FinTech Lending to Borrowers with No Credit History

Laura Chioda Paul Gertler Sean Higgins Paolina Medina*

June 1, 2025

Abstract

Despite the promise of FinTech lending to expand access to credit to populations without a formal credit history, FinTech lenders primarily lend to applicants with a formal credit history and rely on conventional credit bureau scores as an input to their algorithms. Using data from a large FinTech lender in Mexico, we show that alternative data from digital transactions through a delivery app are effective at predicting creditworthiness for borrowers with no credit history. Using account-by-month level data on revenues and costs, a machine learning model predicting profitability generates less profits than a model predicting default.

JEL Codes: G23, G5, O16
Keywords: FinTech; Lending; Machine Learning; Financial Inclusion; Profits

*Chioda: Haas School of Business, UC Berkeley, lchioda@berkeley.edu. Gertler: Haas School of Business, UC Berkeley and NBER, gertler@berkeley.edu. Higgins: Kellogg School of Management, Northwestern University, sean.higgins@kellogg.northwestern.edu. Medina: C.T. Bauer College of Business, University of Houston, pcmedina@bauer.uh.edu. We gratefully acknowledge financial support from USAID and Digital Frontiers (Equitable AI Challenge), CEGA and the Bill & Melinda Gates Foundation (Digital Credit Observatory, DCO), and UC Berkeley's Lab for Inclusive Fintech (LIFT). We thank conference participants at CEGA, LIFT, NBER, Princeton/IFC, and We-Fi/EBRD/IDB for valuable feedback. We thank José Antonio Murillo, Luz Téllez, Roberto Arriaga, Pedro Armengol, and Jacobo Zetune from RappiCard Mexico (https://rappicard.mx/) for generously supporting this research project, providing access to the data, and answering questions. We thank Josh Blumenstock, Nitin Kohli, and Enrique Seira for helpful advice. Luis Roman, Yusuf Abdul, Iñaki Fernández, Jora Li, Lucía Bertoletti, Haoyuan Song, and Victoria Hollingshead provided excellent and unwavering research support. William Karl Thomson and Matthew Gorby offered excellent computing and data service support. The authors declare that they have no financial or material interests in the findings of this paper.

1 Introduction

Online FinTech lenders are an increasingly important source of credit for households and small businesses (Berg, Fuster, and Puri, 2022; Buchak, Matvos, Piskorski, and Seru, 2018; Gopal and Schnabl, 2022). The promise of FinTech lending is that by using alternative data sources to evaluate creditworthiness and reducing other frictions such as travel costs and loan processing time, FinTech lenders can expand access to credit to populations with limited or no credit history—i.e., the financially excluded. In practice, however, while FinTech lenders do improve their default prediction models using alternative data sources, most of their lending algorithms still rely at least partly on conventional credit bureau scores (Johnson, Ben-David, Lee, and Yao, 2023) and do not substantially expand access to credit for those traditionally excluded from the financial system (Fuster, Plosser, Schnabl, and Vickery, 2019).

Using data from a large FinTech lender in Mexico, we show that alternative data—digital transactions data—can be quite effective in predicting creditworthiness *even for borrowers with no credit history*. All applicants in our sample lack a traditional credit score from the credit bureau, because they have either no credit history or at best a limited credit history that the credit bureau deems as insufficient to use to generate a credit score.[1]

Our FinTech partner, RappiCard Mexico, is a joint venture between Banorte, a large bank in Mexico, and Rappi, the leading on-demand delivery platform for food, goods, and services in Latin America. RappiCard Mexico leverages digital footprints and transaction data to inform credit card lending decisions. The company lends to applicants both with and without credit history. When lending to individuals with a credit history and thus a credit score in the Mexican credit bureau, they combine credit bureau data with transaction-level data on delivery orders through the app and use a machine learning algorithm to assess risk.

At the time of our collaboration, when lending to clients with no credit history, RappiCard had not relied on a machine learning algorithm; instead, they used a set of parsimonious rules for various client segments to make their lending decisions. We use data on the subsequent repayment behavior of these borrowers to assess risk. Specifically, we combine the repayment information with transaction-level data on purchases made through the delivery app, data on these applicants' "digital footprints" (Berg, Burg, Gombovi, and Puri, 2020), and other data sources, to build machine learning models to predict creditworthiness.[2]

[1]For conciseness we refer to these borrowers with no credit bureau score as having "no credit history." None of the applicants in our sample had a credit card prior to applying for a credit card from our FinTech partner.

[2]The other data sources include a "no-hit" score—developed by the credit bureau for those with no credit history or an insufficient credit history to report a traditional credit score—and socioeconomic characteristics at the census tract level. The "no-hit" score is reported by the credit bureau for all Mexican citizens with no credit history and thus no traditional credit score; it is independent of (i.e., not comparable to) the traditional credit scores reported for those who do have a credit history, and is based on publicly available data sets aggregated at the local level merged with the

We find that the machine learning model using alternative data predicts creditworthiness with sufficiently high accuracy for our partner to be comfortable lending using this model: our FinTech partner is now piloting a model similar to the model we present in this paper for a random subsample of applicants with no credit history, and in this pilot the model has outperformed their previous method of using parsimonious rules to make lending decisions for applicants with no credit history.

Our benchmark model achieves an area under the receiver operating characteristic curve (AUC) of 0.796. This exceeds the thresholds recommended by Iyer, Khwaja, Luttmer, and Shue (2016) for desirable AUCs of 0.6 in data-scarce environments and 0.7 in data-rich environments, is at the upper end of AUCs estimated using alternative data (even when combined with credit bureau data) in middle-income countries, and exceeds the AUCs for populations with no credit history (see Table 1 for a comparison of samples and AUCs across studies). We also assess model performance using three threshold-dependent measures: precision, recall, and F1 score, where the F1 score is the harmonic mean of precision and recall. To determine the threshold of predicted probability of default that the lender should use to make origination decisions, we use account-by-month level profitability data over the first 12 months since origination of each card to determine the profit-maximizing approval threshold (in the training sample, to avoid overfitting). We find that the profit-maximizing threshold is a 24% probability of default, and using this threshold precision is 0.421, recall is 0.666, and the F1 score is 0.516.

To test how important each group of features is, we estimate the AUC of a model which excludes them and compare it to that of the model trained using all features. We find that the digital footprints data—featuring the same set of variables as in Berg, Burg, Gombovi, and Puri (2020), such as the device type, operating system, and email host of the applicant—and the digital transactions data from the delivery app are associated with the highest marginal contributions to the model's predictive power. The digital footprints data contribute 0.124 to the AUC and the digital transactions data contribute 0.026.[3] We also find that—as expected—the performance of the model is increasing in the richness of the transactions history through the delivery app. Specifically, when we split the sample into quintiles based on the number of transactions they have completed through the app at the time of loan application, we find that the model for the top quintile—who have 29 or more transactions through the delivery app—has an AUC of 0.838 while that for the lowest quintile—who have at most 2 transactions through the delivery app—has an AUC of 0.741. Evaluating the model's performance using precision, recall, or F1 score with the profit-maximizing approval threshold rather than AUC similarly reveals that the most important features are from

location where the applicant lives; see CRIF (2018).

[3]Comparing this to the other data sources, the "no-hit" scores generated by the credit bureau using publicly available geographic data for borrowers with no or limited credit history, combined with credit history data for those with a limited but insufficient credit history to generate a traditional credit bureau score, contribute 0.006 to the AUC. The socioeconomic characteristics at the census tract level contribute 0.001.

the digital footprint and transactions data, and that the model performs best for those with more transactions.

Next, we ask how well models predicting default perform from the perspective of the lender's ultimate objective of maximizing profits. Models that predict default, rather than profits, are used extensively by both traditional and FinTech lenders to make credit decisions (Rajan, Seru, and Vig, 2015; Johnson, Ben-David, Lee, and Yao, 2023). Given that firms—including financial firms—often forgo profitable opportunities (Gertler, Higgins, Malmendier, and Ojeda, 2025; Mishra, Prabhala, and Rajan, 2022), are lenders leaving substantial profits on the table by using default models to make lending decisions? To address this question, we use account-by-month level data on each component of the lender's realized revenues and costs for each originated credit card over time. Revenues to the lender include interest payments, buy now pay later (BNPL) fees, interchange revenue paid to the card issuer each time a purchase is made with the card, late payment fees, and other fees. Costs include charge-offs, cashback rewards, costs of fraudulent transactions, funding costs (calculated by RappiCard as the average daily balance over the month times its monthly cost of capital to borrow that money), and other costs such as the cost of replacing a lost card.

Using these rich data on account-by-month level revenues and costs for the lender, we calculate account-level profits over the first 12 months since card origination and compare a model predicting that a borrower is unprofitable (i.e., generates negative profits over the first 12 months) to our benchmark model predicting default. In both models, we determine the profit-maximizing approval threshold in the training sample and apply this threshold to determine credit allocation based on predicted probabilities in the test sample. Based on AUC, the default model makes more accurate predictions, as the profitability model has a lower AUC of 0.598. Comparing the profits generated for the lender by each of the two models, we find that the model predicting default generates 1.93 times as much profit as the model predicting profitability.

We next explore *why* the model predicting default generates nearly twice as much profit as the model predicting profitability, despite the latter model being trained to predict what should actually matter for the lender's ultimate objective of maximizing profits. First, we plot average profits against the predicted probabilities of default and of negative profits. Profits exhibit an inverted-U relationship to the predicted probability of default: for the lowest-risk borrowers, average profits are low but on average positive, and average profits initially increase with risk of default, but then fall and become negative and continue decreasing as the predicted probability of default increases. The non-monotonic relationship between risk of default and profits is consistent with Agarwal, Chomsisengphet, Mahoney, and Stroebel (2015). However, a key difference is that in our setting, profits for the highest-risk borrowers cross zero and become negative, whereas the highest-risk borrowers approved for credit cards remain profitable in Agarwal, Chomsisengphet, Mahoney, and Stroebel (2015). In contrast to the non-monotonic relationship between predicted probabilities

and profits in our default model, average profits are consistently decreasing with the predicted probability of being unprofitable from our profitability model. In theory, this should favor the profitability model; thus, we turn to comparing borrowers approved by each model to understand why the default model generates higher profits.

While 35% of applicants are approved by both models and 21% are rejected by both models, 33% of applicants are approved only by the default model ("default-approved") and 11% are approved only by the profitability model ("profitability-approved"). Intuitively, the profitability model attempts to identify borrowers who generate substantial interest revenue (or other revenue such as through interchange fees net of rewards) but do not default. It succeeds at identifying users who generate substantial revenue: 76% of profitability-approved borrowers generate interest revenue compared to 67% of default-approved borrowers, average interest revenue of profitability-approved borrowers is 1.6 times as large as that of default-approved borrowers, and 75% of profitability-approved borrowers generate positive interchange fee net of rewards compared to 51% of default-approved borrowers.

However, many of the borrowers who generate substantial interest revenue by carrying revolving debt (and making their interest payments) end up eventually defaulting, and the profitability model fails to predict well which borrowers generating interest revenue end up defaulting: 36% of profitability-approved borrowers are at least 60 days delinquent and 29% incur charge-offs over the first 12 months since origination, compared to only 8% delinquent and 6% with charge-offs among default-approved borrowers. Average charge-offs for profit-approved borrowers (including zeros for those without charge-offs) are 4.3 times as large as average charge-offs for default-approved borrowers. It is not surprising that the profitability model has difficulty distinguishing which of the users generating substantial interest revenue by carrying high revolving debt end up defaulting: default among low-income credit card borrowers in Mexico is largely driven by exogenous shocks such as job loss (Castellanos et al., 2024).

Interchange fee revenue also plays a role. Those approved by both models are relatively high spenders—spending 3,112 Mexican pesos (MXN) per month—and interchange fee revenue from these borrowers is about 16% higher than rewards expenses.[4] The default-approved borrowers are even higher spenders—spending 3,442 MXN per month—but the cost of rewards for these borrowers fully offsets interchange fees. In contrast, the profitability-approved applicants are riskier borrowers who are lower spenders—spending 2,606 MXN per month—but who also have a higher spread between interchange fee revenue and rewards: for these borrowers, interchange fee revenue is 24% higher than rewards. Thus, the profitability model also selects borrowers who generate positive interchange revenue net of rewards, whereas this is not true for borrowers selected only

[4]The purchasing power parity (PPP) adjusted exchange rate was 9.7 MXN per USD at the end of 2022; see `https://www.oecd.org/en/data/indicators/purchasing-power-parities-ppp.html`.

by the default model. The lower-risk borrowers identified by the default model are likely more sophisticated and better at optimizing their shopping behavior to maximize rewards.

We make three main contributions. Our first contribution is to show that alternative data are effective in predicting creditworthiness for borrowers with *no credit history*. Most papers on the use of alternative data for FinTech lending estimate these models on a sample in which all or at least a majority of applicants *do* have a formal credit history and credit score (Table 1), perhaps because FinTech lenders primarily lend to applicants with formal credit histories and thus do not expand access to credit on the extensive margin (Berg, Fuster, and Puri, 2022; Fuster, Plosser, Schnabl, and Vickery, 2019). Some papers focus on subprime borrowers (Di Maggio and Yao, 2021) or borrowers with a thin credit file that is nevertheless sufficient for the credit bureau to generate a credit score (Blattner and Nelson, 2024), and FinTech has the potential to increase the intensive margin of credit access for these borrowers. Nevertheless, the sample in these papers still does have a formal credit history and credit score, and thus improving models to lend to this population will not increase access to credit on the extensive margin. Other papers evaluate machine learning models for samples both with and without credit scores, but in those papers the models for those with no credit history do not perform nearly as well: for example, in Agarwal, Alok, Ghosh, and Gupta (2023), the AUC for those with no formal credit history in India is 0.674, compared to an AUC of 0.738 for those with a formal credit history (when the credit bureau data are included in the model).[5] In contrast, in our sample no applicants have sufficient credit histories for the credit bureau to generate a credit score for them, and we find an AUC of 0.796.

Our second contribution is to test how machine learning models using alternative data to predict *default* for applicants with no credit history compare to machine learning models to predict *profitability*. Given that many lenders predict default rather than profits to make credit allocation decisions (Rajan, Seru, and Vig, 2015; Johnson, Ben-David, Lee, and Yao, 2023), it is important to understand whether they are acting optimally by comparing the profits generated by models predicting default vs. profitability. A number of papers have diagnosed the profitability of different segments of credit card borrowers (e.g., Agarwal, Chomsisengphet, Mahoney, and Stroebel, 2015; Agarwal, Presbitero, Silva, and Wix, 2023; Drechsler et al., 2025; Krivorotov, 2023). Other papers assess the use of alternative data to predict *default*, but lack borrower-level profits data to assess whether lenders would do better by predicting profitability (e.g., Agarwal, Alok, Ghosh, and Gupta, 2023; Berg, Burg, Gombovi, and Puri, 2020; Björkegren and Grissen, 2020, and sev-

[5]Two exceptions are Björkegren and Grissen (2020) and Lee, Yang, and Anderson (2025). In Björkegren and Grissen (2020), the best-performing model using mobile phone data for the 15% of their sample with no credit history has an AUC of 0.766, compared to an AUC of 0.770 for the best-performing model combining mobile phone data and credit bureau data for the 85% of their sample with credit histories. In Lee, Yang, and Anderson (2025), the model for the 18% of the modeling sample with no credit history has an AUC of 0.682, compared to an AUC of 0.677 for the 82% of the sample with credit histories.

eral others detailed in Table 1). By combining alternative data for credit scoring with data not only on defaults but also on account-by-month level revenues and costs for the lender, we compare the profits generated by models predicting default and profitability.

Our third contribution is to assess how the types of borrowers approved by a model predicting profitability compare to those approved by a model predicting default. This is possible in our setting because we observe all components of profits at the account-by-month level, which differs from papers studying credit cards in the US where interchange fee revenue (and, in some cases, rewards) are not observed at the account level and are instead imputed based on market or bank-level averages for implicit interchange rates (and rewards) multiplied by account-level spending (e.g., Agarwal, Chomsisengphet, Mahoney, and Stroebel, 2015; Agarwal, Presbitero, Silva, and Wix, 2023; Krivorotov, 2023; Drechsler et al., 2025).[6] Because we find that borrowers approved by the profitability model are riskier than those approved by the default model, a shift by lenders to models predicting profitability could reduce financial stability in the presence of correlated shocks to default (Krivorotov, 2023).[7] Furthermore, because fewer borrowers are approved by the profitability model than by the default model (when a profit-maximizing threshold is implemented in both models to make lending decisions), shifting to profitability models would reduce credit access.

2 Institutional Context

2.1 Financial Inclusion and Credit Cards

Only 49% of Mexicans have bank accounts, 44% have made or received digital payments, and 12% have credit cards, all significantly below the equivalent rates for countries with similar levels of development. Moreover, women are 14 percentage points (p.p.) less likely than men to have a bank account in Mexico, 11 p.p. less likely to have made or received digital payments, and 8 p.p. less likely to have a credit card. These gender gaps are significantly higher than those of other countries in Latin America and of other OECD countries (Demirgüç-Kunt, Klapper, Singer, and Ansar, 2022).

Credit cards are one of the most common ways for new borrowers to access formal credit in Mexico, where credit cards are the first formal loan type for 74% of all formal sector borrowers (Castellanos et al., 2024). Credit cards in Mexico share several features with credit cards in other countries. Much like in the US, they are both a form of payment and a source of financing. Card

[6] Agarwal, Presbitero, Silva, and Wix, 2023 allow for heterogeneity in interchange rates across card types, assuming interchange rates of 2.75% and 1.25% for reward and non-reward cards, respectively. Our data show that there is still substantial heterogeneity in interchange rates across borrowers within card type (our partner offers only one product) due to variation in interchange rates by merchant type and in borrowers' spending composition across merchants.

[7] In contrast to our paper, Krivorotov (2023) does not observe interchange fees, rewards, or funding costs.

holders use the card to purchase products at merchants that have a point-of-sale terminal. At the end of a 30-day billing cycle, card holders receive a statement showing (among other things) the balance at the end of the cycle (statement balance), the required minimum payment, and a due date, which is typically 20 days after the end of the cycle. To stay current, card holders need to pay at least the minimum payment by the due date. If they pay less than the full statement balance, they incur interest according to the interest rate assigned to the revolving line of credit. If they pay the statement balance in full, they do not incur any interest—effectively getting up to 50 days of free financing. As they make payments, their available credit line (i.e., the difference between their credit limit and outstanding balance) frees up.

Similar to the US, rewards tied to credit card transactions are a common instrument to promote card adoption and compete for consumers (Agarwal, Presbitero, Silva, and Wix, 2023; Wang, 2025). Our FinTech partner, for example, offers cashback of 1% of spending for most transactions and up to 5% of spending for transactions in certain establishments, with an average implicit cashback rate—calculated as cashback costs divided by spending—of 1.15% in the segment of borrowers without credit history.[8] In turn, card issuers receive an interchange fee charged to the merchant for every transaction, which vary by type of merchant and range from from 0% for non-profits and educational institutions to 1.76% for restaurants.[9] Our FinTech partner has an effective interchange rate—calculated as interchange fee revenue divided by spending—of 1.24% in the segment of borrowers without credit history.

