertz, and C. Zhu. (2019a), "Why do individual investors disregard accounting information? The roles of information awareness and acquisition costs." *Journal of Accounting Research* 57: 53-84.
- Blankespoor, E., E. Dehaan, and I. Marinovic (2019b), "Disclosure Processing Costs and Investors' Information Choice: A Literature Review," Working Paper
- Bloomfield, R.J., (2002), "The "incomplete revelation hypothesis" and financial reporting." *Accounting, Horizon*. 16 (3), 233-243.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. (2016), "Stereotypes." *Quarterly Journal of Economics* 131, no. 4: 1753-1794.
- Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer. (2019), "Diagnostic Expectations and Stock Returns." *Journal of Finance* 74, no 6: forthcoming.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. (2018), "Diagnostic Expectations and Credit Cycles." *Journal of Finance* 73, no. 1: 199-227.
- Bordalo, P.; N. Gennaioli; and A. Shleifer. (2012), "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics* 127: 1243–85.
- Bordalo, P.; N. Gennaioli; and A. Shleifer. (2013a), "Salience and Asset Prices." *American Economic Review: Papers and Proceedings* 103: 623–28.
- Bordalo, P.; N. Gennaioli; and A. Shleifer. (2013b), "Salience and Consumer Choice." *Journal of Political Economy* 121: 803–43.
- Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2019). "Memory, attention, and choice". NBER Working Paper.
- Breuer, M., Hombach, K., Muller, M.A., (2017). "How does financial reporting regulation affect firms' banking?" *Review of Financial Studies*. 31, 1265–1297.
- Brown, N., Crowley, R., and Elliott B., (2020). "What are You Saying? Using topic to Detect Financial Misreporting," *Journal of Accounting Research*, Forthcoming
- Bushee, B.J., and H. L. Friedman. (2016), "Disclosure Standards and the Sensitivity of Returns to Mood." *Review of Financial Studies* 29: 787–822
- Bushman, R., J., Gao, J., Pacelli, and X., Martin (2019), "The Influence of Loan Officers on Debt Contract Design and Performance," Working Paper, University of North Carolina

- Campbell, D., M. Loumiotis, and R. Wittenberg-Moerman (2019), "Making sense of soft information: Interpretation bias and loan quality," forthcoming, *Journal of Accounting and Economics*
- Carrizosa, R., and S. G. Ryan. (2017), "Borrower Private Information Covenants and Loan Contract Monitoring." *Journal of Accounting & Economics*, 64: 313–39.
- Cassar, G., Ittner, C.D., Cavalluzzo, K.S., (2015), "Alternative information sources and information asymmetry reduction: evidence from small business debt." *Journal of Accounting & Economics*, 59 (2-3), 242-263.
- Chetty, Raj, Looney, Adam, Kroft, Kory, (2009), "Salience and taxation: Theory and evidence." *The American Economic Review* 99 (4), 1145–1177
- Chiappori, Pierre-Andre and Bernard Salanie, (2001), "Testing for Asymmetric Information in Automobile Insurance," *Journal of Political Economy*, 108(1), 56-78.
- Cole Shawn Kanz Martin Klapper Leora, (2015), "Incentivizing Calculated Risk-Taking: Evidence from an Experiment with Commercial Bank Loan Officers," *Journal of Finance*, 70, 537–575.
- Cortes, K., Duchin, R., Sosyura, D., (2016), "Clouded judgment: the role of sentiment in credit origination." *Journal of Financial Economics*, 121 (2), 392-413.
- Costello, A., A. Down, and M. Mehta (2019), "Machine + Man: A randomized field experiment on the role of private information in lending markets", Working Paper, University of Michigan
- Darmouni, O., and A. Sutherland. (2018), "Learning about Competitors: Evidence from SME Lending." Working Paper, MIT Sloan School of Management
- Dechow, P., W. Ge, and C. Schrand. (2010), "Understanding Earnings Quality: A Review of the Proxies, their Determinants and their Consequences." *Journal of Accounting and Economics* 50: 344–401.
- Dehaan, E., J. Madsen, and J. Piotroski. (2016), "Do weather-induced moods affect the processing of earnings news?." *Journal of Accounting Research*, 55, 509-550
- Dehaan, E.; T. Shevlin; and J. Thornock. (2015), "Market (In)Attention and the Strategic Scheduling and Timing of Earnings Announcements." *Journal of Accounting and Economics* 60: 36–55.
- Dellavigna, S., and J. M. Pollet. (2009), "Investor Inattention and Friday Earnings Announcements." *The Journal of Finance* 64: 709–49.
- Diamond, D. W. (1985). "Optimal release of information by Firms." *Journal of Finance* 40, 1071-1094.
- Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2019), "Measuring Bias in Consumer Lending." National Bureau of Economic Research, Working Paper.
- Duflo, E.; M. Greenstone; R. Pande; and N. Ryan. (2013), "Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India." *Quarterly Journal of Economics* 128: 499– 545.
- Duflo, E., Greenstone, M., Pande, R. and Ryan, N. (2018), "The Value of Regulatory Discretion: Estimates From Environmental Inspections in India." *Econometrica*, 86: 2123-2160.

- Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. (2018), “Predictive modeling of U.S. health care spending in late life.” *Science*. 360(6396):1462-1465.
- Erel, I., Li, Stern, C., Tan, and M., Weisbach. (2019), “Selecting Directors Using Machine Learning.” NBER Working Paper
- Ertan, A., Loumiotis, M. and Wittenberg-Moerman, R., (2017). “Enhancing Loan Quality Through Transparency: Evidence from the European Central Bank Loan Level Reporting Initiative.” *Journal of Accounting Research*, 55(4), pp.877-918.
- Evans, David S., and Boyan Jovanovic, (1989), “An estimated model of entrepreneurial choice under liquidity constraints,” *Journal of Political Economy* 97, 808-827.
- Fahlenbrach, Rüdiger, Robert Prilmeier, René M. Stulz. (2017), “Why Does Fast Loan Growth Predict Poor Performance for Banks?” *Review of Financial Studies* 31(3): 1014–63.
- Fischer, P., and R. Verrecchia, (1999), “Public information and heuristic trade,” *Journal of Accounting and Economics* 27, pp. 89-124.
- Fischer, P. E., and Verrecchia, R. E. (2004). “Disclosure Bias.” *Journal of Accounting and Economics* 38, 223–250.
- Fisman, R., Paravisini D., and Vig V.. (2017), “Cultural proximity and loan outcomes.” *American Economic Review* 107:457–92.
- Floyd, E. and List, J. A. (2016), “Using Field Experiments in Accounting and Finance.” *Journal of Accounting Research*, 54: 437-475.
- Friedman, Jerome H., (2001), “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, 29, 1189–1232.
- Gabaix, Xavier (2014), “A sparsity-based model of bounded rationality, ” *Quarterly Journal of Economics* 129.4, pp. 1661-1710.
- Gabaix, Xavier, (2019), “Behavioral Inattention.” In *Handbook of Behavioral Economics- Foundations and Applications 2*, Volume 2, edited by Douglas Bernheim, Stefano DellaVigna, and David Laibson. Elsevier.
- Gabaix, X., D. Laibson, G. Moloche and S. Weinberg (2006), "Information Acquisition: Experimental Analysis of a Boundedly Rational Model," *American Economic Review*, 96(4): 1043-1068.
- Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer, (2016), “Expectations and investment,” *NBER Macroeconomics Annual* 30, 379–431.
- Gennaioli, Nicola, and Andrei Shleifer, (2010), “What comes to mind,” *Quarterly Journal of Economics* 125, 1399–1433.
- Gennaioli, N. and A. Shleifer (2018), “A crisis of beliefs: Investor psychology and financial fragility.” Princeton University Press.

- Goldstein, Itay, and Liyan Yang (2017). “Information Disclosure in Financial Markets.” *Annual Reviews of Financial Economics*, 9, 101-25
- Graham, J., Hanlon, M., Shevlin, T., N. Shroff. (2017), “Tax rates and corporate decision-making.” *Review of Financial Studies* 30, 3128-3175.
- Granja, João, Christian Leuz, and Rajan. (2019), “Going the Extra Mile: Distant Lending and Credit Cycles.” NBER Working Paper 25196.
- Greenwood, Robin, and Samuel G. Hanson, (2013), “Issuer quality and corporate bond returns,” *Review of Financial Studies* 26, 1483–1525.
- Greenwood, Robin, and Andrei Shleifer, (2014), “Expectations of returns and expected returns,” *Review of Financial Studies* 27, 714–746.
- Grossman, S.J., and J. E. Stiglitz (1980). “On the Impossibility of Informationally Efficient Markets.” *American Economic Review* 70: 393–408.
- Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu (2019), “Empirical Asset Pricing via Machine Learning,” forthcoming, *Review of Financial Studies*
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, (2009), “The Elements of Statistical Learning,” Berlin: Springer.
- Heinle, M.S., and K.C. Smith. (2017), “A Theory of Risk Disclosure.” *Review of Accounting Studies* 22 (4): 1459-1491.
- Hirshleifer, D.; S. S. Lim; and S. H. Teoh. (2009), “Driven to Distraction: Extraneous Events and Underreaction to Earnings News.” *Journal of Finance* 64: 2289–325.
- Hirshleifer, David, Lim, Sonya S, Teoh, Siew Hong, (2011), “Limited investor attention and stock market misreactions to accounting information.” *The Review of Asset Pricing Studies* 1 (1), 35–73.
- Hirshleifer, David, Teoh, Siew Hong, (2003), “Limited attention, information disclosure, and financial reporting.” *Journal of Accounting and Economics* 36 (1), 337–386.
- Hirshleifer D. Shumway T. (2003), “Good day sunshine: Stock returns and the weather.” *Journal of Finance* 58:1009–32.
- Hoffman, M., L. B. Kahn, and D. Li (2018), “Discretion in hiring.” *Quarterly Journal of Economics* 133 (2), 765–800
- Holtz-Eakin, Douglas, David Joulfaian, and Harvey S. Rosen, (1994), “Sticking it out: Entrepreneurial survival and liquidity constraints,” *Journal of Political Economy* 102, 53-75.
- Hughes, J., and S. Pae (2004), “Voluntary disclosure of precision information,” *Journal of Accounting and Economics* 37: 261-289
- Iyer, R., Khwaja, A.I., Luttmer, E.F.P., Shue, K., (2016), “Screening peers softly: inferring the quality of small borrowers.” *Management Science*. 62 (6), 1554-1577.

- Jaffee, Dwight and Thomas Russell, (1976), "Imperfect Information, Uncertainty, and Credit Rationing." *Quarterly Journal of Economics*, 90, November, 651-666.
- Jensen, M., and Meckling, W. (1976), "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics* 3, 305 – 360.
- Kahneman, Daniel, and Amos Tversky. (1972), "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3, no. 3: 430-454.
- Kahneman, Daniel. (2011), "Thinking Fast and Slow." New York: Farrar, Straus and Giroux.
- Karlan, Dean S. and Jonathan Zinman, (2008), "Credit Elasticities in Less-Developed Economies: Implications for Microfinance," *American Economic Review*, 98 (3), 1040–1068
- Karlan, Dean and Jonathan Zinman (2009), "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment," *Econometrica*, Vol. 77, pp. 1993–2008.
- Kim, O., Verrecchia, R.E., (1994), "Market Liquidity and Volume around Earnings Announcements." *Journal of Accounting and Economics* 17, 41-67.
- Kirschenheiter, M., and N. Melumad (2002), "Can 'Big Bath' and earnings smoothing co-exist as equilibrium financial reporting strategies?" *Journal of Accounting Research* 40: 761-796.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, (2018), "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133, 237–293
- Kraft, P. (2015). "Rating agency adjustments to GAAP financial statements and their effect on ratings and credit spreads." *The Accounting Review*, 90, 641-674.
- Lawrence, A., Ryans, J.P., Sun, E. and Laptev, N., 2018. "Earnings Announcement Promotions: A Yahoo Finance Field Experiment." *Journal of Accounting and Economics*, 66(2-3), pp.399-414
- Leuz, C., and P. Wysocki. (2016), "The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research." *Journal of Accounting Research* 54: 525–622.
- Li, F. (2010), "The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research*, 48: 1049-1102
- Libby, R., Bloomfield, R., Nelson, M.W., (2002), "Experimental research in financial accounting." *Accounting, Organization and Society*, 27 (8), 775-810
- Lieberman, Andres, Daniel Paravisini, and Vikram Pathania, (2019), "High-cost debt and perceived creditworthiness: Evidence from the UK," Working paper, New York University.
- Liberti, J. M. and M. A. Petersen (2018), "Information: Hard and soft," *Review of Corporate Finance Studies*, 8, 1–41.
- Lisowsky, P., Minnis, M., Sutherland, A., (2017), "Economic growth and financial verification." *Journal of Accounting Research*, 55 (4), 745-794.

