Artificial Intelligence and Debt Collection: Evidence from a Field Experiment

Qingchen Wang and Yijun Zhou[*]

This Draft: July 2024

[*] Wang: No affiliation; Zhou: Baruch College (email: yijun.zhou@baruch.cuny.edu). We would like to thank Linda Allen, Kerry Back, Tania Babina, Anthony Cookson, Isil Erel, Antonio Gargano, Anna Pavlova, Lin Peng, Laura Veldkamp, S. Vish Viswanathan, Chaojun Wang, Jun Yang, Hong Zhang, Dexin Zhou, as well as the seminar and conference participants at Baruch College, WAPFIN@Stern 2022, NBER SI Digital Economics and Artificial Intelligence 2023, and NFA 2023 for their valuable comments and suggestions.

Artificial Intelligence and Debt Collection: Evidence from a Field Experiment

Abstract

This paper examines the role of artificial intelligence (AI) in facilitating the non-judicial collection process of delinquent consumer debt. Leveraging a randomized field experiment in the Netherlands, we show that algorithmic calling decisions achieve higher repayment rates with fewer collection calls compared with human collection officers. Uncovering the black box of AI, we find that it extracts predictive signals from unstructured notes compiled by collectors. These signals not only predict whether the delinquent borrowers would repay during the non-judicial collection process, but also shed light on the underlying motivations or impediments of delinquent borrowers' repayment behavior.

Keywords: Artificial Intelligence, Alternative Data, Debt Collection, Field Experiment
JEL Codes: D14, G02, G23, G51, O33.

I. Introduction

As of Q4 2023, US aggregate household debt balances have reached $17.5 trillion, with delinquency rates rising for nearly all types of debt (Federal Reserve Bank of New York 2023). While economists have long recognized the importance of preventing debt delinquency and have devoted substantial attention to the ex-ante screening and monitoring, the ex-post resolution of delinquent debt—particularly the debt collection process—has been relatively overlooked. Once debt becomes delinquent, creditors often outsource the collection process to third-party collection agencies. In their efforts to recover debt, these collectors employ a variety of tactics, ranging from early-stage non-judicial actions such as phone calls, letters, and text messages, to later-stage judicial actions such as lawsuits and wage garnishments.[1]

In this paper, we focus on the non-judicial collection process of delinquent consumer debt. While repayment can be enforced through legal actions that carry inherent enforcement power, it is intriguing to consider how non-judicial actions can also prompt repayment from previously delinquent borrowers. Existing theories suggest multiple, non-exclusive economic and behavior mechanisms that could underlie the effectiveness of non-judicial collection actions. In economic models (e.g., Kehoe and Levine 2001 and Chatterjee et al. 2007), households assess the perceived costs and benefits of repayments while subject to liquidity constraints.[2] Even if borrowers initially decide to default, their financial constraints may evolve, and their decisions may be influenced by the non-judicial collection process. Theoretical work dating back to Stigler (1961) has demonstrated that communications by motivated agents, referred to as persuasion, can provide

[1] See Zywicki (2015) for further discussions on the economics of the debt collection industry.

[2] Economically, making repayments involves monetary costs that can exacerbate consumers' financial distress, whereas paying off debt enhances consumers' credit profiles and future access to credit. Additionally, being in delinquency has been shown to incur non-pecuniary costs, such as moral dilemmas and decreased psychological well-being as shown by Brown, Taylor, and Price (2005) and Ong, Theseira, and Ng (2019).

information and result in changed behaviors (e.g., Kamenica and Gentzkow 2011). In this context, communications with collectors may alter delinquent borrowers' perceived costs/benefits of repayment.[3] Additionally, these communications may change consumers' behavior by directly entering their utility functions (e.g., Stigler & Becker 1977; and Becker & Murphy 1993).[4]

There are other factors in play as well. Attention and memory, recognized as limited cognitive resources, can explain many instances of consumers' seemingly irrational behavior such as neglecting key information or risks (e.g., Gabaix 2019; Bordalo, Gennaioli and Shleifer 2020). If inattention or limited memory impedes debt repayment, communications during the debt collection process can act as reminders, refocusing the borrower's attention on repaying the debt. Another factor contributing to borrowers' delinquency is present bias and the self-control problems (e.g., Laibson 1997; and O'Donoghue and Rabin 1999).[5] From this perspective, debt collectors may aid naive borrowers with present biases by enhancing their awareness, helping them set financial goals, and monitor their progress towards repayments via consistent communications.

Despite mechanisms that might motivate some delinquent borrowers to repay during the non-judicial collection process, as previously described, the efficiency of collections is often hindered by information asymmetries between consumers and collectors.[6] Specifically, due to the unobservable states of delinquent borrowers and significant heterogeneities among them, creditors and collectors struggle to distinguish those facing insolvency from those inclined towards

[3] For example, delinquent borrowers may not previously perceive the impact of having delinquent debt on their future credit but may become aware of it through communications. Alternatively, borrowers may update their estimation of the creditor/collector's likelihood of taking judicial actions to enforce repayment through repetitive collection efforts.

[4] The non-judicial collection process may impose social pressure on borrowers and increase their non-pecuniary costs of remaining in delinquency. For example, experimental evidence shows that communications can promote cooperation and pro-social behavior (e.g., Charness and Dufwenberg 2006; Bicchieri and Lev-On 2007).

[5] As modelled by Heidhues and Kőszegi (2010), naive consumers with self-control problems may over-borrow/over-consume, and it has also been shown empirically that consumers with present bias fail to stick to their repayment plans (Kuchler and Pagel 2021).

[6] Theories since Akerlof (1970) and Stiglitz and Weiss (1981) have examined the role of information asymmetries in credit markets. The prevalence and importance of these asymmetries have also been demonstrated empirically (e.g., Adams, Einav and Levin 2009; Karlan and Zinman 2009; Agarwal, Chomsisengphet and Liu 2010; Dobbie and Skiba 2013; Stroebel 2016; Gupta and Hansman 2022; DeFusco, Tang and Yannelis 2022; Vihriälä 2023).

repayment. It is challenging to discern the constraints and impediments faced by borrowers, let alone determine whether communications would make a difference. Consequently, creditors and collectors may exert unnecessary collection efforts on delinquent borrowers who are in financial distress and are unable to repay anyway, resulting in high collection costs and low repayment rates for creditors, as well as potential non-pecuniary costs for delinquent borrowers. Meanwhile, some delinquent borrowers with limited attention or self-control issues might be overlooked, despite the fact that more collection efforts on these borrowers would makes both the creditor and the borrowers better off.

In this paper, we examine the role of artificial intelligence (AI) in facilitating the non-judicial collection process. In the model of Drozd and Serrano-Padial (2017), debt collectors can use information technology to generate more precise signals about borrowers' financial state, thereby better allocating collection efforts. However, micro-level empirical evidence on how debt collectors utilize technology to reduce information asymmetries remains scarce. The emergence of big data in new forms—such as digital footprints (Berg et al. 2020), mobile usage data and social footprints (e.g., Agarwal et al. 2023), and lender comments on borrowers (Costello, Down and Mehta 2020)—introduces both opportunities and challenges in processing information within credit markets. Concurrently, rapid advances in machine learning algorithms have led to marked improvements in solving predictive problems, particularly through their superior handling of high-dimensional covariates in flexible functional forms (Mullainathan and Spiess 2017).[7] This backdrop sets the stage for investigating whether AI can be leveraged to address delinquent consumer debt and what insights AI can offer about delinquent consumers.

[7] Machine learning (ML) is generally considered as a subset of the broader category of artificial intelligence (AI). According to Jordan and Mitchell (2015), machine learning is "one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science".

We start with the historical non-judicial collection process of 36,031 debtors from a medium-size debt collection agency in the Netherlands, which we refer as historical data. These debtors are delinquent borrowers with uncollateralized financial services debt (e.g., credit card debt and buy-now-pay-later debt), and their collection process started from 2019 August and ended after 180 days or upon the borrower's repayment. During the non-judicial collection process, (human) collection officers determined the frequency and timing of calls for each borrower, with (human) calling agents executing these calls. The data contains not only information about the debt and the delinquent borrowers that the collector had, but also the interactions between the collector and borrowers during the informal collection process, including the unstructured notes taken by calling agents that had communications with the borrowers.

Using the historical data, we train machine learning algorithms to predict borrower's repayment likelihoods during the collection process. The algorithms demonstrated strong predictive performance, with AUCs ranging from 0.77 to 0.84.[8] In the out-of-sample testing sample, the decile with the highest predicted repayment likelihoods achieves an aggregate repayment rate of over 70%, while the decile with the lowest predicted repayment likelihoods ends with an aggregate repayment rate of below 10%.

Furthermore, we extend beyond merely predicting repayment likelihoods by using algorithms to decide which borrowers the collector should make follow-up calls with (referred to as *algorithmic calling decisions*). Specifically, we follow Künzel et al (2019) and train machine learning algorithms to estimate the heterogeneous effects of each call on each borrower's repayment.[9] Our validation analysis within the historical data supports the competence of

[8] The machine learning algorithms used here refer to the Gradient Boosted Decision Trees (Friedman 2001). Details regarding the construction of features and training of the predictive algorithms are described in Online Appendix Part B.

[9] The machine learning algorithms here refer to the meta-learners of Künzel et al (2019). Details regarding the use of meta-learners to estimate the heterogeneous effects of calls are discussed in Online Appendix Part C. Essentially, the effect of call is the treatment effect estimated as the difference in repayment with and without the call.

algorithms in predicting repayment likelihoods and identifying calls of high value to improve the repayment rate. Yet, within the confines of the historical data, we are limited to only observing the calls made by human collection officers and the subsequent repayment outcomes. We cannot observe the outcome of calls deemed valuable by algorithms but not executed by humans.

To thoroughly examine the effectiveness of algorithmic calling decisions, we leverage a randomized field experiment conducted by the same debt collection agency. Starting in June 2020, a new cohort of 7,839 delinquent borrowers with uncollateralized financial services debt entered the non-juridical collection process and were randomly assigned to two groups. The first group, consisting of 3,885 borrowers (referred to as the human group), receives calls determined by human collection officers. The second group, comprising of 3,954 borrowers (referred to as the AI group), receives calls based on algorithmic decisions. Specifically, trained algorithms are applied daily to predict the heterogeneous effects of each call on borrowers in the AI group and select those considered high value for execution. To test whether algorithmic decisions could help reduce contact frequency without compromising repayment rates, the collector intentionally imposes stricter limits on the number of calls to the AI group.

The experiment incorporates several features that help us isolate the performance of AI in deciding which borrowers to contact. First, the same team of calling agents, unaware of the experiment, conduct the calls for both groups. Each day, these agents are assigned a list of borrowers to call, without knowing whether the list is generated by human collection officers or algorithms. This ensures uniformity in the nature and contents of the communication between the borrowers and the collector across both groups. Additionally, human collection officers are also unaware of the experiment and their compensation structures have remained unchanged. Moreover, they retain access to all the raw data in the operations management system and can observe all the information about the debt and the borrowers, including the notes taken by calling agents. Therefore, AI is not equipped with more information; rather, we conjecture human officers are constrained by their capacity to process information, highlighting a key advantage of employing AI.

The experiment provides compelling evidence that algorithmic calling decisions lead to higher repayment rates with fewer collection actions. On average, borrowers in the AI group receive 3.50 calls, which is 0.32 fewer than the 3.82 calls received by borrowers in the human group. Despite this reduction in call frequency, the repayment rates—defined as the percentage of borrowers who repay their debt by the end of the collection process—increase to 53.24% in the AI group, compared to 43.14% in the human group. This increase of 10.1 percentage points corresponds to a 23.40% rise in repayment rates among delinquent borrowers. In this analysis, we are comparing the repayment rates of all delinquent borrowers in each group by the end of the non-judicial collection process. Since these borrowers were randomly assigned to either group, the only difference is whether the calls were determined by human collection officers or by algorithms. Consequently, these results suggest that AI is more adept at identifying calls that have a higher impact on borrowers' repayment outcomes than human collection officers.

To further understand the implications of algorithmic calling decisions, we also explore the potential distributional effects. We classify borrowers into five quintiles based on their credit profiles scored by the creditor from the lowest to the highest, and then compare the repayment rates between the AI group and the human group.[10] The results show that while algorithmic calling decisions consistently yield higher repayment rates across each credit score quintile, the impact is

[10] In the Netherlands, unlike in the U.S. where a numerical credit scoring system like FICO is used, individual credit profiles are managed by Bureau Krediet Registratie (BKR), which functions similarly to a credit bureau. BKR maintains records of all types of loans and credit agreements, including details on outstanding loans, overdue payments, defaults, bankruptcies, and foreclosures. The creditor of these borrowers is a financial services company and employs an internal scoring system to rate borrowers' credit profiles.

more pronounced for borrowers in the lowest three credit score quintiles. Specifically, the increase in percentage points resulting from algorithmic decisions follows an inverse U-shaped pattern; whereas, when assessing the relative increase as a percentage of the repayment rates in the human group, the pattern is monotonically decreasing. This finding suggests that AI may particularly excel over human decisions in settling delinquent debt for borrowers with lower credit profiles.[11]

Having established that AI can enhance aggregate repayment rates, we further explore the discrepancy between human and algorithmic calling decisions as the second step. We re-apply the algorithms daily to borrowers in the human group, selecting the same number of calls as the human collection officers, and label these as *AI-identified high-value calls*. We then examine the correlation between calls decided by human collection officers and those identified by AI as high value, finding interesting dynamics over time. While there is significant positive correlation between them, we observe that this correlation diminishes as the collection process progresses. Specifically, decisions made by human collection officers positively correlate with those of the algorithms during the first 90 days of the informal collection process, but this correlation becomes statistically insignificant (and even negative) thereafter. These findings suggest that it may become increasingly challenging for human officers to predict whether a borrower will repay and whether further communications will be effective as the borrower remains longer in the collection process without repayment. On the other hand, as more interactions occur between collectors and borrowers, AI may be able to extract more precise signals about the borrowers and act upon these signals effectively.

[11] On one hand, AI has the potential to assist disadvantaged groups with non-prime credit profiles in resolving delinquent debts, thereby promoting better financial health and broadening their access to future financial services. However, there is also a risk that AI decisions could deplete the liquidity of these individuals and further aggravate their financial hardship.

However, key questions remain: What signals does AI extract, and what do these signals reveal about the delinquent borrowers? During interactions with borrowers, calling agents typically take notes on the discussions, which could be considered as soft information gathered during the collection process. Each of these unstructured notes is transformed into a high-dimensional numerical vector using the bag-of-words method, with each word and phrase treated as an individual feature used by the algorithms. To demystify the black box of AI and gain deeper insights into the motivations or impediments behind borrowers' (non-)repayment behaviors, we analyze these high-dimensional features and identify three important categories.[12]

The first category includes words and phrases such as "difficulties," "job loss," and "unemployed," which we refer to as *expressions of financial hardship*. Empirically, we confirm a negative correlation between debt repayment and expressions of financial hardship. AI has considered this category as an important signal for predicting borrowers' repayment likelihoods. Moreover, when deciding which borrowers to contact, AI is also less likely to choose those noted for financial hardship, especially if the hardship is noted later in the collection process or if the borrower has a better credit profile. In contrast, while human collection managers are also less inclined to select calling borrowers with noted financial hardships, the magnitude of this effect is only half that observed with algorithmic decisions, and human officers do not typically consider the interaction between expressions of financial hardship and the borrower's credit profile.

The second category includes words and phrases such as "promise," "willing," and "commitment," which we consider as *expressions of repayment intent*. We find that borrowers who have expressed repayment intention in past communications, as noted by calling agents, are indeed more likely to repay their debt. Additionally, contacting these borrowers is associated with

[12] The procedures of identifying these three categories of features are discussed in Section IV.

additional positive effects on repayment. It is possible that these borrowers genuinely intend to repay but sometimes fail to do so due to inattention or limited self-control (e.g., Kuchler and Pagel 2021). In such instances, follow-up communications may help direct their attention back to their initial intentions and may have a monitoring effect on their progress towards repayment. Additionally, communications may further reinforce these borrowers' guilt aversion, encouraging them to keep their promises (e.g., Charness and Dufwenberg 2006). Interestingly, both algorithms and human collection officers are acting upon this signal when deciding which borrowers to call. Both are more likely to choose borrowers with noted repayment intents, though the effect of this signal on human collection officers is only one-ninth of that on the algorithms.

The third category of signals includes terms such as "impact," "consequences," and "credit profile," which we categorize to as *discussions of non-repayment consequences*. If borrowers have discussed the consequences of non-repayment during conversations with the collector, they are more likely to repay their debt. However, we find that calling these borrowers is associated with little additional effects on their likelihood of repayment. These findings support the informational role of the non-judicial collection process: once borrowers are informed of the non-repayment consequences through initial communications, further communications do not significantly impact their behavior.