In addition, unlike in the US, UK, and Nordic countries (Di Maggio, Williams, and Katz, 2022; Guttman-Kenney, Firth, and Gathergood, 2023; Laudenbach, Molin, Roszbach, and Sondershaus, 2025), but like in other countries including Turkey, South Korea, and Brazil (Aydin, 2022; Cho and Rust, 2017; de Lima Junior, Silva, Altoé Junior, and Ruhe, 2021), credit card holders in Mexico have access to "Buy Now, Pay Later" (BNPL) installment loans through their credit card. Card holders can finance purchases of specific items from specific merchants with short-term loans that allow payment in installments over a set period, typically ranging from 3 to 12 months and, in some cases, up to 18 months. These installment loans can be interest-free or carry interest rates lower than those applied to revolving balances. If the monthly installment is not fully covered by the payment against the corresponding bill, the unpaid amount is financed through the revolving line of credit, potentially leading to additional interest charges but not resulting in delinquency status.

[8]BNPL transactions do not accrue cashback and as a result, implicit cashback rates are less than 1% for some consumers.

[9]Mexico's Central Bank publishes on its website the levels of interchange fees by merchant type that financial institutions have agreed upon, which have not changed since 2013; see `https://www.banxico.org.mx/servicios/cuotas-de-intercambio-por-el-uso-de-tarjetas-de-cr/cuotas-intercambio-tarjetas-c.html`. For merchants that use point-of-sale (POS) terminals from FinTech payments companies (as in Gertler, Higgins, Malmendier, and Ojeda, 2025), rather than POS terminals from banks, the interchange fee is 1.76% regardless of the type of establishment.

BNPL via credit cards accounts for 47% of credit card balances in Mexico (Banco de México, 2024) and 32% for our FinTech partner. The establishment selling the product bears the financing cost—either fully (for interest-free BNPL loans) or partially (for interest-bearing BNPL loans)—by paying an upfront fee to the credit card issuer, who in turn finances the transaction and bears the risk.

2.2 FinTech Lending

Fostering a dynamic FinTech environment has been part of regulators' strategy to promote financial inclusion in Mexico. In 2018, the Mexican Congress passed a FinTech law and, as of the end of 2023, Mexico is one of the largest FinTech markets in Latin America with 650 FinTech start-ups (CNBV, 2019; Department Of Commerce, 2023). The most active segment of FinTech activity is lending, with 146 companies active in this space, followed by payments and remittances, personal financial management, and crowdfunding (Finnovista, 2023).

One of the main products through which FinTech companies lend is credit cards (CNBV, 2023). Traditionally, the credit card market in Mexico has been dominated by a few large banks: as of December 2021, the top two largest banks controlled 57% of the cards issued by traditional financial institutions and the top five largest banks controlled 87% of them. However, in 2022 one of the main drivers of the growth in consumer credit was credit cards issued by FinTech lenders (CNBV, 2023), with the largest FinTech lender becoming the fifth-largest credit card issuer in the country.[10]

In contrast to the U.S. where regulatory agencies have been actively evaluating conditions in which alternative data can be used in credit origination decisions (Bureau, 2017; Bureau, 2019), and where some lenders have explicitly requested No-Action-Letters to ensure their models are compliant with appropriate regulations before using them (Di Maggio and Ratnadiwakara, 2024), in Mexico there is no regulation explicitly limiting the variables that FinTech lenders can use for credit origination. Both FinTech and traditional lenders increasingly use alternative data in their credit origination decisions (Economista, 2024).

2.3 Delivery Platforms

RappiCard Mexico has access to transaction data from Rappi, the leading on-demand delivery platform of Latin America. An on-demand delivery platform connects customers with couriers via mobile apps or websites for immediate or scheduled deliveries of goods or services to desired locations within set time frames. Rappi provides a variety of services through its mobile

[10] See `https://www.bloomberglinea.com/2023/02/27/neobanco-nu-es-el-quinto-emisor-de-tarjetas-de-credito-en-mexico-moodys/`.

app, including the purchase of groceries, household items, restaurant food, alcoholic beverages, and pharmaceutical products, as well as booking of flights and hotels. It also allows users to request cash withdrawals and the execution of miscellaneous errands. Orders are completed by local couriers, typically within 30 minutes to one hour. Delivery apps are a growing business in Mexico. In the first quarter of 2023, 24% of mobile phone users had at least one delivery app installed on their phone, representing a 142% increase since 2019 (Trecone, 2023). The market is concentrated among three players who, as of the latest counts, operate in approximately 100, 80, and 57 cities in Mexico, respectively.[11]

3 Data

The data for our analysis was provided by RappiCard Mexico. To apply for a credit card, individuals must have an account with Rappi and complete the application through its mobile app. There is no requirement for a minimum number of transactions nor a waiting period after account creation.[12] Applicants need only provide their full name, address, date of birth, and tax identification number, and consent to a credit check.

3.1 Sample

Our data set consists of information from 181,488 credit cards originated between January 2021 and May 2024 for borrowers with no credit history. This sample includes all applicants who were flagged by the Mexican credit bureau as having null or insufficient credit history to have a traditional credit score, had no prior credit card, and applied for and received a RappiCard credit card over this time period.[13] For borrowers with no credit history over this time period, RappiCard did not use a machine learning algorithm; instead, approval decisions were based on a set of parsimonious rules. These parsimonious rules were based on a threshold for the no-hit score (which is based only on where an individual lives and not any individual-specific data) and sometimes a threshold on a second variable.[14]

The 181,488 credit cards originated between January 2021 and May 2024 are the applications that were approved using these parsimonious rules among 1,328,426 applications from borrowers

[11] See `https://www.forbes.com.mx/rappi-ya-rueda-en-100-ciudades-de-mexico-dolores-hidalgo-la-ultima-en-sumarse/`, `https://web.didiglobal.com/mx/conductor/ciudades`, and `https://www.uber.com/es-MX/newsroom/uber-eats-expansion-en-mexico/`.

[12] Burlando, Kuhn, and Prina (2025) study the effects of a digital lender in Mexico imposing a waiting period.

[13] The filter of having no prior credit card removes those who had a credit card prior to applying for the RappiCard but who nevertheless had an insufficient credit history for the credit bureau to generate a credit score for them.

[14] For example, in the past RappiCard purchased scores based on cell phone records which are sold to lenders by independent local providers. RappiCard implemented a threshold using this variable in addition to the no-hit score threshold to determine which applicants received credit cards.

with no credit history and no prior credit card.

For these borrowers, we also have data on balances and repayment from origination through May 2024. To not understate default rates associated with recent or completely inactive card holders, we impose two additional restrictions on the analysis sample: card holders (i) must have had the card for at least twelve months by the end of our repayment data in May 2024 and (ii) must have completed at least one transaction using their credit card within twelve months of origination. This leaves us with an analysis sample of 146,036 borrowers. Figure A.1 shows the number of applications per month and the number of originated contracts per month from the beginning of our sample period through May 2023 (given the restriction that the borrower must have the card for at least twelve months by the end of our repayment data in May 2024, which effectively filters out all applications after May 2023).

During the timing of our sample, from January 2021 to May 2024, Mexico had a relatively stable macroeconomic environment (Figure A.2). Unemployment, measured monthly, shows a slight downward trend over this time period, starting at 4.5% and reaching 2.7% in June 2024, with an average of 3.3% over the period; this can be compared with average unemployment of 4.1% from 2006 (the earliest we have unemployment data) to 2024. Annual inflation shows significant variation starting at 3.5% in January 2021, reaching a maximum of 8.7% in August 2022, and declining to 5.0% in June 2024; average inflation in Mexico was 8.1% from 1995–2024 and 4.5% from 2006–2024. Quarterly GDP growth averages 0.7%, ranging from -1.0% to 1.5% without a distinctive trend, which can be compared to average quarterly GDP growth of 0.6% from 1995–2024 and 0.4% from 2006–2024.

3.2 Data Sources and Measurement

For each approved applicant, we observe the following information:

> *Digital footprint user characteristics*, such as gender, operating system, device model and type, acquisition channel, and email provider, and explicitly including all variables in the digital footprint identified by Berg, Burg, Gombovi, and Puri (2020).[15]
>
> *Transaction-level data from the delivery platform,* including date and time of the order placed, a list of each item purchased, the quantity of each item purchased, its unit price, fees, discounts, tips, and total order cost. The data also include payment method (credit card, debit card, or cash), store name, and geographic identifiers for the store. This is more granular than traditional transaction-level data from credit or debit cards (as used in, e.g.,

[15]The variable “email error" used in Berg, Burg, Gombovi, and Puri (2020) is not applicable in our setting. This variable captures when an email address is invalid. A valid email address is required to have an account with the delivery app. As a result, all credit card applications are associated with a valid email address.

Higgins, 2024), as it allows us to observe not only the shop where the order was placed, but the specific items purchased from that shop.

"No-hit" scores. All of the applicants in our sample are referred to by the credit bureau as the "no-hit segment." This means that they have no formal credit history or too limited of a credit history for the credit bureau to use those data to provide a credit score. For them, the credit bureau issues a flag indicating that the traditional score (built from credit histories) is not applicable. Beginning in 2018, the credit bureau contracted a third party to develop a "no-hit" score for all Mexicans who do not have a traditional credit score. The no-hit score is based on geographic indicators merged with the location where the individual lives. The geographic indicators come from a variety of public records, including demographics, economic activity, public safety, social cohesion, and access to and use of credit at the local level (see CRIF, 2018). Traditional credit scores and no-hit scores are independent from each other, with traditional scores ranging 456 to 760, and no-hit scores ranging from 463 to 735. The no-hit segment is thus distinct from the subprime segment of the traditional market (studied in the US in Di Maggio and Ratnadiwakara, 2024)—identified by low values on the traditional credit score—and by those with thin credit files that are nevertheless sufficient for the credit bureau to generate a credit score (studied in the US in Blattner and Nelson, 2024).

Credit history for those with limited credit history. For borrowers in our sample who do have a credit history—all of whom have an insufficient credit history for the credit bureau to assign a traditional credit score—we observe length of credit history and balances (if any). Our sample excludes anyone who had a credit card prior to applying for a card from our FinTech partner. While the credit bureau's rules on what constitutes a sufficient credit history to generate a credit score are proprietary, these rules are unlikely to differ between Mexico and other countries such as the US since the credit bureau in Mexico is TransUnion.

Socioeconomic characteristics at the census tract level, obtained by combining publicly available information from Mexico's National Institute of Statistics (INEGI) with location information collected by the delivery platform whenever a user logs in.

Furthermore, we observe the transaction-level data and "no-hit scores" for applicants who were rejected by RappiCard based on their parsimonious rules, but despite RappiCard observing these data sources for all applicants, we did not receive the other data sources for rejected applicants.

We also observe the following variables related to credit card terms and use for accepted applicants:

Credit card terms assigned at loan origination, including the interest rate and credit limit.

Account-by-month level data on statement balances, minimum payments, and repayments. The minimum payment is the minimum amount the borrower must pay to avoid their balance being considered past due and, by regulation, must cover the interest charges of the period plus a fraction of the outstanding balance (Medina and Negrin, 2022).

Account-by-day level data on arrears. When a borrower does not pay the minimum balance on time, their statement balance is considered past due. We observe account-by-day records of arrears (tracking the number of days past due).

Transaction-level data on credit card spending, including the date, location, merchant category code of the merchant, interchange rate for the transaction, and the amount spent on the transaction.

For all cards issued during our observation period, we also observe account-by-month information on realized revenues and costs for the FinTech lender. These are the variables used in practice by our FinTech partner to evaluate the profitability of each account.

Interest revenue corresponds to the *realized* revenues from interest payments made by the borrower as part of their monthly payments.

BNPL revenue includes fees paid by merchants who offer installment payments (often interest-free) to their customers to increase demand, as well as interest charges paid by the card holder when the installment plan is not interest-free.

Interchange fee revenue from the interchange fees received by the card issuer for credit card transactions made by the borrower. As in the US, interchange fees vary by type of merchant (GAO, 2009); in our context, they vary from 0% to 1.76%.

Revenue from fees including revenue from late payment fees and revenue from other fees such as card replacement fees and cash advance fees for ATM withdrawals using the credit card. This lender does not charge an annual fee.

Charge-offs, which correspond to balances deemed not collectible by a lender and removed from its balance sheet. This measure captures two distinct regulatory concepts. The first, known as *quitas*, refers to partial reductions (or haricuts) on balances renegotiated with delinquent borrowers. These accounts may be reactivated depending on subsequent payment behavior, but the reduction amount is considered unrecoverable and written off. Renegotiations start at the discretion of the lender or borrower for accounts between 30 and 180 days delinquent. The second concept, *castigos*, refers to balances written-off on accounts that are permanently closed due to failure to repay. The time of write-off is at the discretion of the

lender. The practice of our partner is to close accounts permanently when they reach 180 days in arrears.

Funding costs are defined, for each account-month, as average daily balances multiplied by the lender's cost of capital.[16] Funding costs are non-zero both for revolvers and for transactors—who pay off their balances in full each billing cycle and do not revolve debt—since revolvers carry balances month to month and transactors get up to 50 days of free credit due to the billing structure of credit cards (see Section 2.1).

3.3 Summary Statistics

Table 2 shows descriptive statistics for our modeling sample, i.e., the 146,036 applicants with no credit history and no prior credit card who were approved by RappiCard using their parsimonious approval rules for borrowers with no credit history, and who also had at least twelve months with the card by the end of our data period in May 2024 and made at least one transaction during their first twelve months with the card.[17]

Table 2, Panel A shows a subset of the features that are used in our machine learning models. The applicants in our sample are relatively young: the average user age in our sample is 25. Younger people are less likely to have a credit score (Cookson, Guttman-Kenney, and Mullins, 2025), more likely to use smartphones and delivery apps, and also more likely to consider a FinTech lender as a potential source of credit (Doerr, Frost, Gambacorta, and Qiu, 2022; Krivorotov, 2023), including being more willing to share transactions data with lenders (Armantier et al., 2024). Less than half of the sample (38%) uses an Apple product, which is an important predictor of creditworthiness (Berg, Burg, Gombovi, and Puri, 2020). There is not a lot of variation in the no-hit score, which has a mean of 641, a standard deviation of 14, and an interquartile range of 635 to 649; this is not surprising given that the no-hit score is based only on publicly available geographic-level information merged with the location of the applicant.[18] There is also little variation in the census tract-level variables: for example, the marginality index has a mean of 0.96 and a standard deviation of 0.01.

There is substantially more variation in measures from the transaction-level data. The average number of orders on the app is 24 with a standard deviation of 52 and interquartile range of 3 to 22, the average percent of orders paid in cash is 51% with an interquartile range of 17% to 90%, and the median amount per order is 350 MXN with a standard deviation of 345 pesos. The majority (82%) of purchases are orders from food establishments, while 5% are from supermarkets and 4%

[16] Our partner directly provided account-by-month level funding costs using their internal cost of capital.

[17] In addition, we lose less than 0.5% of the observations due to these borrowers not appearing in the profit components data set due to marginal inconsistencies in the samples included in different data tables.

[18] We refer to this as little variation since no-hit scores range from 463 to 735.

are from pharmacies.

Table 2, Panel B, shows descriptive statistics on credit card terms and use of the cards. The average annual interest rate is 80%. This is not substantially higher than the average interest rate across all credit cards in Mexico, which is 64% based on data from Mexico's Central Bank (and it is not surprising that RappiCard's interest rate for borrowers with no credit history would be higher than the average interest rate across all credit cards in Mexico). The lender has used different interest rates over time, but for the segment with no credit history has broadly followed a policy of a single interest rate at any point in time, except during periods of transition from one interest rate to another.[19] During the early part of our sample period in 2021, nearly all cards were originated with a 72% annual interest rate, while they were originated with an 87% annual interest rate in the first half of 2022 and an 80% annual interest rate in the second half of 2022 and throughout 2023 (Figure A.3). It is not uncommon for other FinTech lenders to similarly engage in limited or no risk-based pricing.[20]

The average credit limit at origination is 5,647 MXN. In terms of card use, average spending is 3,103 MXN per month, the average statement balance 3,774 MXN, the average minimum payment 460 MXN, and the average repayment 2,963 MXN.

Table 2, Panel C, shows account-by-month level averages of the various revenues and costs incurred by the lender on that card. The average revenues from interest payments and BNPL are 65 pesos and 7 MXN per card per month, respectively, while the average amount lost to charge-offs is 61 MXN per card per month. Combining these, average interest and BNPL revenue net of charge-offs is 11 MXN per card per month. Average interchange fee revenue is 39 MXN per card per month, while average rewards on each card are 36 MXN per month, leaving 3 MXN per card per month in interchange fee net of rewards. Other revenues for the lender include late payment fees (13 MXN per card per month on average) and other fees (3 MXN), while other costs include costs from fraudulent transactions (0.5 MXN per card per month) and other costs (12 MXN). Finally, funding costs are 31 MXN per card per month on average.

[19] In contrast, the lender does use risk-based pricing for borrowers who do have a formal credit history; our sample does not include borrowers with a formal credit history, however.

[20] In practice, while many FinTech lenders do some form of risk-based pricing (as our partner does for loans to borrowers with a formal credit history), the variation in interest rates is often explained mostly by the applicant's formal credit score (Johnson, Ben-David, Lee, and Yao, 2023). It is also not uncommon for FinTech lenders to charge the same interest rate to all borrowers (as our partner does for loans to borrowers without a formal credit history): for example, the FinTech lenders in Berg, Burg, Gombovi, and Puri (2020), Choi et al. (2025), and Yang (2025) charge the same rate for all borrowers. Furthermore, in the UK, even banks keep credit card interest rates nearly constant across consumers of varying default risk (Matcham, 2025).

3.4 Target Default Variable

We define the target variable for the machine learning model predicting default as overdue for more than 60 days at any point during the first twelve months since origination, which we refer to throughout the paper as "default." We compare the definitions of default used by various papers in Table A.1, column 2, which shows that there is no standard definition across studies. This likely also reflects a large amount of heterogeneity across the definitions of default used by lenders in their credit scoring models.

Two choices must be made for the measure of delinquency: how many days overdue must a borrower be, and over what time period should delinquency be measured? We choose 60 days overdue because this is the threshold used by RappiCard in their credit scoring models, as they consider this to be an early warning of future charge-offs. Furthermore, 60 days was the regulatory definition of default for credit cards in Mexico through January 2022 (Banco de México, 2021). Table A.1 shows that 60 and 90 days delinquent are both commonly-used thresholds for credit cards and other consumer credit products in other studies.