- Lopez-Salido, David, Jeremy Stein, and Egon Zakrajsek, (2017), "Credit-market sentiment and the business cycle," *Quarterly Journal of Economics* 132, 1373–1426
- Loughran, T. and McDonald, B. (2016), "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research*, 54: 1187-1230
- Michels, J., (2012), "Do unverifiable disclosures matter? Evidence from peer-to-peer lending." *Accounting Review*. 87 (4), 1385-1413.
- Minnis, M., (2011), "The value of financial statement verification in debt financing: evidence from private U.S. firms." *Journal of Accounting Research*, 49 (2), 457-506.
- Minnis, M. and Sutherland, A. (2017), "Financial Statements as Monitoring Mechanisms: Evidence from Small Commercial Loans." *Journal of Accounting Research*, 55: 197-233.
- Mullainathan, S. (2002), "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117(3): 735-774.
- Mullainathan, S. and Obermeyer, Z (2019), "Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error," NBER Working Paper No. w26168
- Mullainathan, Sendhil, and Jann Spiess, (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.
- Petersen, Mitchell A., and Raghuram G. Rajan. (1994), "The Benefits of Lending Relationships: Evidence from Small Business Data." *Journal of Finance* 49: 3-37
- Petersen, Mitchell A., and Raghuram G. Rajan, (1995), "The effect of credit market competition on lending relationships." *Quarterly Journal of Economics*. 110 (2), 407-443.
- Petersen, Mitchell A., and Raghuram G. Rajan, (2002), "Does distance still matter? The information revolution in small business lending," *Journal of Finance* 57, 2533–2570
- Rajan, R. (1992), "Insiders and outsiders: The choice between informed investors and arm's length debt." *Journal of Finance* 47:1367–1400
- Rajan, R. (2019), "The Third Pillar: How Markets and the State Leave the Community Behind," New York: Penguin.
- Roychowdhury, S., N. Shroff, and R. Verdi. (2019), "The Effects of Financial Reporting and Disclosure on Corporate Investment: A Review." Forthcoming, *Journal of Accounting and Economics*
- Simon, Herbert A. (1955), "A behavioral model of rational choice". *Quarterly Journal of Economics* 69.1, pp. 99-118.
- Sims, Christopher A. (2003), "Implications of rational inattention". *Journal of Monetary Economics* 50.3, pp. 665-690.
- Stein, Jeremy, (2002), "Information Production and Capital Allocation: Decentralized versus Hierarchical Firms," *Journal of Finance*, 57, 1891–1921

Subramanyam, K.R. (1996), "Uncertain precision and price reactions to information", *The Accounting Review* 71: 207-219.

Sutherland, A. (2018), "Does credit reporting lead to a decline in relationship lending? Evidence from information sharing technology." *Journal of Accounting and Economics* 66 (1):123-141.

Stango, Victor and Jonathan Zinman (2009), "Exponential Growth Bias and Household Finance," *Journal of Finance*, 64 (6), 2807–49.

Stiglitz, Joseph, and Andrew Weiss, (1981), "Credit rationing in markets with imperfect information," *American Economic Review* 71, 393–410.

Tversky, Amos and Daniel Kahneman (1974), "Judgment under uncertainty: Heuristics and biases". In: *science* 185.4157, pp. 1124-1131.

Varian, Hal R. (2014), "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28(2): 3–28.

Verrecchia, R. E. (1982), "Information acquisition in a noisy rational expectations economy," *Econometrica*, 50, 1415-1430.

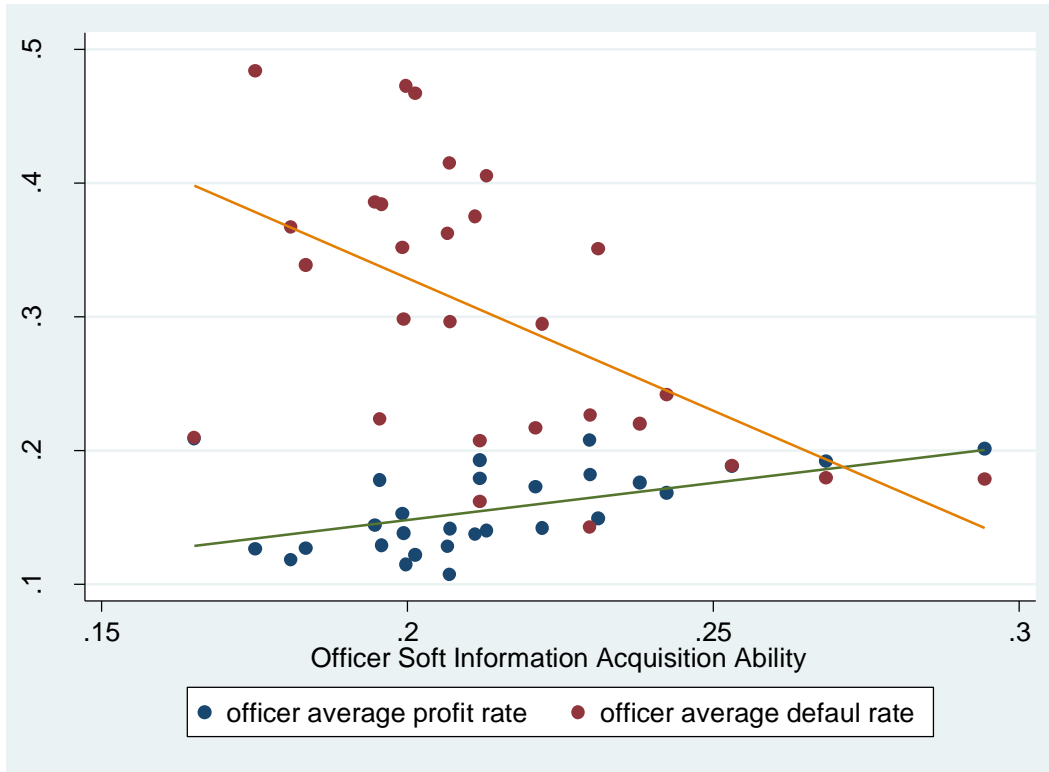
Wagenaar, Willem A. and Sabato D. Sagaria (1975), "Misperception of Exponential Growth," *Attention, Perception & Psychophysics*, 18 (5), 416–22.

Wagenaar, Willem A. and Hans Timmers (1978), "Extrapolation of Exponential Time Series Is Not Enhanced by Having More Data Points," *Perception & Psychophysics*, 24 (2), 182–84.

Wagenaar, Willem A. and Hans Timmers (1979), "The Pond-and-Duckweed Problem: Three Experiments on the Misperception of Exponential Growth," *Act Psychologica*, 43 (3), 239–51

Figure 7

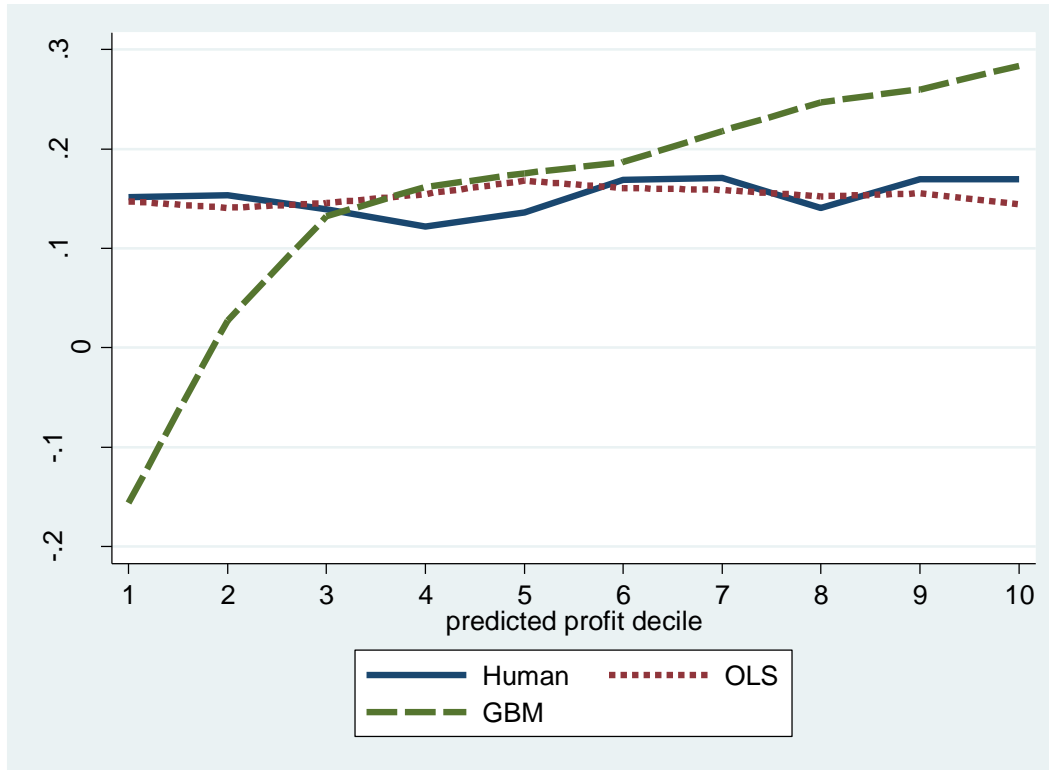
The Relation between Soft Information and Loan Performance (Officer Level)



Note: The figure presents the relation between an officer-level measure of soft information acquisition and loan outcome. The vertical axis measures average profit (red) or average default rate (yellow) for each loan officer. Profit is defined as $(\text{total repayment} - \text{loan size})/\text{loan size}$. Default is an indicator variable that equals one if a borrower fails to fully repay all her loan obligation. The horizontal axis measures each loan officer's soft information acquisition ability, defined as the standard deviation of loan-level soft information, s_{ijt} , for each loan officer.