Algorithmic decisions exhibit a consistent pattern. Discussions of non-repayment consequences do not significantly increase the likelihood that AI will decide to call these borrowers; in fact, they make AI less likely to call them if they have low credit profiles. In contrast, human collection officers are significantly more likely to call borrowers with noted discussions of non-repayment consequences. This finding indicates that while human collection officers recognize the positive correlation between past discussions on non-repayment consequences and future repayment likelihoods, they may not realize that calling these borrowers have little further impact. Under these circumstances, the excessive follow-up calls decided by human collection managers may be both economically and socially costly.

In summary, this final part of our analysis provides suggestive evidence on the mechanisms underlying the enhanced repayment achieved through algorithmic calling decisions. We demonstrate the capabilities of AI in extracting and utilizing information during the non-judicial debt collection process. Moreover, these signals not only capture the heterogeneity among delinquent borrowers but also shed light on the economic and behavioral motivations or impediments to their repayment behavior. While human collection officers also utilize these signals, significant differences exist between AI and human decisions in both the magnitude and direction of responsiveness to these signals, which may contribute to the higher repayment rates achieved by AI.

This paper connects to several strands of the literature. First, this paper relates to the line of literature that examines consumer credit markets, in particular, the repayment of debt by consumers (e.g., Agarwal, Liu and Souleles 2007; Bertrand and Morse 2011; Cadena and Schoar 2011; Karlan, Morten, and Zinman 2015; Laudenbach and Siegel 2018; Bursztyn et al. 2019; Gathergood et al. 2019; Keys and Wang 2019; Argyle, Nadauld and Palmer 2020; Cookson et al. 2022; Agarwal et al. 2023; Choi et al. 2024). In this paper, we focus on the repayment of delinquent debt in the ex-post collection process by third-party collectors. Few recent empirical papers examine the debt collection industry. Fedaseyeu (2020), Romeo and Sandler (2021), and Fonseca (2023) show that stricter regulations on debt collection reduces consumers' access to credit. A more micro analysis is provided by Cheng, Severino, and Townsend (2021) that link the court records of debt collection lawsuits in Missouri with credit data and they show that settlements increase financial distress relative to going to court. In contrast, this paper examines the non-judicial collection before any legal actions are initiated and we focus on the role of algorithms in making calling decisions during this process.[13] We provide micro-level evidence for Drozd and Serrano-Padial (2017)'s model in which information technology can reduce informational asymmetries and help collectors allocate collection efforts.

This paper also adds to the behavioral literature showing that individuals have limited attention, memory and self-control, and their behavior may be changed via reminders, persuasion, and social pressures (see DellaVigna 2009 for a review). Previous studies (e.g., Karlan et al. 2016) have shown that reminders are effective in increasing household savings while there are mixed results on the effects of reminders on repayment.[14] This paper shows that borrowers are heterogenous and calls from the collector may have varying impacts on them, which can help reconcile the mixed results of reminders on debt repayment. Moreover, our findings imply that algorithms are useful in capturing individual level heterogeneity and can potentially be applied to other behavioral interventions.

Last but not least, this paper contributes to the rising literature that examines artificial intelligence and big data in economics and finance (see Mullainathan and Spiess 2017, Goldstein, Spatt and Ye 2021, and Kelly and Xiu 2023, for reviews). Algorithms excel at uncovering patterns from data and their applications have been examined in various settings such as selecting human

[13] Choi et al. (2024) also examine the use of artificial intelligence in the collection of delinquent consumer debt. However, their focus is on the role of AI in executing phone calls to borrowers, and they find that AI callers perform worse than human callers in recovering debt. In contrast, this paper focuses on the role of AI in deciding which borrowers to call, with the actual calls being executed by human callers. We find that AI can improve calling decisions by learning from the decisions made by humans in the past.

[14] Cadena and Schoar (2011) show that reminders are as effective as financial incentives to increase loan repayment, and Medina (2021) also shows that reminders for upcoming credit card payments before the due date can reduce late payment. Laudenbach and Siegel (2018) find that phone calls made by bank agents to borrowers can help resolve delinquent loan. On the contrary, Karlan, Morten and Zinman (2015) find that reminders only increase repayment when they include the account officer’s name and only for clients serviced by the account officer previously, and Bursztyn et al. (2019) show that reminders sent to late-paying credit card holders eight days after missing their due date do not increase repayment if the reminders do not contain moral incentives. In a similar setting as ours, Holzmeister et al (2022) examine the collection of debt by a third-party collector, but find no effects of nudging interventions, including descriptive social norm nudges and (non-)deterrent information nudges in letters.

capital (Chalfin et al. 2016 and Erel et al. 2021), analyzing corporate financial information (Cao et al. 2021), and predicting consumers' creditworthiness (Agarwal et al. 2019; Berg et al. 2020; Tantri 2021; Fuster et al. 2022; Di Maggio, Ratnadiwakara, and Carmichael 2022). Focusing on the post-delinquency stage of credit markets, this paper employs algorithms to both make predictions regarding delinquent borrowers' repayment likelihoods and estimate the heterogenous treatment effects from follow-up calls. The algorithms are then compared with human collection officers in a field experiment, generating compelling evidence that algorithmic decisions lead to higher repayment with fewer collection actions. Furthermore, in alignment with Goldstein, Spatt and Ye's (2021) call for insights into the "psychology of machines," this paper explores the black box of AI and examines the signals used by algorithms. Finally, echoing the perspectives of Mullainathan and Obermeyer (2022) and Ludwig and Mullainathan (2024), we demonstrate that algorithms have the potential to help us understand human behavior.

The rest of the paper is organized as follows: Section II describes the setting and AI framework. Sections III presents the design of the field experiment and the main results. Section IV analyzes algorithmic calling decisions and human behavior. Finally, the paper concludes in Section V.

II. Setting and AI Framework

A. Debt Collection Industry

In Q4 2023, US aggregate household debt balances reached \$17.5 trillion, with 3.1% of outstanding debt in some stage of delinquency (Federal Reserve Bank of New York, 2023).[15] In

[15] According to the Household Debt and Credit Report by the Federal Reserve Bank of New York, delinquency transition rates increased for all debt types except for student loans in 2023Q4. Notably, about 8.5% of credit card balances (annualized) transitioned into delinquency.

the European Union, data from CEIC shows that household debt amounted to $7.16 trillion USD as of 2024 February, with delinquency rates varying significantly by debt type and country.[16] While creditors can use internal collection departments to collect delinquent debt, many choose to outsource the collection to third-party collectors as a common practice (Zywicki 2015). The third-party collector comprises the debt collection industry, which mainly deals with uncollateralized debt, such as healthcare, student loans, telecom and utilities debt, and financial services debt (e.g., credit card debt and buy-now-pay-later debt), and they are usually more than 90 days past due.

The debt collection process is usually divided into two main phases: the non-judicial phase and the judicial phase. During the non-judicial phase, the creditor or the debt collection agency attempts to recover the debt by contacting the borrower via phone calls, letters, and text messages to directly negotiate repayment. If the non-judicial phase does not result in repayment, the creditor or collection agency may proceed to the judicial phase by filing lawsuits and submitting claims to the court. If the court rules in favor of the creditor, the court may issue actions such as wage garnishment, seizure of assets, or other enforcement actions.[17,18]

The debt collectors in different regions are subject to the regulation of corresponding authorities and are generally prohibited from "inappropriate" collection actions, especially in the

[16] S&P Global reported that the total credit card delinquency rate in Continental Europe stood at 1.79% in the fourth quarter of 2023. A report by the European Banking Authority shows that 5.5% of EU banks' stock of consumer loans were non-performing as of September 2019.

[17] In the Netherlands, there is no numerical credit score system (such as FICO in the US). Instead, individuals have their credit profiles managed by Bureau Krediet Registratie (BKR), which functions similarly to a credit bureau. BKR maintains records of all types of loans and credit agreements for each individual, including outstanding loans, overdue payments, defaults, bankruptcies, and foreclosures. Negative records are typically retained for five years after the debt has been settled. However, once a debt is settled, the registration is updated to reflect this change, thereby potentially improving the individual's creditworthiness.

[18] If borrowers find themselves unable to repay their debt in the Netherlands, they can file for bankruptcy as in the US, and the court will then determine whether the borrower is indeed unable to meet their financial obligations. Once the bankruptcy is declared, a trustee will be appointed by the court to liquidate the bankrupt individual's assets to repay the debt and the individuals may be relieved of their remaining debt (but some types of debt may not be discharged). The public record of bankruptcy will also affect the individual's ability to obtain future credit. Alternatively, borrowers who are unable to pay their debts but wish to avoid the full impact of bankruptcy can apply for debt relief/restructuring programs under the Dutch law WSNP (Wet schuldsanering natuurlijke personen). These programs involve repayment plans approved by the court, during which the borrower must live on a budget determined by the trustee and the remainder income will be allocated towards the debt repayment. After a certain period (usually three years), the borrower may be discharged from the remaining debt.

US and Europe. For example, in the US, the Consumer Financial Protection Bureau (CFPB) is responsible for regulating the debt collection industry under the Fair Debt Collection Practices Act (FDCPA), which states that debt collectors cannot harass, oppress, or abuse you or anyone else they contact. The European Parliament's Committee on Economic and Monetary Affairs Draft Report in 2019 also includes provisions related to debt collection and aims to create a transparent and regulated environment for debt collection activities. In the Netherlands, the Dutch Civil Code (Burgerlijk Wetboek) is the primary legal framework that governs the debt collection. It includes requirements such as notifying the debt details and consequences to the borrower, and the principles of reasonableness and fairness on the collection procedures. The regulations are designed to ensure that the collection practices are conducted fairly, without undue harassment or intimidation, and borrowers also have the right to challenge the validity of the debt and they are protected from being charged of unreasonable collection costs.

While debt collection agencies operate under strict regulations, they continue to receive numerous consumer complaints, with excessive telephone calls during the collection process being a particularly common issue. Therefore, debt collectors have incentives to reduce the frequency of calls for both economic efficiency and compliance reasons.

B. Observational Data & Collection Process

In this paper, we collaborate with a debt collection company in the Netherlands, which we refer to as the debt collector. The collector first provides us with a set of 36,031 accounts (delinquent borrowers) from a single creditor, a global financial services company. The creditor delegated the collection of this batch of debt to the debt collector, who received a commission-based collection fee determined by the total amount of debt recovered, without considering the time value of money during the collection period. Additionally, the collector was not authorized to offer discounts on this batch of debt and did not charge additional fees to the borrowers during the collection process.

As agreed between the creditor and the collector, the collection process began with the collector sending a written notice to the borrowers. Subsequently, the collector could contact borrowers via phone calls, exercising discretion over the frequency and timing of these contacts. The collection process for this batch of debt began in August 2019 and was scheduled to last for 180 days from the first contact day, although it could end earlier if the borrower made a repayment.[19] During each day, the (human) collection officers at the agency determined which borrowers to contact, and the team of (human) calling agents carried out these calls.[20] The agents were incentivized to maximize the aggregate repayment amounts by their commission-based compensation structures. Throughout the process, calling agents followed up with borrowers primarily to remind them about their debt repayment, inform them of the consequences of non-repayment, and answer any questions regarding the specifics of the debt and repayment.

We have information about the debt and the entire 180-day collection process of the 36,031 borrowers.[21] The validity and accuracy of the debt have been verified. As reported in Table I Panel A, the average amount of debt among these borrowers was equivalent to $471.75 USD, with a median amount of $116.28 USD. The majority of the debt was comprised of small amounts, while the distribution was right-skewed. By the end of the 180-day collection process, 15,153 borrowers

[19] Repayment by the borrower, whether it be a full repayment or a partial repayment as part of a payment plan, immediately ended the collection process. Borrowers who have made a full repayment or entered into payment plans are subsequently managed by a post-payment team.

[20] In this collection agency, the total number of phone calls to be made to each batch of borrowers each day is first determined by considering the details of each batch of debt and the capacity of the calling team. Then, for each batch, one or more collection officers will select specific delinquent borrowers to call each day, depending on the size of the batch, with rotations of collection officers over time. Since we do not have information on the identity of each collection officer, we consider them as one representative collection officer.

[21] Debt collection agencies in Europe are regulated by the General Data Protection Regulation (GDPR) when handling personal data and are required to protect the privacy of individuals. The data compliance officer in the debt collection agency is responsible for overseeing the agency's adherence to GDPR requirements, including ensuring that all data processing activities are lawful, transparent, and secure. The data shared with us have undergone a process of anonymization, removing any personally identifiable information such as age, gender, race, income, or actual addresses.

(i.e., 42.55% out of all the 36,031 borrowers) had repaid their debt. On average, each borrower received 3.68 calls from the calling agents during the collection process.[22] This dataset is referred to as the historical data that we use to train machine learning algorithms.

C. AI Framework (1): Using Algorithms to Predict Repayment

As the first step, we employ algorithms to predict delinquent borrowers' repayment likelihoods. The technical details of selecting, training, and evaluating the predictive algorithms are provided in the online appendix. In brief, we use the historical data to train a Gradient Boosted Decision Trees algorithm (Friedman 2001) on daily basis during the collection process. The algorithms predict the likelihood of a delinquent borrower repaying the debt by the end of the 180-day period, based on information available to collectors as of each training day.

The list of features incorporated by the algorithms is detailed in online appendix Table A.I and can be classified into two groups. The first group contains *hard-information* features such as debt amount in collection, debt age, debt type, contact information type, the borrower's relationship length with the creditor, and their credit profile (provided by the creditor). These features are established at the start of the collection process and usually remain unchanged throughout. Additionally, we incorporate more features that are constructed during the collection process, such as the number of phone calls made, the duration of past calls, and borrower website logins. These features have also been integrated into the operations management system of the collection agency.

The second group comprises *soft-information* features. During interactions with borrowers, calling agents typically keep notes about the discussions. These notes, considered as soft information, are displayed as raw text in the operations management system. We employ the bag-

[22] It is possible that a borrower received multiple calling attempts on the same day if previous calls were not answered. The calling agents can make up to three attempts per day until the call is answered, but multiple attempts on the same day are considered as one call.

of-words method to transform each text into a high-dimensional numerical vector of word and phrase counts, with each word or phrase encoded as a distinct feature in the algorithms. While the bag-of-words approach ignores syntax and word order, it preserves the frequency of occurrences in the notes; we retain the 5,000 most commonly used words and phrases in the texts. There are various methods to extract information from unstructured text data (see Gentzkow, Kelly and Taddy 2019). Compared to approaches that convert the text into a single measure such as the sentiment score, the bag-of-words method retains a broader range of unprocessed information from the raw text, which can be further leveraged by the algorithms; Compared to more advanced techniques like Word2vec and large language models (LLMs) that transform text into high-dimensional numerical vectors, the bag-of-words approach maintains a one-to-one mapping from words and phrases to features in the vector space. This mapping allows us to later investigate which words or phrases are deemed important by the algorithms.

In training our models, we split the historical data into a training and test set to guard against overfitting and achieve good prediction performance, with out-of-sample AUCs ranging from 0.77 to 0.84.[23] In out-of-sample testing, the decile with the highest predicted repayment likelihoods achieves an aggregate repayment rate of over 70%, whereas the decile with the lowest predicted repayment likelihoods has an aggregate repayment rate below 10%. Although this paper does not aim to disentangle the power of machine learning algorithms from the power of data employed, our findings suggest the importance of soft-information features. Specifically, when repeating the training process with two versions of the algorithms—one incorporating soft-information features and the other only hard-information features—we observe a notably higher AUC for the model

[23] AUC (Area Under the Curve) is a widely used metric for evaluating predictive performance, such as in predicting credit scoring within finance literature (e.g. Berg et al. 2020). Generally sparking, an AUC of 0.6 is considered as satisfactory in environments with limited information, while in information-rich settings, targets of 0.7 or higher are desirable (Iyer et al. 2016).

that includes soft-information features. This indicates that soft information, when encoded in this manner, provides valuable signals that enhance the predictive accuracy of the algorithms.

D. AI Framework (2): Using Algorithms to Make Calling Decisions

Rather than solely predicting the repayment likelihoods of borrowers over time, we employ machine learning to determine which borrowers the collector should follow up with.[24] Specifically, we follow Künzel et al. (2019) and train meta-learners to estimate heterogeneous treatment effects. In this context, the "treatment" is equivalent to receiving the call on each day and the treatment effect is the difference in repayment likelihood with and without the call on that particular day. Detailed descriptions of the meta-learners method are available in the online appendix. By using this approach, we can estimate the heterogeneous effect of each call on an individual borrower's repayment likelihood on daily basis, which in turn can be leveraged to select the high-value calls for execution in the collection process. Since the total number of phone calls to be made to each batch of borrowers each day is determined by considering the specifics of each batch of debt and the capacity of the calling team, we take this as given and do not examine it in this paper due to the lack of information on other debts the collector is handling. Instead, we focus on the decision-making process for selecting borrowers within each batch each day.