Regarding the length of time over which delinquency is measured, we note that unlike for installment loans where there is a clear time period over which to measure default (i.e., the maturity of the loan), for credit cards there is no clear time period over which to measure default. Furthermore, there is a trade-off in the length of time we use: for shorter periods of time, we may have substantial measurement error as we label some borrowers who will default in the future as non-defaulters. On the other hand, if we use a long time period, we limit the sample that can be included in the model. (For example, since our data end in May 2024, if we used 24 months as the relevant time window, we would only be able to include those who applied by May 2022 in the model.) This trade-off may be particularly acute for FinTech lenders, as they are relatively recent entrants to the market and may not have a sufficiently large sample to measure default over a longer time horizon. An additional drawback of using a longer time window is that it lengthens the cycle of model deployment.

Figure A.4 illustrates this trade-off. In panel (a), we plot the cumulative proportion of borrowers who are at least 60 days delinquent as of x months after card origination. Because this cumulative proportion continues increasing over time since origination, any threshold of x months since origination will have the drawback of mislabeling some borrowers who default after month x as non-defaulters, and the shorter the time period considered, the greater the extent of this mislabeling. On the other hand, panel (b) shows the size of the modeling sample based on different time periods. To determine the size of the modeling sample, we impose two restrictions based on the threshold of x months since origination: (i) the account must be observed for x months by the end of our data period in May 2024, and (ii) the card must have at least one transaction within x months since origination. The number of observations in the modeling sample initially increases because

the second constraint dominates: some borrowers do not make a transaction until they have had the card for a few months. After four months since origination, the number of observations in the modeling sample begins decreasing because the first constraint dominates: for a given value of x, the sample can only include those who applied for a card at least x months before our data end in May 2024.[21]

We use a threshold of 12 months since origination, which leaves us with a modeling sample of 146,036 borrowers. With this definition, 20% of the modeling sample is delinquent; Table A.1, column 3, shows how this compares to delinquency rates in other studies.

3.5 Target Profitability Variable

To measure profitability, we use the data on account-by-month level profitability described in Section 3.2 to measure the profits obtained by the lender on each card, restricted to the first 12 months since card origination to be comparable to the default model. Consistent with the practice of our FinTech partner, we define realized profits at the account-by-month level as the sum of interest revenue, BNPL revenue, interchange fee revenue, revenue from late-payment fees, and revenue from other fees, minus charge-offs, cost of rewards, funding costs, cost of fraudulent transactions, and other costs. This definition is consistent with Agarwal, Chomsisengphet, Mahoney, and Stroebel (2015) and broadly consistent with Guttman-Kenney and Shahidinejad (2023).[22] We then aggregate information of the first 12 months since card origination to compute account-level profits as

$$\begin{aligned}\text{Profits}_i = \sum_{t=1}^{12} &\text{Interest and BNPL revenue net of charge-offs}_{it} + \text{Interchange revenue net of rewards}_{it} \\ &+ \text{Late fee and other fee revenue}_{it} - \text{Funding costs}_{it} - \text{Other costs}_{it}, \end{aligned} \tag{1}$$

where t defines months relative to card origination.

We then create a binary variable equal to 1 if the borrower is unprofitable, i.e., if they generate negative profits for the lender. We create this binary variable so that we can use the same type of model to predict profitability as we use to predict default. We leave to future research the question

[21] An alternative would be to consider defaults over the entire time period for which we observe data, which varies by card. For example, for a card originated in November 2022, we would use the 18 months of data on default, while for a card originated in November 2023, we would use the 6 months of available data on default. The drawback of this method is that we observe default over different time periods for different cohorts of applicants, and the type of people applying for cards in November 2022 might differ from the type of people applying for cards in November 2023. Nevertheless, we conduct a robustness test using this definition of default, and estimate a slightly lower AUC of 0.776 (Table A.2). (When using this alternative method of defining the time period over which we measure default, we impose a restriction that card holders must have had their card for at least 4 months, but measure default over however many months there are between the card's month of origination and May 2024.)

[22] In contrast to Guttman-Kenney and Shahidinejad (2023) we do not subtract acquisition costs, since those are sunk at the time of origination and, as a result, are not relevant in this setting.

of whether higher profits can be achieved by predicting the continuous measure of profits or using a more sophisticated machine learning model of profits—such as one separately predicting each component of profits and determining how to optimally weight each component (see Section 7.2).

4 Machine Learning Methods

4.1 Algorithm Details

We use data on default and profitability to train machine learning models using extreme gradient boosting, or XGBoost (Chen and Guestrin, 2016). Like random forests (Breiman, 2001), XGBoost is an ensemble learner. Ensemble learning is a process that combines several base predictors to produce improved accuracy or stability (Yin and Li, 2022). However, XGBoost and random forests differ in the way they merge predictions from multiple weak models to produce more accurate predictions. Random forests train multiple independent models in parallel and combine the results of multiple classifiers modeled on different subsamples of the data.[23] XGBoost, like other boosting methods, adds new models into the ensemble sequentially, where each subsequent model attempts to correct the errors of the previous one. In particular, with boosting methods, the training data for each subsequent classifier increasingly focuses on instances misclassified by previously generated classifiers.

XGBoost has become the standard in industry and academic settings due to its scalability and accuracy. It has been shown to outperform other machine learning algorithms in many predictive modeling tasks (Mienye and Sun, 2022).[24] While manual hyperparameter tuning is essential and time-consuming in many machine learning algorithms, it is especially so in XGBoost. We use Bayesian optimization to tune hyperparameters (for both our default and profitability models), relying on sequential model-based optimization as in Bergstra, Yamins, and Cox (2013). Bayesian optimization is more efficient than grid or random search because it attempts to balance exploration and exploitation of the search space. It is also well-suited for cases with a large number of hyperparameters and large search space. Details on the search space we adopt can be found in Table A.3. The Bayesian optimization algorithm was implemented with the aid of 3-fold cross-validation.

XGBoost is the algorithm of choice in other recent work that relies on machine learning to predict creditworthiness (Agarwal, Alok, Ghosh, and Gupta, 2023; Blattner and Nelson, 2024; Blattner, Nelson, and Spiess, 2024; Lee, Yang, and Anderson, 2024; Lee, Yang, and Anderson, 2025). Contributions not using XGBoost opt for random forests (Björkegren and Grissen, 2020;

[23] Breiman's (1996) bagging (bootstrap aggregation) instances selected to train individual classifiers are bootstrapped replicas of the training data, with each instance having equal chance of being in each training set (Yin and Li, 2022).

[24] The combination of ensemble learning, gradient descent optimization, and regularization techniques are some of the elements that explain XGBoost's performance and popularity.

Butaru et al., 2016; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2022; Huang et al., 2023; Netzer, Lemaire, and Herzenstein, 2019; Rishabh, 2024) or other methods such as logistic regression (Berg, Burg, Gombovi, and Puri, 2020) and deep neural networks (Sadhwani, Giesecke, and Sirignano, 2021).[25] Table 1 presents an overview of papers that employ machine learning to predict creditworthiness, including country, target populations (and in particular the fraction of the target population with a conventional credit score from the credit bureau), data, and methods.

Our models learn on a training set and are evaluated on a testing set. The training set corresponds to 80% of the modeling data set and is a random sample of the modeling data, stratified by gender and target variable (default).[26] Stratification guarantees that the incidence of each class (default and no default) is preserved in both sets. The testing set—i.e., the remaining 20% of the modeling data—permits us to assess model performance on data unseen by the algorithm, as well as to guard against overfitting.

Our models are trained by minimizing *log-loss* or, equivalently, *cross-entropy loss*. Intuitively, log-loss measures how close the predicted probability is to the corresponding actual/true value (0 or 1 in the case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value. Log-loss penalizes highly confident incorrect predictions; it takes into account the quality of the predicted probabilities, not only the predicted class labels. As such, log-loss is well-suited to problems for which probability estimates are an object of interest—as they are for a lender that will allocate credit to everyone with predicted probabilities of default below a threshold (or credit scores above a threshold). In addition, log-loss allows for a more nuanced evaluation of uncertainty. This contrasts to specifying that the model's objective function is to maximize metrics such as the area under the receiver operating characteristic curve (AUC-ROC) or the area under the precision-recall curve (AUC-PR), as these metrics solely focus on the models ability to discriminate between defaulters and non-defaulters, without penalizing miscalculations in predicted probabilities.

Log-loss also has the advantage of delivering predicted probabilities which are already calibrated, especially in our setting where the label is not subject to misclassification and in combination with 3-fold cross-validation, regularization, and careful data splitting to avoid overfitting. We show calibration curves comparing predicted and true frequencies of the positive label in Figure A.5, which show that the model is well-calibrated.[27]

[25]Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022) use random forests as their main method, but also use XGBoost in robustness tests.

[26]To accurately compare performance and predictions for the default and profitability models and ensure that observed differences are not driven by differences in training-testing splits, we always use the training-testing split for the default models, which is stratified by gender and default.

[27]Calibration techniques like isotonic regression (Zadrozny and Elkan, 2002) and platt-scaling (Böken, 2021) require a validation set, typically 10% of the modeling sample, which can be costly in terms of sample size and can result in less efficient use of the data with possible implications for model performance. As a robustness check, we

Finally, in line with recent recommendations, we do not correct for class imbalance. The observed imbalance in our data arises naturally in the context of our problem and is not related to biased sampling and/or incorrect labels (i.e., misclassification).[28] Furthermore, our default rate/imbalance is 20%, which can be considered moderate compared to settings with very low default rates such as mortgages (e.g., Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2022). Finally, recent work has pointed to the potential unintended consequences of class imbalance corrections for risk prediction models: these corrections can yield poorly calibrated models, where the probability of belonging to the minority class is strongly overestimated, without delivering higher AUC-ROCs compared to models trained without class imbalance correction (van den Goorbergh, van Smeden, Timmerman, and Van Calster, 2022; Piccininni et al., 2024).

4.2 Model Performance Measures

We report four measures of model performance: AUC-ROC, precision, recall, and F1 score. The AUC-ROC measures the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate for all thresholds. Thus, the AUC-ROC, often abbreviated simply as AUC, is a threshold-free measure. An AUC of 0.5 implies that the model performs no better than random guessing, while an AUC of 1 implies that the model makes perfect predictions.

We use AUC as our main performance metric, as it is reported in nearly all studies and is threshold-free. Even when considering performance across settings with different proportions of the positive class (which in our context means different default rates) and even in the presence of *class imbalance* (i.e., when the incidence of one class is significantly higher than the other, such as far fewer defaulters than non-defaulters), comparisons using AUC remain meaningful due to its mathematical and empirical properties. While AUC does not depend on the proportion of the positive class in theory based on its mathematical definition (Flach and Kull, 2015), some have argued that in practice AUC could mask performance issues when dealing with highly or extremely imbalanced data (e.g., Davis and Goadrich, 2006), such as in settings in which the prevalence of the positive class is below 10%. Richardson et al. (2024) carefully analyze this point and show empirically, via simulation and using real-world data, that AUC is a robust and appropriate comparison metric in many scenarios with different class imbalance, and argue that

ran a version of our models in which we split the data into 80% training, 10% validation, and 10% testing sets and used platt-scaling to calibrate probabilities. Our calibration curves in these robustness tests were largely identical to the calibration curves shown in Figure A.5.

[28] In other words, because we observe the ground truth of whether someone is at least 60 days delinquent over the first 12 months since card origination, there is no misclassification. In contrast, in other contexts (e.g., cancer presence based on imaging, mental health condition based on self-reported symptoms, content meditation, and sentiment analysis) labels can be subject to measurement error misclassification.

AUC provides fair comparisons of models even across contexts with different default probabilities (Richardson et al., 2024).

Furthermore, our default rate of 20% does not place us in a regime typically considered highly or extremely imbalanced and, as such, we are less at risk of metric distortion (e.g., AUC inflation) or model instability.

We also assess model performance using precision, recall, and F1 score. Because these measures depend on the approval threshold, however, comparisons across papers and across models are potentially misleading if the classification thresholds differ. Precision measures the proportion of predicted positive cases that are actually true positives (in our context, the proportion of predicted defaulters who actually default); this is calculated as the true positives divided by the sum of true positives and false positives. Recall measures the proportion of actual positive cases (actual defaulters) that were correctly identified by the model, calculated as true positives divided by the sum of true positives and false negatives. The F1 score is the harmonic mean of precision and recall.

We take an economic approach to determining the threshold to use for the threshold-dependent performance metrics (as well as for confusion matrices and the comparison of individuals approved and rejected by the default vs. profitability models). In particular, we use detailed account-by-month level data on profits and combine predicted probabilities of default *in the training sample* (to avoid overfitting) with the realized profits from each account to determine via grid search the profit-maximizing approval thresholds: these thresholds are a 24% predicted probability of default and a 47% predicted probability of negative profits. We then evaluate approval decisions *in the testing sample* using these profit-maximizing thresholds determined in the training sample to estimate the threshold-dependent performance metrics and confusion matrices, to compare borrowers approved and rejected by each model, and to estimate the overall profits earned by the lender with each model.

5 Results

5.1 Default Model

The benchmark model, which uses all available features, achieves an out-of-sample AUC of 0.796 (Table 3, Panel A). Studies in highly data-rich environments, in which credit scores are often included in the algorithm, obtain AUCs typically in the 0.66 to 0.88 range (e.g., Blattner and Nelson, 2024; Blattner, Nelson, and Spiess, 2024; Di Maggio and Ratnadiwakara, 2024; Meursault, Moulton, Santucci, and Schor, 2024; Netzer, Lemaire, and Herzenstein, 2019). In contrast, AUCs estimated by studies in middle-income countries are lower, typically in the 0.61 to 0.76 range (e.g., Agarwal, Alok, Ghosh, and Gupta, 2023; Frost et al., 2019; Gambacorta, Huang, Qiu, and Wang,

2024; Lee, Yang, and Anderson, 2024; Rishabh, 2024). The AUC of our model predicting default is at the upper end of those from middle-income settings—which include traditional credit scores as inputs in their algorithm, unlike ours since our sample has no formal credit bureau score (Table 1).

Our model predicting default also performs well using threshold-dependent metrics (precision, recall, F1 score), and AUC-PR (reported in Table A.1). Table 3, Panel A, reports the threshold-dependent metrics for our model, where we use the profit-maximizing threshold (estimated in the training data) for the threshold-dependent performance metrics. The model has a precision of 0.421, recall of 0.666, and an F1 score of 0.516. Estimates of these metrics from other studies, albeit with different approval thresholds that complicate comparing performance across studies, are shown in Table A.1.

Table 3, Panel B, shows a confusion matrix for the default model, again using the profit-maximizing approval threshold. We report the percent of observations in the testing sample based on whether they would be approved or rejected by the default model and whether they ultimately default or not. The proportion approved by the default model who do not default (true negatives, since "default" is the positive class in our model) is 30.8%, while the proportion approved by the model who do default (false negatives) is 4.0%. The proportion rejected by the model who do not default (false positives) is 49.0%, while the proportion rejected by the model who do default (true positives) is 16.2%.

We next assess the importance of each data source by comparing the AUC and other performance metrics of our benchmark model using all of the data sources to that of separate models trained with all features but one data source (Table 4). The digital footprint user characteristics, which include the same set of features as in Berg, Burg, Gombovi, and Puri (2020), have the largest marginal contribution to the AUC: the AUC of a model with all data sources except the digital footprint user characteristics is 0.672, a reduction in AUC of 0.124 compared to the benchmark model. The transaction-level data from the delivery platform is the second most-important data set, as a model without that information has an AUC of 0.770, a reduction of 0.026. Omitting the no-hit score and limited credit history or census tract-level socioeconomic characteristics lead to smaller AUC reductions of 0.006 and 0.001, respectively. The finding that the digital footprint and transactions data are the most important data sources for the model's performance is robust to using precision, recall, or F1 to measure predictive accuracy.

Given the importance of the transaction-level data in our FinTech lender's competitive advantage over other lenders, as well as its high marginal contribution to the AUC relative to other data sources (except the digital footprint data), we next assess how the predictive accuracy of the model varies by the "thickness" of a user's transaction history. The number of transactions made through the app may be analogous to a formal credit history in the sense that the model might perform

more poorly for those with a "thin" transaction history (few transactions) compared to those with a "thick" transaction history (many transactions). To assess this, we segment the data into quintiles by number of transactions and estimate separate machine learning models for each quintile.[29] Indeed, the predictive power of the models is broadly increasing in transaction history: for those in the first quintile with only 2 or fewer transactions, the AUC is 0.741, while for those in middle quintile with 7–13 transactions the AUC is 0.772, and for those in the fifth quintile with 29 or more transactions the AUC is 0.838 (Table 5).

5.2 Profitability Model

We next consider a model which uses profitability as target variable. Based on our main performance metric that is not threshold-dependent, the AUC, the profitability model does not perform nearly as well as the model predicting default: the profitability model has an AUC of 0.598 (Table 3, Panel A). The profit-maximizing threshold for the profitability model is a 47% probability of negative profits. Using this threshold, the profitability model is substantially more restrictive than the default model, approving only two-thirds as many applicants. In the confusion matrix (Table 3, Panel B), the proportion approved by the profitability model who are actually profitable (true negatives, since a binary measure of "negative profits" is the positive class in our model) is 20.5%, while the proportion approved by the model who generate negative profits (false negatives) is 14.3%. The proportion rejected by the model who are actually profitable (false positives) is 30.7%, while the proportion rejected by the model who generate negative profits (true positives) is 34.5%.

When we turn to threshold-dependent metrics, the profitability model has recall of 0.519, precision of 0.583, and an F1 score of 0.549 (Table 3, Panel A). We do not directly compare these to the corresponding metrics for the default model, given that the two models use different approval thresholds. In particular, the profit-maximizing threshold of the profitability model leads to approving fewer applicants than are approved by the default model with a profit-maximizing threshold, and for a given model, applying a more selective threshold will mechanically reduce recall and increase precision.

For the lender, the metric that ultimately matters is profits. Conceptually, from the perspective of a lender trying to maximize profits, we can think of two types of errors made by a model predicting default: statistical errors and economic errors. Statistical errors occur when the model mispredicts a user's true label/status (i.e., whether the user defaults or is unprofitable, depending on the model). Economic errors occur only for the model predicting default, because a borrower may not default but still be unprofitable or may default but still be profitable. However, the default model may still generate higher profits despite making economic errors if it makes fewer statistical

[29] Table A.4 shows summary statistics by quintile of number of transactions.

errors than the profitability model, and the comparison of the models' predictive performance above suggests that the default model indeed makes fewer statistical errors than the profits model.