Figure 8

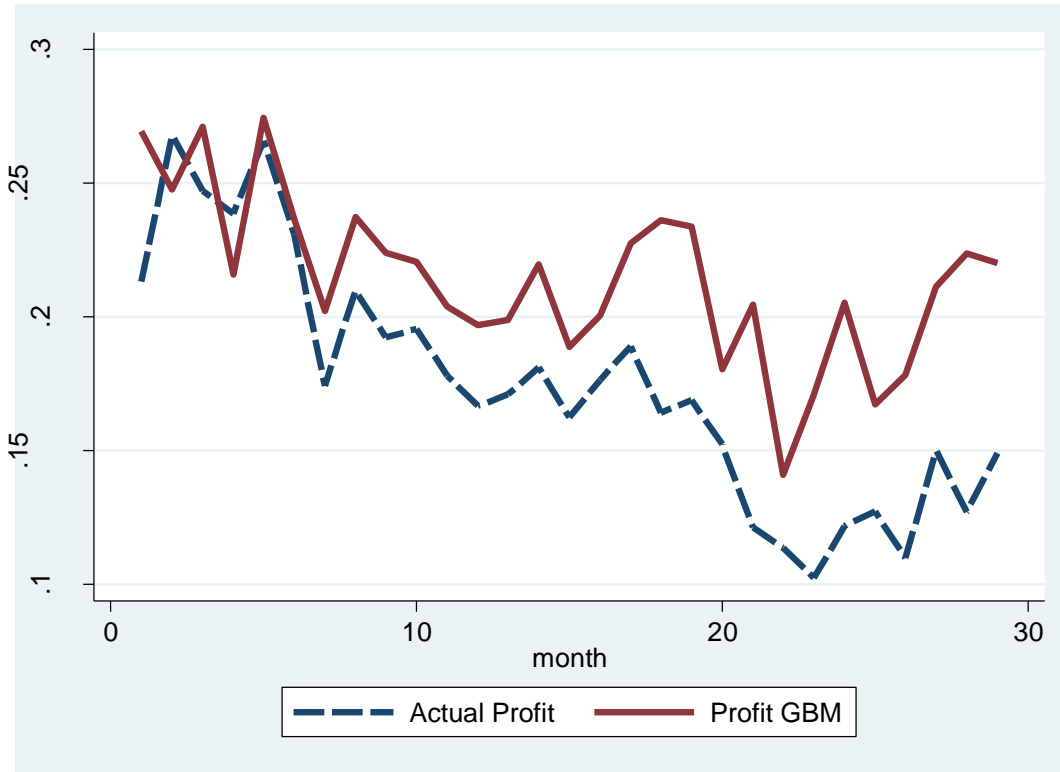
Observed Profit across Predicted Profit Deciles



Note: The figure presents the relation between the average observed profit of loans in each predicted profit decile as predicted by each model (human, OLS, and GBM). A model performs better if loans it predicts to be more profitable (and thus ranks higher) are indeed more profitable, indicating the model ranks borrowers closer to borrowers' true ranking. Therefore, a better performing model should have a more positive slope in the graph. To obtain the GBM (OLS) curve, I first predict a borrower's loan repayment (i.e., profit) using the $M(X)$ trained by GBM (OLS). Next, for each month, I sort all borrowers in the hold-out sample into deciles by their GBM (OLS) predicted profit and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed profit. The GBM (OLS) curve tracks average observed profit for each predicted profit decile as predicted by GBM (OLS). To obtain the Human curve, I first regress observed loan size on the requested loan amount and keep the residual. Next, for each month, I sort all borrowers in the hold-out-sample into deciles by this residual and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed profit. The observed profit rate is defined as (total repayment – loan size)/loan size.

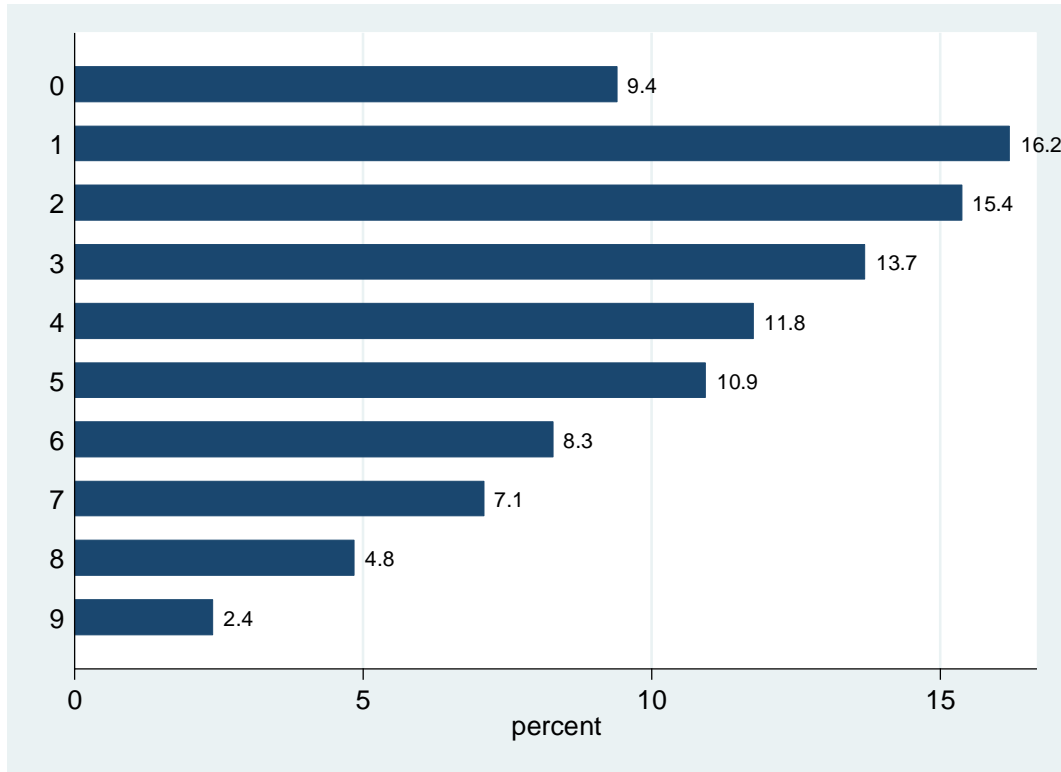
Figure 9

Monthly Performance Comparison between Machine Learning and Loan Officers



Note: This figure compares loan officers' performance with the machine learning model (GBM)'s performance month by month. The solid curve is the observed monthly profit of loans (i.e., aggregate loan profit rate for each month). Profit is defined as $(\text{total repayment} - \text{loan size})/\text{loan size}$. The dashed curve is the out-of-sample monthly profit of the machine learning model (GBM) in month t trained with data up to month $t-1$.

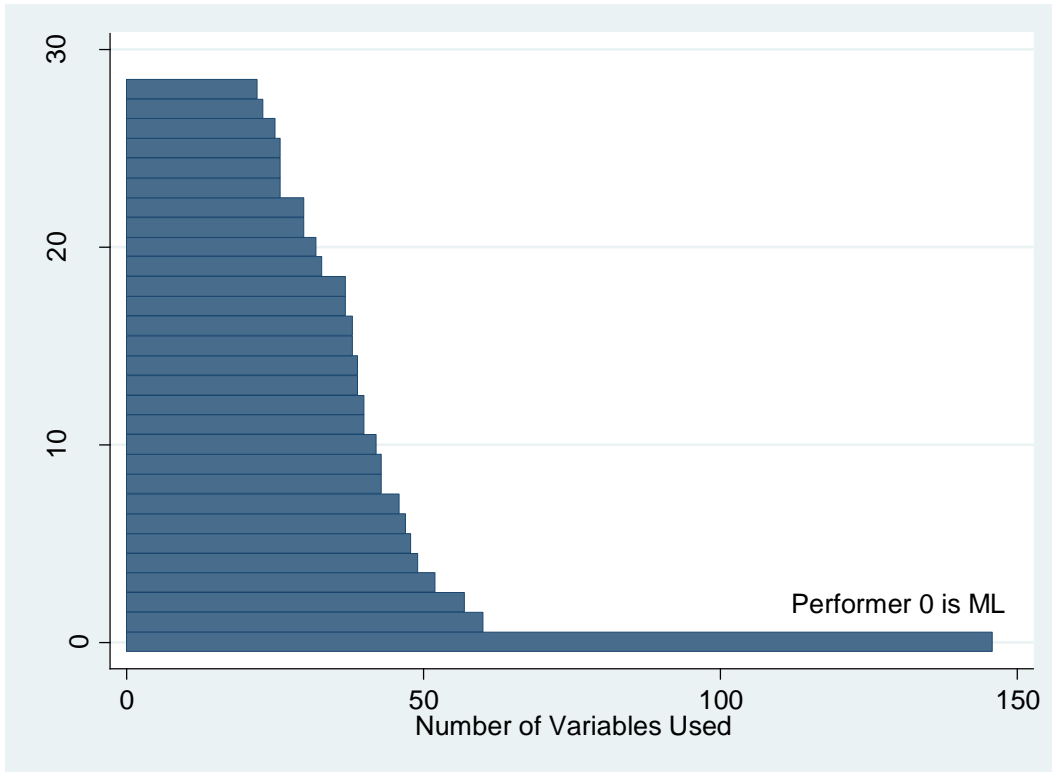
Figure 10
Distribution of Human and Machine Disagreement



Note: This figure presents the distribution of disagreement between loan officers and the machine learning model in their ranking of borrowers in the hold-out sample. Bar X represents the percentage of borrowers that loan officers rank X deciles away from the machine learning model's ranking. To obtain the decile ranking of borrowers by loan officers, I first regress observed loan size on the requested loan amount and keep the residual. Next, for each month, I sort all borrowers in the hold-out sample into deciles by this residual and then pool borrowers in the same decile across months together. To obtain the decile ranking of borrowers by the machine learning model, I first predict a borrower's loan repayment using the $M(X)$ trained by GBM. Next, for each month, I sort all borrowers in the hold-out sample into deciles by their GBM predicted repayment and then pool borrowers in the same decile across months together.

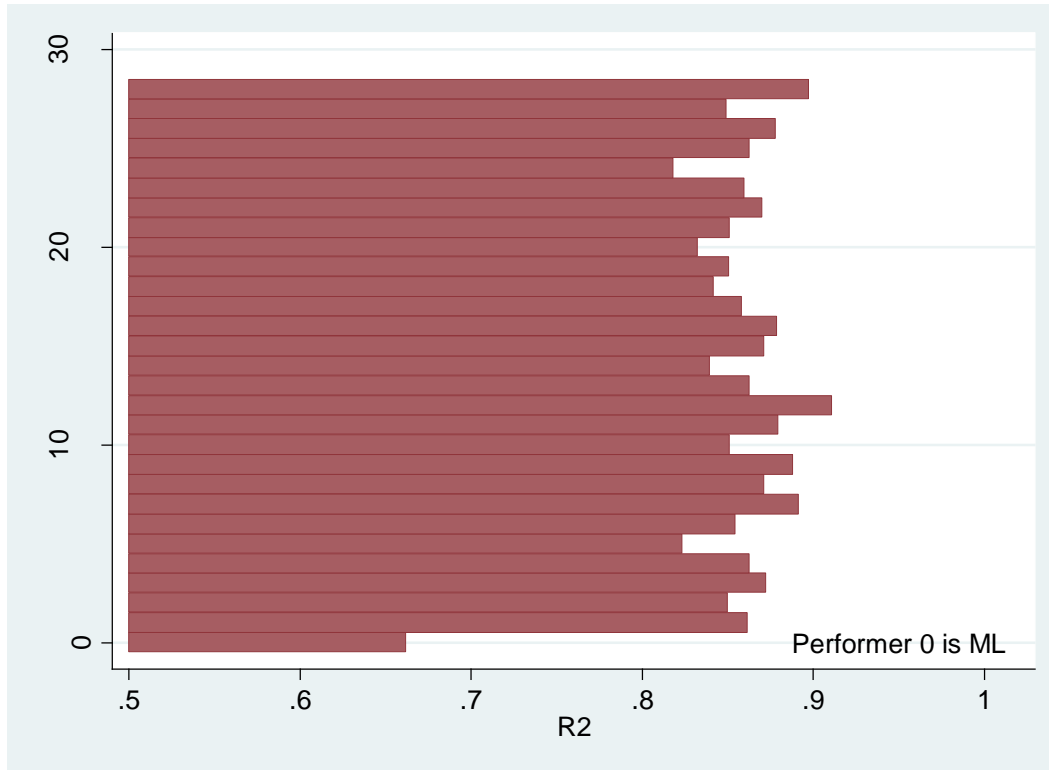
Figure 11

Number of Variables Used by the Machine Learning Model and Each Loan Officer



Note: This figure presents whether the machine learning model uses far more variables to predict loan outcomes than each loan officer uses in her decision. I regress the fitted values of $M(X)$ and each $H_j^M(X)$ on the entire set of borrower characteristics X using an OLS forward stepwise selection procedure. This procedure begins with an empty model and adds in variables in X one by one. In each step, a variable that has the lowest p-value is added. The procedure stops if no variable left has a p-value smaller than 5% if added. Among a total of 205 borrower characteristics, the machine learning model (performer 0 on the vertical axis) finds 147 useful to predict loan outcomes. In contrast, officers use 22 to 56 of these variables in their decisions, suggesting that cognitive constraint is an important factor explaining their underperformance.