According to Künzel et al. (2019), meta-learners are applicable for estimating heterogeneous treatment effect in observational studies and have shown comparably, if not better, performance

[24] The decision on whom to call differs significantly from predicting who will repay. For example, consider various types of borrowers. Some, upon receiving the initial contact at the start of the collection process, are likely to repay regardless of further communication—either because their attention has been drawn to repaying, or because they perceive the costs of non-repayment as high and the benefits of repayment as favorable. Conversely, other borrowers may remain unlikely to pay despite additional contact due to factors like binding liquidity constraints or a low perceived cost of delinquency. Continuing to call these two groups of borrowers can be economically inefficient and costly for both collectors and borrowers. However, for another group, additional communications may direct their attention to the debt or heighten their perceived benefits of repaying, thereby increasing their likelihood of repayment.

to the widely used causal forest method (Athey and Imbens 2016). It is important to clarify that this paper does not intend to compare the efficacy of different machine learning algorithms for estimating the heterogeneous treatment effects; rather, we use meta-learners as a representative of advanced machine learning techniques.

Nevertheless, due to the unobservable nature of the treatment effects, it is difficult to establish whether algorithms or human collection officers make better calling decisions. We discussed the challenges in more details in the online appendix, where we also present a simple validation analysis that supports the capability of machine learning to identify calls of high value on borrowers' repayment. In the following section, we leverage a randomized field experiment to evaluate the effect of algorithms in facilitating the debt collection process.

III. Field Experiment

A. Experimental Sample and Design

Starting in June 2020, the debt collection agency conducted a randomized field experiment with a new cohort of 7,839 delinquent borrowers from the same creditor (the same financial services company as in the historical data). As presented in Table I Panel B, the average amount of debt among these borrowers amounts to $675.48, with a median amount of $129.61. In line with the historical data, the distribution exhibits right-skewness, with the majority of the debt consisting of small amounts. The general collection process and agreed-upon collection fees between the creditor and the collector echo those in the historical data. The debt collector was delegated to contact delinquent borrowers via phone calls in the 180-day non-judicial collection process, during which the collector can contact borrowers via phone calls and exercise discretion concerning the frequency and timing of contact. The average number of contact calls made is 3.66, which is also similar to that in the historical data.

To test the effectiveness of the algorithms in making decisions on which borrowers to contact, the debt collector randomly assigned these 7,839 borrowers into two groups. The first group, consisting of 3,885 borrowers (referred to as the human group), receives calls determined by human collection officers as in the historical data. The second group, comprising of 3,954 borrowers (referred to as the AI group), receives calls based algorithmic decisions during the whole collection process. The algorithms, pre-trained using the meta-learners method with the historical data, predicted the heterogeneous effects of each call on borrowers in the AI group each day and selected those calls with the highest predicted value for execution.

To test whether algorithmic calling decisions could help reduce contact frequency without compromising repayment rates, the collector intentionally imposes stricter limits on the number of calls to the AI group. As indicated in Table I Panel C, the debt in the human group has an average balance of $680.46 and the debt in the AI group has an average amount of $670.58, with no significant difference between the two. Due to random assignment, there are no significant differences in other characteristics between the borrowers in two groups, such as their credit scores, the age of their debt, and so forth.

In this randomized experiment, the "treatment" is defined not merely by whether borrowers receive a call each day but by who decide those calls: borrowers in the AI group receive calls based on algorithmic decisions, whereas those in the human group (serving as the "control" group) receive calls determined by human collection officers. The experiment also incorporates several key features to mitigate other confounding factors. First, contacting/calling in both groups is made by the same team of calling agents who are not aware of the experiment. Each day, the calling agents are given a list of borrowers to contact, without the knowledge of whether the list is compiled by human collection officers or algorithms. This setup ensures that the nature and contents of the communication between the borrowers and the collector remain consistent across both groups. Additionally, human collection officers are also unaware of the experiment, and their compensation structures remain the same since the previous year. They continue to have access to all the raw data within the operations management system and can observe all the information about the debt and the borrowers, including the notes taken by calling agents. Therefore, AI is not equipped with more information compared with human collection officers, ensuring a balanced comparison.

B. Repayment Rates

Our primary outcome of interest is debt repayment. We observe whether and when the borrower repays his/her debt during the 180-day collection process, and examine whether algorithmic decisions can result in higher repayment. Given the random assignment of borrowers to receive collection calls decided by either human collection officers or algorithms, our identification strategy is straightforward using Equation (1):

$$Repayment_i = \alpha + \beta \times AI\ indicator_i + \gamma' \boldsymbol{X_i} + \epsilon_i \qquad (1)$$

Here, *Repayment*$_i$ is an indicator for borrower i having repaid his/her debt during the 180-day collection process. The AI indicator equals one for borrowers in the AI group subjected to AI-decided collection calls and zero for the borrowers in the human group receiving collection calls decided by human officers. The variables $\boldsymbol{X_i}$ include the characteristics of the debt and the borrower. The debt amount refers to the amount of debt assigned by the creditor for collection from the borrower during this collecting period. The credit score is an internal metric used by the creditor to measure the borrower's creditworthiness based on their credit profiles, with values ranging from

1 to 5. A higher score indicates better creditworthiness. The relationship score, also generated by the creditor, assesses the significance of the creditor's relationship with each borrower. The score is typically based on the duration of their relationship, the number of associated accounts, and their balances, ranging from 1 to 3. A higher score reflects a more valued relationship by the creditor.

Table II presents the results from estimating Equation (1). The dependent variables are indicators for repayment, which are equal to one if the borrower repaid the debt during the 180-day collection process and zero otherwise. The estimations in Columns (1) and (2) are based on the full sample. We start by presenting results from a regression without controls, representing raw repayment rates. Here, the repayment rate is 10.10 percentage points (pp hereafter) higher for borrowers in the AI group who are subject to AI-decided collection, compared to a baseline repayment rate of 43.14% for borrowers in the human group with human-decided collection. The difference in repayment rates is significant at the 1 percent level (p-value < 0.001). Compared to human-decided collection, AI-decided collection leads to higher repayment rates in the AI group, and the increase is both statistically and economically significant.

Upon adding control variables in Column (2), the results remain similar. Additionally, we find that the repayment rate is inversely correlated with the amount of debt, suggesting that delinquent borrowers with larger amount of debt are less likely to repay. On the other hand, the internal metrics generated by the creditor—the credit score and relationship score—are positively correlated with the repayment, indicating that borrowers with higher credit scores and relationship scores are more likely to repay their debt.

In Columns (3) to (6) of Table II, we analyze the repayment rates between human and AI groups based on the number of calls received by borrowers. We divide the borrowers into two subsamples: those who receive calls at or below the median number, three, and those who receive more than three calls. Specifically, Columns (3) and (4) present results for borrowers who receive three or fewer calls, while Columns (5) and (6) detail outcomes for those who receive four or more calls. The results show that borrowers subjected to AI-based calling decisions are more likely to repay in both subsamples. We also observe that the effect of AI decisions on increasing repayment rates is more pronounced among borrowers receiving more than three calls (11.29 percentage points in Column 5) compared to those receiving fewer calls (8.22 percentage points in Column 3). This difference is significant, especially considering that the repayment rate for the human group in the subsample receiving more than three calls stands at only 27.65%. These findings suggest that the likelihood of repayment significantly decreases if the borrower remains longer in the collection process without making a payment. In such scenarios, human collection officers may find it increasingly challenging to determine which borrowers to call effectively, whereas AI can better identify those for whom additional calls could effectively enhance repayment.

We also find suggestive evidence for the higher effectiveness of calls decided by algorithms, as presented in online appendix Table A.III. Specifically, we examine the correlation between collection calls received by borrowers each day and their final repayment in the human group and AI group separately. We find that daily calls, whether decided by human collection officers or AI, are positively correlated with the likelihood of repayment at the 1% significance level. Although these correlational coefficients cannot be interpreted as causal effects, the findings suggest that both human collection officers and algorithms are capable of predicting repayment and deciding collection calls to facilitate repayment. Moreover, the coefficients of the call indicator are larger in the AI group, indicating that algorithmic calling decisions are more effective than those made by human collection officers.

Additionally, given the substantial heterogeneity in debt amounts within our sample, we explored whether the effectiveness of algorithmic decisions varies by the size of the debt. We divide the borrowers into five quintiles based on their debt amounts, from the lowest and the highest, and analyze the difference between the AI group and human group within each quintile. We first observe that repayment rates generally decrease as the debt amount increases, irrespective of whether the collection calls were decided by AI or human officers. Specifically, the average repayment rate is 59.82% for debt in the lowest quintile with an average amount of $45.12; in contrast, the average repayment rate drops to 20.93% for debt in the highest quintile, with an average amount of $2818. More importantly, as shown in Figure I, borrowers in the AI group exhibit higher repayment rates across all quintiles of debt amounts, in comparison with those in the human group. This result indicates that the success of algorithmic collection decisions is not solely attributable to focusing on debt with smaller amounts; rather, AI appears to enhance repayment across a broad range of debt amounts.

Another interesting aspect explored is the time it takes for borrowers to repay in each group. Figure II represents the percentages of borrowers who have repaid their debt by days in collection, throughout the 180-day collection process. The blue line represents the cumulative repayment rates for the borrowers in the AI group, subject to AI-decided collection calls. The red line represents the cumulative repayment rates of borrowers in the human group, who are subject to collection calls decided by human collection officers. Figure II clearly reveals that the repayment rates are higher in the AI group than in the human group, and more importantly, the significant higher repayment rates in the AI group are obtained over time. Initially, there is no significant difference among the borrowers between the two groups within the first five days of collection. For example, 379 out of the 3954 borrowers in the AI group have repaid their debt by then compared to 369 out of the 3885 in the human group. However, from day 9 onwards, notable differences emerge. By the 10th day, 16.64% of borrowers in the AI group have repaid their debt, versus 15.65% in the human group.

After interactions between the borrowers and the collector increase, repayment rates in the AI group begin to significantly exceed those in the human group, amounting to over 10 percentage points difference by the end of the collection process. This pattern indicates that algorithmic collection decisions do not simply accelerate debt repayment by allocating more calls to those borrowers who are likely to repay earlier. Rather, AI-decided collection appears to facilitate more borrowers to repay over time by calling those borrowers with whom further contacts can yield higher repayment. We also notice that both lines begin to flatten after 100 days into the collection process, indicating a diminished likelihood of repayment as the collection period extends.

The dynamics of repayment align with findings from the online appendix Figure B.III, which shows that algorithms incorporating soft-information features outperform those without, and that the outperformance increases over time as more soft information is obtained and incorporated into the algorithms. Even though we do not differentiate between the power of the machine learning algorithms and the power of data employed, the initial insignificant difference in repayment rates between the human and AI groups and the significant divergence in repayment rates accumulated over time provide indicative evidence of the value of soft information collected by calling agents during interactions with borrowers. This also highlights the ability of AI in utilizing such information to improve debt recovery outcomes.

C. Calling Frequency

Next, we analyze the frequency of calls in both groups. As previously noted, the experiment design compels AI to select fewer collection calls. Table III presents the results of estimating Equation (2) to examine the number of calls allocated to borrowers in each group:

(2) $$Number\ of\ calls_i = \alpha + \beta \times AI\ indicator_i + \gamma' \boldsymbol{X_i} + \epsilon_i$$

In this equation, *Number of calls$_i$* represents the total number of collection calls a borrower *i* receives from the collector during the 180-day collection process. The AI indicator equals one for borrowers in the AI group who are subjected to AI-decided collection calls and zero for the borrowers in the human group who receive collection calls decided by human collection officers. The results show that borrowers in the AI group receive an average of 3.5 calls, compared to 3.82 calls for those in the human group.

To explore how algorithmic decisions manage to reduce the number of calls while simultaneously increasing repayment rates, we divide the borrowers into two subsamples based on whether the borrowers have repaid the debt during the collection process. Columns (3) and (4) report results from the subsample of borrowers who have not repaid their debt by the end of the collection (i.e., borrowers without repayment). Here, the coefficients of the AI indicator are negative and significant; borrowers without repayment receive 0.32 fewer calls under algorithmic decisions, compared to an average of 4.63 calls decided by human collection officers. Conversely, Columns (5) and (6) analyze the subsample of borrowers who have repaid their debt within the collection process (i.e., borrowers with repayment). The results show that these borrowers receive a similar number of calls under algorithmic decisions as the human collection officers. These results suggest that AI particularly excels at identifying the subset of borrowers who are unlikely to repay despite additional calls. Consequently, AI reduces the number of calls for these borrowers, thus lowering the associated costs without negatively impacting the overall repayment.

D. Distributional Effects

Having demonstrated that algorithmic calling decisions can increase repayment rates with fewer collection calls in the non-judicial process, we next explore potential distributional effects, considering that the impact of algorithms may vary across different groups of borrowers. Since the credit profile of households measures their current financial health and affects future credit access, we classify borrowers into five quintiles based on their credit profiles scored by the creditor from the lowest to the highest, and re-estimate Equation (1) within each quintile to compare repayment rates between the AI and the human groups.

The results are presented in Table IV. Panel A shows estimations without control variables, while Panel B shows estimations with control variables. We can first observe that repayment rates in the human group increase monotonically across the credit score quintiles—8.26%, 27.56%, 50.48%, 60.1%, and 68.64%, respectively. Meanwhile, we find that the coefficients of the AI indicator are positive and significant across all quintiles, showing that algorithmic call decisions consistently yield higher repayment rates compared with decisions made by human collection officers. The magnitude of increased repayment rates, in percentage points, follows an inverse U-shaped pattern: 5.29pp, 12.70pp, 13.92pp, 9.7pp, and 8.4pp from the lowest to highest credit score quintiles. When assessed as a relative increase compared to the repayment rates in the human group, the pattern is monotonically decreasing, with AI decisions enhancing repayment rates by 64.09%, 46.08%, 27.58%, 16.14%, and 12.23% from quintile 1 to quintile 5. This evidence suggests that AI is particularly effective at improving repayment outcomes for borrowers with lower credit profiles, outperforming human decisions more significantly among these groups.

Several hypotheses, not necessarily mutually exclusive, might explain the pronounced improvement of algorithmic collection decisions for borrowers with non-prime credit profiles. One possibility is the greater heterogeneity among these borrowers in terms of their likelihood of repayment and the responsiveness to collection calls, while algorithms are particularly adept at capturing this individual-level variability (Mullainathan and Obermeyer 2022). Additionally, non-prime borrowers often have less comprehensive credit information, posing challenges for human collection officers, while AI can process diverse information more effectively, thereby better identifying non-prime borrowers who are more likely to repay when contacted. Another potential explanation is the presence of biases—conscious or unconscious—in human collection officers towards borrowers with non-prime credit profiles, whereas AI, presumably impartial to such biases, can make collection decisions for non-prime borrowers with greater economic efficiency.

IV. Black box of AI

In the above analyses, we compare the repayment rates of all delinquent borrowers in the human and AI groups by the end of the non-judicial collection process. Since these borrowers were randomly assigned to either group, the only difference is whether the calls were determined by human collection officers or by algorithms. Consequently, these results suggest that AI is more adept at identifying calls that have a higher impact on borrowers' repayment outcomes than human collection officers.

Importantly, we are not claiming that AI is universally superior to all humans in making calling decisions within the debt collection process; and we are not claiming our approach is the only or most effective solution to improve human collection decisions. Rather, we demonstrate that algorithms, trained on historical data where calling decisions were made by the same group of human collection officers, can make better decisions compared to this specific group. This highlights the potential of algorithms to learn from past human decisions and subsequently make improvements in deciding future collection calls. The follow-up question then becomes: what are the differences between the decisions made by algorithms and those made by human collection officers?

A. AI-Identified High-Value Call

To explore the discrepancy between human and algorithmic decisions, we re-apply the algorithms daily to borrowers in the human group, requiring the algorithm to select the same number of calls as the human collection officers did. This ensures a fair comparison between algorithmic and human decisions. We label these calls selected by algorithms as *AI-identified high-value calls*, as they are predicted by the algorithms to have high value on borrowers' repayment. We then examine the correlation between calls decided by human collection officers and those identified by algorithms as high value by estimating the following Equation (3):

$$Human\ decided\ call_{i,t} = \alpha + \beta \times AI\ identified\ call_{i,t} + \gamma' \boldsymbol{X_{i,t}} + \epsilon_{i,t} \tag{3}$$

Human-decided call is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise. AI-identified call is an indicator equal to one if algorithms identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. Control variables include the number of days that the borrower has been in the non-judicial collection process, the amount of their debt, and the borrower's credit score and relationship score with the creditor.

The results are presented in Table V. Columns (1) and (2) are estimated using the full sample of daily-borrower observations. We find that the coefficient of AI-identified calls is positive and significant, with a t-statistic of 6.76. This indicates a significant correlation between the decisions of human officers and algorithms. When AI identifies a call as high value, the likelihood that human collection officers also select the call increases by 1.91 percentage points. This is economically significant, considering the unconditional likelihood of human officers selecting the call each day is approximately 2.2 percentage points.