We thus compare the total profits generated by the default and profitability models. In Figure 1, we sum the profits obtained by the lender across all individuals who would be approved based on a given approval threshold (i.e., whose predicted probability of default or negative profits is below the threshold), and repeat this exercise for each approval threshold. As expected, total profits are highest near the profit-maximizing approval threshold (though the peak of the total profits curves do not correspond exactly to the profit-maximizing thresholds since the curves are based on lending decisions made in the testing sample, while the profit-maximizing thresholds are determined in the training sample). Total profits in the figure are normalized such that the most profitable model across all approval thresholds and across both the default and profitability models is equal to 1. Using the profit-maximizing thresholds, the profitability model generates only 51.9% as much profits as the default model.[30]

To understand the difference in total profits from the two models, we first plot average profits across bins of predicted probabilities for each model in Figure 2. In the default model, the average profits of the lowest-risk (lowest predicted probability of default) borrowers are positive but close to zero. These are users who do not carry revolving debt on the card, and thus are very unlikely to default but also do not generate interest revenue for the lender; they are also sophisticated about optimizing their shopping to generate rewards and thus do not generate interchange fee revenue net of rewards. Average profits increase as the predicted probability of default increases among the lowest-risk borrowers due primarily to more borrowers carrying revolving debt and making their interest payments. However, after the probability of default increases above about 10%, average profits for borrowers with predicted probabilities of default above that threshold begin to fall, as these borrowers become more likely to default; thus, a subset of these borrowers generate charge-off costs for the lender. Average profits fall below zero for predicted probabilities of default above around the profit-maximizing threshold of a 24% predicted probability of default.

In contrast, the average profits by bin of predicted probability of generating negative profits are consistently decreasing. Profits are positive and large for the borrowers with the lowest probability of negative profits, and steadily decrease until they cross zero and become negative around the profit-maximizing threshold of a 47% probability of being unprofitable. While the broadly monotonic relationship between average profits and predicted probabilities for the profitability

[30] We test the robustness of this finding in two ways. First, we conduct 10,000 bootstrap draws of the testing sample and, for each bootstrap sample, calculate the profits from each model and the ratio of profits from the profitability model to that of the default model. The 10,000 bootstrap estimates are shown in Figure A.6. The 95% confidence interval for the ratio of total profits from the profitability model to that of the default model is $[0.419, 0.618]$. Second, we estimate the ratio using an alternative testing-training split with a test sample that is disjoint from our benchmark test sample, and find a ratio of 0.43, which is within the 95% confidence interval of the ratio for our main testing-training split.

model—compared to the non-monotonic relationship for the default model—should in theory favor the profitability model, the lower average profits in bins with positive profits for the profitability model also indicate that even for the bins where profits are positive, a number of borrowers generating negative profits are receiving loans and pulling down the average within the bin.

We next compare the predicted probabilities of default and negative profits for each borrower, and the types of borrowers approved by each model. Figure 3 shows the predicted probabilities from the two models for each borrower in the test sample. The non-monotonic relationship between risk of default and profits (now measured by the predicted probability of negative profits) is again evident in this figure, consistent with Agarwal, Chomsisengphet, Mahoney, and Stroebel (2015): many of the lowest-risk borrowers tend to have a relatively high predicted probability of negative profits, but these probabilities then tend to initially decrease as the predicted probability of default increases. Then for predicted probabilities of default above about 10%, the predicted probabilities of negative profits tend to increase as the probability of default increases.

We also show the profit-maximizing approval thresholds in Figure 3, which are a 24% predicted probability of default and a 47% predicted probability of negative profits. The fraction of observastions in each quadrant is shown in Panel C of Table 3. The lower-left quadrant shows applicants who would be approved by both models (34.8% of the test sample), the upper-right quadrant shows applicants who would be rejected by both models (33.2%), the upper-left quadrant shows applicants who would be rejected by the profitability model but approved by the default model (21.1%), and the lower-right quadrant shows applicants who would be rejected by the default model but approved by the profitability model (10.8%).

We next explore how the borrowers in these four quadrants compare. Table A.5 shows summary statistics for each of these groups. Focusing on differences between borrowers approved by only the default model ("default-approved") and borrowers approved by only the profitability model ("profitability-approved"), default-approved borrowers are less likely to be women, are more likely to use an iPhone, have made more orders on the app, and are less likely to pay orders in cash. In terms of loan terms, they are awarded higher credit limits and spend more, but have lower statement balances and make larger payments, reflecting that they roll-over a smaller fraction of their balances from month to month. Based on these measures, the default-approved borrowers are lower-risk and higher-income (within this population of people with no credit history) than the profitability-approved borrowers.

Figure 4 and Table A.5, Panel C, compare the account-level revenues and costs for borrowers in these four quadrants.[31] Applicants approved only by the default model generate less in-

[31] Figure 4 includes the fraction of accounts with positive interest or BNPL in a single bar. As a complement to that figure: 67% of applications approved only by the default model have interest revenue, while 76% of those approved only by the profitability model have interest revenue.

terest revenue than those approved only by the profitability model (53 vs. 84 MXN per card per month). However, the default-approved applicants are much less likely to have charge-offs (8.1% vs. 28.6%), and average charge-offs (including zeros for those who do not charge off) are 30 MXN per card per month for default-approved applicants and 105 MXN per card per month for profitability-approved applicants. The net effect is that average interest and BNPL revenues net of charge-offs are 38 MXN per card per month for default-approved borrowers and -16 MXN per card per month for profitability-approved borrowers.

Thus, it appears that the profitability model attempts to identify borrowers that generate substantial interest revenue (by carrying revolving debt and making their interest payments) and who do not generate charge-offs by defaulting. While it succeeds at predicting the former, it fails at predicting the latter. Why is the model predicting profits unable to predict which of these high-interest-revenue applicants end up defaulting? Note that predicting who does not generate charge-offs by defaulting *among those who generate substantial interest revenues* is a different prediction problem than the one undertaken by the default model, which is to predict who does not default among the full sample. The default model accomplishes this by approving borrowers who carry less debt and thus generate lower interest payments but also default less.

The applicants approved by each model also differ in the interchange revenue net of rewards that they generate for the lender. However, interchange revenue net of rewards is small relative to interest revenue net of charge-offs: for example, for those approved by both models, interchange net of rewards is 6 MXN per card per month, compared to 52 MXN per month for interest revenues net of charge-offs. Applicants approved by both models spend 3,113 MXN per month on average and generate 42 MXN per month in interchange revenue, which is about 16% higher than their rewards expenses of 37 MXN per month. The default-approved borrowers spend even more (3,442 MXN per month), but the cost of rewards for these borrowers (50 MXN per month) fully offsets interchange fees (49 MXN per month). The profitability-approved borrowers, on the other hand, spend only 2,606 MXN per month, but who also have a higher 24% spread between interchange fee revenue (25 MXN per month) and rewards (20 MXN per month).

The differences in the spread between interchange revenues and rewards between those approved by each model suggests that the lower-risk borrowers identified by the default model are likely more sophisticated and better at optimizing their shopping behavior to maximize rewards. We use the account-level data on interchange fee revenue and rewards expenses to test this in Figure A.7, and find that lower-risk borrowers indeed generate higher rewards *per dollar of spending*. While these borrowers also shop at stores that generate higher interchange revenue for the lender (given that interchange fees vary by merchant type), their sophistication about rewards dominates. Interchange revenue net of rewards is negative for the lowest-risk borrowers and increases as risk increases.

To further understand why the profitability models generates lower profits than the default model, we graph the components of revenues and profits by quartile of predicted probabilities of default and of negative profits in Figure A.8. The default model succeeds at identifying users who end up defaulting and generating charge-offs, as can be seen by the steep gradient in charge-offs across quartiles of predicted probability of default; furthermore, charge-offs are a large driver of negative profits. The profitability model, in contrast, attempts to identify users generating substantial interest revenue, as can be seen by the high interest revenues in the first quartile of the predicted probability of negative profits. However, a number of these individuals end up defaulting: average charge-offs in the first quartile of the predicted probability of negative profits are much larger than in the first quartile of the predicted probability of default.

6 Robustness

All results in the previous section discuss predictions and performance in the context of an out-of-sample evaluation using a random split of the modeling sample, ignoring the time dimension. However, machine learning models are not static objects: their performance is constantly re-assessed, as models interact with new data and macroeconomic and technological conditions change over time (Ben-David, Johnson, and Stulz, 2025).

While the monitoring and maintenance of ML models have become integral parts of any deployment system to ensure their effectiveness and reliability, the choice of target variable and the relevant loan performance window (in our context, 12 months) has implications for the recency of the data entering the model, as well as for how quickly model performance can be assessed and models retrained. Concretely, our model trained on data up to the end of May 2023 using a 12-month default window could have been deployed, at the earliest, in June 2024. By then, the 2023 data might no longer be representative of the 2024 environment, because of changes in macroeconomic conditions, technologies, the pool of applicants, or the relationship between features and predictions.

Reassuringly, macroeconomic indicators in Mexico during the period January 2021 to June 2024 were relatively stable, with modest economic growth and mildly fluctuating levels of unemployment and inflation (Figure A.2). Moreover, as discussed in Section 3.1, the levels of these variables from 2021–2024 were not unusual compared to the levels in previous decades going back to 1995. We further assess whether changes over time in Mexico may affect our model using an out-of-sample/out-of-time test below.

6.1 Out-of-Sample/Out-of-Time Test

To assess whether a model trained in the beginning of our sample period still provides reliable predictions at the end of the sample period, we implement an out-of-sample/out-of-time test, in the spirit of Berg, Burg, Gombovi, and Puri (2020). We train a new model that splits the modeling sample of 146,036 borrowers into two periods based on date of origination: January 1st, 2021 to August 31st, 2022 and September 1st, 2022 to May 31st, 2023.[32] The first subperiod represents our training set; the second corresponds to the out-of-sample/out-of-time testing set. We chose the cut-offs to achieve close to an 80%/20% training-testing split, as in our benchmark model, and the two subperiods account for 79% and 21% of the modeling sample, respectively.

Table A.2, Panel B, shows performance metrics for the out-of-sample/out-of-time test. The model trained on the earlier part of our sample period and tested on the later part of our sample period performs nearly as well as the benchmark model with a random training-testing split. In particular, the AUC is 0.781 (compared to 0.796 for the benchmark model), precision is 0.393 (compared to 0.421), recall is 0.654 (compared to 0.666), and the F1 score is 0.491 (compared to 0.516). We conclude that the out-of-time model shows marginal decreases in performance (as expected), but that the XGBoost model is reasonably robust when performing an out-of-sample/out-of-time test. This suggests that our model is reasonably robust to potential changes over the course of our sample period in both macroeconomic conditions and the applicant pool.

6.2 Data Drift

Not only do we not observe a significant degradation of the model's performance over time (model decay), but we also do not detect any systematic data drift, meaning a shift in covariates where the distribution of the input data changes over time due to a change in characteristics of the applicant pool (Lu et al., 2018). For every quarter starting in January 2021, we test whether the data from a new, adjacent quarter's contracts present a statistically significant shift in the distribution of any of the covariates/features relative to all previous quarters. We conduct this test across all binary, categorical, and numerical covariates and all quarters, resulting in 1,863 comparisons; across all of these tests, only 73 comparisons (3.9%) are significant at the 5% level, which is consistent with what would be expected by chance.[33]

[32] As in our benchmark model, we can only include applicants through May 2023 because we measure the target variable over a 12-month window, and have performance data through May 2024.

[33] The period with the largest number of significant comparisons (15) is Q3-2021. This quarter is toward the beginning of the sample and is a period of rapid growth and likely experimentation. Details on data drift testing procedures for numeric (continuous and discrete), binary, and categorical variables are provided in Table A.6. Each group of variables is tested using three testing procedures. A comparison is deemed significant if at least two of the three tests yield a statistically significant drift for each group of variables.

6.3 Performance across Geographies with Different Economic Conditions

We also assess the robustness of the model performance across geographies with heterogeneous economic conditions. We use predictions from our benchmark model (see Section 4) on the original test sample—not from models estimated separately for each state—and compute state-level AUCs for borrowers, who are assigned to a state based on the address provided at the time of application.

Figure 5 shows scatter plots of state-level AUCs (vertical axis) against state-level GDP per capita (horizontal axis). We find no systematic relationship between economic activity and AUCs, suggesting that the performance of the model is stable across areas with different economic environments. Panel A uses GDP per capita for 2021 (the beginning of our analysis period), Panel B uses GDP per capita for 2023 (the latest year for which state-level GDP is available), and Panel C uses the change in GDP per capita between 2023 and 2021.[34] Some states with high GDP per capita have narrower confidence intervals—due to a larger presence of our lender and hence a larger sample size. We note that pooling observations from states with low GDP per capita would lead to more comparable sample sizes and confidence intervals while preserving the null relation between economic activity and AUCs.

6.4 Penalized Linear Models

A potential concern with highly flexible algorithms is linked to their ability to uncover unseen and intricate data patterns which do not generalize well in unseen data. In the context of XGBoost, this behavior could arise due to non-robust data (see Sections 6.1 and 6.2), outlier sensitivity, overfitting, and other issues. The best-known and most extreme version of this phenomenon has been studied in the context of neural networks which extract their own features, some of which may be non-robust even when derived from patterns in the data distribution (Ilyas et al., 2019; Biggio et al., 2013; Athalye, Engstrom, Ilyas, and Kwok, 2018).

To further examine the stability and robustness of our results, we implemented two regularized logistic regression models, using L1 regularization (LASSO; Tibshirani, 1996) and elastic net regularization (GLMNET; Friedman, Hastie, and Tibshirani, 2010), with penalty parameters tuned via cross-validation. We compared LASSO and GLMNET with our original XGBoost implementation using a consistent evaluation pipeline, including performance metrics with bootstrapped confidence intervals.[35] Empirically, we do not see significant differences between the two regu-

[34] In Figure A.9 we replicate the analysis using GDP per capita without revenue from oil. This is a common adjustment to state-level GDP in Mexico since revenue from oil is appropriated by the federal government and not by the states. Since default is measured up to May 2024 but the latest measurement available of state-level GDP is 2023, in Figure A.10 we replicate the analysis with a quarterly indicator of economic activity available for the first half of 2024. As before, there is no clear pattern between economic activity and model performance during that period.

[35] The performance of logistic regression classifiers coupled with L1/L2 norms is affected by the scale of features. In order to allow for an adequate comparison across features with different scales and improve the classifier perfor-

larization methods, with GLMNET regularization behaving almost like a pure LASSO, assigning 90% weight to the L1 regularization parameter and 10% to L2; for all four performance metrics, the performance of LASSO and GLMNET is identical up to the third decimal place.[36]

Table A.2, Panel C, compares the two penalized linear models with our benchmark XGBoost (XGB) model. While XGBoost outperforms both penalized linear models, the performance of all three models is broadly consistent. XGBoost has a moderately higher AUC (XGB: 0.796 vs. LASSO/GLMNET: 0.768). XGBoost also achieves a better F1 score (XGB: 0.516 vs. LASSO/GLMNET: 0.486) by better balancing precision and recall, with higher precision (XGB: 0.421 vs. LASSO/GLMNET: 0.382) and therefore fewer false positive predictions. The three models identify an almost identical proportion of true defaulters, as proxied by recall (XGB: 0.666 vs. LASSO/GLMNET: 0.667). These results offer additional confidence that the XGBoost model is uncovering relevant nonlinear data patterns that deliver better precision and sharper class separation, while matching recall of the logistic regression models—which are notoriously more cautious and tend to favor recall by smoothing predictions towards the observed default rate and generating more inclusive classification boundaries.

Overall, the results across the three algorithms also offer additional evidence on the stability of our benchmark model. The model's robustness can be explained by choices made when training XGBoost. Not only does our implementation include L1 and L2 regularization (with the tuned relative importance giving more prominence to the L1 relative to the L2 norm, in line with the above GLMNET results), but the tuned hyperparameters result in additional layers of regularization along several dimensions. That is, stochastic training (50% features, 75% sample), conservative learning (1.9% learning rate), and structural constraints prevent individual trees from becoming too complex (allowing for many trees with shallow depth, and penalizing too many splits). By design, our modeling approach had the goals of avoiding overly precise but non-generalizable predictions and of guarding against possible brittleness.

7 Discussion

7.1 Reject Inference for Borrowers without Credit History

A limitation of our data is that we only observe repayment and default outcomes for applicants approved for credit by our FinTech partner, based on their risk appetite and parsimonious approval

mance, features were centered and scaled to have zero mean and unit variance (separately in the training/testing sets) to ensure coefficients could be directly compared in magnitude and to improve model convergence. Standardization is not needed for XGBoost models, since tree splits depend on the order or rank of values, not on the magnitude or scale of inputs.

[36] L1 regularization tends to be more aggressive in feature selection but can be unstable when features are correlated. Elastic net, which combines L1 (LASSO) and L2 (Ridge), tends to offer stability in terms of feature selection.

rules intended to identify borrowers with low probability of default (see Section 3). As a result, there is a substantial portion of applicants whose default behavior is not observed and who may significantly differ from our modeling sample.[37]

This selection problem is known as *reject inference* in the credit scoring literature (FinRegLab, 2023b; Caro and Nelson, 2024), and it is analogous to the problem of *selective labels* in the machine learning literature (Lakkaraju et al., 2017). As noted in Blattner and Nelson (2024) and FinRegLab (2023a), a frequent solution is to simply drop rejected applicants and train and evaluate models in the sample of approved applicants. Nevertheless, there are a few methods to address this problem in practice. One alternative is to infer repayment behavior on the product of interest using *credit bureau proxies* (FinRegLab, 2023b). For example, when modeling mortgage repayment, for rejected applicants (for whom mortgage repayment is not observed), one could use repayment behavior on other types of loans, and/or in other periods of time, using data from credit bureaus (Blattner and Nelson, 2024; Caro and Nelson, 2024).[38] However, by definition, the use of credit bureau proxies is not feasible for individuals without credit history.

One solution that is available for individuals without credit history—albeit at a significantly higher cost for lenders than using credit bureau proxies—is to allocate credit to a random sample of those whom the model says to reject. By doing so, future models can be trained on a data set that represents the entire pool of potential applicants, including not only those who are accepted by the status-quo model but also a random sample of those rejected by it. This is consistent with the practice of some lenders of periodically lending to a small subset of individuals below their approval threshold to learn about the false positive rate of their models (Meursault, Moulton, Santucci, and Schor, 2024). When implementing our model in the future, our lender plans to lend to a small random subset of rejected applicants for the purpose of conducting this form of reject inference.

7.2 More Sophisticated Profits Model

Our profitability model uses a binary variable of whether the borrower generates positive profits for the lender as the target. We use a binary target to use the same type of model to predict profitability as we use to predict default, enabling us to directly compare the performance of the two models using classification metrics.

[37] While we did not receive the full set of data sources for rejected applications (see Section 3 for details), and we did not receive any information from the riskiest of the rejected applications—as per our lenders' assessment—in the data and sample of rejected borrowers we have access to, the median no-hit score of rejected applications is 625, compared to 638 for approved applicants.

[38] There are other methods that involve imputing the value of the (missing) dependent variable of rejected applicants based on their observable characteristics and the repayment behavior of approved applicants (FinRegLab, 2023b). However, those methods do not address selection on unobservables and, in some instances, have been found to be of limited use (Crook and Banasik, 2004).

One alternative would be to predict a continuous measure of profits and then implement an approval threshold on predicted profits. Typically within the machine learning literature, there is a preference to use a binary classifier when the downstream decision is binary, as in our case of a lender making an approval decision. This is because binary classifiers optimize classification metrics while models predicting a continuous outcome optimize mean squared error (MSE), without regard for the classification implications of that MSE. Nevertheless, it remains an empirical question whether such a model would perform better or worse than predicting a binary measure of profitability, and we leave this question to future research.