Figure 12
The Proportion of Variation (R^2) Explained by Linear Projection



Note: This figure presents if loan officers systematically fail to incorporate nonlinear signals in their lending decisions. It reports the R-squared of the stepwise OLS regressions in Figure 11. The 147 variables that the machine learning model (Performer 0 on the vertical axis) uses explain 66% of the variation in its predictions. The much smaller sets of variables that loan officers use in their decisions explain a much larger portion of the variation in their decisions, ranging from 83% to 92%.

Figure 13
Loan Officer Soft Information Acquisition Skill



Note: The figure plots soft information acquisition ability for each loan officer. To measure this ability, I regress loan repayment rate (i.e., profit) on the loan-level measure of soft information $s_{i,t}$ for each officer j with a full set of controls and time fixed effects. This graph plots the coefficient of $s_{i,t}$ in these regressions for each officer j with its 95% confidence interval. For each officer, the coefficient of $s_{i,t}$ measures the average ability of her acquired soft information to predict loan profit, and thus captures her soft information acquisition skill.

Table 1
List of Variables

| Continuous Variables |
|---|
| Loan Contracts |
| loan size (yuan) |
| loan maturity (months) |
| Demographic Information |
| Age |
| number of children |
| number of months living in the current address |
| monthly earning (yuan) |
| number of months working in the current company |
| Accounting Information - Credit |
| history max loan amount |
| history loan overdue period |
| history loan overdue amount |
| history number of mortgage loans |
| history number of other types of loans |
| number of credit cards |
| number of credit card issuing banks |
| number of loan issuing banks |
| Number of outstanding loans |
| Accounting Information - Cash Flow |
| annual profit (self-reported) |
| monthly revenue from bank statement |
| monthly rent payment from bank statement |
| monthly short-term debt payment from bank statement |
| monthly long-term debt payment from bank statement |
| monthly total cash flows from bank statement |

Note: This table reports a full list of codified variables available to the loan officers through the lender's internal system. To preserve the confidentiality of the lender that provided the data, I do not report the summary statistics for these variables.

Table 1 (cont'd)
List of Variables

| Discrete Variables | | |
|---------------------------|-------------------------------------|---|
| Demographic Info | Accounting Info – Credit | Accounting Info - Current Assets/Liabilities |
| gender | if have secured loan | Buyer |
| male | Yes | have >= 2 Fortune Global 500 buyers |
| female | No | have >= 3 Fortune Chinese 500 buyers |
| education | if have a business loan | have >=5 buyers |
| college or above | Yes | have <5 buyers |
| 3-year college | No | no buyer information provided |
| high school | if have a mortgage loan | Supplier |
| junior high or below | Yes | have >= 2 Fortune Global 500 suppliers |
| marriage status | No | have >= 3 Fortune Chinese 500 suppliers |
| married | | have >=5 suppliers |
| unmarried | Accounting Info – Industry | have <5 suppliers |
| other | Restaurant | no supplier information provided |
| residential type | food processing | account receivable |
| self-owned | Retail | <50% of revenue, average age <30 days |
| family-owned | hotel and accommodation | <50% of revenue, average age <45 days |
| rental | finance and insurance | <50% of revenue, average age <60 days |
| dorm | ICT | <50% of revenue, average age <90 days |
| reside with | wholesale trade | >=50% of revenue |
| self | retail trade | no account receivable information provided |
| with parents | resident services | account payable |
| with spouse | textile and garment manufacturing | <50% of revenue, average age <30 days |
| with co-workers | Entertainment | <50% of revenue, average age <45 days |
| with relatives | scientific and polytechnic services | <50% of revenue, average age <60 days |
| with others | Leasing | <50% of revenue, average age <90 days |
| job status | Construction | >=50% of company revenue |
| self-employed | Manufacturing | no account payable information provided |
| employee | | Inventory |
| other | Accounting Info – Location | >10% of company revenue |
| if house owner | 23 cities | >20% of company revenue |
| yes | | >30% of company revenue |
| no | | >50% of company revenue |
| if have insurance | | no inventory information provided |
| yes | | |
| no | | |
| if have automobile | | |
| yes | | |
| no | | |

Table 2
The Relation between Soft Information and Loan Outcome

| Dependent = Profit | | | | |
|--------------------|--------------------|--------------------|--------------------|--------------------|
| | (1) | (2) | (3) | (4) |
| Soft Info | 0.144*** (7.58) | 0.148*** (7.48) | 0.129*** (7.24) | 0.128*** (7.22) |
| Control | N | N | Y | Y |
| Time Fixed | N | Y | Y | Y |
| Officer Fixed | N | Y | N | Y |
| R2 | 0.003 | 0.016 | 0.055 | 0.056 |
| N | 33,879 | 33,879 | 33,879 | 33,879 |

Note: This table reports if loan-level measure of soft information, $s_{i,t}$, predicts loan outcome. The dependent variable is loan profit, defined as the repayment ratio = (total repayment – loan size)/loan size. Soft Info is $s_{i,t}$ estimated in equation (3) using the decomposition method in section 3.2.2. Control variables include all codified borrower characteristics. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table 3
Causal Relation between Loan Size and Loan Outcome

| Dependent = Repay Ratio | | | | |
|-------------------------|-----------------|--------------------|--------------------|-------------------|
| | OLS | | 2SLS | |
| | (1) | (2) | (3) | (4) |
| | | | IV=Leniency | IV=Leniency |
| Loan Size | 0.010 (1.66) | 0.053*** (7.34) | 0.522*** (7.54) | 0.613** (1.97) |
| | | | First-Stage | |
| | | | 0.561*** (5.17) | 0.154*** (4.6) |
| Controls | N | Y | N | Y |
| Time Fixed | N | Y | N | Y |
| N | 33,879 | 33,879 | 33,879 | 33,879 |

| Dependent = Repay Dollar | | | | |
|--------------------------|---------------------|---------------------|---------------------|--------------------|
| | OLS | | 2SLS | |
| | (1) | (2) | (3) | (4) |
| | | | IV=Leniency | IV=Leniency |
| Loan Size | 0.167*** (11.78) | 0.223*** (15.30) | 0.544*** (13.42) | 0.714** (2.55) |
| | | | First-Stage | |
| | | | 0.561*** (5.17) | 0.154*** (4.60) |
| Controls | N | Y | N | Y |
| Time Fixed | N | Y | N | Y |
| N | 33,879 | 33,879 | 33,879 | 33,879 |

Note: This table reports regression of repayment ratio (Panel A) or repayment dollar (Panel B) on loan size. Column (3) and (4) use officer leniency as an instrument for loan size. Officer leniency is the average loan size for each officer. Control variables include all codified borrower characteristics. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table 4**Model Performance: Human vs. Machine Learning**

| Model Performance | (1) | (2) | (3) | (4) |
|------------------------------|--------|----------|---------|---------|
| (Profit Rate) | Human | Human | ML(OLS) | ML(GBM) |
| | $H(X)$ | $H(X)+S$ | $M(X)$ | $M(X)$ |
| Hold-out Sample (n=6,776) | 0.140 | 0.156 | 0.135 | 0.215 |

Note: This table reports profit generated by each model aggregated across all loans in the hold-out sample. Profit for each loan is defined as (total repayment – total loan size)/total loan size. In column (1), $H(X)$ represents the loan officers’ decision rule based on hard information only. In column (2), $H(X)+S$ represents loan officers’ observed decisions based on both hard and soft information. So 15.6% in column (2) is the lender’s actual profit. Column (3) and (4) are generated by machine decisions $M(X)$ trained with OLS and GBM, respectively.

Table 5**Model Performance: Human vs. Machine in the Tails**

| Predicted Profit Percentile | Average Observed Loan Profit Rate | | |
|-----------------------------|-----------------------------------|---------|---------|
| | Human | ML(OLS) | ML(GBM) |
| 10% | 0.15 | 0.15 | -0.16 |
| 20% | 0.15 | 0.14 | 0.03 |
| 30% | 0.14 | 0.15 | 0.13 |
| 80% | 0.14 | 0.15 | 0.25 |
| 90% | 0.17 | 0.16 | 0.26 |
| 100% | 0.17 | 0.14 | 0.28 |

Note: The table presents the average observed profits in each tail decile of predicted profit as predicted by each of the three models. To obtain the GBM (OLS) deciles of predicted profit, I first predict a borrower's loan repayment using the $M(X)$ trained by GBM (OLS). Next, for each month, I sort all borrowers in the hold-out sample into deciles by their GBM (OLS) predicted repayment and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed repayment rate (i.e. profit). To obtain the Human predicted profit, I first regress observed loan size on the requested loan amount. Next, for each month, I sort all borrowers in the hold-out-sample into deciles by this residual and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed repayment rate (i.e. profit). Observed repayment rate (i.e. profit) is defined as (total repayment – loan size)/loan size.

Table 6
Explaining Human Mis-ranking

| Dependent = Misranking (H(X) vs. M(X)) | | |
|--|--------------------|--------------------|
| | Probit | |
| | (1) | (2) |
| Salience | 0.275*** (6.71) | 0.279*** (6.72) |
| R2 | 0.006 | 0.011 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |

Note: This table reports results from Probit regressions of the loan officer’s mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 1 decile away from her ranking by the machine learning model $M(X)$. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table 7
Explaining Human Mis-ranking (2)

| Dependent = High Misranking (H(X) vs. M(X)) | | |
|---|---------------------|---------------------|
| | Probit | |
| | (1) | (2) |
| Saliency | 0.391*** (12.30) | 0.392*** (12.49) |
| R2 | 0.014 | 0.020 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |

Note: This table reports results from Probit regressions of the loan officer's mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 5 deciles away from her ranking by the machine learning model $M(X)$. *Saliency* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table 8
Human Mis-ranking by Experience

| Dependent = Misranking (H(X) vs. M(X)) | | | | |
|--|--------------------|--------------------|--------------------|--------------------|
| | Low Experience | | High Experience | |
| | (1) | (2) | (3) | (4) |
| Salience | 0.189*** (5.32) | 0.192*** (5.16) | 0.372*** (5.39) | 0.377*** (5.47) |
| R2 | 0.003 | 0.011 | 0.011 | 0.013 |
| Officer Fixed | N | Y | N | Y |
| N | 6,776 | 6,776 | 6,776 | 6,776 |

| Dependent = High Misranking (H(X) vs. M(X)) | | | | |
|---|--------------------|--------------------|---------------------|---------------------|
| | Low Experience | | High Experience | |
| | (1) | (2) | (3) | (4) |
| Salience | 0.328*** (6.41) | 0.331*** (6.44) | 0.456*** (16.64) | 0.454*** (16.61) |
| R2 | 0.010 | 0.019 | 0.019 | 0.022 |
| Officer Fixed | N | Y | N | Y |
| N | 6,776 | 6,776 | 6,776 | 6,776 |