However, we observe that these correlations decrease as the collection process progresses. As shown in Column (2), there is a negative interaction between AI-identified calls and the number of days in collection. Furthermore, we conduct estimations on two separate subsamples: Column (3) is based on daily-borrower observations from the first 90 days, while Column (4) is based on observations from the later 90 days. That is, decisions made by human collection officers significantly correlate with those of the algorithms during the first 90 days of the informal collection process, but this correlation becomes statistically insignificant thereafter.

Furthermore, consistent with our previous findings on the heterogeneous effects of AI on borrowers with different credit profiles, we observe that the correlations between algorithmic and human calling decisions also vary with the borrower's credit profile. Table VI presents the results from re-estimating Equation (3) for borrowers in each credit score quintile. The coefficients of AI-identified calls exhibit a monotonically increasing pattern from quintile 1 to quintile 4, indicating that the correlation between algorithmic and human decisions is higher for borrowers with better credit profiles. Specifically, the correlation for borrowers in credit score quintile 4 is three times larger than that for borrowers in quintile 2. A striking finding is that this correlation is negative for borrowers in quintile 1, suggesting significant divergence between the calling decisions made by AI and those made by human collection officers for borrowers with the lowest credit profiles. Possible explanations for this divergence include greater heterogeneity among non-prime borrowers, less comprehensive credit information, or the presence of biases—conscious or unconscious—in human decision-making.

B. AI Extracted Signals

The above analyses show that while there is an overall significant correlation between decisions made by algorithms and human collection officers, this correlation diminishes as the collection process progresses over time. One possibility is that predicting whether a borrower will repay and whether further collection calls will be effective becomes inherently more challenging as the borrower remains in the collection process without repayment. As shown in online appendix Figure B.III, the performance of algorithms without incorporating soft-information features also decreases over time, while the performance of algorithms that incorporate soft-information features gradually improves. These soft-information features are constructed from the unstructured notes taken by calling agents on past discussions.[25, 26] This suggests that as more interactions occur between collectors and borrowers, human collection officers and algorithms may utilize the newly available information differently, with algorithms appearing to use it more effectively.

However, key questions remain: What signals does AI extract, and what do these signals reveal about the delinquent borrowers? To demystify the "black box" of AI and gain deeper insights into the motivations or impediments behind borrowers' (non-)repayment behaviors, we analyze the soft-information features in three steps. First, we estimate the importance scores for these soft-information features.[27] While the complete list of words and phrases is not shared due to data privacy issues, we are provided with the top 100 words and phrases from the bag-of-words

[25] Indeed, these notes are based on the calling agents' judgments, which we cannot verify. However, our primary interest lies in understanding whether and how AI utilizes this information, particularly in comparison to human collection officers.

[26] As discussed in the online appendix part B, there are various methods to extract information from unstructured text data (see Gentzkow, Kelly and Taddy 2019). We convert each of the unstructured notes into a high-dimensional numerical vector of word and phrase counts using the bag-of-words method, with each word and phrase treated as an individual feature. Compared to approaches that convert the text into a single measure such as the sentiment score, the bag-of-words method retains a broader range of unprocessed information from the raw text, which can be further leveraged by algorithms; Compared to more advanced techniques like Word2vec and large language models (LLMs) that transform text into high-dimensional numerical vectors, the bag-of-words approach maintains a one-to-one mapping from words and phrases to features in the vector space. This mapping allows us to later investigate which words or phrases are deemed important by the algorithm.

[27] Gradient boosted decision trees measure the importance of features by "weight" (the number of times a feature appears in the trees), "gain" (the average gain of splits which use the feature), and "cover" (the average coverage of splits which use the feature).

vocabulary that algorithms consider most important. These 100 words and phrases are then converted into numerical vectors of semantic embeddings via fastText, a machine learning algorithm developed by Bojanowski et al. (2017). The vectorized words and phrases are further grouped into thematic clusters based on their semantic similarities using the spherical k-means algorithm of Dhillon and Modha (2001).[28]

After analyzing the thematic clusters identified by spherical k-means, three important categories of features emerge.[29] The first category includes words and phrases such as "difficulties," "job loss," and "unemployed," which we refer to as *expressions of financial hardship*. The second category includes words and phrases such as "promise," "willing," and "commitment" which we consider as *expressions of repayment intent*. The third category of signals includes terms such as "impact," "consequences," and "credit profile," which we categorize to as *discussions of non-repayment consequences*.[30]

We consider these as signals extracted by algorithms and first examine their relationship with debt repayment by borrowers. Specifically, we estimate Equation (4):

$$\text{(4)} \qquad Repayment_{i,t+} = \alpha + \beta_1 \times AI\ extracted\ signal_{i,t} + \beta_2 \times Human\ decided\ call_{i,t} + \beta_3 \times ML\ extracted\ signal_{i,t} \times Human\ decided\ call_{i,t} + \gamma' \boldsymbol{X_{i,t}} + \epsilon_{i,t}$$

$Repayment_{i,t+}$ is an indicator for whether borrower i has repaid their debt after day t within the 180-day collection process. $AI\ extracted\ signal_{i,t}$ is an indicator equal to one if the specific category of signals was noted for borrower i before day t, and zero otherwise. In Table VII, the

[28] The two-step approach of using algorithms to convert words into numerical vectors and group vectorized words into clusters based on semantic similarities has been applied in recent papers in Economics (e.g., Decarolis and Rovigatti 2021). The fastText algorithm, developed by Facebook's AI Research, can generate vector representations for words for different languages, including Dutch. The spherical k-means algorithm is also a widely used clustering algorithm that partitions data into k distinct, non-overlapping subsets or clusters, aiming to minimize the variance within each cluster. The result is k clusters where data within each cluster are more similar to each other than to those in other clusters.

[29] Following the "elbow method," we evaluated k across a range from 2 to 10, and found $k = 4$ to be optimal. Specifically, within cluster sum of squared distance to decrease rapidly up to $k = 4$, and flatten thereafter.

[30] The fourth cluster contains other words and phrases such as "credit card " and "account" that can hardly be interpreted without the specific context.

AI-extracted signals correspond to expressions of financial hardship in Column (1), expressions of repayment intent in Column (2), and discussions of non-repayment consequences in Column (3). Human-decided call is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise.

As shown in Column (1), there is a negative correlation between debt repayment and expressions of financial hardship. If a borrower has expressed financial hardship during past conversations, as noted by calling agents, the borrower is less likely to repay the debt. Additionally, expressions of financial hardship are associated with a negative, though not significant, impact on the positive effects of human-decided calls on repayment. This is intuitive because if a borrower is unable to repay due to a liquidity constraint, communications cannot change this constraint.[31]

In contrast to financial hardship, Column (2) shows that signals of repayment intent are positively associated with future repayment. Specifically, borrowers who have expressed repayment intent in past communications, as noted by the calling agents, are indeed more likely to repay their delinquent debt in the future. This supports the ability of calling agents to gather soft information when communicating with the borrowers, as seen in other settings (e.g., Hertzberg, Liberti and Paravisini 2010). This finding is also consistent with theories of social norms and behavioral compliance (e.g., Bicchieri 2005), indicating that individuals tend to fulfill their promises under social pressure. Furthermore, we find that calling these borrowers is associated with additional positive effects on repayment. It is possible that these borrowers genuinely intend to repay but sometimes fail to do so due to inattention or limited self-control (e.g., Kuchler and Pagel 2021). In such instances, follow-up communications may help direct their attention back to their initial intentions and may have a monitoring effect on their progress towards repayment.

[31] Alternatively, a sophisticated delinquent borrower might strategically express financial hardship to deceive calling agents and avoid future collection efforts; in such cases, communications are unlikely to change the sophisticated borrower's behavior either

Additionally, communications may further reinforce these borrowers' guilt aversion, encouraging them to keep their promises (e.g., Charness and Dufwenberg 2006).

In Column (3), the AI extracted signals relate to the discussion of non-repayment consequences during the collection process. We find that if a borrower has discussed the consequences of non-repayment in conversations with calling agents, the borrower is more likely to repay the debt. This aligns with the economic models of Kehoe and Levine (2001) and Chatterjee et al. (2007), where borrowers trade off the benefits and costs of non-repayment: if the borrower perceives the costs of being in delinquency to be higher, they are more likely to repay. Meanwhile, we find that calling these borrowers who discussed non-repayment consequences during past communications is associated with little additional effects on their repayment likelihoods. These findings support the informational role of communications between (at least some) delinquent borrowers and collectors: once borrowers are aware of the non-repayment consequences, additional communications have limited impact.

C. AI Extracted Signals and Calling Decisions of Algorithms and Human Officers

Finally, we compare how these signals correlate with the decisions of both algorithms and human collection officer. Specifically, we examine Equation (5):

$$Calling\ decision_{i,t} = \alpha + \beta_1 \times AI\ extracted\ signal_{i,t} + \gamma' \boldsymbol{X_{i,t}} + \epsilon_{i,t} \tag{5}$$

When examining decisions of algorithms, the calling decision is the previously defined AI-identified high-value call, which is an indicator equal to one if algorithms identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. When examining decisions of human collection officers, the calling decision is the previously defined human-decided call, which is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise. The AI-extracted signal is an indicator equal to one if the specific category of signals was noted for the borrower before that day, and zero otherwise.

Table VIII reports results from estimating Equation (5), where AI-extracted signal corresponds to expressions of financial hardship. The dependent variable is the AI-identified high-value call in Columns (1) to (3) and the human-decided call in Columns (4) to (6).

As shown in Column (1), the coefficient of expressions of financial hardship is negative and significant, with a t-statistic of −4.24. When deciding which borrowers to contact, the likelihood that AI identifies the call as high value decreases by 0.7 percentage points if the borrower has noted financial hardship. Considering the unconditional likelihood of human officers selecting the call is approximately 2.2 percentage points each day, this decrease is economically significant. Compared with the results in Column (4), we find that noted expressions of financial hardship are also negatively correlated with the likelihood that human collection officers decide to call the borrower each day, but the likelihood only decreases by 0.31 percentage points. The results suggest that if some of these delinquent borrowers with noted financial hardship are indeed in financial distress, AI would direct fewer collection efforts toward them, reducing both the collection costs for creditors and potentially non-pecuniary costs on delinquent borrowers, compared to human collection officers.

Table VIII Columns (2) and (5) include interactions between expressions of financial hardship and days in collection (i.e., the number of days the borrower has been in the collection process without repayment). The interaction effects are positive and significant, indicating that the negative correlations between noted expressions of financial hardship and calling decisions of both AI and human officers decrease over time. In Columns (3) and (6), the interactions between expressions of financial hardship and the borrower's credit score are included. The results show

that the negative correlation between noted expressions of financial hardship and AI's calling decisions reduces if the borrower has a higher credit score. In contrast, the coefficient of the interaction term is not significant in Column (6), suggesting that human collection officers do not typically consider the interaction between expressions of financial hardship and the borrower's credit profile when deciding whom to contact.

Regarding the control variables, we find that the debt amount significantly increases the likelihood that algorithms consider a call as high value, while the effect on human collection officers is positive but not significant. One interpretation is that algorithms adhere to the goal of maximizing total repayment amounts and are therefore more likely to consider calls associated with larger debt amounts as high value. In contrast, human collection officers may be deterred by the lower repayment likelihood associated with large debts. Furthermore, we observe that the borrower's credit score is positively correlated with calls decided by human officers, whereas the correlation with AI-identified high-value calls is not significant. This indicates that human collection officers' decisions are more directly influenced by borrowers' credit profiles compared to algorithmic decisions.

The second category of AI-extracted signals corresponds to expressions of repayment intent. As discussed earlier, borrowers who expressed repayment intent are more likely to repay their debt and contacting these borrowers is associated with additional positive effects on repayment. Table IX reports results from estimating Equation (5) with this category of signals. Columns (1) and (4) demonstrate that algorithms are more likely to identify calls to borrowers with noted repayment intents as high value, and human collection officers are also more likely to decide to call these borrowers, but to a much weaker extent. Economically, the effect of this noted repayment intent on human collection officers' calling decisions is only one-ninth that of algorithms. This suggests that human collection officers may either miss some of these signals or incorporate them to a lesser extent compared with AI. If some delinquent borrowers genuinely intend to repay but later fail due to limited attention or other behavioral factors, more follow-up calls as decided by AI may not only improve repayment rates for creditors but also help delinquent borrowers adhere to their repayment intent.

Furthermore, we consider whether the effects of expressions of repayment intent vary with time the borrower has been in the collection process without repayment. Table IX Column (2) shows a negative and significant coefficient for the interaction term between noted expressions of repayment intent and days in collection, suggesting that algorithms are more likely to identify calls to borrowers with noted repayment intents as high value, especially if the intent is expressed early in the collection process, while the effects decrease over time. In contrast, Column (5) shows that human collection officers do not differentiate between whether the intent is mentioned early or later in the collection process. On the other hand, the decisions of AI and human collection officers exhibit different patterns towards borrowers with varying credit profiles. As shown in Columns (3) and (6), the impact of expressions of repayment intent on the calling decisions of algorithms does not significantly vary with the borrower's credit score, while the positive effect on the calling decisions of human collection officers decreases as the borrower's credit score increases.

The third category of AI-extracted signals corresponds to discussions of non-repayment consequences. Table X examines the correlations between this category of signals and the calling decisions of algorithms and human collection officers. Column (1) shows that noted discussions of non-repayment consequences do not significantly increase the likelihood that AI will decide to call these borrowers on average. Considering the time-varying effects, as shown in Column (2), the correlations between noted discussions of non-repayment consequences and AI-identified high-value calls are positive and significant, but only in the early days of the collection process. Moreover, the results in Column (3) show that noted discussions of non-repayment consequences are negatively correlated with AI-identified high-value calls for borrowers with low credit scores.

In contrast, Columns (4) to (6) show that human collection officers are significantly more likely to call borrowers with noted discussions of non-repayment consequences, with an average effect of 1.91 percentage points each day. This magnitude is economically significant, considering the unconditional likelihood of 2.2 percentage points. The correlations between noted discussions of non-repayment consequences and human-decided calls decrease over time and decrease (though not significantly) with borrowers' credit scores. As previously discussed, if borrowers have discussed the consequences of non-repayment during conversations with the collector, they are more likely to repay their debt, but calling these borrowers is associated with little additional effects on their likelihood of repayment. These findings indicate that while human collection officers recognize the positive correlation between past discussions on non-repayment consequences and future repayment likelihoods, they may not realize that calling these borrowers has little further impact. Under these circumstances, the excessive follow-up calls decided by human collection officers may be both economically and socially costly.

Taken together, the analysis above shows that AI can extract informative signals from past interactions between the collector and borrowers, revealing significant heterogeneities among delinquent borrowers. Consequently, communications from the collector may serve entirely different roles for different delinquent borrowers. While human collection officers also utilize these signals, we observe significant differences between AI and human decisions in both the magnitude and direction of their responsiveness to these signals. These differences may help explain the higher repayment rates achieved by AI in the experiment.

V. Conclusion

In this paper, we examine the role of artificial intelligence (AI) in facilitating the non-judicial collection of delinquent consumer debt. While various mechanisms may motivate delinquent borrowers to repay during the non-judicial collection process, the efficiency of collection is often hindered by information asymmetries between consumers and collectors. In the model of Drozd and Serrano-Padial (2017), debt collectors can use information technology to generate more precise signals about borrowers and better allocate collection efforts. However, micro-level empirical evidence on how debt collectors utilize technology to reduce information asymmetries remains scarce.

This paper provides such evidence. Specifically, we show how algorithms can be trained with historical data to not only predict repayment likelihood but also estimate the heterogeneous effects of collection calls on borrowers' repayment behavior. The effectiveness of AI in making calling decisions during the non-judicial collection process has been established through a randomized field experiment: algorithmic calling decisions lead to higher repayment rates with fewer collection actions.

Furthermore, we explore the discrepancy between human and algorithmic calling decisions. We find that during the early stages of the collection process, there is a significant correlation between calling decisions made by human collection officers and those made by AI. However, this correlation diminishes as the collection process progresses, which may help explain the differences in the resulting repayment outcomes under human and algorithmic calling decisions. Unpacking the black box of AI, we show that AI can extract informational signals from unstructured notes compiled by calling agents during the process. These signals not only capture the heterogeneity among delinquent borrowers but also shed light on the economic and behavioral motivations or impediments to their repayment behavior.

Reference

Adams, William, Liran Einav, and Jonathan Levin, 2009, Liquidity constraints and imperfect information in subprime lending, *American Economic Review* 99, 49−84.

Agarwal, Sumit, Chunlin Liu, and Nicholas S. Souleles, 2007, The reaction of consumer spending and debt to tax rebates—evidence from consumer credit data, *Journal of political Economy* 115, 986−1019.