A alternative would be to implement a more sophisticated profits model. For example, the more sophisticated algorithm might separately try to predict each component of profits (charge-offs, interest revenue, interchange fee revenue, rewards expenses, etc.) and then determine optimal weights to assign to each of these components, with regularization of the weights. We also leave to future research the question of whether such a model could generate higher profits than our benchmark model predicting default.

We also note that while these more sophisticated models could potentially produce higher profits than a default model over the first 12 months since origination—which is what they optimize on—a default model might still produce higher profits over the lifetime of each card. This is because, empirically, those who default always generate negative lifetime profits for the lender, while those who generate substantial interest revenue and profits over the first 12 months since origination (because they carry a substantial revolving balance but do not default) may experience a shock later on and default, making them unprofitable over the lifetime of the card. Because FinTech lenders are relatively new entrants, however, they face a binding trade-off between the sample size used in their models and the length of time over which they can measure profits; in contrast, longstanding traditional lenders are able to implement models that seek to maximize a customer's lifetime value (Cowan, Mercuri, and Khraishi, 2023; Ekinci, Uray, and Ülengin, 2014).

A borrower who has a low probability of default but also low profits generated by the credit card—who will thus be approved by the default model but not the profitability model—may also have a higher lifetime value to the lender due to cross-selling of other new or existing products in the future (Basten and Juelsrud, 2023; Li, Sun, and Montgomery, 2011). On the other hand, a borrower who carries high revolving debt and generates high interest revenue but has a higher probability of default—who will be approved by the profitability model but not the default model—will likely be unprofitable even after accounting for cross-selling if they end up defaulting. This is an additional reason that, even if a more sophisticated profits model appeared to generate higher profits from the credit card, a profit-maximizing lender might nevertheless implement a default model.

7.3 Endogenous Interest Rates

In our models, the lender makes approval decisions, but we do not explore endogenizing interest rates, primarily because our FinTech partner generally charges the same interest rate to all approved borrowers with no formal credit history at a given point in time (Figure A.3). Endogenizing interest rates would require strong assumptions, as changes in interest rates could cause changes in demand and default, and thus on profitability. The changes in default caused by a change in interest rates can occur both through a change in selection and through a causal effect of interest rates on credit card use and default (Karlan and Zinman, 2009).

The lender would thus need estimates of the elasticities of demand and default with respect to interest rates, which would be difficult to obtain without running randomized experiments. It is an open question whether or not FinTech lenders are running these types of experiments to measure these elasticities and incorporate them into their credit scoring models (Berg, Fuster, and Puri, 2022). Evidence to date suggests that the causal effect of changes in interest rates on credit card default is relatively small (Castellanos et al., 2024). Furthermore, for borrowers without credit history, who have access to fewer borrowing options, we would expect demand to be relatively inelastic to interest rates. Whether or not endogenizing interest rates in credit scoring models has an impact on model performance and lender profits remains an open question. In contrast to interest rates, the causal effect of changes to credit limits on credit card use appears substantial (Aydin, 2022). This suggests that incorporating the impact of credit limits into credit scoring models and endogenizing credit limits as part of the lender's decision could be a promising opportunity.

8 Conclusion

Traditional financial institutions such as banks typically do not lend to borrowers without formal financial history, and banks' past attempts to expand credit access to first-time formal borrowers with no credit history have often failed (Castellanos et al., 2024). Meanwhile, online FinTech lenders have rapidly proliferated around the world (Berg, Fuster, and Puri, 2022), and proponents argue that FinTech lending promises to expand access to credit and increase financial inclusion by using alternative data sources to evaluate creditworthiness. In other words, if alternative data sources such as call logs, social media interactions, and retail transactions can accurately predict credit *on their own* for people with no credit history, these potential borrowers would no longer necessarily be excluded from credit markets.

Many FinTech companies indeed use these alternative data sources in models to predict creditworthiness, and several academic studies have evaluated the predictive accuracy of these alternative data sources in assessing credit risk. However, most FinTech lending algorithms still rely at least partly on conventional credit scores (Johnson, Ben-David, Lee, and Yao, 2023), and in these stud-

ies all or at minimum a majority of applicants do have formal credit histories and conventional credit scores reported by the credit bureau. When FinTech companies rely on the credit bureau score as one input to their credit scoring algorithm, and in practice only approve applicants who do have traditional credit scores, they do not fulfill FinTech's promise of expanding access to credit on the extensive margin.

We train machine learning models to assess credit risk for a population in which no one has a conventional credit score in the credit bureau, either because they have no credit history or an insufficient credit history for the credit bureau to generate a credit score. We show that a model trained on alternative data sources for this population with no credit history is effective at predicting default. In particular, the predictive accuracy of our model is at the upper end of studies in middle-income countries (and is also higher than that of some studies in more data-rich environments such as the US), despite the models in other studies being estimated for populations that are already more financially included in the sense that they already have conventional credit scores at the time of loan application, and despite those models using credit bureau scores as an input to the model.

When we compare a model predicting default—as used in practice by many traditional and FinTech lenders—to a model predicting profitability, we find that the model predicting default generates more profits than the model predicting profitability. Comparing the types of borrowers approved by each model, we find that the superior performance of the model predicting default is largely due to the complexity of predicting profits. In particular, the most profitable users are those who revolve a large amount of debt and make their interest payments but do not default. (While high-spending transactors who do not revolve debt may in theory also be profitable due to interchange fee revenue, we find that, for our lender, these card holders do generate substantial interchange fee revenue but are also sophisticated about maximizing rewards, and thus do not generate much revenue in interchange fees net of rewards.)

The model predicting profitability succeeds at identifying borrowers who generate substantial interest revenues, but fails to effectively screen out those who initially generate interest revenues but ultimately default. This may reflect that within this set of high-risk borrowers who revolve substantial amounts of debt, default is largely driven by relatively exogenous shocks such as job loss (Castellanos et al., 2024), which is hard to predict. In contrast, the default model screens out the full set of high-risk borrowers, including both those who are ex-post profitable because they do not default and those who are ex-post unprofitable because they default.

Our findings have two potential implications.

First, our results suggest that alternative data can be used to expand credit to underserved populations. In countries where alternative data is regularly used, this can have a direct impact on the strategy of FinTech companies considering serving this segment. In countries where regulatory agencies are evaluating the costs and benefits of the use of alternative data for credit origination,

this can inform the discussion by showing that predicting repayment behavior of populations without credit history is feasible (Bureau, 2017). The distributional consequences of more granular predictions, privacy and fairness considerations, and the net effect of all of these forces on welfare require further research (Berg, Fuster, and Puri, 2022; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2022; Armantier et al., 2024).

Second, a shift by lenders from models predicting default to models predicting profitability or profits could have implications for aggregate credit access, financial stability, and redistribution. A model predicting profitability lends to a smaller number of borrowers (when the lender uses a profit-maximizing threshold to make approval decisions for both the default and profitability models), reducing aggregate credit access. Because a model predicting profitability also lends to riskier borrowers who generate high interest revenue for the lender, shifting to a profitability model could reduce financial stability in the presence of correlated shocks across borrowers. However, because the riskier borrowers approved by the profitability model are lower-income and more likely to be women, there are also redistributive effects of shifting to a profitability model. Nevertheless, the welfare effects of increasing lending to these more vulnerable borrowers who would be approved only by a profitability model is ambiguous, given that many end up defaulting ex post, which will negatively impact their new credit score and future access to credit.

Table 1: Comparison of studies that predict creditworthiness

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)
This paper	Mexico	FinTech credit card	0%	Delivery app transactions data, digital footprints, credit history for those with limited credit history (but no credit scores)	XGBoost	0.796
Agarwal, Alok, Ghosh, and Gupta (2023)	India	FinTech loan	81%	Digital data from mobile phones; call logs; demographics, address, bank statements, salary slips; traditional credit score (CIBIL)	Random forest, XGBoost, logit	0.738 for sample with credit history, 0.674 for sample without credit history
Albanessi and Vamossy (2024)	US	Credit card	100%	Credit bureau files and credit scores	Hybrid deep neural network/gradient boosting	0.906
Berg, Burg, Gombovi, and Puri (2020)	Germany	FinTech loan	94%	Digital footprints (device type, operating system, email service provider, writing style, etc.), credit scores	Logit	0.734
Björkegren and Grissen (2020)	A middle-income South American country	Mobile phone airtime credit	85%	Mobile phone call logs and text data, history of phone bill payment, credit bureau data	Random forest, logit	0.772

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Blattner and Nelson (2024)	US	Mortgage	100%	TransUnion consumer credit report data and public and Infutor data on consumers' mortgage transactions, socio-economic characteristics, and lenders' information, Vantage credit scores	XGBoost, random forest, logit	0.840 for minority and 0.887 for non-minority sample
Blattner, Nelson, and Spiess (2024)	US	Credit card	100%	Credit bureau files and credit scores	XGBoost, random forest, logit, elastic net, neural net	0.867
Butaru et al. (2016)	US	Credit card	100%	Account-level credit card data from 6 major commercial banks, macroeconomic variables, credit bureau data including credit score	Random forest, logit	Not reported
Caire and Vidal (2024)	India	FinTech loan to microenter-prises	95%	Data from partner gig worker platforms on earnings, working hours, driver ratings, and other measures	Logit	0.710
De Cnudde et al. (2019)	Philippines	Microfinance loan	Not reported	Facebook data (sociodemographics, likes, comments, social network)	Linear support vector machine	0.825
Di Maggio and Ratnadiwakara (2024)	US	FinTech loan	100%	Age, annual income, debt-to-income ratio, FICO credit score	Random forest	0.659
Duarte, Fonseca, Kohli, and Reif (2025)	US	Mortgage, student loans, credit card	100%	Payment history data, debt and collections, credit utilization and credit limits, credit history length, recent credit activity	XGBoost	0.712
Frost et al. (2019)	Argentina	FinTech SME loan	100%	Sales data and internal rating from e-commerce platform, credit score	Logit, XGBoost	0.764

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022)	US	Mortgage	100%	Income, loan-to-value (LTV) ratio, origination amount, FICO credit score, etc.	Random forest, XGBoost, logit	0.861
Gambacorta, Huang, Qiu, and Wang (2024)	China	FinTech loan	100%	Call data including frequencies, duration, etc., app use data, credit history, default history, frequency of credit card usage, credit scores produced by FinTech based on formal credit history	Logit	0.607
Hair, Howell, Johnson, and Matsumoto (2025)	US	Small business loans	100%	FICO score, business age, number of employees, requested loan amount, state, industry, region, loan type	RF, logit, OLS	0.663
Huang et al. (2023)	China	FinTech SME loan	100%	Asset data such as housing property, gender, age, and business type, data on provincial and municipal economy, MYbank credit histories and credit scores	Random forest	0.841
Iyer, Khwaja, Luttmer, and Shue (2016)	US	FinTech P2P loan	100%	Borrower income, number of past delinquencies, maximum interest rate borrower is willing to pay, picture and text description in loan application, Experian credit score	OLS	0.714
Jagtiani and Lemieux (2019)	US	FinTech P2P loan	100%	Personal installment loan-level data from LendingClubs unsecured consumer platform, similar loan-level data from traditional lenders, FICO credit scores	Logit	0.689
Johnson, Ben-David, Lee, and Yao (2023)	US	FinTech loan	100%	Income, requested loan amount, loan purpose, credit bureau data, FICO credit score	Logit	0.665

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Khandani, Kim, and Lo (2010)	US	Credit card	100%	Customer transactions data and account balance data from a major commercial bank, credit bureau data, credit scores	Generalized classification and regression trees	0.952
Lee, Yang, and Anderson (2024)	Multiple countries in Asia	Credit card	50%	Supermarkets loyalty card data and credit card spending and payment history, sociodemographic data, credit scores	XGBoost	0.679 for sample with credit history, 0.647 for sample without credit history
Lee, Yang, and Anderson (2025)	Peru	Credit card	82%	Self reported socioeconomic characteristics, Equifax RP2 scores, credit history and retail transaction data	XGBoost	0.682 for sample without credit history, 0.677 for sample with credit history
Meursault, Moulton, Santucci, and Schor (2024)	US	Bank loan	100%	Credit bureau records, credit score	XGBoost, logit	0.883
Netzer, Lemaire, and Herzenstein (2019)	US	FinTech P2P loan	100%	Textual data from loan requests on Prosper, a FinTech P2P lending platform, plus financial and demographic information	Random forest, logit	0.726
Rishabh (2024)	India	Bank loans and FinTech loan	95% for banks, 90% for FinTech	Payment history data, demographic data, TransUnion credit scores	Random forest, logit	0.7 for banks, 0.68 for FinTech
Sadhwani, Giesecke, and Sirignano (2021)	US	Mortgage	100%	Loan data and monthly performance records, local and national economic data from Zillow and the Federal Housing Administration (FHA), FICO credit scores	Deep learning neural network, logit	0.700

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)
San Pedro, Proserpio, and Oliver (2015)	A Latin American country	Credit card	100%	Mobile phone usage logs from a telecommunications company, digital footprints, sociodemographics, credit bureau data	Regularized logit, support vector machines, gradient boosted trees	0.725

This table reports the country, loan type, percent with credit score, data sources, machine learning methods, and predictive performance (proxied by AUC) of other studies using machine learning models to predict creditworthiness. Agarwal, Alok, Ghosh, and Gupta (2023) use both random forest (RF) and XGBoost; we report the AUC from their best-performing model, which is RF. Agarwal, Alok, Ghosh, and Gupta (2023) do not report an overall AUC for the full sample including those with and without credit scores. For Berg, Burg, Gombovi, and Puri (2020), we report the out-of-sample AUC using credit bureau scores, digital footprints, and fixed effects. Björkegren and Grissen (2020) use both RF and logistic regression; we report the AUC from their best-performing model which is logistic regression. For Blattner and Nelson (2024) we report the AUC of the XGBoost baseline model. For Blattner, Nelson, and Spiess (2024), we report the AUC of the XGBoost model. For Caire and Vidal (2024), we use the AUC for KarmaLife borrowers; we do not use the AUC for Fundfina borrowers since prior loan repayment with the same lender was used as an input to the Fundfina model, and this input would not be available to lenders seeking to lend to new applicants rather than repeat borrowers. For De Cnudde et al. (2019), we report the AUC from the best-performing model, which is an ensemble using a network-only link-based classifier to process the Facebook network data. Di Maggio and Ratnadiwakara (2024) report the AUC of the FinTech platform's model for the full sample as well as those with subprime and prime credit scores; we use their AUC for the full sample. For Frost et al. (2019), we report the AUC for the XGBoost model. For Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022), we report the AUC for RF with race as a variable. For Gambacorta, Huang, Qiu, and Wang (2024), the "% with credit score" is based on the percent with a credit score produced by the FinTech based on formal borrowing histories, as the paper does not have access to credit bureau data (though the sample is likely to have a credit score in the credit bureau). The AUC we report for Gambacorta, Huang, Qiu, and Wang (2024) is the one for the baseline model using all information except the interest rate, as the interest rate would not be available at the time of loan application. For Hair, Howell, Johnson, and Matsumoto (2025), we report the AUC for their preferred default model using FICO and cash flow data. For Huang et al. (2023), the "% with credit score" is based on the presumed percent with MYBank credit scores based on formal credit histories, as the paper does not have access to credit bureau data (though the sample is likely to have a credit score in the credit bureau). For Iyer, Khwaja, Luttmer, and Shue (2016), we report the AUC combining all data. For Jagtiani and Lemieux (2019), we report the AUC for the best-performing model, which uses rating grades and other control factors. Khandani, Kim, and Lo (2010) report a range of AUCs without additional detail (and do not report if they are estimated in-sample or out-of-sample); we report the upper end of the range they report. For Lee, Yang, and Anderson (2024), we report the AUC for the best-performing model, which uses all data sources predicting ever-delinquent. For Lee, Yang, and Anderson (2025), we report the AUC for the model using baseline and retail data. For Meursault, Moulton, Santucci, and Schor (2024), we report the AUC for the overall XGBoost model, averaged over all years. For Netzer, Lemaire, and Herzenstein (2019), we report the AUC of the model with text, financial, and demographic data. For Rishabh (2024), we report the AUC of the model using "traditional hard information" and granular payments data. Sadhwani, Giesecke, and Sirignano (2021) report AUCs for going from each potential state this month to each potential state next month, where the potential states are current, 30 days delinquent, 60 days delinquent, 90 days delinquent, and foreclosure; we use the AUC for predicting transitioning from 60 days delinquent to 90 days delinquent in their best-performing model. San Pedro, Proserpio, and Oliver (2015) do not report the "% with credit score", but the authors report an AUC using credit bureau

data, so we assume it is 100%. For San Pedro, Proserpio, and Oliver (2015), we report the AUC for default at 90 days using all data sources. AUC = area under the receiver operating characteristic curve; FICO = Fair Isaac Corporation; P2P = peer-to-peer.