Note: The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by officers more than 1 decile away (Panel A) or 5 deciles away (Panel B) from her ranking by the machine learning model $M(X)$. Experience is measured by the total loan applications processed. A borrower falls into a High (Low) Experience sample if the application is processed by an officer with the above (below) median experience. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table 9
Salience and Soft Information Acquisition

| Dependent = Profit | | | | | | |
|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | Whole Sample | | Non-salient Sample | | Salient Sample | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Soft Info | 0.126*** (7.21) | 0.127*** (7.22) | 0.120*** (6.52) | 0.121*** (6.65) | 0.129*** (4.64) | 0.131*** (4.86) |
| Control | Y | Y | Y | Y | Y | Y |
| Time Fixed | Y | Y | Y | Y | Y | Y |
| Officer Fixed | N | Y | N | Y | N | Y |
| R2 | 0.055 | 0.056 | 0.056 | 0.057 | 0.102 | 0.102 |
| N | 33,879 | 33,879 | 26,113 | 26,113 | 7,766 | 7,766 |

Note: This table reports results from regressions of loan profit on soft information. The dependent variable is the loan repayment ratio, a measure of profit, defined as (total repayment – loan size)/loan size. Soft Info is $s_{i,t}$ estimated in equation (3) using the decomposition method in section 3.3.2. Control variables include all codified borrower characteristics. Column (1) and (2) are taken from Table 2. Column (3) and (4) are results on a subsample in which borrowers have no salient information. Column (5) and (6) are results on a subsample in which borrowers have at least one salient accounting variable. An accounting variable is defined as salient if its value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table 10
Combining Human Strength and Machine Strength

| Performance of $M(X)$ | | | | |
|-----------------------|--------------|-------------|---------------|------------|
| | (1) | (2) | (3) | (4) |
| | Whole Sample | Soft = High | Soft = Medium | Soft = Low |
| Profit | 0.215 | 0.22 | 0.201 | 0.189 |
| Training Sample | 33,879 | 10,293 | 11,864 | 11,722 |
| Testing Sample | 6,766 | 6,766 | 6,766 | 6,766 |

Note: This table presents the total profit for each model aggregated across all borrowers in the hold-out sample. Profit is defined as (total repayment – total loan size)/total loan size. Column (1) is taken from Table 4. Column (2) through (4) report profits from $M(X)$ trained on three subsamples divided by loan officers’ soft information production ability. Officer’s soft information production ability is measured by the coefficient of $s_{i,j,t}$ in an officer-by-officer regression that regresses loan repayment ratio on $s_{i,j,t}$, a full set of controls and time fixed effects (Figure 13).

Appendix to

**Assessing Human Information Processing in Lending Decisions: A Machine
Learning Approach**

Miao Liu

The University of Chicago, Booth School of Business

December 2019

Appendix A: Additional Figures and Tables

Appendix B: Results on Gender Bias

Appendix C: Proof of Proposition 1

Appendix D: Machine Learning Models in Details

D.1. OLS

D.2. Tree Models

D.2.1. Boosting

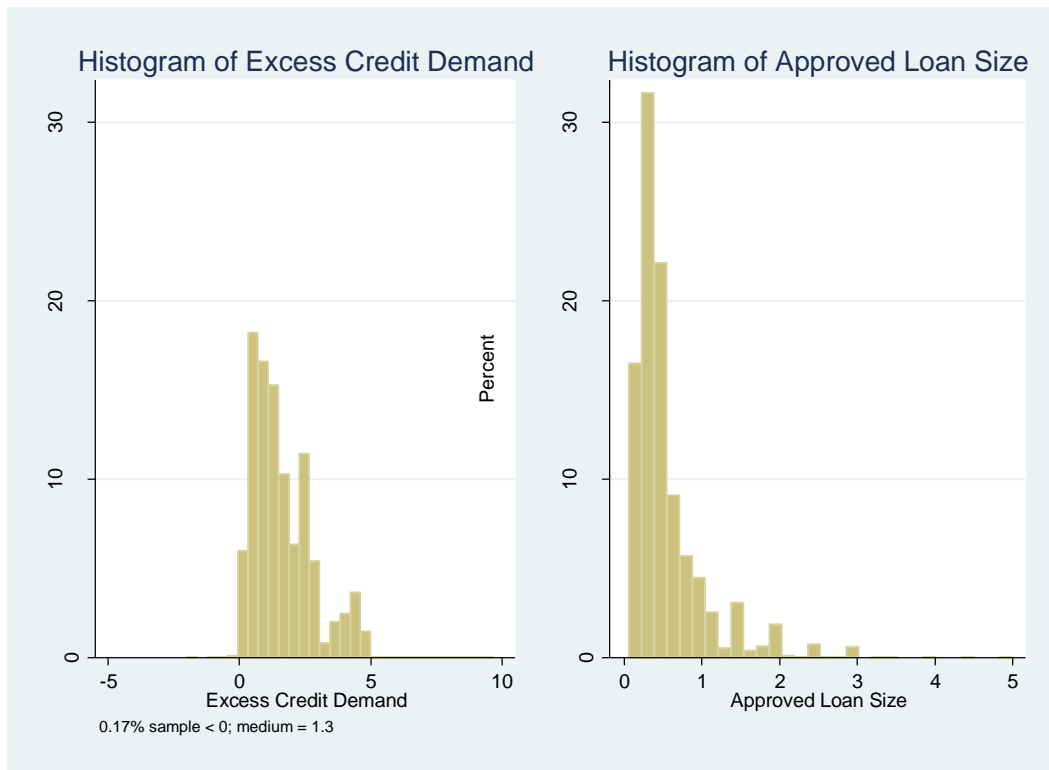
D.2.2. Random Forest

D.3. Neural Network

D.4. LASSO

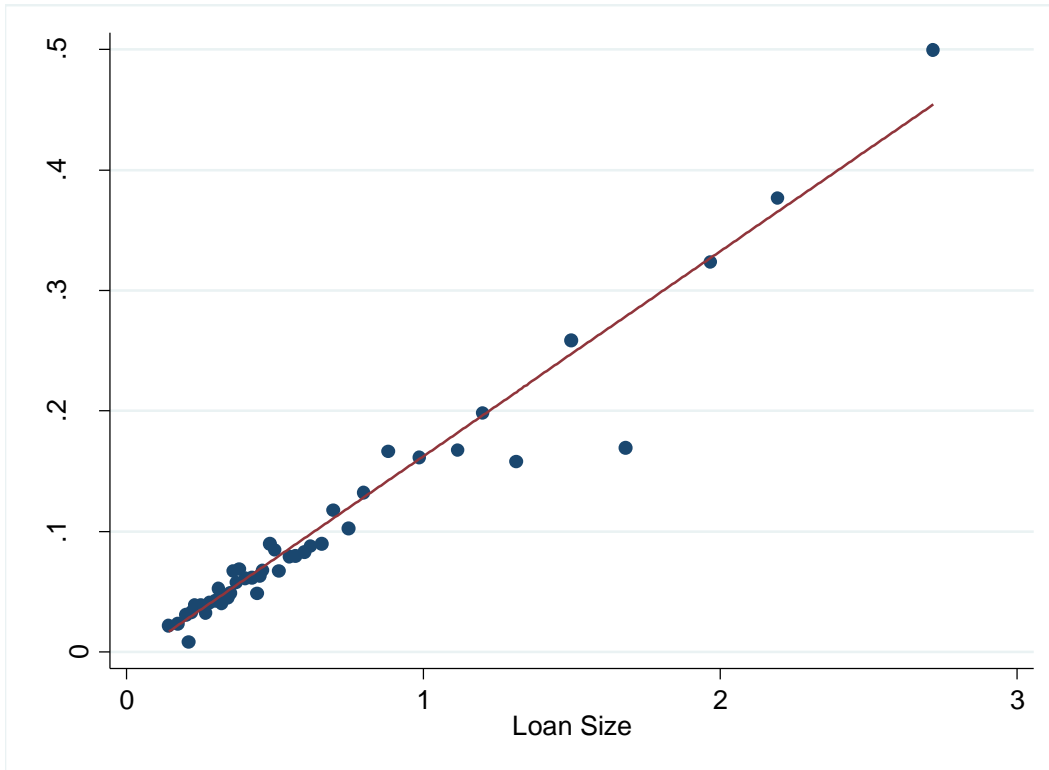
Appendix A: Additional Figures and Tables

Figure A1
Approved Loan Size and Excess Loan Demand



Note: This figure reports a histogram of excess demand, defined as the difference between requested and approved loan size, together with a histogram of actual approved loan size. The median excess credit demand (1.3) is greater than the 93rd percentile of the approved loan size.

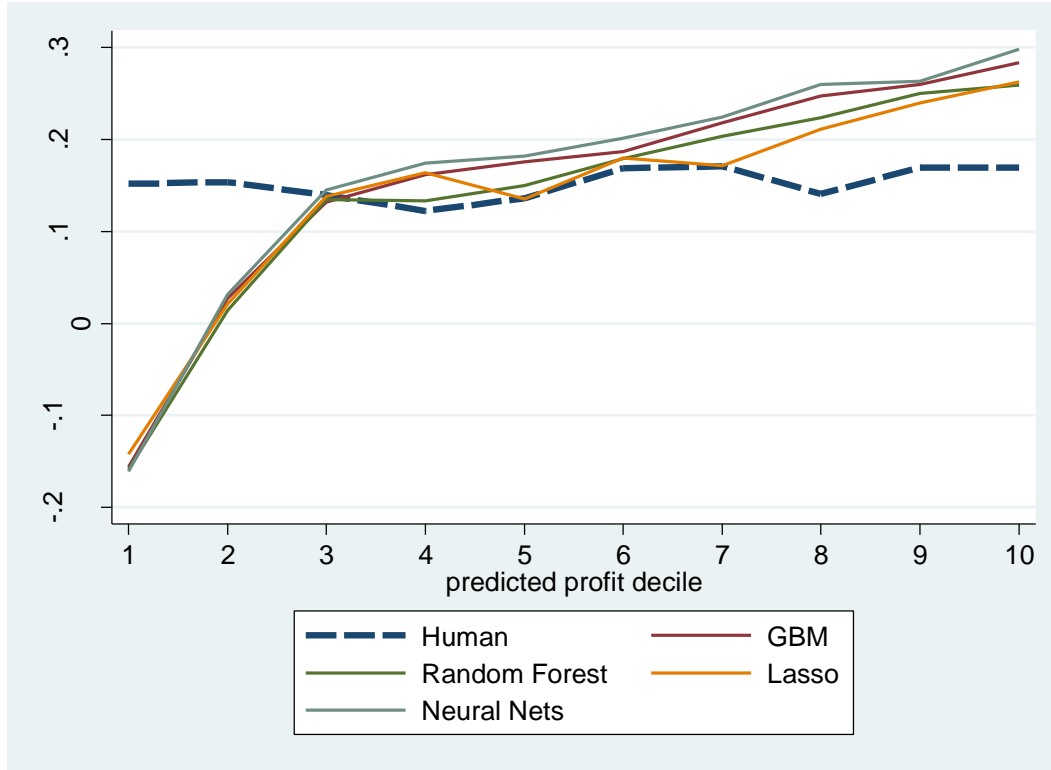
Figure A2
Linear Relation between Loan Size and Loan Profit



Note: This figure reports a binscatter of dollar profit against loan size. The dollar profit rate is defined as total loan repayment in dollars. It shows that the relationship between loan size and dollar profit, as specified in (4), is linear throughout the range of loan size.