Agarwal, Sumit, Muris Hadzic, Changcheng Song, and Yildiray Yildirim, 2023, Liquidity constraints, consumption, and debt repayment: Evidence from macroprudential policy in Turkey, *Review of Financial Studies* 36, 3953−3998.

Agarwal, Sumit, Shashwat Alok, Pulak Ghosh, and Sudip Gupta, 2019, Financial inclusion and alternate credit scoring: Role of big data and machine learning in Fintech, *Indian School of Business.*

Agarwal, Sumit, Souphala Chomsisengphet, and Chunlin Liu, 2010, The importance of adverse selection in the credit card market: Evidence from randomized trials of credit card solicitations, *Journal of Money, Credit and Banking* 42, 743−754.

Akerlof, George, 1970, The arket for" lemons": qualitative uncertainty and the market mechanism, *Quarterly Journal of economics* 84, 488−500.

Argyle, Bronson S., Taylor D. Nadauld, and Christopher J. Palmer, 2020, Monthly payment targeting and the demand for maturity, *Review of Financial Studies* 33, 5416−5462.

Athey, Susan, and Guido Imbens, 2016, Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences* 113, 7353−7360.

Becker, Gary S., and Kevin M. Murphy, 1993, A simple theory of advertising as a good or bad, *Quarterly Journal of Economics* 108, 941−964.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri, 2020, On the rise of fintechs: Credit scoring using digital footprints, *Review of Financial Studies* 33, 2845−2897.

Bertrand, Marianne, and Adair Morse, 2011, Information disclosure, cognitive biases, and payday borrowing, *Journal of Finance* 66, 1865−1893.

Bicchieri, Cristina, and Azi Lev-On, 2007, Computer-mediated communication and cooperation in social dilemmas: an experimental analysis, *Politics, Philosophy & Economics* 6, 139−168.

Bicchieri, Cristina, 2005, *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017, Enriching word vectors with subword information, *Transactions of the association for computational linguistics* 5, 135−146.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, 2020, Memory, attention, and choice, *Quarterly Journal of Economics* 135, 1399−1442.

Brown, Sarah, Karl Taylor, and Stephen Wheatley Price, 2005, Debt and Distress: Evaluating the Psychological Cost of Credit, *Journal of Economic Psychology* 26 (5): 642–63.

Bursztyn, Leonardo, Stefano Fiorin, Daniel Gottlieb, and Martin Kanz, 2019, Moral incentives in credit card debt repayment: Evidence from a field experiment, *Journal of Political Economy* 127, 1641–1683.

Cadena, Ximena, and Antoinette Schoar, 2011, Remembering to pay? Reminders vs. financial incentives for loan payments, *NBER Working Paper* w17020.

Cao, Sean, Wei Jiang, Junbo L. Wang, and Baozhong Yang, 2021, From man vs. machine to man+ machine: The art and AI of stock analyses, *NBER Working Paper* w28800.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan, 2016, Productivity and selection of human capital with machine learning, *American Economic Review* 106, 124–127.

Charness, Gary, and Martin Dufwenberg, 2006, Promises and partnership, *Econometrica* 74, 1579–1601.

Chatterjee, Satyajit, Dean Corbae, Makoto Nakajima, and José-Víctor Ríos-Rull, 2007, A quantitative theory of unsecured consumer credit with risk of default, *Econometrica* 75, 1525–1589.

Cheng, Ing-Haw, Felipe Severino, and Richard R. Townsend, 2021, How do consumers fare when dealing with debt collectors? Evidence from out-of-court settlements, *Review of Financial Studies* 34, 1617–1660.

Choi, James J., Dong Huang, Zhishu Yang, and Qi Zhang, 2024, Better than human? Experiments with AI debt collectors.

Cookson, J. Anthony, Erik P. Gilje, and Rawley Z. Heimer, 2022, Shale shocked: Cash windfalls and household debt repayment, *Journal of Financial Economics* 146, 905–931.

Costello, Anna M., Andrea K. Down, and Mihir N. Mehta, 2020, Machine+ man: A field experiment on the role of discretion in augmenting AI-based lending models, *Journal of Accounting and Economics* 70, 101360.

Decarolis, Francesco, and Gabriele Rovigatti, 2021, From mad men to maths men: Concentration and buyer power in online advertising, *American Economic Review* 111, 3299–3327.

DeFusco, Anthony A., Huan Tang, and Constantine Yannelis, 2022, Measuring the welfare cost of asymmetric information in consumer credit markets, *Journal of Financial Economics* 146, 821–840.

DellaVigna, Stefano, 2009, Psychology and economics: Evidence from the field, *Journal of Economic literature* 47, 315–372.

Dhillon, Inderjit S., and Dharmendra S. Modha, 2001, Concept decompositions for large sparse text data using clustering, *Machine learning* 42, 143–175.

Di Maggio, Marco, Dimuthu Ratnadiwakara, and Don Carmichael, 2022, Invisible primes: Fintech lending with alternative data. *NBER Working Paper* w29840.

Dobbie, Will, and Paige Marta Skiba, 2013, Information asymmetries in consumer credit markets: Evidence from payday lending, *American Economic Journal: Applied Economics* 5, 256–282.

Drozd, Lukasz A., and Ricardo Serrano-Padial, 2017, Modeling the revolving revolution: the debt collection channel, *American Economic Review* 107, 897–930.

Erel, Isil, Léa H. Stern, Chenhao Tan, and Michael S. Weisbach, 2021, Selecting directors using machine learning, *Review of Financial Studies* 34, 3226–3264.

Fedaseyeu, Viktar, 2020, Debt collection agencies and the supply of consumer credit, *Journal of Financial Economics* 138, 193–221.

Fonseca, Julia, 2023, Less mainstream credit, more payday borrowing? Evidence from debt collection restrictions, *Journal of Finance* 78, 63–103.

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

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, 2022, Predictably unequal? The effects of machine learning on credit markets, *Journal of Finance* 77, 5–47.

Gabaix, Xavier, 2019, Behavioral inattention, In *Handbook of behavioral economics: Applications and foundations 1*, vol. 2, pp. 261-343. North-Holland.

Gathergood, John, Neale Mahoney, Neil Stewart, and Jörg Weber, 2019, How do individuals repay their debt? The balance-matching heuristic, *American Economic Review* 109, 844–875.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as data, *Journal of Economic Literature* 57, 535–574.

Goldstein, Itay, Chester S. Spatt, and Mao Ye, 2021, Big data in finance, *Review of Financial Studies* 34, 3213–3225.

Gupta, Arpit, and Christopher Hansman, 2022, Selection, leverage, and default in the mortgage market, *Review of Financial Studies* 35, 720–770.

Heidhues, Paul, and Botond Kőszegi, 2010, Exploiting naivete about self-control in the credit market, *American Economic Review* 100, 2279–2303.

Hertzberg, Andrew, Jose Maria Liberti, and Daniel Paravisini, 2010, Information and incentives inside the firm: Evidence from loan officer rotation, *Journal of Finance* 65, 795–828.

Holzmeister, Felix, Jürgen Huber, Michael Kirchler, and Rene Schwaiger, 2022, Nudging debtors to pay their debt: Two randomized controlled trials, *Journal of Economic Behavior & Organization* 198, 535–551.

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

Jordan, Michael I., and Tom M. Mitchell, 2015, Machine learning: Trends, perspectives, and prospects, *Science* 349, 255–260.

Kamenica, Emir, and Matthew Gentzkow, Bayesian persuasion, *American Economic Review* 101, 2590–2615.

Karlan, Dean, and Jonathan Zinman, 2009, Observing unobservables: Identifying information asymmetries with a consumer credit field experiment, *Econometrica* 77, 1993–2008.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman, 2016, Getting to the top of mind: How reminders increase saving, *Management Science* 62, 3393–3411.

Karlan, Dean, Melanie Morten, and onathan Zinman, 2015, A personal touch in text messaging can improve microloan repayment, *Behavioral Science & Policy* 1, 25–31.

Kehoe, Timothy J., and David K. Levine, 2001, Liquidity constrained markets versus debt constrained markets, *Econometrica* 69, 575–598.

Kelly, Bryan, and Dacheng Xiu, 2023, Financial machine learning, *Foundations and Trends® in Finance* 13, 205–363.

Keys, Benjamin J., and Jialan Wang, 2019, Minimum payments and debt paydown in consumer credit cards, *Journal of Financial Economics* 131, 528–548.

Kuchler, Theresa, and Michaela Pagel, 2021, Sticking to your plan: The role of present bias for credit card paydown, *Journal of Financial Economics* 139, 359–388.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu, 2019, Metalearners for estimating heterogeneous treatment effects using machine learning, *Proceedings of the National Academy of Sciences* 116, 4156–4165.

Laibson, David, 1997, Golden eggs and hyperbolic discounting, *Quarterly Journal of Economics* 112, 443–478.

Laudenbach, Christine, and Stephan Siege, 2018, Personal communication in an automated world: Evidence from loan repayments, *Journal of Finance, forthcoming*.

Ludwig, Jens, and Sendhil Mullainathan, 2024, Machine learning as a tool for hypothesis generation, *Quarterly Journal of Economics* 139, 751–827.

Mullainathan, Sendhil, and Jann Spiess, 2017, Machine learning: an applied econometric approach, *Journal of Economic Perspectives* 31, 87–106.

Mullainathan, Sendhil, and Ziad Obermeyer, 2022, Diagnosing physician error: A machine learning approach to low-value health care, *Quarterly Journal of Economics* 137, 679–727.

O'Donoghue, Ted, and Matthew Rabin, 1999, Doing it now or later, *American Economic Review* 89, 103–124.

Ong, Qiyan, Walter Theseira, and Irene YH Ng, 2019, Reducing debt improves psychological functioning and changes decision-making in the poor, *Proceedings of the National Academy of Sciences* 116, 7244–49.

Romeo, Charles, and Ryan Sandler, 2021, The effect of debt collection laws on access to credit, *Journal of Public Economics* 195, 104320.

Stigler, George J, 1961, The economics of information, *Journal of Political Economy* 69, 213–225.

Stigler, George J., and Gary S. Becker, 1977, De gustibus non est disputandum, *American Economic Review* 67, 76–90.

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

Stroebel, Johannes, 2016, Asymmetric information about collateral values, *Journal of Finance* 71, 1071–1112.

Tantri, Prasanna, 2021, Fintech for the poor: Financial intermediation without discrimination, *Review of Finance* 25, 561–593.

Vihriälä, Erkki, 2023, Self-imposed liquidity constraints via voluntary debt repayment, *Journal of Financial Economics* 150, 103708.

Zinman, Jonathan, 2015, Household debt: Facts, puzzles, theories, and policies, *Annual Review of Economics* 7, 251–276.

Zywicki, Todd J, 2015, The law and economics of consumer debt collection and its regulation, *Loy. Consumer L. Rev.* 28: 167.

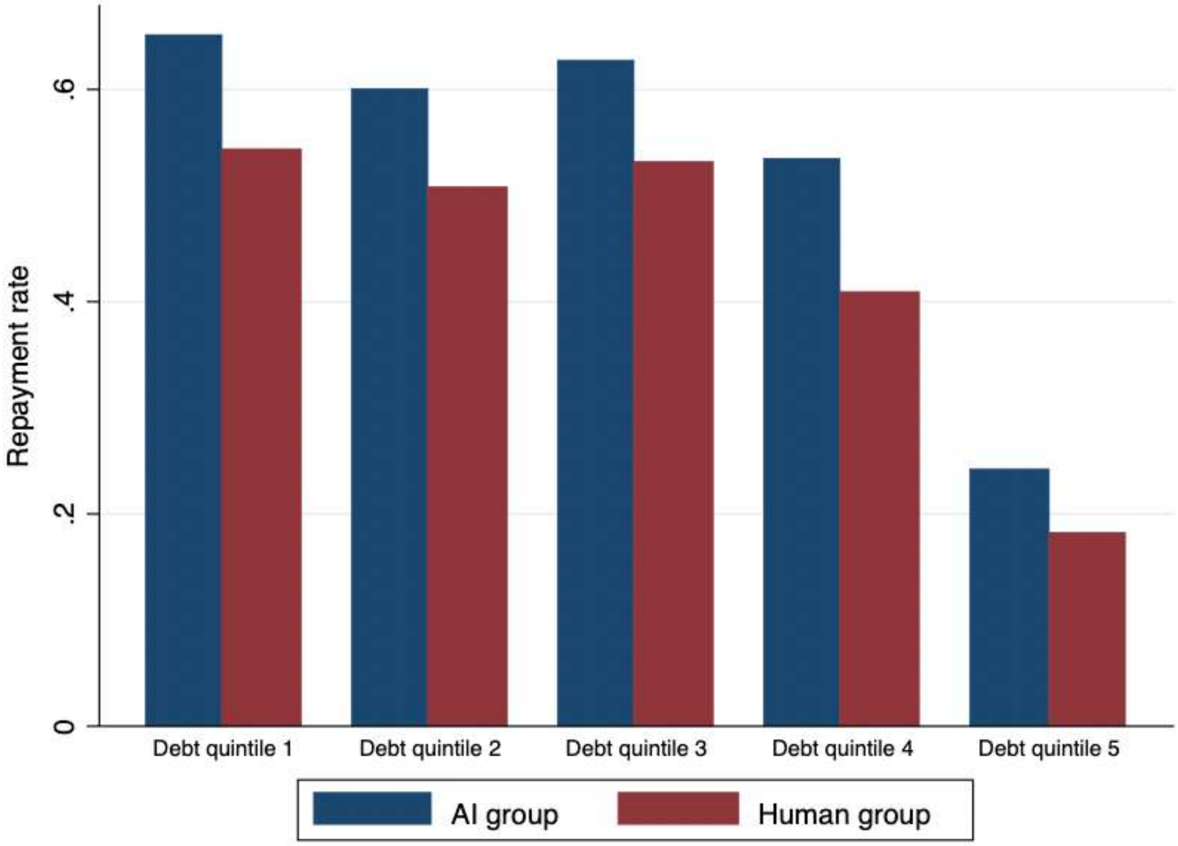

Figure I: Repayment by debt amount

This figure presents the percentages of borrowers who have repaid their debt across quintiles, ordered by increasing debt amounts. Blue bars indicate repayment rates for borrowers in the AI group, subject to collection calls decided by algorithms, while red bars show repayment rates for borrowers in the human group, subject to collection calls decided by human collection officers.

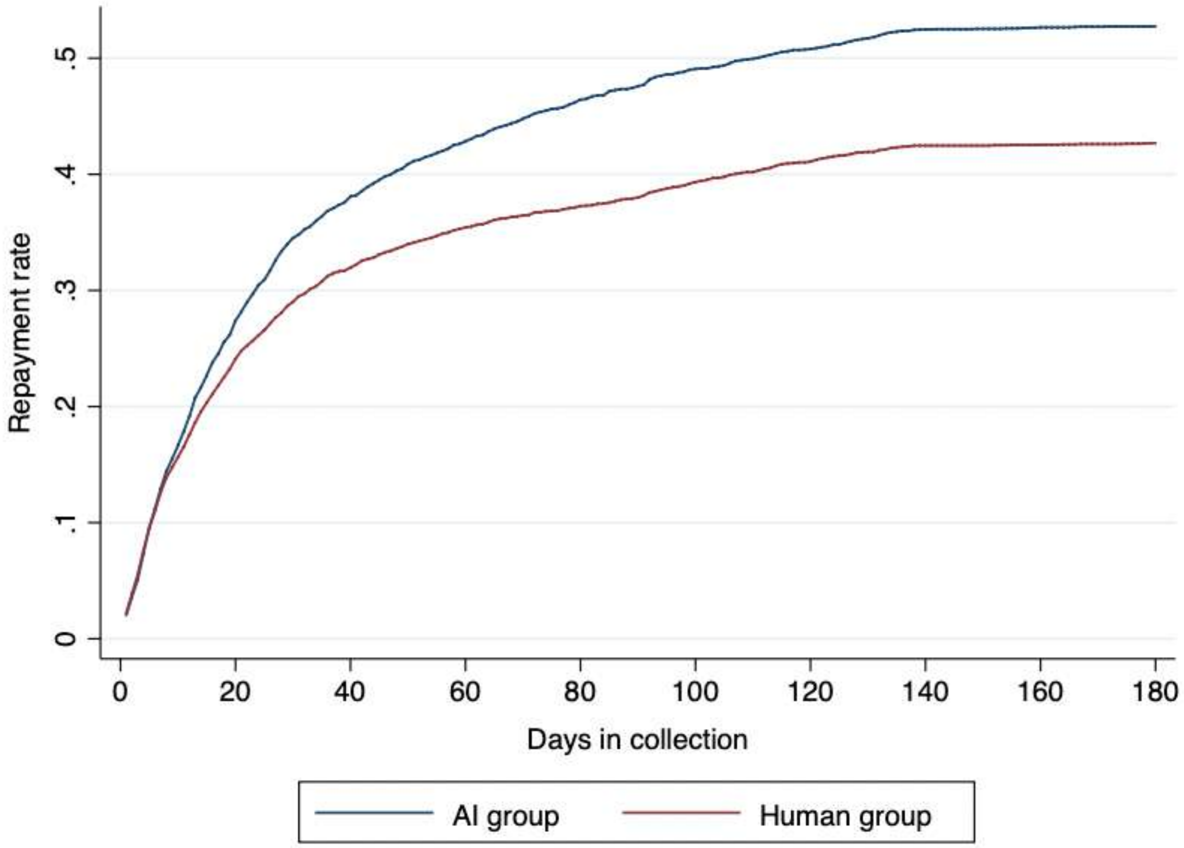

Figure II: Repayment by Days in Collection

This figure represents the percentages of borrowers who have repaid their debt, based on the number of days they have been in the 180-day collection process (days in collection). The blue line represents the cumulative repayment rates of borrowers in the AI group, who are subject to collection calls decided by algorithms. The red line represents the cumulative repayment rates of borrowers in the human group, who are subject to collection calls decided by human collection officers.