Table 2: Summary Statistics for Modeling Sample

	Mean (1)	Std dev (2)	25th perc. (3)	Median (4)	75th perc. (5)
Panel A: Subset of Features					
Woman - dummy	0.39	0.49	0.00	0.00	1.00
User age	25.11	8.01	20.00	23.00	27.00
User iOS (Apple) operating system - dummy	0.38	0.49	0.00	0.00	1.00
No-hit score	640.65	14.00	635.00	642.00	649.00
Number of orders on app	24.05	52.36	3.00	9.00	22.00
Proportion orders paid in cash	0.51	0.37	0.17	0.50	0.90
Median amount per order (MXN)	349.84	344.87	173.00	250.00	403.00
Proportion orders at supermarkets	0.05	0.14	0.00	0.00	0.04
Proportion orders at pharmacies	0.04	0.12	0.00	0.00	0.01
Proportion orders at food establishments	0.82	0.26	0.73	0.94	1.00
Marginality (SES) index of census tract	0.96	0.01	0.96	0.97	0.97
Years of schooling among age 15+ in census tract	12.40	1.66	11.33	12.43	13.62
Proportion households own a motor vehicle in census tract	0.64	0.17	0.53	0.64	0.76
Panel B: Credit Card Terms and Use					
Interest rate	0.80	0.09	0.72	0.87	0.87
Credit limit (MXN)	5,647.40	1,921.28	5,000.00	5,000.00	6,000.00
Spending (MXN)	3,103.32	2,151.42	1,677.10	2,563.92	3,937.25
Statement balance (MXN)	3,773.97	2,270.77	2,211.71	3,610.73	4,973.13
Minimum payment (MXN)	459.58	649.00	69.64	148.85	479.85
Repayment (MXN)	2,962.79	2,593.41	1,229.58	2,321.88	3,901.89
Delinquency - dummy	0.20	0.40	0.00	0.00	0.00
Panel C: Profit Components (MXN per Month)					
Interest revenue	65.02	85.65	0.00	25.83	106.40
BNPL revenue	7.31	15.81	0.00	0.00	7.42
Charge-offs	61.43	156.99	0.00	0.00	0.00
Interest and BNPL revenue net of charge-offs	10.97	189.62	0.00	30.08	107.69
Interchange fee revenue	38.71	39.04	11.41	28.56	53.51
Cost of rewards	35.98	38.03	10.32	26.29	49.51
Interchange fee revenue net of rewards	2.90	17.53	-2.28	2.21	8.87
Late payment fee revenue	13.07	30.68	0.00	0.00	5.21
Funding costs	30.65	16.30	19.25	30.40	41.08
Costs from fraudulent transactions	0.53	9.72	0.00	0.00	0.00
Other fee revenue	2.57	7.41	-0.01	0.00	0.00
Other costs	12.39	13.36	3.46	8.58	16.38
Negative profits - dummy	0.49	0.50	0.00	0.00	1.00

This table shows summary statistics for the sample that we use in our machine learning modeling, which consists of applicants with no credit history and no prior credit card who were approved by RappiCard, had at least twelve months with the card by the end of our data period in May 2024, and made at least one transaction during their first 12 months with the card. Observations are at the user level, and the full modeling sample of $N = 146{,}036$ is included. Panel A shows summary statistics for a small subset of the features used by our machine learning model. For non-binary variables, upper-tail winsorization was performed at the 99.9th percentile consistent with the way features were winsorized for modeling. Census tract for each user is inferred based on login activity on the delivery app. The marginality (SES) index is a summary measure of economic vulnerability at the census track level, which takes values between 0 (high marginality) and 1 (low marginality). Panel B shows summary statistics on the terms and use of the credit cards. Interest rate and credit limit are measured at origination. Statement balance, minimum payment, and repayment are averages over the first 12 monthly statements (measured in MXN per month). Delinquency is measured as at least 60 days delinquent at any point over the first 12 months since origination. Panel C shows summary statistics for the average monthly revenues obtained and costs incurred by the lender for each card, where borrower-level monthly averages are computed over the first 12 months since card origination. For panels B and C, non-binary variables are winsorized at the 1st and 99th percentiles, with the exception of "costs from fraudulent transactions," as positive values occur for this variable for less than 1% of users; borrower-level profits are computed using non-winsorized revenue and cost variables. Std. dev. = standard deviation; perc. = percentile; iOS = Apple device operating system; MXN = Mexican pesos; SES = socioeconomic status; BNPL = buy now pay later.

Table 3: Predictive performance, confusion matrices and matrix of model disagreement for default and profitability models

	(1)	(2)	(3)	(4)
Panel A: Performance Metrics				
	AUC	Precision	Recall	F1
Default model	0.796	0.421	0.666	0.516
	[0.790, 0.802]	[0.411, 0.431]	[0.654, 0.678]	[0.506, 0.526]
Profitability model	0.598	0.583	0.519	0.549
	[0.592, 0.605]	[0.575, 0.592]	[0.511, 0.527]	[0.542, 0.556]
Panel B: Confusion Matrices				
	Default model		Profitability model	
	Did not default	Defaulted	Profitable	Unprofitable
Approved	30.8%	4.0%	20.5%	14.3%
Rejected	49.0%	16.2%	30.7%	34.5%
Panel C: Matrix of Model Disagreement				
	Default model			
Profitability model	Approved	Rejected		
Rejected	33.2%	21.1%		
Approved	34.8%	10.8%		

This table shows out-of-sample AUC, precision, recall, F1 scores, confusion matrices and matrix of model disagreement for the default model and the profitability model. The results use $N = 146{,}036$ users, split into training data to train the machine learning models and testing data to calculate out-of-sample measures of model performance. Bootstrapped 95% confidence intervals are included in square brackets, using 10,000 bootstrap samples. Threshold-dependent measures and confusion matrices use the profit-maximizing thresholds of a 24% probability of default and a 47% probability of negative profits. AUC = area under the receiver operating characteristic curve. F1 score = the harmonic mean of precision and recall.

Table 4: Marginal contribution of each data source to AUC

Feature set	AUC	AUC reduction	Precision	Precision reduction	Recall	Recall reduction	F1	F1 reduction
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All	0.796	0	0.421	0	0.666	0	0.516	0
	[0.790, 0.802]		[0.411, 0.431]		[0.654, 0.678]		[0.506, 0.526]	
All, but digital footprint	0.672	0.124	0.320	0.101	0.489	0.177	0.387	0.129
	[0.665, 0.680]		[0.310, 0.329]		[0.477, 0.502]		[0.377, 0.396]	
All, but transactions	0.770	0.026	0.397	0.024	0.651	0.015	0.493	0.023
	[0.764, 0.777]		[0.387, 0.406]		[0.638, 0.663]		[0.483, 0.502]	
All, but no-hit score	0.790	0.006	0.413	0.008	0.668	-0.002	0.511	0.005
	[0.783, 0.796]		[0.404, 0.423]		[0.655, 0.680]		[0.501, 0.520]	
All, but socioeconomic	0.795	0.001	0.422	-0.001	0.670	-0.004	0.517	-0.001
	[0.789, 0.801]		[0.412, 0.432]		[0.657, 0.682]		[0.508, 0.527]	

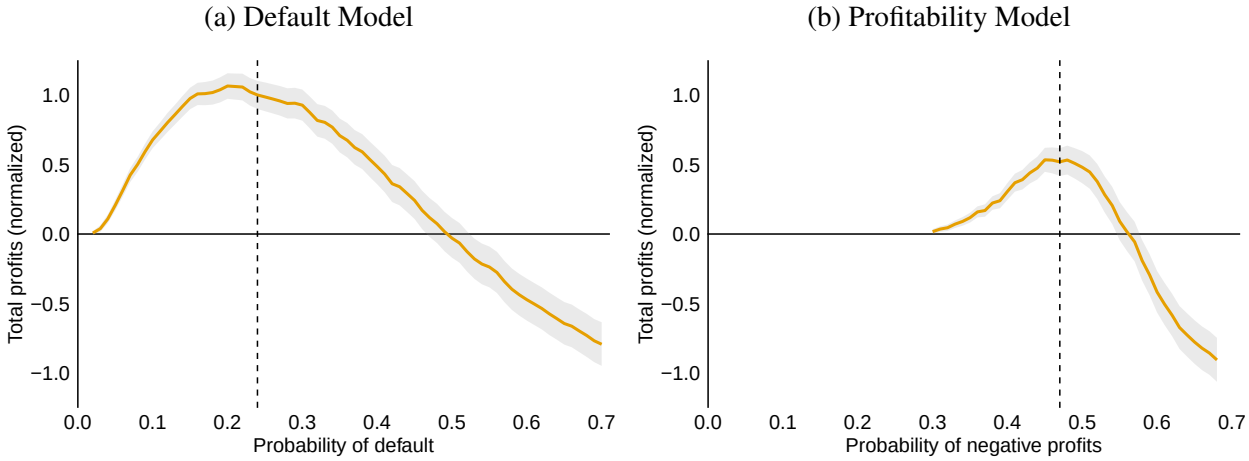
This table shows the differences in model performance (measured by out-of-sample AUC, precision, recall, and F1 score) between a model trained with all features and a separate model trained with features from all but one data source. The "reduction" columns (even-numbered columns) show the difference in model performance between each model and the model using all data sources. The results use $N = 146{,}036$ users, split into training data to train the machine learning models and testing data to calculate out-of-sample performance metrics. Bootstrapped 95% confidence intervals are included in square brackets, using 10,000 bootstrap samples. Threshold-dependent measures use the profit-maximizing threshold of a 24% probability of default. AUC = area under the receiver operating characteristic curve. F1 score = the harmonic mean of precision and recall.

Table 5: AUC by quintile of number of transactions through delivery platform

Quintile	Number of transactions (1)	AUC (2)	Precision (3)	Recall (4)	F1 (5)
1	2 or fewer	0.741	0.393	0.664	0.494
		[0.727, 0.756]	[0.373, 0.413]	[0.639, 0.689]	[0.474, 0.513]
2	3-6	0.775	0.403	0.680	0.506
		[0.761, 0.788]	[0.383, 0.423]	[0.656, 0.706]	[0.487, 0.526]
3	7-13	0.772	0.384	0.653	0.483
		[0.757, 0.787]	[0.362, 0.406]	[0.626, 0.681]	[0.462, 0.505]
4	14-28	0.804	0.428	0.633	0.511
		[0.789, 0.818]	[0.403, 0.454]	[0.602, 0.662]	[0.486, 0.535]
5	29 or more	0.838	0.477	0.644	0.548
		[0.825, 0.850]	[0.450, 0.504]	[0.614, 0.673]	[0.524, 0.571]

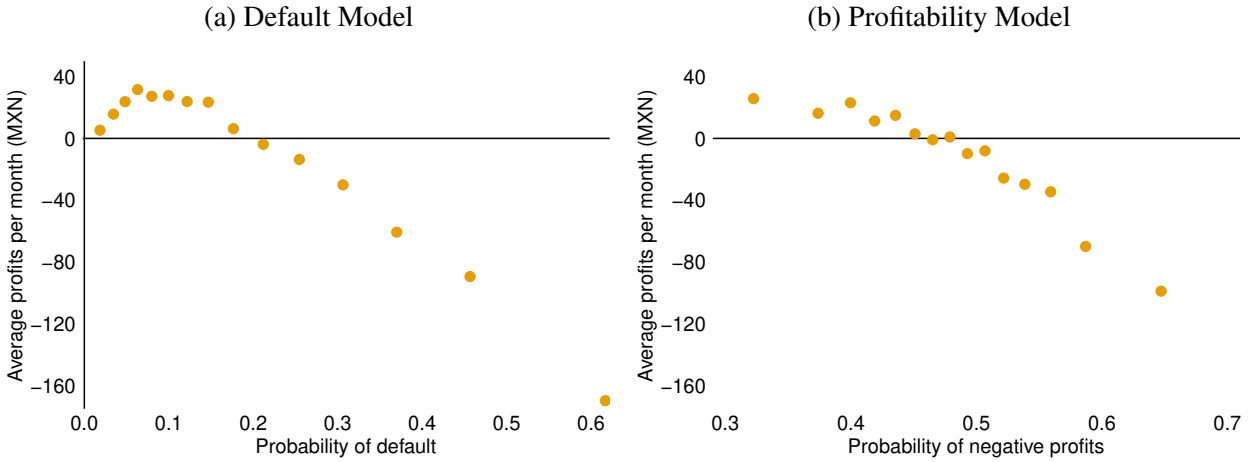
This table shows out-of-sample AUCs, precision, recall, and F1 scores for separate models estimated for each quintile of the distribution of number of transactions made through the delivery platform. Data are split into quintiles of the full modeling sample; machine learning models are then trained on the training data for each quintile and performance metrics are calculated on the testing data for each quintile. The results use $N = 146{,}036$ users, split into five quintiles based on number of transactions through the delivery platform, then into training data to train the machine learning model for that quintile and testing data to calculate out-of-sample performance metrics for that quintile. Bootstrapped 95% confidence intervals are included in square brackets, using 10,000 bootstrap samples. Threshold-dependent measures use the profit-maximizing threshold of a 24% probability of default (we do not use separate profit-maximizing thresholds by quintile). AUC = area under the receiver operating characteristic curve. F1 score = the harmonic mean of precision and recall.

Figure 1: Profits Generated by Default and Profitability Models across Approval Thresholds



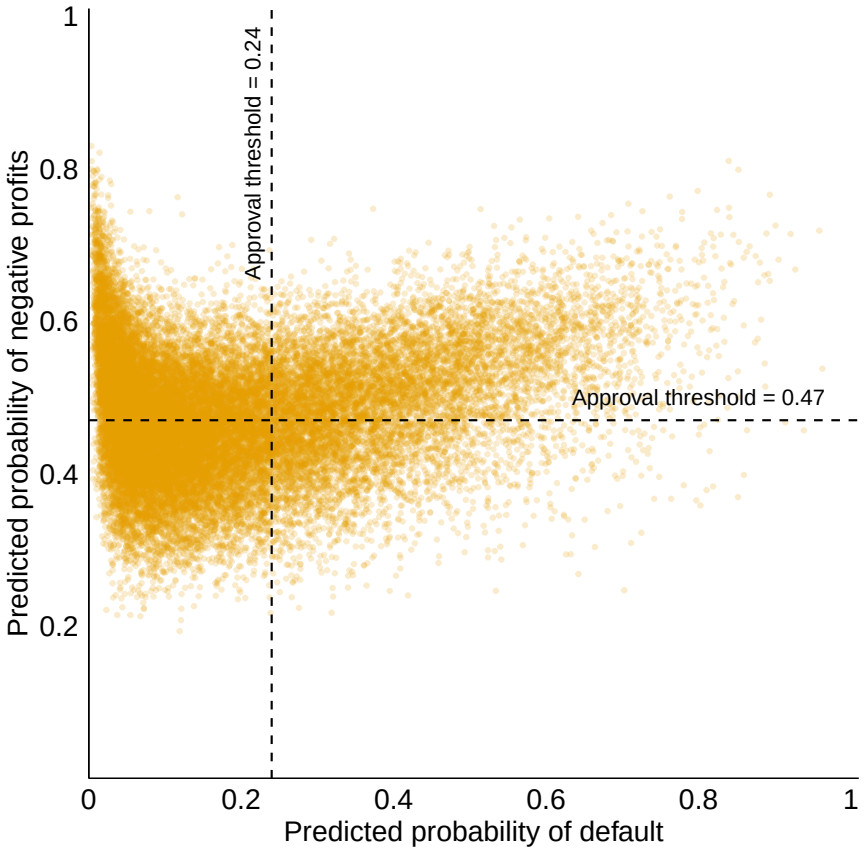
This figure shows the total profits obtained by the default and profitability models when making approval decisions in the testing data for different approval thresholds. It is calculated by summing the profits obtained by the lender across all individuals in the testing data who would be approved based on a given approval threshold (i.e., whose predicted probability of default or or negative profits is below the threshold). The results use $N = 146{,}036$ users, randomly split into training and testing data. Profit-maximizing thresholds, estimated in the training data, are shown as dashed vertical lines (these do not necessarily correspond to the peak of the curves shown in the figures since the curves are based on lending decisions made in the testing data, while the profit-maximizing thresholds are determined in the training data to avoid over-fitting). Total profits are normalized such that the most profitable model across all approval thresholds and across both the default and profitability models is normalized to 1. Using the profit-maximizing thresholds, the profitability model generates 51.9% as much profits as the default model. Bootstrapped 95% confidence intervals are shaded in gray, using 10,000 bootstrap samples. Approval thresholds below the 1st percentile and above the 99th percentile of predicted probabilities are omitted from the graph for legibility.

Figure 2: Average Realized Profits by Predicted Probabilities



This figure plots a binscatter of average realized profits in the testing sample across 15 bins in the predicted probability of default or of negative profits, with approximately equal number of observations in each bin. The results use $N = 146{,}036$ users, randomly split into training and testing data. Profits are constructed by summing non-winsorized revenues and costs variables at the account level, then winsorizing the account-level profits variable at the 1st and 99th percentiles.

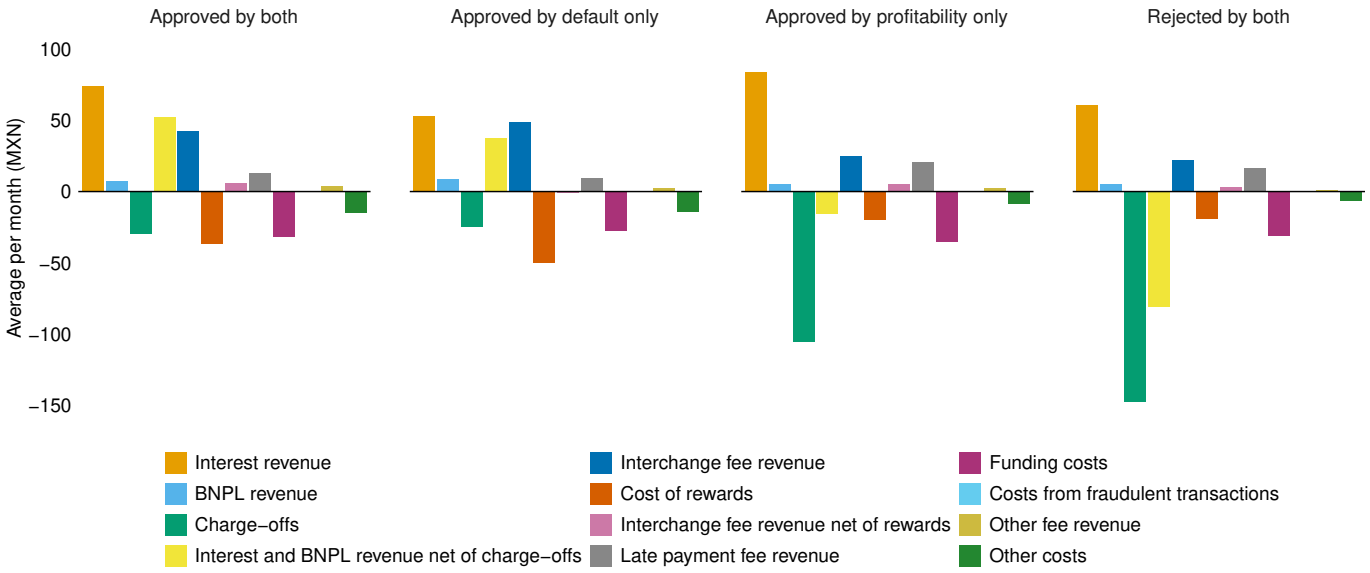
Figure 3: Predicted probabilities in default and profitability models



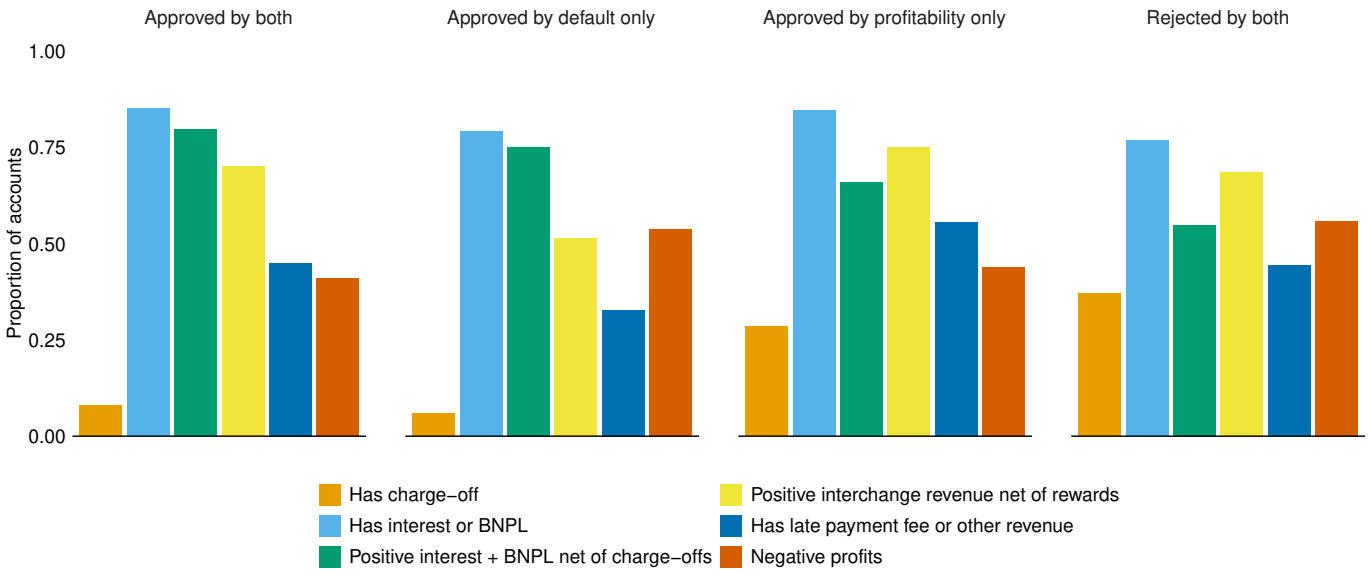
This figure shows the predicted probabilities of default and negative profits for each observation in our out-of-sample testing data. The results use $N = 146{,}036$ users, split into training data to train the machine learning models and testing data to calculate out-of-sample predicted probabilities which are shown in the figure. Fixing the approval threshold to the profit maximizing thresholds (24% predicted probability of default and 47% predicted probability of negative profits), the lower-left quadrant shows applicants who would be approved by both models, the upper-right quadrant shows applicants who would be rejected by both models, the upper-left quadrant shows applicants who would be rejected by the profitability model but approved by the default model, and the lower-right quadrant shows applicants who would be rejected by the default model but approved by the profitability model.