Figure A3

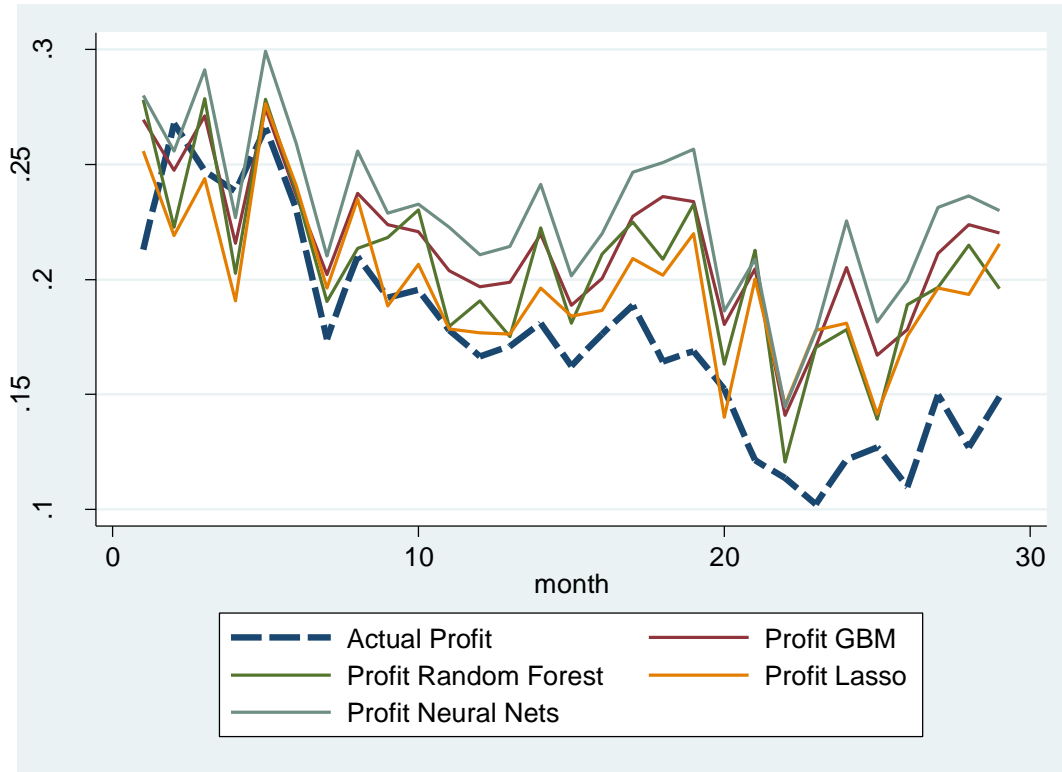
Observed Profit across Predicted Profit Deciles for Each Model



Note: The figure presents the relation between the average observed profit of loans in each predicted profit decile as predicted by each model. A model performs better if loans it predicts to be more profitable (and thus ranks higher) are indeed more profitable, indicating the model ranks borrowers closer to borrowers' true ranking. Therefore, a better performing model should have a more positive slope in the graph. For each of the four machine learning model (GBM, Random Forest, LASSO, and Neural Nets), I first predict a borrower's loan repayment (i.e., profit) using the $M(X)$ trained by the model. Next, for each month, I sort all borrowers in the hold-out sample into deciles by their model predicted profit and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed profit. The curve then tracks the average observed profit for each predicted profit decile as predicted by the model. To obtain the Human curve, I first regress observed loan size on the requested loan amount and keep the residual. Next, for each month, I sort all borrowers in the hold-out-sample into deciles by this residual and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed profit. The observed profit rate is defined as (total repayment – loan size)/loan size. Details of each machine learning model are given in appendix B.

Figure A4

Monthly Performance Comparison between Machine Learning and Loan Officers



Note: This figure compares loan officers' performance with each of the four machine learning model's performance month by month. The dashed curve is the observed monthly profit of loans (i.e., aggregate loan profit rate for each month). Profit is defined as $(\text{total repayment} - \text{loan size}) / \text{loan size}$. The solid curves are the out-of-sample monthly profits of the four machine learning models in month t trained with data up to month $t-1$. Details of each machine learning model are given in appendix B.

Table A1
The relation between Loan Size and Loan Offer Turndown

| Dependent = Turndown Indicator | | | |
|---------------------------------------|-----------------------|------------------------|-----------------------|
| | Probit | | |
| | (1) | (2) | (3) |
| Approved Loan Size (100,000 yuan) | -0.271*** (-10.76) | -0.292*** (-11.27) | -0.355*** (-12.75) |
| Requested Loan Size (100,000 yuan) | 0.164*** (19.33) | 0.173*** (20.38) | 0.150*** (17.58) |
| Controls | N | N | Y |
| Time Fixed | N | Y | Y |
| Officer Fixed | N | Y | Y |
| R2 | 0.02 | 0.03 | 0.05 |
| N | 40,087 | 40,087 | 40,087 |

Note: This table represents the results of the linear probability regression of a turndown indicator on approved loan size and requested loan size. Turndown equals one if a loan is approved but turned down by the borrower. Approved Loan Size is the loan amount approved by loan officers. Requested Loan Size is the loan amount requested by the borrower in the loan application. T-stats are in parentheses. Standard errors are clustered by the loan officer.

*** significance at 1% level

** significance at 5% level

Table A2**Model Performance: Human vs. Machine Learning**

| Model Performance | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|------------------------------|--------|----------|--------|--------|--------|--------|--------|
| (Profit Rate) | Human | Human | OLS | GBM | RF | LASSO | NNet |
| | $H(X)$ | $H(X)+S$ | $M(X)$ | $M(X)$ | $M(X)$ | $M(X)$ | $M(X)$ |
| Hold-out Sample (n=6,776) | 0.140 | 0.156 | 0.135 | 0.215 | 0.208 | 0.192 | 0.228 |

Note: This table reports profit generated by each model aggregated across all loans in the hold-out sample. Profit for each loan is defined as (total repayment – total loan size)/total loan size. In column (1), $H(X)$ represents the loan officers’ decision rule based on hard information only. In column (2), $H(X)+S$ represents loan officers’ observed decisions based on both hard and soft information. So 15.6% in column (2) is the lender’s actual profit. Column (3)-(7) are generated by machine decisions $M(X)$ trained with OLS, GBM, Random Forest, LASSO, and Neural Nets, respectively. Details of each machine learning model are given in appendix B.

Table A3**Model Performance: Human vs. Machine in the Tails**

| Predicted Profit Percentile | Average Observed Loan Profit Rate | | | | | |
|-----------------------------|-----------------------------------|------|-------|-------|-------|-------|
| | Human | OLS | GBM | RF | LASSO | NNet |
| 10% | 0.15 | 0.15 | -0.16 | -0.16 | -0.14 | -0.16 |
| 20% | 0.15 | 0.14 | 0.03 | 0.01 | 0.02 | 0.03 |
| 30% | 0.14 | 0.15 | 0.13 | 0.13 | 0.14 | 0.14 |
| 80% | 0.14 | 0.15 | 0.25 | 0.22 | 0.21 | 0.26 |
| 90% | 0.17 | 0.16 | 0.26 | 0.25 | 0.24 | 0.26 |
| 100% | 0.17 | 0.14 | 0.28 | 0.26 | 0.26 | 0.30 |

Note: The table presents the average observed profits in each tail decile of predicted profit as predicted by loan officers (Human) and each machine learning model(OLS, GBM, Random Forest, LASSO, and Neural Nets). For each of the machine learning models, to obtain the deciles of model predicted profit, I first predict a borrower’s loan repayment using the $M(X)$ trained by the model. Next, for each month, I sort all borrowers in the hold-out sample into deciles by their model predicted repayment and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed repayment rate (i.e. profit). To obtain the loan officer's (Human) predicted profit, I first regress observed loan size on the requested loan amount. Next, for each month, I sort all borrowers in the hold-out-sample into deciles by this residual and then pool borrowers in the same decile across months together. Finally, for each decile, I compute its average observed repayment rate (i.e. profit). Observed repayment rate (i.e. profit) is defined as (total repayment – loan size)/loan size. Details of each machine learning model is given in appendix B.

Table A4
Explaining Human Mis-ranking

| Panel A: Dependent = Over-ranking (H(X) vs. M(X)) | | |
|--|-----------------------|-----------------------|
| | Probit | |
| | (1) | (2) |
| Salience | -1.281*** (-23.83) | -1.288*** (-23.35) |
| R2 | 0.099 | 0.105 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |
| Panel B: Dependent = Under-ranking (H(X) vs. M(X)) | | |
| | Probit | |
| | (1) | (2) |
| Salience | 1.298*** (26.86) | 1.306*** (26.30) |
| R2 | 0.138 | 0.143 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |

Note: This table reports results from probit regressions of the loan officer's over-ranking (Panel A) and under-ranking (Panel B) of borrowers on behavioral factors. The dependent variable in Panel A is an indicator of over-ranking that equals one if a borrower is ranked by loan officers more than 1 decile above from her ranking by the machine learning model $M(X)$. The dependent variable in Panel B is an indicator of under-ranking that equals one if a borrower is ranked by loan officers more than 1 decile below from her ranking by the machine learning model $M(X)$. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table A5
Explaining Human High Mis-ranking

| Panel A: Dependent = High Over-ranking (H(X) vs. M(X)) | | |
|---|-----------------------|-----------------------|
| | Probit | |
| | (1) | (2) |
| Salience | -1.395*** (-11.47) | -1.413*** (-11.25) |
| R2 | 0.071 | 0.082 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |
| Panel B: Dependent = High Under-ranking (H(X) vs. M(X)) | | |
| | Probit | |
| | (1) | (2) |
| Salience | 1.238*** (25.89) | 1.252*** (25.82) |
| R2 | 0.171 | 0.177 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |

Note: This table reports results from probit regressions of the loan officer's high over-ranking (Panel A) and high under-ranking (Panel B) of borrowers on behavioral factors. The dependent variable in Panel A is an indicator of high over-ranking that equals one if a borrower is ranked by loan officers more than 5 decile above from her ranking by the machine learning model $M(X)$. The dependent variable in Panel B is an indicator of high under-ranking that equals one if a borrower is ranked by loan officers more than 5 decile below from her ranking by the machine learning model $M(X)$. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table A6**Explaining Human Mis-ranking Using Alternative Machine Learning Model (Random Forest)**

| Dependent = Misranking (Officer vs. ML(RF)) | | |
|---|--------------------|--------------------|
| | Probit | |
| | (1) | (2) |
| Salience | 0.292*** (7.11) | 0.296*** (7.12) |
| R2 | 0.006 | 0.011 |
| Officer Fixed | N | Y |
| | N | 6,776 |
| | 6,776 | 6,776 |

Note: This table reports results from probit regressions of the loan officer's mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 1 decile away from her ranking by the machine learning model $M(X)$. $M(X)$ is trained using Random Forest. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table A7**Explaining Human Mis-ranking Using Alternative Machine Learning Model (LASSO)**

| Dependent = Misranking (Officer vs. ML(Lasso)) | | |
|--|--------------------|--------------------|
| | Probit | |
| | (1) | (2) |
| Salience | 0.228*** (6.22) | 0.230*** (6.26) |
| R2 | 0.005 | 0.010 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |

Note: This table reports results from probit regressions of the loan officer's mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 1 decile away from her ranking by the machine learning model $M(X)$. $M(X)$ is trained using LASSO. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table A8**Explaining Human Mis-ranking Using Alternative Machine Learning Model (Neural Nets)**

| Dependent = Misranking (Officer vs. ML(Neural Nets)) | | |
|--|--------------------|--------------------|
| | Probit | |
| | (1) | (2) |
| Salience | 0.327*** (7.83) | 0.337*** (7.87) |
| R2 | 0.006 | 0.012 |
| Officer Fixed | N | Y |
| N | 6,776 | 6,776 |

Note: This table reports results from probit regressions of the loan officer's mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 1 decile away from her ranking by the machine learning model $M(X)$. $M(X)$ is trained using Neural Nets. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Appendix B: Results on Gender Bias

The theory of representativeness predicts that loan officers approve loan sizes too small for borrower groups with characteristics representative of high risk, because such characteristics exaggerate officers' perception of the risks. I focus on borrowers with salient information as one such group in the main body of the paper. Another borrower group potentially subject to representativeness bias consists of female borrowers. According to the Chinese Private Enterprises Survey (CPES), a national survey from 1993 to 2018, female entrepreneurs represent only around 10% of their sample. If gender stereotypes affect loan officers' expectations, as in the model of Bordalo et al. (2016), such that they do not perceive females as "typical entrepreneurs," they may underestimate the potential profitability of lending to women.