Table I: Summary Statistics

Panel A of this table provides summary statistics of the amount of debt and the number of calls made to each borrower in the historical data. Panel B provides summary statistics of the amount of debt and the number of calls made to each borrower in the experiment data. Panel C presents the summary statistics of the amounts of debt, the number of calls, and borrowers' internal credit scores in the human group and the AI group, separately. The credit score is an internal metric used by the creditor to measure the borrower's creditworthiness based on their credit profiles, with values ranging from 1 to 5. A higher score indicates better creditworthiness. Borrowers in the human group receive collection calls determined by human collection officers, while borrowers in the AI group receive collection calls based on algorithmic decisions.

Panel A: Historical data

	N	Mean	SD	P25	P50	P75
Amount of debt in collection ($)	36031	471.75	2659.21	60	116.28	292.63
Number of calls made	36031	3.68	3.76	1	2	5

Panel B: Experiment data

	N	Mean	SD	P25	P50	P75
Amount of debt in collection ($)	7839	675.48	2573.00	65.63	129.61	377.12
Number of calls made	7839	3.66	3.17	1	3	5

Panel C: Human group vs. AI group

Group	Amount of debt in collection ($)					
	N	Mean	SD	P25	P50	P75
Human group	3885	680.46	2361.78	66.56	132.93	386.39
AI group	3954	670.58	2765.15	64.28	128.13	365.35
p-Value		0.87				

Group	Number of calls made					
	N	Mean	SD	P25	P50	P75
Human group	3885	3.8196	3.3654	1	3	5
AI group	3954	3.4985	2.9540	1	2	5
p-Value		0.00				

Group	Internal credit score					
	N	Mean	SD	P25	P50	P75
Human group	3885	2.9995	1.4010	2	3	4
AI group	3954	3.0000	1.4002	2	3	4
p-Value		0.98				

Table II: Debt Repayment

This table represents the results from Ordinary Least Squares regressions of debt repayment on the AI indicator. The dependent variables are indicators for repayment, which are equal to one if the borrower repaid the debt during the 180-day collection process and zero otherwise. The AI indicator equals one for borrowers in the AI group who received AI-decided collection calls, and zero for those in the human group receiving collection calls decided by human officers. The estimations in Columns (1) and (2) are based on the full sample. Columns (3) and (4) are based on the subsample of borrowers who receive three or fewer calls, while Columns (5) and (6) are based on subsample of borrowers who receive four or more calls. Control variables include the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Robust standard errors are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	Repayment					
	Full sample		Number of calls <=3		Number of calls >=4	
	(1)	(2)	(3)	(4)	(5)	(6)
AI group	0.1010***	0.0993***	0.0822***	0.0823***	0.1129***	0.1122***
	(8.99)	(9.83)	(5.77)	(6.54)	(6.67)	(7.05)
Debt amount		-0.0149***		-0.0174***		-0.0074
		(-3.40)		(-2.93)		(-1.18)
Credit score		0.1441***		0.1523***		0.1124***
		(31.77)		(26.63)		(15.46)
Relationship score		0.0230***		0.0193*		0.0223
		(2.67)		(1.88)		(1.52)
Mean of dependent variable - human group	0.4314		0.5368		0.2765	
Observations	7839	7839	4774	4774	3065	3065
R^2	0.010	0.200	0.007	0.228	0.014	0.133

Table III: The Number of Calls

This table represents the results from Ordinary Least Squares regressions of the number of calls on the AI indicator. The dependent variables are the total number of calls a borrower *i* receives from the collector during the 180-day collection process. The AI indicator equals one for borrowers in the AI group who received AI-decided collection calls, and zero for those in the human group receiving collection calls decided by human officers. Columns (1) and (2) use the entire sample. Columns (3) and (4) restrict the sample to borrowers that have not repaid by the end of the 180-day collection process. Columns (5) and (6) restrict the sample to borrowers that have repaid within the 180-day collection process. Control variables include the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Robust standard errors are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	Number of calls					
	Full sample		Borrowers without repayment		Borrowers with repayment	
	(1)	(2)	(3)	(4)	(5)	(6)
AI group	-0.3211***	-0.3153***	-0.3206***	-0.3122***	0.0365	0.0183
	(-4.49)	(-4.44)	(-2.84)	(-2.76)	(0.48)	(0.24)
Debt amount		0.1267***		0.1061**		0.2015***
		(3.60)		(2.24)		(4.34)
Credit score		-0.2290***		0.1356**		-0.1359***
		(-7.29)		(2.57)		(-3.78)
Relationship score		0.0090		0.1396		-0.0212
		(0.17)		(1.30)		(-0.41)
Mean of dependent variable - human group	3.82		4.63		2.76	
Observations	7839	7839	4058	4058	3781	3781
R^2	0.003	0.022	0.002	0.004	0.000	0.018

Table IV: Debt Repayment and Credit Score

This table represents the results from Ordinary Least Squares regressions of debt repayment on the AI indicator. The dependent variables are indicators for repayment, which are equal to one if the borrower repaid the debt during the 180-day collection process and zero otherwise. The AI indicator equals one for borrowers in the AI group who received AI-decided collection calls, and zero for those in the human group receiving collection calls decided by human officers. Borrowers are classified into five quintiles based on their credit scores, from lowest to highest. Columns (1) to (5) report the regression results within each quintile. Panel A shows estimations without control variables, while Panel B shows estimations with control variables, including the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Robust standard errors are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Panel A

Dependent variable:	Repayment				
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
	(1)	(2)	(3)	(4)	(5)
AI group	0.0529***	0.1270***	0.1392***	0.0970***	0.0840***
	(3.34)	(5.30)	(5.84)	(4.00)	(3.71)
Mean of dependent variable - human group	0.0826	0.2756	0.5048	0.6010	0.6864
Control	Yes	Yes	Yes	Yes	Yes
Observations	1538	1537	1690	1537	1537
R^2	0.007	0.018	0.020	0.010	0.009

Panel B

Dependent variable:	Repayment				
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
	(1)	(2)	(3)	(4)	(5)
AI group	0.0447***	0.1262***	0.1400***	0.0967***	0.0830***
	(2.88)	(5.27)	(5.87)	(3.99)	(3.67)
Debt amount	-0.0361***	0.0146	0.0144	0.0322*	0.0328*
	(-5.77)	(1.55)	(1.16)	(1.75)	(1.75)
Relationship score	0.0985***	0.0037	0.0361	0.0067	0.0181
	(3.01)	(0.16)	(1.64)	(0.39)	(1.34)
Mean of dependent variable - human group	0.0826	0.2756	0.5048	0.6010	0.6864
Control	Yes	Yes	Yes	Yes	Yes
Observations	1538	1537	1690	1537	1537
R^2	0.044	0.020	0.022	0.013	0.012

Table V: Algorithmic Judgements and Human Collection Officers' Decisions

This table represents the results from Ordinary Least Squares regressions of human collection officers' decisions on algorithmic judgements. The dependent variable is the human-decided call, which is an indicator set to one if human collection officers decided to call the borrower on that day, and zero otherwise. AI-identified call is an indicator equal to one if algorithms identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. Columns (1) and (2) are estimated using the full sample of daily-borrower observations. Column (3) is based on daily-borrower observations from the first 90 days, while Column (4) is based on observations from the later 90 days. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	Human-decided call			
	Full Sample		First 90 days	Second 90 days
	(1)	(2)	(3)	(4)
AI-identified call	0.0191***	0.0507***	0.0159***	-0.0009
	(6.76)	(11.02)	(5.38)	(-0.35)
Days in collection	-0.0004***	-0.0004***	-0.0008***	-0.0001***
	(-59.19)	(-57.93)	(-48.61)	(-11.17)
AI-identified call × Days in collection		-0.0009***		
		(-14.67)		
Debt amount	0.0003	0.0004	-0.0000	0.0007***
	(0.97)	(1.24)	(-0.07)	(4.64)
Credit score	0.0008**	0.0008**	0.0008	0.0003*
	(2.38)	(2.20)	(1.39)	(1.76)
Relationship score	0.0009	0.0009	0.0013	0.0004
	(1.24)	(1.24)	(1.14)	(1.14)
Control	Yes	Yes	Yes	Yes
Observations	441679	441679	239974	201705
R^2	0.021	0.022	0.012	0.001

Table VI: Human Collection Officers' Decisions by Credit Quintiles

This table represents the results from Ordinary Least Squares regressions of human collection officers' decisions on algorithmic judgements. The dependent variable is the human-decided call, which is an indicator set to one if human collection officers decided to call the borrower on that day, and zero otherwise. AI-identified call is an indicator equal to one if algorithms identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. Borrowers are classified into five quintiles based on their credit scores, from lowest to highest. Columns (1) to (5) report the regression results within each quintile. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	Human-decided call				
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
	(1)	(2)	(3)	(4)	(5)
AI-identified call	-0.0142***	0.0169***	0.0324***	0.0486***	0.0427***
	(-5.70)	(2.70)	(4.18)	(5.59)	(4.94)
Days in collection	-0.0003***	-0.0004***	-0.0004***	-0.0004***	-0.0004***
	(-24.78)	(-30.04)	(-28.63)	(-27.20)	(-23.60)
Debt amount	-0.0011**	0.0018***	0.0035***	0.0045***	0.0048**
	(-2.07)	(2.78)	(4.01)	(3.17)	(2.58)
Relationship score	0.0035	0.0015	-0.0010	-0.0005	-0.0000
	(1.49)	(0.79)	(-0.57)	(-0.41)	(-0.03)
Control	Yes	Yes	Yes	Yes	Yes
Observations	127588	104732	86036	68091	55232
R^2	0.017	0.022	0.024	0.024	0.024

Table VII: AI-Extracted Signals

This table represents the results from Ordinary Least Squares regressions of debt repayment on AI-extracted signals. The dependent variable is an indicator for repayment, which equals one if the borrower repaid the debt during the 180-day collection process, and zero otherwise. The AI-extracted signal is an indicator equal to one if the specific category of signals was noted for the borrower before that day, and zero otherwise. The AI-extracted signals correspond to expressions of financial hardship in Column (1), expressions of repayment intent in Column (2), and discussions of non-repayment consequences in Column (3). The human-decided call is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	Repayment		
	Financial hardship	Repayment intent	Non-repayment consequences
	(1)	(2)	(3)
AI-extracted signal	-0.0273***	0.0295**	0.0213***
	(-3.45)	(2.24)	(2.93)
Human-decided call	0.0555***	0.0473***	0.0438***
	(7.82)	(6.91)	(5.18)
Human-decided call × AI-extracted signals	-0.0229	0.0605***	0.0096
	(-1.37)	(3.21)	(0.88)
Days in collection	-0.0018***	-0.0018***	-0.0018***
	(-35.62)	(-35.67)	(-35.65)
Debt amount	0.0024	0.0025	0.0029
	(0.93)	(1.00)	(1.13)
Credit score	0.0507***	0.0505***	0.0505***
	(13.38)	(13.29)	(13.33)
Relationship score	0.0185**	0.0190**	0.0189**
	(2.01)	(2.06)	(2.05)
Control	Yes	Yes	Yes
Observations	441679	441679	441679
R^2	0.163	0.163	0.163

Table VIII: Expressions of Financial Hardship

This table represents the results from Ordinary Least Squares regressions of calling decisions of AI and human collection officers on expressions of financial hardship. The dependent variable is the AI-identified high-value call in Columns (1) to (3), which is an indicator equal to one if AI identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. The dependent variable is the human-decided call in Columns (4) to (6), which is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise. The expressions of financial hardship, labeled as "hardship" in the table, is an indicator equal to one if this category of signals was noted for the borrower before that day, and zero otherwise. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	AI-identified high-value call			Human-decided call		
	(1)	(2)	(3)	(4)	(5)	(6)
Hardship	-0.0070***	-0.0191***	-0.0156***	-0.0031***	-0.0102***	-0.0020
	(-4.24)	(-4.17)	(-4.06)	(-3.76)	(-4.57)	(-1.10)
Hardship × Days in collection		0.0001***			0.0001***	
		(3.99)			(4.87)	
Hardship × Credit score			0.0033**			-0.0004
			(2.24)			(-0.62)
Days in collection	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***
	(-26.95)	(-25.72)	(-26.94)	(-60.77)	(-57.57)	(-60.74)
Debt amount	0.0028***	0.0029***	0.0028***	0.0003	0.0003	0.0003
	(4.18)	(4.20)	(4.09)	(1.00)	(1.03)	(1.03)
Credit score	0.0011	0.0011	0.0006	0.0009**	0.0008**	0.0009**
	(1.48)	(1.47)	(0.79)	(2.40)	(2.38)	(2.39)
Relationship score	-0.0018	-0.0018	-0.0018	0.0009	0.0009	0.0009
	(-1.25)	(-1.25)	(-1.25)	(1.16)	(1.16)	(1.16)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Observations	441679	441679	441679	441679	441679	441679
R^2	0.021	0.022	0.022	0.021	0.021	0.021

Table IX: Expressions of Repayment Intent

This table represents the results from Ordinary Least Squares regressions of calling decisions of AI and human collection officers on expressions of repayment intent. The dependent variable is the AI-identified high-value call in Columns (1) to (3), which is an indicator equal to one if AI identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. The dependent variable is the human-decided call in Columns (4) to (6), which is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise. The expressions of repayment intent, labeled as "intent" in the table, is an indicator equal to one if this category of signals was noted for the borrower before that day, and zero otherwise. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	AI-identified high-value call			Human-decided call		
	(1)	(2)	(3)	(4)	(5)	(6)
Intent	0.0287***	0.0703***	0.0269***	0.0033**	0.0053*	0.0085***
	(6.58)	(6.80)	(2.85)	(2.55)	(1.81)	(3.02)
Intent × Days in collection		-0.0005***			-0.0000	
		(-6.81)			(-1.11)	
Intent × Credit score			0.0007			-0.0019**
			(0.21)			(-2.19)
Days in collection	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***
	(-26.75)	(-24.95)	(-26.75)	(-60.50)	(-57.45)	(-60.48)
Debt amount	0.0027***	0.0028***	0.0027***	0.0003	0.0003	0.0003
	(4.02)	(4.09)	(4.02)	(1.07)	(1.08)	(1.08)
Credit score	0.0008	0.0007	0.0007	0.0008**	0.0008**	0.0010***
	(1.09)	(1.04)	(1.03)	(2.34)	(2.33)	(2.71)
Relationship score	-0.0015	-0.0015	-0.0015	0.0009	0.0009	0.0009
	(-1.06)	(-1.08)	(-1.05)	(1.24)	(1.24)	(1.22)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Observations	441679	441679	441679	441679	441679	441679
R^2	0.024	0.026	0.024	0.021	0.021	0.021

Table X: Discussions of Non-Repayment Consequence

This table represents the results from Ordinary Least Squares regressions of calling decisions of AI and human collection officers on discussions of non-repayment consequences. The dependent variable is the AI-identified high-value call in Columns (1) to (3), which is an indicator equal to one if AI identified calling the borrower on that day as high value and decided to make the call, and zero otherwise. The dependent variable is the human-decided call in Columns (4) to (6), which is an indicator equal to one if human collection officers decided to call the borrower on that day, and zero otherwise. The discussions of non-repayment consequences, labeled as "consequences" in the table, is an indicator equal to one if this category of signals was noted for the borrower before that day, and zero otherwise. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	AI-identified high-value call			Human-decided call		
	(1)	(2)	(3)	(4)	(5)	(6)
Consequences	0.0025	0.0128***	-0.0090**	0.0191***	0.0471***	0.0209***
	(1.51)	(3.14)	(-2.36)	(25.64)	(28.45)	(12.56)
Consequences × Days in collection		-0.0001***			-0.0003***	
		(-4.14)			(-27.02)	
Consequences × Credit score			0.0043***			-0.0007
			(3.43)			(-1.18)
Days in collection	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0003***	-0.0004***
	(-26.83)	(-19.98)	(-26.80)	(-58.58)	(-41.76)	(-58.60)
Debt amount	0.0030***	0.0029***	0.0028***	0.0005	0.0003	0.0005*
	(4.35)	(4.28)	(4.15)	(1.59)	(1.18)	(1.66)
Credit score	0.0011	0.0010	-0.0004	0.0005*	0.0005	0.0008**
	(1.47)	(1.45)	(-0.50)	(1.68)	(1.55)	(2.34)
Relationship score	-0.0017	-0.0017	-0.0018	0.0010	0.0011*	0.0010
	(-1.20)	(-1.17)	(-1.23)	(1.52)	(1.65)	(1.53)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Observations	441679	441679	441679	441679	441679	441679
R^2	0.021	0.022	0.022	0.024	0.028	0.024

Online Appendix

Artificial Intelligence and Debt Collection: Evidence from a Field Experiment

This Online Appendix consists of three parts. Part A tabulates the additional empirical analyses discussed in the main text. Part B presents the details of selecting, training, and evaluating the algorithms in predicting repayment likelihoods. Part C presents the technical steps of using algorithms to make calling decisions and a validation analysis.