Figure 4: Revenues, Costs, and Profits by Model Disagreement

(a) Continuous Measures of Revenues, Costs, and Profits



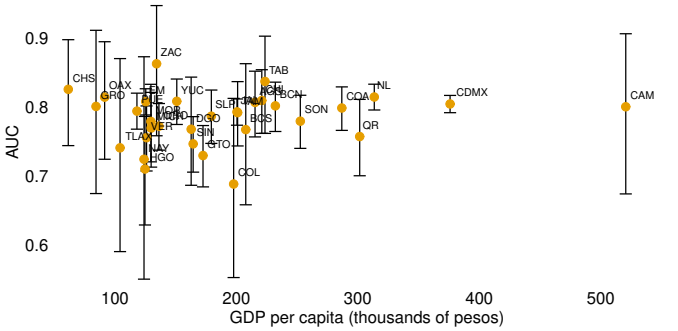
(b) Binary Measures



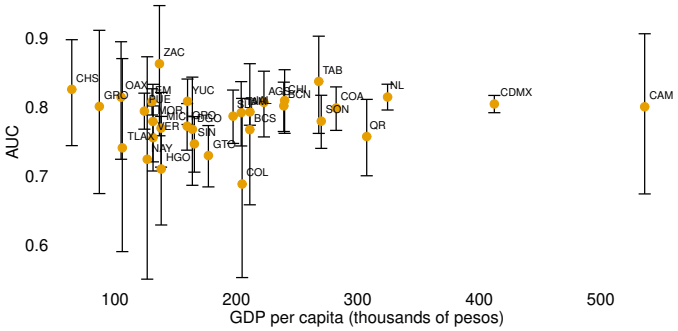
This figure shows the components of revenues, costs, and profits by model disagreement between the default and profitability models for the testing sample: accepted by both models, accepted by default model only, accepted by profitability model only, and rejected by both models. The results use $N = 146{,}036$ users, randomly split into training and testing data. In panel (a), variables are winsorized at the 1st and 99th percentiles, with the exception of "costs from fraudulent transactions," as positive values occur for this variable for less than 1% of users. In panel (b), borrower-level profits used to define the proportion with negative profits are computed using non-winsorized revenue and cost variables.

Figure 5: State-level AUCs and economic activity

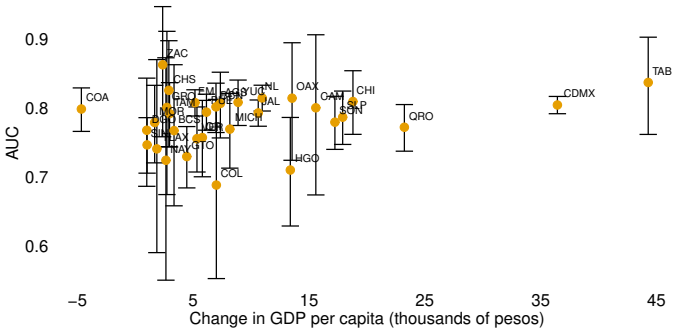
(a) 2021



(b) 2023



(c) 2023 - 2021



This figure shows scatterplots of state-level AUCs from our benchmark default model (vertical axis) against levels (or changes) in GDP per capita for each state (horizontal axis, in thousands of pesos from 2018). Borrowers are assigned to states based on the address provided at the time of credit card application. The results use $N = 146{,}036$ users, randomly split into training data and testing data. AUCs are computed on the testing data considering only observations from borrowers in the corresponding state. Predictions come from the default model of Section 4 (not from separate models by state). GDP per capita is calculated by dividing the state-level GDP, published by Mexico's National Statistical Institute (INEGI), by the population in each state and year, as reported by Mexico's National Population Agency (CONAPO). Panel A uses GDP per capita for 2021. Panel B uses GDP per capital for 2023. Panel C uses the change in GDP per capita between 2023 and 2021. Vertical lines represent 95% confidence intervals for AUCs, obtained with 10,000 bootstrap repetitions. Labels correspond to state names.

References

Agarwal, Sumit, Shashwat Alok, Pulak Ghosh, and Sudip Gupta (2023). "Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech."

Agarwal, Sumit, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel (2015). "Regulating Consumer Financial Products: Evidence from Credit Cards." *The Quarterly Journal of Economics* 130(1), 111–164.

Agarwal, Sumit, Andrea Presbitero, André F. Silva, and Carlo Wix (2023). "Who Pays for Your Rewards? Redistribution in the Credit Card Market." SSRN Scholarly Paper. Rochester, NY.

Albanessi, Stefania and Domonkos Vamossy (2024). "Credit Scores: Performance and Equity."

Armantier, Olivier, Sebastian Doerr, Jon Frost, Andreas Fuster, and Kelly Shue (2024). "Nothing to Hide? Gender and Age Differences in Willingness to Share Data." *SSRN Electronic Journal*.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok (2018). "Synthesizing Robust Adversarial Examples." *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 284–293.

Aydin, Deniz (2022). "Consumption Response to Credit Expansions: Evidence from Experimental Assignment of 45,307 Credit Lines." *American Economic Review* 112(1), 1–40.

Banco de México (2021). "Adopción de Criterios Contables IFRS9."

Banco de México (2024). "Indicadores Basicos de Tarjetas de Credito."

Basten, Christoph and Ragnar Juelsrud (2023). "Cross-Selling in Bank-Household Relationships: Mechanisms and Implications for Pricing." *Review of Financial Studies*, hhad062.

Ben-David, Itzhak, Mark J. Johnson, and René M. Stulz (2025). "Models Behaving Badly: The Limits of Data-Driven Lending." *Review of Finance* 29(3), 711–745.

Berg, Tobias, Valentin Burg, Ana Gombovi, and Manju Puri (2020). "On the Rise of FinTechs: Credit Scoring Using Digital Footprints." *The Review of Financial Studies* 33(7), 2845–2897.

Berg, Tobias, Andreas Fuster, and Manju Puri (2022). "FinTech Lending." *Annual Review of Financial Economics* 14(1), 187–207.

Bergstra, James, Dan Yamins, and David Cox (2013). "Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms." *Python in Science Conference*. Austin, Texas, 13–19.

Biggio, Battista, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli (2013). "Evasion Attacks against Machine Learning at Test Time." Vol. 7908, 387–402.

Björkegren, Daniel and Darrell Grissen (2020). "Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment." *The World Bank Economic Review* 34(3), 618–634.

Blattner, Laura and Scott Nelson (2024). "How Costly Is Noise? Data and Disparities in Consumer Credit."

Blattner, Laura, Scott Nelson, and Jann Spiess (2024). "Unpacking the Black Box: Regulating Algorithmic Decisions." *Proceedings of the 23rd ACM Conference on Economics and Computation*. Boulder CO USA: ACM, 559–559.

Böken, Björn (2021). "On the Appropriateness of Platt Scaling in Classifier Calibration." *Information Systems* 95, 101641.

Breiman, Leo (1996). "Bagging Predictors." *Machine Learning* 24(2), 123–140.

Breiman, Leo (2001). "Random Forests." *Machine Learning* 45(1), 5–32.

Buchak, Greg, Gregor Matvos, Tomasz Piskorski, and Amit Seru (2018). "Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks." *Journal of Financial Economics* 130(3), 453–483.

Bureau, Consumer Financial Protection (2017). "Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process."

Bureau, Consumer Financial Protection (2019). "Interagency Statement on the Use of Alternative Data in Credit Underwriting."

Burlando, Alfredo, Michael A. Kuhn, and Silvia Prina (2025). "Too Fast, Too Furious? Digital Credit Delivery Speed and Repayment Rates." *Journal of Development Economics* 174, 103427.

Butaru, Florentin, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, and Akhtar Siddique (2016). "Risk and Risk Management in the Credit Card Industry." *Journal of Banking & Finance* 72, 218–239.

Caire, Dean and Maria Fernandez Vidal (2024). "Leveraging Transactional Data for Micro and Small Enterprise (MSE) Lending."

Caro, Spencer and Scott Nelson (2024). "The Arity of Disparity: Updating Disparate Impact for Modern Fair Lending."

Castellanos, Sara G., Diego Jiménez Hernández, Aprajit Mahajan, Eduardo Alcaraz Prous, and Enrique Seira (2024). "Contract Terms, Employment Shocks, and Default in Credit Cards." Working Paper.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, 785–794.

Cho, SungJin and John Rust (2017). "Precommitments for Financial Self-Control? Micro Evidence from the 2003 Korean Credit Crisis." *Journal of Political Economy* 125(5), 1413–1464.

Choi, Darwin, Zhenyu Gao, Sing-Sen Lam, Tian Li, and Wenlan Qian (2025). "Scared Away: Credit Demand Response to Expected Motherhood Penalty in the Labor Market." SSRN Scholarly Paper. Rochester, NY.

CNBV (2019). “Boletin Trimestral de Inclusion Financiera.”

CNBV (2023). “Panorama Anual de Inclusion Financiera.”

Cookson, J. Anthony, Benedict Guttman-Kenney, and William Mullins (2025). “Immigration and Credit in America.”

Cowan, Greig, Salvatore Mercuri, and Raad Khraishi (2023). “Modelling Customer Lifetime-Value in the Retail Banking Industry.”

CRIF (2018). “CRIF Desarrolla Nuevo Score No Hit Para Buró de Crédito En México.”

Crook, Jonathan and John Banasik (2004). “Does Reject Inference Really Improve the Performance of Application Scoring Models?” *Journal of Banking and Finance*.

Davis, Jesse and Mark Goadrich (2006). “The Relationship between Precision-Recall and ROC Curves.” *Proceedings of the 23rd International Conference on Machine Learning - ICML ’06*. Pittsburgh, Pennsylvania: ACM Press, 233–240.

De Cnudde, Sofie, Julie Moeyersoms, Marija Stankova, Ellen Tobback, Vinayak Javaly, and David Martens (2019). “What Does Your Facebook Profile Reveal about Your Creditworthiness? Using Alternative Data for Microfinance.” *Journal of the Operational Research Society* 70(3), 353–363.

De Lima Junior, João Manoel, Gabriela Borges Silva, José Egidio Altoé Junior, and Ana Paula Ruhe (2021). *Repercussões jurídicas e econômicas do mercado de cartões de crédito*.

Demirgüç-Kunt, Asli, Leora Klapper, Dorothe Singer, and Saniya Ansar (2022). *Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19*. Washington, D.C.: World Bank Group.

Department Of Commerce (2023). “Mexico - Financial Technologies (Fintech) Industry.”

Di Maggio, Marco and Dimuthu Ratnadiwakara (2024). “Invisible Primes: Fintech Lending with Alternative Data.” SSRN Scholarly Paper 3937438.

Di Maggio, Marco, Emily Williams, and Justin Katz (2022). “Buy Now, Pay Later Credit: User Characteristics and Effects on Spending Patterns.” Working Paper.

Di Maggio, Marco and Vincent Yao (2021). “Fintech Borrowers: Lax Screening or Cream-Skimming?” *Review of Financial Studies* 34(10), 4565–4618.

Doerr, Sebastian, Jon Frost, Leonardo Gambacorta, and Han Qiu (2022). “Population Ageing and the Digital Divide.” *SUERF Policy Brief*.

Drechsler, Itamar, Hyeyoon Jung, Weiyu Peng, Dominik Supera, and Guanyu Zhou (2025). “Credit Card Banking.” Staff Reports (Federal Reserve Bank of New York). Federal Reserve Bank of New York.

Duarte, Victor, Julia Fonseca, Divij Kohli, and Julian Reif (2025). “The Effects of Deleting Medical Debt from Consumer Credit Reports.”

Economista, El (2024). “Datos Alternativos Cada Vez Mas Presentes Para Otorgar Un Credito.”

Ekinci, Yeliz, Nimet Uray, and Füsun Ülengin (2014). "A Customer Lifetime Value Model for the Banking Industry: A Guide to Marketing Actions." *European Journal of Marketing* 48(3/4), 761–784.

Finnovista (2023). "Fintech Radar Mexico 2023."

FinRegLab (2023a). "Explainability and Fairness in Machine Learning for Credit Underwriting." https://finreglab.org/research/explainability-fairness-in-machine-learning-for-credit-underwriting-policy-analysis/.

FinRegLab (2023b). "Machine Learning Explainability & Fairness: Insights from Consumer Lending."

Flach, Peter and Meelis Kull (2015). "Precision-Recall-Gain Curves: PR Analysis Done Right." *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc.

Friedman, Jerome H., Trevor Hastie, and Rob Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33, 1–22.

Frost, Jon, Leonardo Gambacorta, Yi Huang, Hyun Song Shin, and Pablo Zbinden (2019). "BigTech and the Changing Structure of Financial Intermediation." *Economic Policy* 34(100), 761–799.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther (2022). "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *The Journal of Finance* 77(1), 5–47.

Fuster, Andreas, Matthew Plosser, Philipp Schnabl, and James Vickery (2019). "The Role of Technology in Mortgage Lending." *The Review of Financial Studies* 32(5), 1854–1899.

Gambacorta, Leonardo, Yiping Huang, Han Qiu, and Jingyi Wang (2024). "How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm." *Journal of Financial Stability* 73, 101284.

GAO (2009). "Credit Cards." https://www.gao.gov/assets/gao-10-45.pdf.

Gertler, Paul, Sean Higgins, Ulrike Malmendier, and Waldo Ojeda (2025). "Do Behavioral Frictions Prevent Firms from Adopting Profitable Opportunities?" Working Paper.

Gopal, Manasa and Philipp Schnabl (2022). "The Rise of Finance Companies and FinTech Lenders in Small Business Lending." *The Review of Financial Studies* 35(11). Ed. by Gregor Matvos, 4859–4901.

Guttman-Kenney, Benedict, Chris Firth, and John Gathergood (2023). "Buy Now, Pay Later (BNPL) ...on Your Credit Card." *Journal of Behavioral and Experimental Finance* 37, 100788.

Guttman-Kenney, Benedict and Andres Shahidinejad (2023). "Unraveling Information Sharing in Consumer Credit Markets." *SSRN Electronic Journal*.

Hair, Christopher M, Sabrina T Howell, Mark J Johnson, and Siena Matsumoto (2025). "Modernizing Access to Credit for Younger Entrepreneurs: From FICO to Cash Flow."

Higgins, Sean (2024). "Financial Technology Adoption: Network Externalities of Cashless Payments in Mexico." *American Economic Review* 114(11), 3469–3512.

Huang, Yiping, Zhenhua Li, Han Qiu, Sun Tao, Xue Wang, and Longmei Zhang (2023). "BigTech Credit Risk Assessment for SMEs." *China Economic Review* 81, 102016.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry (2019). "Adversarial Examples Are Not Bugs, They Are Features."

Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo F. P. Luttmer, and Kelly Shue (2016). "Screening Peers Softly: Inferring the Quality of Small Borrowers." *Management Science* 62(6), 1554–1577.

Jagtiani, Julapa and Catharine Lemieux (2019). "The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform." *Financial Management* 48(4), 1009–1029.

Johnson, Mark J., Itzhak Ben-David, Jason Lee, and Vincent Yao (2023). "FinTech Lending with LowTech Pricing." SSRN Scholarly Paper. Rochester, NY.

Karlan, Dean and Jonathan Zinman (2009). "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment." *Econometrica* 77(6), 1993–2008.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo (2010). "Consumer Credit-Risk Models via Machine-Learning Algorithms." *Journal of Banking & Finance* 34(11), 2767–2787.

Krivorotov, George (2023). "Machine Learning-Based Profit Modeling for Credit Card Underwriting - Implications for Credit Risk." *Journal of Banking & Finance* 149, 106785.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2017). "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, 275–284.

Laudenbach, Christine, Elin Molin, Kasper Roszbach, and Talina Sondershaus (2025). "Buy Now Pay (Less) Later: Leveraging Private BNPL Data in Consumer Banking." SSRN Scholarly Paper. Rochester, NY.

Lee, Jung Youn, Joonhyuk Yang, and Eric Anderson (2024). "Using Grocery Data to Predict Credit Card Payments." *Management Science*.

Lee, Jung Youn, Joonhyuk Yang, and Eric Anderson (2025). "Who Benefits from Alternative Data for Credit Scoring? Evidence from Peru."

Li, Shibo, Baohong Sun, and Alan L Montgomery (2011). "Cross-Selling the Right Product to the Right Customer at the Right Time." *Journal of Marketing Research* 48(4), 683–700.

Lin, J (1991). "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37(1), 145–151.

Lu, Jie, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang (2018). "Learning under Concept Drift: A Review." *IEEE Transactions on Knowledge and Data Engineering*, 1–18.

Massey Jr., Frank J. (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association* 46(253), 68–78.

Matcham, William (2025). "Risk-Based Borrowing Limits in Credit Card Markets."

Medina, Paolina C. and Jose L. Negrin (2022). "The Hidden Role of Contract Terms: The Case of Credit Card Minimum Payments in Mexico." *Management Science* 68(5), 3856–3877.

Meursault, Vitaly, Daniel Moulton, Larry Santucci, and Nathan Schor (2024). "One Threshold Doesn't Fit All: Tailoring Machine Learning Predictions of Consumer Default for Lower-income Areas." *Journal of Policy Analysis and Management*, 1–24.