In this appendix, I test whether gender stereotypes together with salient information explain loan officers' mis-ranking of borrowers in the following regression. It is important to note that, as discussed in Dobbie et al. (2019), this regression is not powerful enough to distinguish between gender bias due to representativeness as in Bordalo et al. (2016) and that due to taste-based bias as in Becker (1957, 1993).

$$Mistranking_{ijt}^K = \beta_1 Gender_{ijt} + \beta_2 Salience_{ijt} + \varepsilon_{it}$$

$Mistranking_{ijt}^K$ is a 0-1 indicator variable turning on if officer j ranks borrower i more than K deciles away from machine's ranking. I consider $K=1$ and $K=5$. $Gender$ is an indicator that equals 1 if the borrower is a male. $Salience$ is an indicator equals 1 if the borrower has at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. I examine the relative importance of these factors in explaining human decision errors by testing their relative ability to explain variation in mis-ranking. Table B1 (for $K=1$) and Table B2 (for $K=5$) summarize the results.

Table B1 suggests that a male borrower is 6.8% less likely to be mis-ranked. This result is consistent with findings in other credit markets (Campbell et al. 2019; Dobbie et al. 2019). Salience is a much stronger factor to explain mis-ranking. Loan officers are 28.1% more likely to mis-rank when they observe salient

information in borrowers' accounting. In addition, salience is the dominant factor in explaining the R-squared of the variation in mis-ranking. Table B2 changes the outcome variable from $K=1$ to $K=5$ (i.e., a borrower is ranked by loan officers at least five deciles away from her machine ranking). All results are similar but more pronounced.

Table B3 tests if biases disappear with experience. As in Table 8, I split the sample into loans that are processed by officers with above and below median experience (i.e., 3,793 applications) and test whether experience affects biases. Similar to salience bias, gender bias does not disappear with experience.

In Table B4, I divide the sample by borrower's gender and test whether soft information acquisition is hampered by gender stereotypes. As benchmarks, I report in Columns (1) and (2) whole-sample results, taken from Table 2. Columns (3) to (6) show that loan officers acquire less soft information when the borrower is a female, suggesting that gender stereotypes impede soft information acquisition. Intuitively, talking to a borrower but starting with an incorrect belief might make the conversation less effective.

Table B1
Explaining Human Mis-ranking

| Dependent = Misranking (H(X) vs. M(X)) | | | | |
|--|---------------------|--------------------|---------------------|---------------------|
| | Probit | | | |
| | (1) | (2) | (3) | (4) |
| Gender=Male | -0.065** (-2.07) | | -0.069** (-2.19) | -0.068** (-2.18) |
| Salience | | 0.275*** (6.71) | 0.276*** (6.66) | 0.281*** (6.72) |
| R2 | 0.000 | 0.006 | 0.007 | 0.012 |
| Officer Fixed | N | N | N | Y |
| N | 6,776 | 6,776 | 6,776 | 6,776 |

Note: This table reports results from Probit regressions of the loan officer's mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 1 decile away from her ranking by the machine learning model $M(X)$. *Gender* is an indicator that equals 1 if the borrower is a male. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table B2
Explaining Human Mis-ranking (2)

| Dependent = High Misranking (H(X) vs. M(X)) | | | | |
|---|----------------------|--------------------|----------------------|----------------------|
| | Probit | | | |
| | (1) | (2) | (3) | (4) |
| Gender=Male | -0.109*** (-3.77) | | -0.121*** (-4.07) | -0.114*** (-3.85) |
| Salience | | 0.391*** (12.3) | 0.394*** (12.53) | 0.395*** (12.65) |
| R2 | 0.001 | 0.014 | 0.015 | 0.021 |
| Officer Fixed | N | N | N | Y |
| N | 6,776 | 6,776 | 6,776 | 6,776 |

Note: This table reports results from Probit regressions of the loan officer's mis-ranking of borrowers on behavioral factors. The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by loan officers more than 5 deciles away from her ranking by the machine learning model $M(X)$. *Gender* is an indicator that equals 1 if the borrower is a male. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Table B3
Human Mis-ranking by Experience

| Dependent = Misranking (H(X) vs. M(X)) | | | | |
|--|--------------------|--------------------|----------------------|----------------------|
| | Low Experience | | High Experience | |
| | (1) | (2) | (3) | (4) |
| Gender=Male | -0.050 (-0.92) | -0.046 (-0.87) | -0.090*** (-2.65) | -0.092*** (-2.57) |
| Salience | 0.188*** (5.27) | 0.192*** (6.71) | 0.375*** (5.39) | 0.379*** (5.46) |
| R2 | 0.004 | 0.011 | 0.012 | 0.013 |
| Officer Fixed | N | Y | N | Y |
| N | 6,776 | 6,776 | 6,776 | 6,776 |

| Dependent = High Misranking (H(X) vs. M(X)) | | | | |
|---|----------------------|----------------------|---------------------|---------------------|
| | Low Experience | | High Experience | |
| | (1) | (2) | (3) | (4) |
| Gender=Male | -0.128*** (-4.14) | -0.117*** (-3.88) | -0.114** (-2.14) | -0.111** (-2.05) |
| Salience | 0.331*** (6.60) | 0.334*** (6.62) | 0.460*** (15.70) | 0.458*** (15.73) |
| R2 | 0.011 | 0.020 | 0.021 | 0.023 |
| Officer Fixed | N | Y | N | Y |
| N | 6,776 | 6,776 | 6,776 | 6,776 |

Note: The dependent variable is an indicator of mis-ranking that equals one if a borrower is ranked by officers more than 1 decile away (Panel A) or 5 deciles away (Panel B) from her ranking by the machine learning model $M(X)$. Experience is measured by the total loan applications processed. A borrower falls into a High (Low) Experience sample if the application is processed by an officer with the above (below) median experience. *Gender* is an indicator that equals 1 for male borrowers. *Salience* is an indicator equals 1 if at least one accounting variable whose value falls into 5% in the left tail of the distribution of that variable across all borrowers. T-stats are in parentheses. Standard errors are clustered by loan officers.
*** significance at 1% level

** significance at 5% level

Table B4
Gender and Soft Information Acquisition

| Dependent = Profit | | | | | | |
|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | Whole Sample | | Borrower = Male | | Borrower = Female | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Soft Info | 0.129*** (7.24) | 0.128*** (7.22) | 0.153*** (7.74) | 0.135*** (7.09) | 0.112*** (3.64) | 0.104*** (3.69) |
| Control | Y | Y | Y | Y | Y | Y |
| Time Fixed | Y | Y | Y | Y | Y | Y |
| Officer Fixed | N | Y | N | Y | N | Y |
| R2 | 0.055 | 0.056 | 0.004 | 0.060 | 0.002 | 0.074 |
| N | 33,879 | 33,879 | 26,075 | 26,075 | 7,804 | 7,804 |

Note: This table reports results from regressions of loan profit on soft information. The dependent variable is the loan repayment ratio, a measure of profit, defined as (total repayment – loan size)/loan size. Soft Info is $s_{i,t}$ estimated in equation (3) using the decomposition method in section 3.2.2. Control variables include all codified borrower characteristics. Column (1) and (2) are taken from Table 2. Column (3) and (4) are results on a subsample with male borrowers. Column (5) and (6) are results on a female borrower subsample. T-stats are in parentheses. Standard errors are clustered by loan officers.

*** significance at 1% level

** significance at 5% level

Appendix C: Proof of Proposition 1

Denote $\beta = \frac{1}{1+\sigma_\varepsilon^2}$, and $\bar{\beta} = E(\beta) = \frac{1}{2H+1} + \frac{1}{2L+1}$. First, consider the case where the loan officer does not pay the cost C to resolve uncertainty in σ_ε^2 . In this case, the conditional mean of θ is

$$a = E(\theta | x)$$

By Law of Iterated Expectation, we have

$$\begin{aligned} a &= E(\theta | x) = E\{E(\theta | x, \sigma_\varepsilon^2) | x\} \\ &= E(\beta x | x) = \bar{\beta}x \end{aligned}$$

The loan officer's expected quadratic loss is

$$\begin{aligned} L &= E[(\theta - a)^2 | x] \\ &= E\{E[(\theta - a)^2 | x, \sigma_\varepsilon^2] | x\} \\ &= E\{E(\theta^2 - 2\theta\bar{\beta}x + \bar{\beta}^2x^2 | x, \sigma_\varepsilon^2) | x\} \\ &= E\{E(\beta^2x^2 + \text{Var}(\theta | x, \sigma_\varepsilon^2) - 2\bar{\beta}^2x + \bar{\beta}^2x^2 | x, \sigma_\varepsilon^2) | x\} \\ &= E\{\text{Var}(\theta | x, \sigma_\varepsilon^2)\} + x^2\text{Var}(\beta) \end{aligned}$$

Next, consider the loan officer's expected loss if she pays cost C to resolve uncertainty in σ_ε^2 .

$$\begin{aligned} L &= \frac{1}{2}E\left\{\left(\theta - \frac{1}{H+1}x\right)^2 | x\right\} + \frac{1}{2}E\left\{\left(\theta - \frac{1}{L+1}x\right)^2 | x\right\} + C \\ &= \frac{1}{2}\text{Var}(\theta | x, H) + \frac{1}{2}\text{Var}(\theta | x, L) + C \\ &= E\{\text{Var}(\theta | x, \sigma_\varepsilon^2)\} + C \end{aligned}$$

The proposition follows by comparing the expected losses for these two cases.

Appendix D: Machine Learning Models in Details

This section describes the collection of machine learning models that I use. I aim to provide a sufficiently in-depth description of the models so that a reader having no machine learning background can understand the basic model structure.

In its most general form, a borrower i 's future repayment can be written as an additive prediction error model:

$$Repay_{i,t+1} = E_t(Repay_{i,t+1}) + error_{i,t+1} \quad (D1)$$

where

$$E_t(Repay_{i,t+1}) = M(X_{i,t}) \quad (D2)$$

The objective is to find a representation of $E_t(Repay_{i,t+1})$ as a function of borrower characteristics $X_{i,t}$ that maximizes the out-of-sample explanatory power for realized $Repay_{i,t+1}$. Various machine learning models are deployed to estimate such representation $M()$. In the machine learning language, variables in $X_{i,t}$ are called predictors. As a useful benchmark, I start with the least complex method, the simple linear predictive regression model estimated via ordinary least squares (OLS).

D.1. OLS

The OLS model imposes that conditional expectations $M()$ can be approximated by a linear function of borrower characteristics $X_{i,t}$ and the parameter vector θ

$$M(X_{i,t}; \theta) = X'_{i,t}\theta \quad (D3)$$

This model does not allow for nonlinear effects or interactions between predictors. Denote N as the total

number of observations in the training sample, estimation of model D3 uses a standard least squares, or “L2”, objective function:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N (\text{Repay}_{i,t+1} - M(X_{i,t}; \theta))^2 \quad (\text{D4})$$

Minimizing $L(\theta)$ yields the familiar pooled OLS estimator.

D.2. Tree Models

Model (D3) does not captures individual borrower characteristics’ nonlinear and interactive impacts on $\text{Repay}_{i,t+1}$. One way to add non-linear and interactive effects is to include higher-order and interactive terms of $X_{i,t}$. However, multi-way interactions increase the parameterization combinatorially, leading the number of parameters to estimate quickly surpass the number of observations. Without a priori assumptions for which interactions to include, the linear model becomes computationally infeasible.