Online Appendix Part A

Table A.I: List of Features

Features	Descriptions
Debt amount	The amount of debt assigned by the creditor to collect
Debt type	The detailed type of the financial-service debt, including the platform where the debt was originated
Timestamps of debt	The origination date (when the debt was initially incurred); The delinquency date (when the debt become delinquent due to non-repayment past due)
Contact information	The type of available contact information
Zip code	4-digits zip code
Relationship with creditor	The duration of the relationship between the borrower and the creditor; The total number of accounts the borrower holds with the creditor; The balances of accounts associated with the borrower.
Relationship score	An internal metric used by the creditor to evaluate the importance of their relationship with each borrower, ranging from 1 to 3. A higher score indicates that the creditor places a greater value on the relationship
Credit score	The creditor's internal metric to rate the borrower's creditworthiness based on their credit profiles, ranging from 1 to 5, with higher scores indicating better creditworthiness
Number of calls made	The number of outbound calls from the collector to the borrower during different time periods (e.g., past 1 day, past 3 days, past one week)
Length of calls made	The length of outbound calls from the collector to the borrower during different time periods
Number of calls received	The number of inbound calls from the inbound to the collector during different time periods
Length of calls received	The length of inbound calls from the inbound to the collector during different time periods
Website footprints	Footprints of the individual in the collector's website, including the login device and operating systems, the time and frequency of logins
Soft-information features	We employ the bag-of-words method to transform each note taken by calling agents into a high-dimensional numerical vector of word and phrase counts, with each word or phrase encoded as a distinct feature in the algorithms.

Table A.II: Categories of AI-Extracted Signals

This table represents examples of soft-information features that are included in each category of AI-extracted signals. Specifically, categories of AI-extracted signals are identified with the following steps. First, we estimate the importance scores for all the soft-information features used by the algorithms and obtain the top 100 words and phrases from the bag-of-words vocabulary that algorithms consider most important. Second, these 100 words and phrases are then converted into numerical vectors of semantic embeddings via fastText, a machine learning algorithm developed by Bojanowski et al. (2017). Thirdly, the vectorized words and phrases are further grouped into thematic clusters based on their semantic similarities using the spherical k-means algorithm of Dhillon and Modha (2001). After analyzing the thematic clusters identified by spherical k-means, three important categories of features emerge, with each category containing thematically clustered soft-information features that algorithms consider most important.

Categories of AI extracted signals	Examples of included soft-information features
Expressions of financial hardship	baanverlies (job loss) ontslag (dismissal) werkloos (unemployed) moeilijkheden (difficulties) Financiële moeilijkheden (financial difficulties)
Expressions of repayment intent	beloofd (promised) bereid (willing) toezegging (commitment) afgesproken (agreed) gemist (missed)
Discussions of non-repayment consequence	impact (impact) consequenties (consequences) bkr kredietwaardigheid (creditworthiness) kredietprofiel (credit profile) kredietlimieten (credit limits)

Table A.III: Correlation between Repayment and Contacting

This table represents the results from Ordinary Least Squares regressions of debt repayment on the daily call indicator. The dependent variables are indicators for repayment, which are equal to one if the borrower repaid the debt during the 180-day collection process and zero otherwise. The call indicator equals one if the borrower receives the call on that day, and zero otherwise. The estimations in Columns (1) to (3) are based on the borrowers in the human group, who are subject to collection calls decided by human collection officers. Columns (3) and (4) are based on the borrowers in the AI group, who are subject collection calls decided by algorithms. Control variables include the days in collection, the debt amount assigned for collection, the borrower's credit score reflecting their creditworthiness, and the relationship score indicating the creditor's rating of their relationship with the borrower. Standard errors clustered at the borrower level are reported in parentheses, and superscripts of *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Dependent variable:	Repayment indicator					
		Human group			AI group	
	(1)	(2)	(3)	(4)	(5)	(6)
Call indicator	0.1558***	0.0566***	0.0536***	0.2112***	0.0797***	0.0740***
	(20.18)	(8.23)	(8.00)	(25.13)	(10.88)	(10.49)
Days in collection		-0.0019***	-0.0018***		-0.0026***	-0.0025***
		(-34.83)	(-35.67)		(-42.80)	(-42.87)
Debt amount		-0.0267***	0.0028		-0.0388***	-0.0016
		(-13.14)	(1.10)		(-14.71)	(-0.49)
Credit score			0.0509***			0.0671***
			(13.41)			(14.07)
Relationship score			0.0187**			0.0126
			(2.03)			(1.23)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Observations	441679	441679	441679	394044	394044	394044
R^2	0.005	0.127	0.162	0.007	0.181	0.224

Part B: Using Algorithms to Predict Repayment

We considered several machine learning algorithms, including LASSO, support vector machines, random forest, gradient boosted decision trees, and neural networks; Among them, gradient boosted decision trees turn out to have the best predictive performance, followed by random forest. Therefore, we choose the gradient boosted decision trees algorithm as employed by Kleinberg et al. (2018).

We partition our sample (i.e., historical data) into a training set and a test set randomly, where the algorithms are trained based on observations in the training set and the predictive performance is measured based on the test set (i.e., the out-of-sample fit)[1]. Specifically, based on the training sample, we train a gradient boosted decision trees algorithm (Friedman 2001) on each day of the collection process to predict delinquent borrower's likelihood that the borrower will repay the debt by the end of the 180-day collection process, using the information that is available to the collectors by the corresponding training day. The algorithm is a sequential ensemble of multiple decision trees where the depth of each tree, the number of trees, the fraction of observations used per tree, the fraction of features used per tree, and the learning rate in the sequence are selected via 5-fold cross validation to maximize the out-of-sample fit. We feed the algorithm with a large set of potential features (explanatory variables), and the algorithm decides which features to be used in each tree and incorporates them in flexible function forms, similar to the high order interaction terms in linear regressions.

The list of features incorporated by the algorithm is detailed in online appendix Table IN1 and can be classified into two groups. The first group contains *hard-information* features such as debt

[1] We allocated 75% of the historical data to the training set and 25% of the historical data to the test set.

amounts in collection, debt age, debt type, contact information type, the borrower's relationship length with the creditor, and their credit profile (provided by the creditor). These features are established at the start of the collection process and usually remain unchanged throughout. Additionally, we incorporate more features that are constructed during the collection process, such as the number of phone calls made, the duration of past calls, inbound calls received, and borrower website logins. These features have also been integrated into the operations management system of the collection agency.

The second group comprises *soft-information* features. During interactions with borrowers, calling agents typically keep notes about the discussions. These notes, considered as soft information, are displayed as raw text in the operations management system. We employ the bag-of-words method to transform each text into a high-dimensional numerical vector of word and phrase counts, with each word or phrase encoded as a distinct feature in the algorithms. While the bag-of-words approach ignores syntax and word order, it preserves the frequency of occurrences in the notes; we convert the text to lists of unigrams and bigrams, and retain the 5,000 most commonly used words (unigrams) and phrases (bigrams) remaining in the text.

There are various methods to extract information from unstructured text data (see Gentzkow, Kelly and Taddy 2019). Compared to approaches that convert the text into a single measure such as the sentiment score, the bag-of-words method retains a broader range of unprocessed information from the raw text, which can be further leveraged by algorithms; Compared to more advanced techniques like Word2vec and large language models (LLMs) that transform text into high-dimensional numerical vectors, the bag-of-words approach maintains a one-to-one mapping from words and phrases to features in the vector space. This mapping allows us to later investigate which words or phrases are deemed important by the algorithm.

We use the held-out test set to guard against overfitting, and we obtain good prediction performance. The prediction performance or prediction accuracy is measured by AUC and cross-entropy loss. Specifically, AUC represents the area under the receiver operating characteristic (ROC) curve and provides an aggregate measure of the algorithm’s ability to accurately predict borrowers’ repayment likelihoods. AUC ranges in value from 0 to 1 and a higher value represents better prediction performance. One way of interpreting AUC is as the probability that the algorithm will assign a higher rank to a randomly selected positive example than to a randomly selected negative example. Cross-entropy loss, also known as log loss, measures the divergence between the predicted probabilities of repayment and the actual repayment outcomes. Essentially, it quantifies the algorithm's "surprise" when it encounters a mismatch between prediction and reality. Lower values of cross-entropy loss indicate better prediction accuracy, as they represent lesser divergence from the actual outcomes, but cross-entropy loss can be sensitive to class imbalance in the data.

We present the performance of two algorithms (Gradient Boosted Decision Trees as in blue line and random forest as in red line) in Figure B.I, by days in collection, defined as the number of days the borrowers have been in the 180-day collection process. Panel A displays the AUC of each algorithm, which shows significant increases during the early stages of the collection process. Since the 40th day, AUC of Gradient Boosted Decision Trees is mostly above 0.82 and is mostly above 0.80 for Random Forest. In Panel B, we report the cross-entropy loss of each algorithm. During the first 10 days of the collection process, both algorithms exhibit a cross-entropy loss of approximately 0.5, but the loss notably decreases to about 0.3 by the 45th day. Both algorithms demonstrate commendable predictive performance, but Gradient Boosted Decision Trees marginally outperforms Random Forest. Consequently, we elect to use Gradient Boosted Decision Trees as the primary algorithm in this study.

As a complementary visualization of the algorithms' predictive performance, we categorize all borrowers into 10 deciles based on their average predicted repayment likelihood, as estimated by the algorithm (Gradient Boosted Decision Trees) during the collection process. Figure B.II displays the percentage of borrowers who have repaid their debt within each decile, ordered from lowest to highest predicted repayment likelihoods. We observe that the decile with the highest predicted repayment likelihoods concludes with an aggregate repayment rate of over 70%, while the decile with the lowest predicted likelihoods ends with a rate below 10%.

While we have no intention of disentangling the inherent capabilities of machine learning algorithms the alternative data employed —since they are inseparable components—our analysis provides suggestive evidence for the importance of soft-information features. Specifically, we repeat the training procedures but employ two algorithms: one algorithm as before, and a new algorithm with the soft-information features excluded. The AUC of both variants of algorithms are presented in Figure B.III: the red line signifies the AUC of algorithms with soft-information features, while the blue line represents those without. As demonstrated in the figure, the AUC for algorithms that utilize soft-information features experiences a significant enhancement during the initial 60 days of the collection process, thereafter maintaining a relatively stable score above 0.81. Conversely, the AUC for algorithms that exclude soft-information features experiences a pronounced decline from approximately 0.78 to 0.69 during the first 80 days, with a slight continued decrease thereafter. This pattern suggests that without the additional signals embedded in the soft information collected by calling agents when communicating with borrowers, predicting future repayment becomes increasingly challenging if a borrower has not yet repaid. Furthermore, the widening performance gap between the two algorithm versions robustly affirms the potency of soft information in predicting repayment during the collection process.

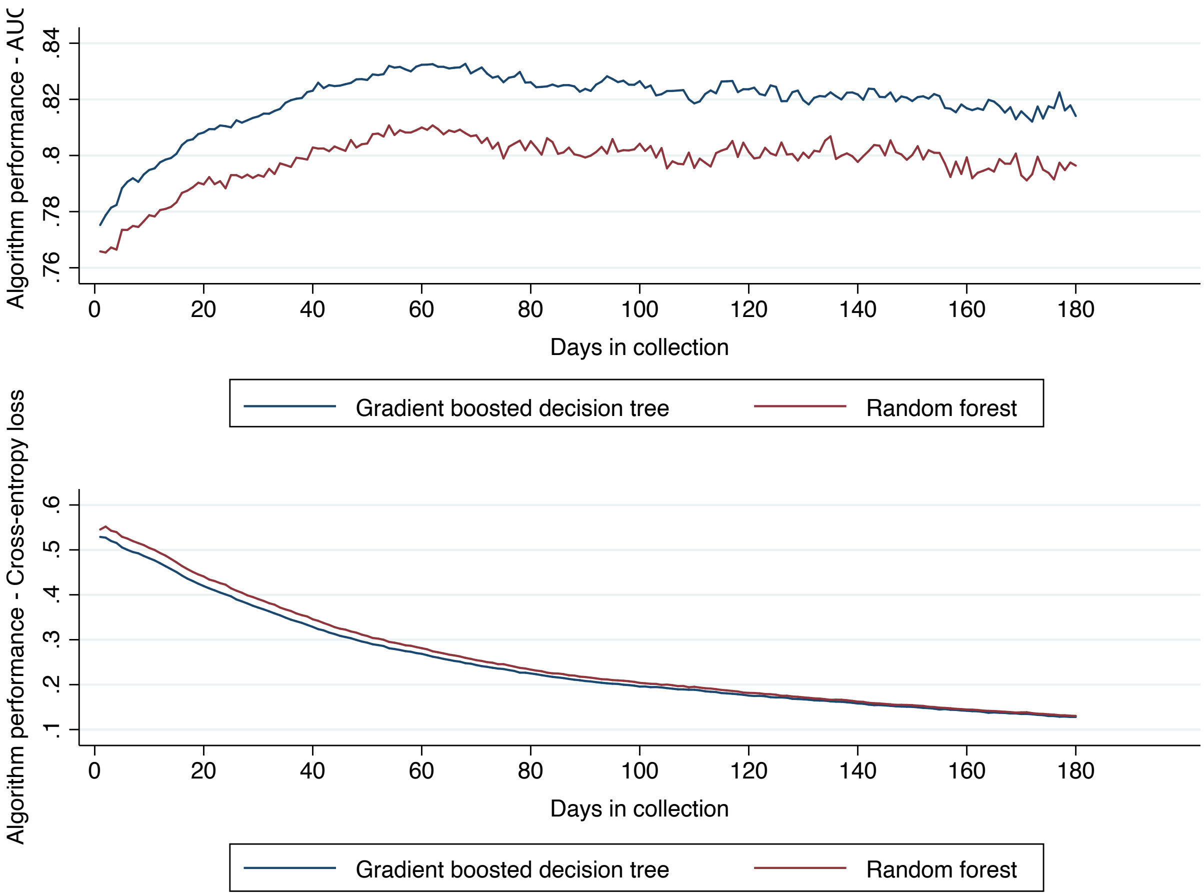

Figure B.I: Performance of the Predictive Algorithms

This figure reports the performance of two algorithms (Gradient Boosted Decision Trees as in blue line and Random Forest as in red line) by days in collection, where days in collection is the number of days borrowers have been in the 180-day collection process. On each day of the collection process, the algorithms are trained to predict the likelihood of repayment for each borrower and evaluated on the test set. Panel A displays the area under the receiver operating characteristic curve (AUC), highlighting the ability of each algorithm to accurately predict borrowers' repayment likelihoods on each day. Panel B illustrates the cross-entropy loss of each algorithm, indicating the divergence of the predicted probabilities from the true outcomes on each day.