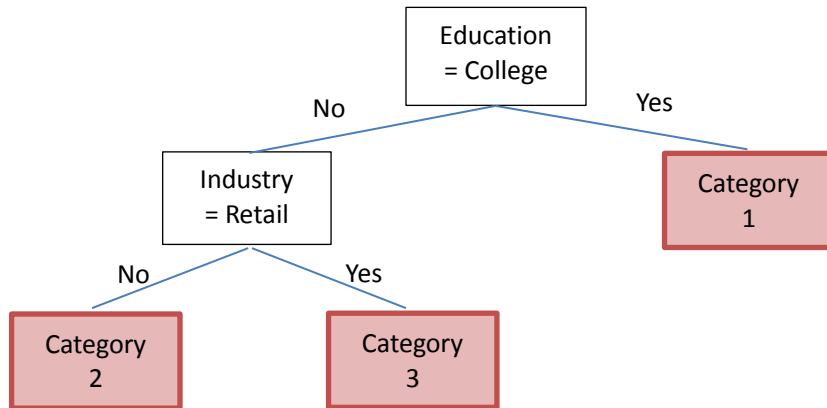


Figure D1: Regression Tree Example

Regression trees have become a popular machine learning approach for incorporating multi-way predictor interactions. Tree models are fully nonparametric. Figure 5, reproduced here, illustrates a basic building block of a decision tree, in which the data are divided through a sequence of binary splits. To predict

$Repay_{i,t+1}$, the first split might be a borrower's education (whether the borrower has a college degree). In the next step, we can split each of the two nodes created by that first split by different variables, such as industry (whether the borrower is in retail). Based on these two characteristics, the sample of borrowers is divided into three categories. The predicted repayment of a new borrower in the hold-out sample who falls into category i is set to equal to the average repayment of borrowers in the training sample in category i . This sequential procedure allows for a high degree of interactivity in the data, a key advantage of machine learning models over OLS.

Formally, the prediction of a tree, T , with K leaves (terminal nodes), and depth L (number of branches), can be written as

$$M(X_{i,t}; \theta, K, L) = \sum_{k=1}^K \theta_k \mathbf{1}_{\{X_{i,t} \in C_k(L)\}} \quad (D3)$$

where $C_k(L)$ is one of the K partitions of the data. Each partition is a product of up to L indicator functions of the predictors. The constant associated with partition k (denoted θ_k) is defined to be the sample average of outcomes within the partition. In the example of Figure D1, the prediction equation is

$$M(X_{i,t}; \theta, K, L) = \theta_1 \mathbf{1}_{\{\text{edu} \neq \text{college}\}} \mathbf{1}_{\{\text{ind} \neq \text{retail}\}} + \theta_2 \mathbf{1}_{\{\text{edu} \neq \text{college}\}} \mathbf{1}_{\{\text{ind} = \text{retail}\}} + \theta_3 \mathbf{1}_{\{\text{edu} = \text{college}\}}$$

To grow a tree is to find partitions that best discriminate among the potential outcomes. The specific predictor upon which a branch is based (education and industry in Figure D1), and the specific value where the branch is split (college degree and retail in Figure D1), is chosen to minimize forecast error. I follow the literature and use the algorithm of Breiman et al. (1984) with a L2 loss for each branch of the tree:

$$L(\theta, C) = \frac{1}{|C|} \sum_{X_{i,t} \in C} (Repay_{i,t+1} - \theta)^2 \quad (D4)$$

$|C|$ denotes the number of observations in partition C . Given C , the optimal choice of θ is $\theta =$

$\frac{1}{|C|} \sum_{x_{i,t} \in C} \text{Repay}_{i,t+1}$ (i.e., the average value of realized repayment in that branch).

The machine learning literature is much more concerned with over-fitting than the standard statistics or econometrics literature. Researchers attempt to select flexible models that fit well, but not so well that out-of-sample prediction is compromised. Regularization is an essential technique used in all machine learning models to achieve high performance in out-of-sample prediction by putting restrictions on model flexibility. That is, instead of directly optimizing an objective function, say minimizing the sum of squared residuals (e.g., $L(\theta, C)$ in (D4)), a term is added to the objective function to penalize the complexity of the model. The amount of regularization is governed by one or a handful of parameters, called tuning parameters, in the penalization term. The values of the tuning parameters are determined explicitly by the out-of-sample predictive performance using a data-driven model selection technique called cross-validation.

Decision trees are known to be prone to overfit, and therefore must be heavily regularized. I consider two popular regularizers that combine forecasts from many different trees into a single forecast.

D.2.1. Boosting

The first regularization method is “boosting”, which recursively combines forecasts from many over-simplified trees (Friedman et al, 2000; Friedman 2001). The boosting procedure, usually referred to as gradient boosted models (GBM), starts by fitting a shallow tree (e.g., with depth $L=1$). This over-simplified tree is a weak predictor with large bias in the training sample. Next, a second simple tree (with the same depth) is used to fit the prediction residuals from the first tree. Forecasts from these two trees are added together to form an ensemble prediction of the outcome, but the forecast component from the second tree is shrunk by a factor $v \in (0,1)$ to help prevent the model from overfitting the residuals. At each new step n , a shallow tree is fitted to the residuals from the model with $n-1$ trees, and its residual

forecast is added to the total with a shrinkage weight of v . This iteration is performed until there are a total of N trees in the ensemble. The final output is therefore an additive model of shallow trees with three tuning parameters (L, v, N). I choose these parameters using 5-fold cross-validation. Details of the implementation algorithm are described below.

Algorithm: Gradient Boosted Tree

Initialize the predictor as $M_0 = 0$;

for n from 1 to N do

 Compute for each $i=1,2,\dots,N$, the negative gradient of the loss function $L()$:

$$\varepsilon_{i,t+1} \leftarrow - \frac{\partial L(\text{Repay}_{i,t+1}, M)}{\partial M} \Big|_{M=M_{n-1}(X_{i,t})}$$

 Grow a (shallow) regression tree of depth L with dataset $\{(X_{i,t}, \varepsilon_{i,t+1}) : \forall i, \forall t\}$

$$f_n(\cdot) \leftarrow M(X_{i,t}; \theta, L)$$

 Update the model by

$$M_n(\cdot) \leftarrow M_{n-1}(\cdot) + v f_n(\cdot),$$

 where v is a tuning parameter that controls the step length

end

Result: The final model output is

$$M_N \left((X_{i,t}; v, N, L) \right) = \sum_{n=1}^N v f_n(\cdot).$$

D.2.2. Random Forest

Random forest is another ensemble method that combines forecasts from many different simplified trees (Brieman, 2001). The baseline procedure draws N different bootstrap samples of the data, fits a separate regression tree to each, then averages their forecasts. Trees for individual bootstrap samples tend to be deep and overfit, making their individual predictions inefficiently variable. Averaging over multiple

predictions reduces this variation, thus stabilizing the trees' predictive performance. The intuition is the same as in “wisdom of the crowds” where independent noises coming from each individual model are washed out in aggregate. Depth L of the trees and number of bootstrap samples N are the tuning parameters. I choose these parameters using 5-fold cross-validation. Details of the implementation algorithm are described below.

Algorithm: Random Forest

for n from 1 to N do

Generate Bootstrap samples $\{(X_{i,t}, \text{Repay}_{i,t+1}) : (i, t) \in \text{Bootstrap}(n)\}$ from the training dataset, for which a tree is grown. At each step of splitting, use only a random subsample of all features. Write the resulting n th tree as:

$$M_n(X_{i,t}, L, \theta_n) = \sum_{k=1}^{2^L} \theta_n^k \mathbf{1}_{\{X_{i,t} \in C_k(L)\}}$$

end

Result: The final model output is

$$M_N(X_{i,t}; N, L) = \frac{1}{N} \sum_{n=1}^N M_n(X_{i,t}, L, \theta_n).$$

D.3. Neural Networks

A neural network model consists of an input layer of raw predictors, one or more hidden layers that interact and nonlinearly transform the predictors, and an output layer that aggregates hidden layers into an ultimate outcome prediction. Analogous to axons in a biological brain, layers of the networks represent groups of “neurons” with each layer connected by “synapses” that transmit signals among neurons of different layers. Figure D2 shows two illustrative examples.

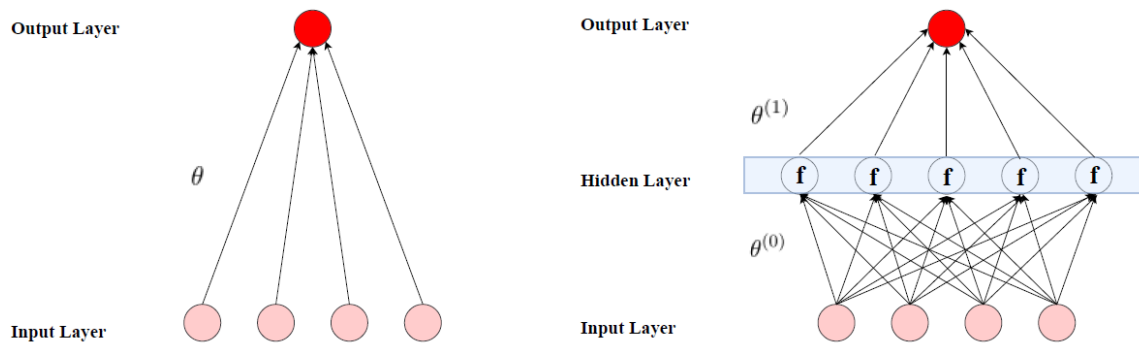


Figure D2: Neural Network Example

The number of units in the input layer is equal to the dimension of the predictors, which is set to four in this example. The left panel shows the simplest possible network that has no hidden layers. Each of the predictor signals is amplified or attenuated according to a five-dimensional parameter vector θ , that includes an intercept and one weight parameter per predictor. The output layer aggregates the weighted signals into the forecast $\theta_0 + \sum_{j=1}^4 X_j \theta_j$. That is, the simplest neural network is equivalent to a linear regression model.

The model can incorporate more flexible functional forms by adding hidden layers between the inputs and output. The right panel of Figure D2 shows an example with one hidden layer that contains five neurons. Each neuron draws information linearly from all of the input units, just as in the simple network on the left. Then, each neuron applies a nonlinear “activation function” f to its aggregated signal before sending its output to the next layer. For example, the second neuron in the hidden layer transforms inputs into an output as $X_2^{(1)} = f(\theta_{2,0}^{(0)} + \sum_{j=1}^4 X_j \theta_{2,j}^{(0)})$. Lastly, the results from each neuron are linearly aggregated into an ultimate output forecast:

$$M(X_{i,t}; \theta) = \theta_0^{(1)} + \sum_{j=1}^5 X_j^{(1)} \theta_j^{(1)}$$

Thus, in the right panel of Figure D2, there are a total of $31 = (4+1)*5+6$ parameters.

D.4. LASSO

The simple OLS model is bound to fail in the presence of many predictors. When the number of predictors approaches the number of observations, the linear model becomes inefficient or even inconsistent. It begins to overfit noise rather than extracting signal. A common machine learning approach for imposing parameter parsimony is to append a penalty to the OLS objective function in order to favor more parsimonious specifications. LASSO is one popular such penalized linear model with a loss function:

$$L(\theta; \cdot) = L(\theta) + \varphi(\theta; \cdot)$$

LASSO specifies a penalty function $\varphi(\theta; \cdot) = \lambda \sum_{j=1}^p |\theta_j|$. The tuning parameter λ is chosen using 5-fold cross-validation.

Reference

Becker, Gary S. 1957. *The Economics of Discrimination*. University of Chicago Press

Becker, Gary S. 1993. "Nobel Lecture: The Economic Way of Looking at Behavior." *Journal of Political Economy*, 101(3): 385–409.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics*, 131(4): 1753–1794.

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

Campbell, D., M. Loumiotis, and R. Wittenberg-Moerman (2019), "Making sense of soft information: Interpretation bias and loan quality," forthcoming, *Journal of Accounting and Economics*

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2019), "Measuring Bias in Consumer Lending." National Bureau of Economic Research, Working Paper.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, 2000, Additive logistic regression: A statistical view of boosting, *Annals of Statistics* 28, 337-374.

Friedman, Jerome H, 2001, Greedy function approximation: a gradient boosting machine, *Annals of statistics* 1189-1232.