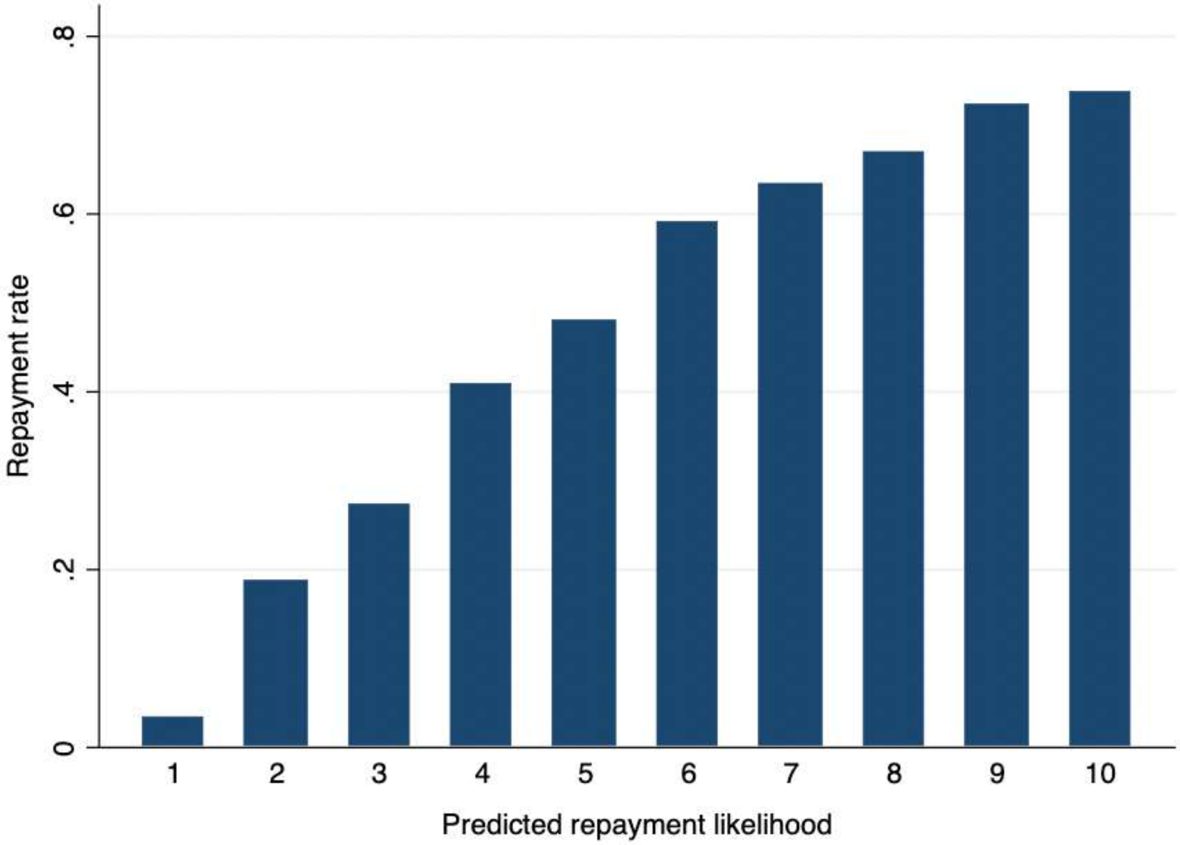

Figure B.II: Repayment by Algorithmic Predictions

This figure represents the percentages of borrowers who have repaid their debt within each decile, ordered by predicted repayment likelihoods from low to high. Specifically, we categorize borrowers on the test set into 10 deciles based on their corresponding average predicted repayment likelihood estimated by the algorithms (Gradient Boosted Decision Trees) during the collection process.

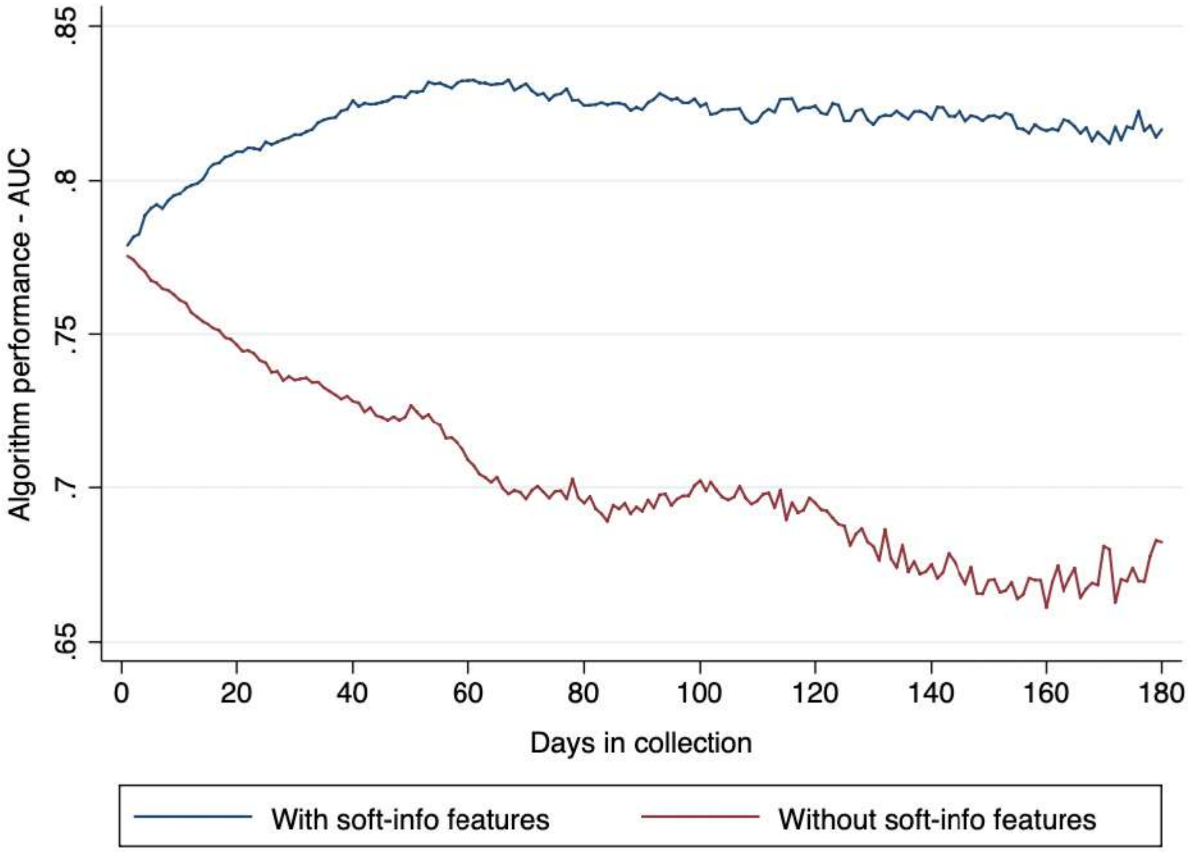

Figure B.III: Performance of the Predictive Algorithms--Impact of Soft-Information Features

This figure reports the AUC of algorithms by days in collection, where days in collection is the number of days borrowers have been in the 180-day collection process. On each day of the collection process, algorithms are trained to predict the likelihood of repayment for each borrower and evaluated on the test set. AUC stands for the area under the receiver operating characteristic curve, highlighting the ability of the algorithms to accurately predict borrowers' repayment likelihoods. The red line signifies the AUC of algorithms with soft-information features, while the blue line represents those without.

Part C: Using Algorithms to Make Calling Decisions

We apply the meta-learners method of Künzel et al. (2019) to estimate the heterogeneous treatment effects. To illustrate this method, let's define the CATE (conditional average treatment effect) function as:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

where Y_i *(1)* is the potential outcome of individual i when i receives the treatment; Y_i *(0)* is the potential outcome of individual i when i receives no treatment; and X is a high-dimensional covariate of feature vector. In this paper, the treatment is equivalent to receiving the call on each day, and the outcome is whether the borrower will repay the debt by the end of the 180-day collection process. We conduct the following three-step estimations for each day of the collection process separately, using the information that is available to the collectors by the corresponding day. The feature vector includes both the hard-information features and the soft-information features as described before.

The first step involves estimating the outcomes (i.e., debt repayment) for the group of borrowers that received calls on the day, and the group of borrowers that did not receive calls on that day. This is done by fitting two supervised machine learning algorithms, referred as the base learners for the first stage:

$$\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]$$

$$\mu_1(x) = \mathbb{E}[Y(1) \mid X = x]$$

The second step first computes the imputed treatment effects. We compute the imputed treatment effects for the individuals in the treated group (receiving calls on the day), based on the control-outcome estimator; and we compute the treatment effects for the individuals in the control group (not receiving calls on the day), based on the treatment-outcome estimator:

$$\widetilde{D}_i^1(x) = Y_i^1 - \hat{\mu}_0(X_i^1)$$

$$\widetilde{D}_i^0(x) = \hat{\mu}_1(X_i^0) \ - Y_i^0$$

Then we model the treatment effects $\tau(x)$ by using the imputed treatment effects as the response variable. This is done by fitting two supervised machine learning algorithms, referred as the base learners for the second stage:

$$\tau_0(x) = \mathbb{E}\left[\widetilde{D}_i^0 \,|\, X = x\right]$$

$$\tau_1(x) = \mathbb{E}\left[\widetilde{D}_i^1 \,|\, X = x\right]$$

The third step defines the CATE (conditional average treatment effect) as the weighted average of the two estimates in stage 2, where the weight g is the unconditional likelihood of receiving calls on the day:

$$\hat{\tau}(x) = g(x)\,\hat{\tau}_0(x) + (1 - g(x))\,\hat{\tau}_1(x)$$

The meta-algorithms can be built on any base algorithms. Künzel et al. (2019) employ Random Forest and Bayesian Additive Regression Trees as examples. In this study, we adopt Gradient Boosted Decision Trees as the base algorithms to keep consistency. Following the three steps above, the algorithms can provide an output estimate for the value of each call, calculated as the predicted heterogeneous treatment effect multiplied by the debt amount.

Due to the unobservable nature of treatment effects, validating the estimated effects of calls is challenging, and comparing the effectiveness of calling decisions made by machine learning and human collection officers is complex. We assume that both human collection officers and AI make calling decisions to maximize the total amount of debt recovered. Figure C.I presents a 2x2 matrix that compares the decisions of human collection officers and AI within the historical data.

In a simplified scenario, a collection call is classified as high-value or low-value by human collection officers (left column) and by AI (top row). Cell (a) [top left] and cell (b) [top right] denote actions perceived as high-value by both parties and high-value by human managers but low-value by AI, respectively. Conversely, cell (c) [bottom left] and cell (d) [bottom right] signify actions perceived as high-value by AI but low-value by human collection officers, and actions deemed low-value by both parties, respectively. Given that the historical data is derived from past decisions and actions of human collection officers—who were supposed to select calls they regard as high-value—the comparison between human collection officers and AI's assessments is limited to the outcomes from cells (a) and (b). Cells (c) and (d) correspond to scenarios considered low-value by human collection officers and cannot be observed.

In this context, we validate the AI estimation within the historical sample by comparing the outcomes of cells (a) and (b) as illustrated in Figure C.I. We first employ the algorithms to estimate the value of all collection calls selected by human collection officers, and calculate the average estimated value of the collection calls received by each borrower during the collection process. Then we categorize borrowers into two groups based on the average value of calls and compare their repayment rates. The results are presented in Figure C.II. The first group includes borrowers who, on average, are the recipients of collection calls classified as low-value by the AI (represented by the blue line), while the second group encompasses borrowers who are on the receiving end of calls that the AI considers high-value on average (illustrated by the red line). We graph the percentage of borrowers in each group who have repaid their debt according to the number of days they have been in the collection process.

As shown in Figure C.II, the cumulative repayment rate rises steeply during the initial 20 days of the collection process, particularly for the group of borrowers receiving calls deemed low-value by AI. Intriguingly, the repayment rate for this group surpasses that of the high-value call recipients until day 44 of the collection process. At first glance, this might imply a discrepancy in AI's estimation. However, a more nuanced examination of the data post day 44 reveals a sustained upward trend for the high-value call group, surpassing 0.38 by the 60th day and culminating at 0.49 by the end of the collection process. This trend aligns with AI's estimations. Conversely, for the group receiving low-value calls, the repayment rate markedly decelerates post day 30. Therefore, this higher initial repayment rate for borrowers receiving low-value calls could suggest that these calls are not necessarily driving increased repayment. Instead, they may merely be accelerating the repayment of certain borrowers who would have repaid even without additional contact, thus validating AI's low-value classification.

This analysis within the historical data affirms AI's competence in identifying high-value calls to improve the aggregate repayment. However, this comparison largely pertains to outcomes from cells (a) and (b) in Figure C.I, with cell (c)—indicative of actions deemed low-value by human managers but high-value by AI—remaining unobservable.

		What AI considers as	
		High-value call	Low-value call
What human considers as	High-value call	(a): observable in historical data	(b): observable in historical data
	Low-value call	(c): unobservable in historical data	(d): unobservable in historical data

Figure C.I: Evaluating High-value Calls with Historical Data

This figure presents a 2x2 matrix, outlining the comparative framework between human collection officers and AI within the test set of historical data. In a simplified scenario, a collection call is either designated as high-value or low-value by human collection officers (represented by the left column) and AI (depicted by the top row). The cells (a) [top left] and (b) [top right] of the matrix illustrates the observable space for comparing the assessments made by human collection officers and AI. In particular, cell (a) signifies collection calls mutually deemed high-value by both entities, whereas cell (b) represents actions perceived as high-value by human collection officers but low-value by AI. Given that the historical data is derived from past decisions and actions of human collection officers—who typically select calls they regard as high-value—the comparison between human collection officers and AI's assessments in the historical data is limited to the outcomes from cells (a) and (b). Cells (c) and (d) which correspond to scenarios considered low-value by human collection officers cannot be observed in the historical data.

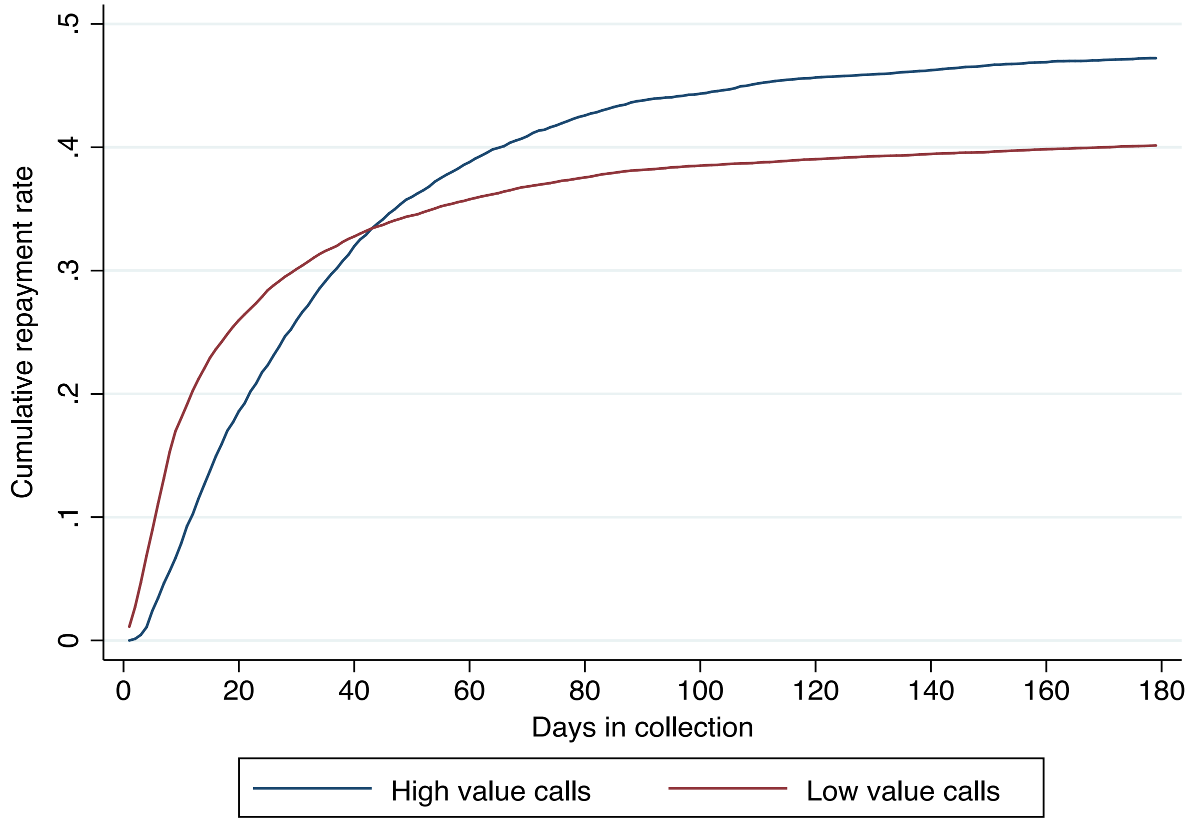

Figure C.II: Validation of High-value Call with Historical Data

This figure presents the validation of the algorithmic calling decisions within the test set of historical data by contrasting cell (a) and cell (b) as delineated in Figure IV. In particular, we use the algorithms to estimate the effect of all collection calls selected by human collection officers within this historical sample and subsequently categorize borrowers into two groups, determined by the average estimated value of the collection calls received by each borrower. The first group includes borrowers who, on average, are the recipients of collection calls classified as low-value by AI (represented by the blue line), while the second group encompasses borrowers who are the recipients of collection calls that AI considers as high-value on average (illustrated by the red line). We then illustrate the percentage of borrowers in each group who have repaid their debt according to the number of days they have been in the collection process.