Mienye, Ibomoiye Domor and Yanxia Sun (2022). "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects." *IEEE Access* 10, 99129–99149.

Mishra, Prachi, Nagpurnanand Prabhala, and Raghuram G Rajan (2022). "The Relationship Dilemma: Why Do Banks Differ in the Pace at Which They Adopt New Technology?" *Review of Financial Studies* 35(7), 3418–3466.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2019). "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications." *Journal of Marketing Research* 56(6), 960–980.

Pearson, Karl (1900). "On the Criterion That a given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50(302), 157–175.

Piccininni, Marco, Maximilian Wechsung, Ben Van Calster, Jessica L. Rohmann, Stefan Konigorski, and Maarten van Smeden (2024). "Understanding Random Resampling Techniques for Class Imbalance Correction and Their Consequences on Calibration and Discrimination of Clinical Risk Prediction Models." *Journal of Biomedical Informatics* 155, 104666.

Rajan, Uday, Amit Seru, and Vikrant Vig (2015). "The Failure of Models That Predict Failure: Distance, Incentives, and Defaults." *Journal of Financial Economics* 115(2), 237–260.

Richardson, Eve, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters (2024). "The Receiver Operating Characteristic Curve Accurately Assesses Imbalanced Datasets." *Patterns* 5(6), 100994.

Rishabh, Kumar (2024). "Beyond the Bureau: Interoperable Payment Data for Loan Screening and Monitoring." *SSRN Electronic Journal*.

Sadhwani, Apaar, Kay Giesecke, and Justin Sirignano (2021). "Deep Learning for Mortgage Risk*." *Journal of Financial Econometrics* 19(2), 313–368.

San Pedro, Jose, Davide Proserpio, and Nuria Oliver (2015). "MobiScore: Towards Universal Credit Scoring from Mobile Phone Data." *User Modeling, Adaptation and Personalization*. Ed. by Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless. Cham: Springer International Publishing, 195–207.

Siddiqi, Naaem (2017). *Intelligent Credit Scoring*. John Wiley & Sons, Ltd.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 267–288.

Trecone (2023). "La evolución del sector de delivery en México: un vistazo al pasado, presente y futuro."

van den Goorbergh, Ruben, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster (2022). "The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression." *Journal of the American Medical Informatics Association* 29(9), 1525–1534.

Wang, Lulu (2025). "Regulating Competing Payment Networks."

Yang, Jinpu (2025). "The Unintended Consequence of Interest Rate Caps on Small Dollar Loans."

Yin, Jiangning and Nan Li (2022). "Ensemble Learning Models with a Bayesian Optimization Algorithm for Mineral Prospectivity Mapping." *Ore Geology Reviews* 145, 104916.

Zadrozny, Bianca and Charles Elkan (2002). "Transforming Classifier Scores into Accurate Multiclass Probability Estimates." *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton Alberta Canada: ACM, 694–699.

Internet Appendix

A Appendix Tables and Figures

Table A.1: Default definition, default rates, and performance metrics for studies that predict creditworthiness

Citation (1)	Default Definition (2)	Default Rate (3)	AUC (4)	Precision (5)	Recall (6)	F1 (7)	AUC-PR (8)
This paper	60 days or more delinquent at any point during the first 12 months after origination	20%	0.796	0.421	0.666	0.516	0.513
Agarwal, Alok, Ghosh, and Gupta (2023)	Not specified	4%	0.738 for sample with credit history, 0.674 for sample without credit history	0.113 for sample with credit history, 0.115 for sample without credit history	0.348 for sample with credit history, 0.356 for sample without credit history	0.171 for sample with credit history, 0.174 for sample without credit history	Not reported
Albanessi and Vamossy (2024)	90 days or more past due on any debt within 8 quarters	18%	0.906	Not reported	Not reported	Not reported	Not reported
Berg, Burg, Gombovi, and Puri (2020)	Unpaid purchase after 3 reminders after 14 days	3%	0.734	Not reported	Not reported	Not reported	Not reported
Björkegren and Grissen (2020)	More than 15 days overdue on the phone bill	11%	0.772	Not reported	Not reported	Not reported	Not reported
Blattner and Nelson (2024)	At least 90 days delinquent on the loan 24 months after the application	8%	0.840 for minority and 0.887 for non-minority sample	0.250	0.540	0.342	Not reported

Citation (1)	Default Definition (2)	Default Rate (3)	AUC (4)	Precision (5)	Recall (6)	F1 (7)	AUC-PR (8)
Blattner, Nelson, and Spiess (2024)	Default of any severity up to 24 months after origination	14%	0.867	Not reported	Not reported	Not reported	Not reported
Butaru et al. (2016)	90 days delinquent over the next 2, 3 or 4 quarters	From 1.36% to 4.36% (varies by bank)	Not reported	Many reported	Many reported	Many reported	Not reported
Di Maggio and Ratnadiwakara (2024)	90 days delinquent on a loan 12 months after origination	8%	0.659	Not reported	Not reported	Not reported	Not reported
Duarte, Fonseca, Kohli, and Reif (2025)	90 days or more past due within the next 18-24 months	13%	0.712	0.736	0.448	0.557	Not reported
Frost et al. (2019)	Loss rate: volume of outstanding credit that is 30 days or more past due over origination amount	1%	0.764	Not reported	Not reported	Not reported	Not reported
Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022)	90 days or more delinquent at some point over the first 3 years after origination	1%	0.861	Not reported	Not reported	Not reported	0.062
Gambacorta, Huang, Qiu, and Wang (2024)	Not specified	16%	0.607	Not reported	Not reported	Not reported	Not reported

Citation (1)	Default Definition (2)	Default Rate (3)	AUC (4)	Precision (5)	Recall (6)	F1 (7)	AUC-PR (8)
Hair, Howell, Johnson, and Matsumoto (2025)	60 days past due, charged-off loan, or the borrower has received forbearance and a modified the loan	17%	0.663	Not reported	Not reported	Not reported	0.266
Huang et al. (2023)	Nonperforming loan ratio	2%	0.841	Not reported	Not reported	Not reported	Not reported
Iyer, Khwaja, Luttmer, and Shue (2016)	3 or more months late as of 3 years after the loan is initiated	31%	0.714	Not reported	Not reported	Not reported	Not reported
Jagtiani and Lemieux (2019)	At least 60 days past due within 24 months after origination	From 5% to 35%	0.689	Not reported	Not reported	Not reported	Not reported
Johnson, Ben-David, Lee, and Yao (2023)	Having a delinquent payment within 1 year since origination	8%	0.665	Not reported	Not reported	Not reported	Not reported
Khandani, Kim, and Lo (2010)	90 days or more delinquent on any credit card account	2%	0.952	0.839	0.688	0.756	Not reported
Lee, Yang, and Anderson (2024)	2 months delinquent	7%	0.679 for sample with credit history, 0.647 for sample without credit history	Not reported	Not reported	Not reported	Not reported

Citation (1)	Default Definition (2)	Default Rate (3)	AUC (4)	Precision (5)	Recall (6)	F1 (7)	AUC-PR (8)
Lee, Yang, and Anderson (2025)	At least 60 days delinquent on any consumer loan with any lender	7.6% for sample with credit history, 9.8% for sample without credit history	0.682 for sample without credit history, 0.677 for sample with credit history	Not reported	Not reported	Not reported	Not reported
Meursault, Moulton, Santucci, and Schor (2024)	90 or more days past due on at least one of the accounts within two years	22%	0.883	Not reported	0.860	Not reported	Not reported
Netzer, Lemaire, and Herzenstein (2019)	Loan status is charge-off, bankruptcy, or delinquency	35%	0.726	Not reported	Not reported	Not reported	Not reported
Rishabh (2024)	90 days or more delinquent, write-offs, or lender classifications indicating a loss	9% for banks, 12% for FinTech	0.7 for banks, 0.68 for FinTech	Not reported	Not reported	Not reported	0.200
Sadhwani, Giesecke, and Sirignano (2021)	Transition states	34%	0.700	Not reported	Not reported	Not reported	Not reported
San Pedro, Proserpio, and Oliver (2015)	90 days or more delinquent within 9 months since card activation	13%	0.725	Not reported	Not reported	Not reported	0.292

This table reports the default definition, default rate, and predictive performance (proxied by AUC, Precision, Recall, F1 and AUC-PR) of other studies using machine learning models to predict creditworthiness. For Blattner and Nelson (2024) we report a weighted average default rate constructed using the default rate for accepted loans (2.2%) and the default rate for rejected loans (11.5%). For Blattner and Nelson (2024) we calculated F1 score from reported recall and precision measures. For Blattner, Nelson, and Spiess (2024) we report a weighted average default rate constructed using the default rate for accepted loans (1.2%) and the

default rate for rejected loans (24.2%). For Frost et al. (2019), we report the default rate as the weighted average of loss rate across internal ratings. For Sadhwani, Giesecke, and Sirignano (2021) we report as the default rate the percentage of loans transitioning from 60 days delinquent to 90 days delinquent. AUC = area under the receiver operating characteristic curve; AUC-PR = area under the precision-recall curve; F1 = harmonic mean of precision and recall; FICO = Fair Isaac Corporation; P2P = peer-to-peer.

Table A.2: Robustness Tests and Alternative Models

Model	AUC (1)	Precision (2)	Recall (3)	F1 (4)
Panel A: Benchmark XGBoost Model				
Default model	0.796 [0.790, 0.802]	0.421 [0.411, 0.431]	0.666 [0.654, 0.678]	0.516 [0.506, 0.526]
Panel B: Robustness				
Alternative definition of target variable	0.776 [0.768, 0.783]	0.484 [0.477, 0.491]	0.850 [0.844, 0.857]	0.617 [0.610, 0.623]
Out-of-time	0.781 [0.775, 0.788]	0.393 [0.383, 0.402]	0.654 [0.642, 0.666]	0.491 [0.481, 0.500]
Panel C: Penalized Logistic Regression Models				
LASSO	0.7684 [0.7590, 0.7782]	0.3820 [0.3680, 0.3957]	0.6670 [0.6498, 0.6850]	0.4857 [0.4722, 0.4998]
GLMNET	0.7681 [0.7587, 0.7772]	0.3821 [0.3686, 0.3951]	0.6672 [0.6505, 0.6835]	0.4859 [0.4722, 0.4987]

This table shows out-of-sample AUC, precision, recall, and F1 scores for robustness tests and alternative models for predicting default. Alternative time window indicates a test using an alternative time window from origination to the end of the data period, such that the time window varies by user, rather than being fixed across users at the first 12 months since origination. Out-of-time is an alternative testing-training split where the training data are those who applied during the earlier part of the period and the testing data are those who applied during the later part of the period. Penalized logistic regression models use L1 regularization (LASSO) and elastic net regularization (GLMNET). The results use $N = 146{,}036$ users, split into training data to train the machine learning models and testing data to calculate out-of-sample measures of model performance. Bootstrapped 95% confidence intervals are included in square brackets, using 10,000 bootstrap samples. Threshold-dependent measures and confusion matrices use the profit-maximizing thresholds of a 24% probability of default. AUC = area under the receiver operating characteristic curve. F1 score = the harmonic mean of precision and recall. Performance metrics are evaluated only on the test set.

Table A.3: Search space used in machine learning algorithm

Panel A: XGBoost Classifier	
Evaluation metric	log-loss
Tuning	hyperopt, max eval 1,250
Panel B: Hyperparameter Space	
Tree-specific hyperparameters	
max_depth	hp.quniform('max_depth', 1, 100, 1)
min_child_weight	hp.loguniform('min_child_weight', -2, 3)
subsample	hp.uniform('subsample', 0.5, 1),
colsample_bytree	hp.uniform('colsample_bytree', 0.5, 1),
n_estimator	hp.quniform('n_estimators', 100, 1000, 1)
Learning and Regularization hyperparameters	
eta, learning rate	hp.loguniform('learning_rate', -9, 0),
gamma	hp.loguniform('gamma', -10, 10),
alpha (L1)	hp.loguniform('reg_alpha', -10, 10),
lambda (L2)	hp.loguniform('reg_lambda', -10, 10),

This table shows the search space used for hyperparameters in our XGBoost machine learning algorithm. XGBoost = extreme gradient boosting.

Table A.4: Summary statistics by quintile of number of transactions

	Quintile 1 (1)	Quintile 2 (2)	Quintile 3 (3)	Quintile 4 (4)	Quintile 5 (5)
Panel A: Subset of Features					
Woman - dummy	0.37	0.38	0.39	0.40	0.40
User age	26.04	25.06	24.60	24.44	25.35
User iOS (Apple) operating system - dummy	0.28	0.32	0.37	0.42	0.53
No-hit score	640.50	640.18	639.78	640.18	642.62
Number of orders on app	1.19	4.33	9.59	19.67	87.50
Proportion orders paid in cash	0.66	0.59	0.51	0.46	0.35
Median amount per order (MXN)	387.26	347.49	332.19	328.66	350.39
Proportion orders at supermarkets	0.05	0.04	0.05	0.06	0.07
Proportion orders at pharmacies	0.04	0.04	0.04	0.04	0.03
Proportion orders at food establishments	0.83	0.84	0.83	0.81	0.79
Marginality (SES) index of census tract	0.96	0.96	0.96	0.97	0.97
Years of schooling among age 15+ in census tract	11.83	12.12	12.34	12.60	13.18
Proportion households own a motor vehicle in census tract	0.60	0.62	0.64	0.66	0.70
Panel B: Credit Card Terms and Use					
Interest rate	0.80	0.80	0.81	0.81	0.81
Credit limit (MXN)	5,760.67	5,510.58	5,562.61	5,621.17	5,786.52
Spending (MXN)	2,899.36	2,915.69	2,996.00	3,156.64	3,576.43
Statement balance (MXN)	3,832.70	3,689.26	3,696.52	3,747.68	3,905.77
Minimum payment (MXN)	490.03	466.06	456.99	445.54	436.51
Repayment (MXN)	2,572.44	2,676.60	2,799.14	3,031.34	3,765.54
Delinquency - dummy	0.22	0.21	0.21	0.19	0.17
Panel C: Profit Components (MXN per Month)					
Interest revenue	60.92	61.10	62.94	65.98	74.73
BNPL revenue	8.05	7.37	7.46	7.27	6.34
Charge-offs	70.82	64.21	61.73	57.01	52.39
Interest and BNPL revenue net of charge-offs	-1.76	4.31	8.73	16.31	28.76
Interchange fee revenue	33.33	34.86	36.76	40.16	49.08
Cost of rewards	29.91	30.95	33.25	36.89	49.63
Interchange fee revenue net of rewards	3.65	4.07	3.68	3.38	-0.37
Late payment fee revenue	12.47	12.15	12.54	12.99	15.30
Funding costs	29.54	30.00	30.27	31.07	32.48
Costs from fraudulent transactions	0.46	0.48	0.56	0.62	0.54
Other fee revenue	1.90	2.22	2.38	2.67	3.75
Other costs	10.00	11.11	11.78	12.98	16.32
Negative profits - dummy	0.49	0.49	0.49	0.49	0.49

This table shows means for the sample that we use in our machine learning modeling, which consists of applicants with no credit history and no prior credit card who were approved by RappiCard, had at least twelve months with the card by the end of our data period in May 2024, and made at least one transaction during their first 12 months with the card. The mean is calculated separately by quintile of number of transactions in the delivery app. Observations are at the user level, and the testing sample of $N = 29{,}208$ is included. Panel A shows summary statistics for a small subset of the features used by our machine learning model. For non-binary variables, upper-tail winsorization was performed at the 99.9th percentile consistent with the way features were winsorized for modeling. Census tract for each user is inferred based on login activity on the delivery app. The marginality (SES) index is a summary measure of economic vulnerability at the census track level, which takes values between 0 (high marginality) and 1 (low marginality). Panel B shows summary statistics on the terms and use of the credit cards. Interest rate and credit limit are measured at origination. Statement balance, minimum payment, and repayment are averages over the first 12 monthly statements (measured in MXN per month). Delinquency is measured as at least 60 days delinquent at any point over the first 12 months since origination. Panel C shows summary statistics for the average monthly revenues obtained and costs incurred by the lender for each card, where borrower-level monthly averages are computed over the first 12 months since card origination. For panels B and C, non-binary variables are winsorized at the 1st and 99th percentiles, with the exception of "costs from fraudulent transactions," as positive values occur for this variable for less than 1% of users; borrower-level profits are computed using non-winsorized revenue and cost variables. Std. dev. = standard deviation; perc. = percentile; iOS = Apple device operating system; MXN = Mexican pesos; SES = socioeconomic status; BNPL = buy now pay later.

Table A.5: Summary statistics by model disagreement

	Approved by both (1)	Approved by default only (2)	Approved by profitability only (3)	Rejected by both (4)
Panel A: Subset of Features				
Woman - dummy	0.42	0.37	0.42	0.35
User age	25.02	24.24	27.40	25.35
User iOS (Apple) operating system - dummy	0.43	0.41	0.34	0.28
No-hit score	640.79	644.33	636.11	637.24
Number of orders on app	25.69	28.18	23.93	13.48
Proportion orders paid in cash	0.53	0.39	0.64	0.62
Median amount per order (MXN)	356.74	360.76	348.20	327.23
Proportion orders at supermarkets	0.05	0.06	0.05	0.06
Proportion orders at pharmacies	0.04	0.03	0.05	0.04
Proportion orders at food establishments	0.84	0.82	0.82	0.80
Marginality (SES) index of census tract	0.96	0.97	0.96	0.96
Years of schooling among age 15+ in census tract	12.48	12.64	12.14	12.01
Proportion households own a motor vehicle in census tract	0.66	0.66	0.63	0.60
Panel B: Credit Card Terms and Use				
Interest rate	0.83	0.78	0.82	0.78
Credit limit (MXN)	5,531.41	6,032.74	5,204.56	5,475.85
Spending (MXN)	3,112.84	3,441.83	2,606.49	2,748.23
Statement balance (MXN)	3,419.58	3,411.79	4,274.62	4,625.11
Minimum payment (MXN)	336.32	255.22	715.71	845.77
Repayment (MXN)	3,146.44	3,478.81	2,243.36	2,008.58
Delinquency - dummy	0.11	0.08	0.36	0.45
Panel C: Profit Components (MXN per Month)				
Interest revenue	74.17	53.18	84.02	60.61
BNPL revenue	7.66	8.94	5.28	5.51
Charge-offs	29.51	24.58	105.21	147.24
Interest and BNPL revenue net of charge-offs	52.42	37.62	-15.91	-81.01
Interchange fee revenue	42.31	49.00	25.04	22.12
Cost of rewards	36.51	50.28	20.22	19.38
Interchange fee revenue net of rewards	5.80	-1.19	5.10	3.14
Late payment fee revenue	12.73	9.57	20.56	16.46
Funding costs	31.92	27.30	35.30	30.78
Costs from fraudulent transactions	0.53	0.49	0.33	0.51
Other fee revenue	3.57	2.41	2.39	1.39
Other costs	15.17	14.06	8.95	6.29
Negative profits - dummy	0.41	0.54	0.44	0.56

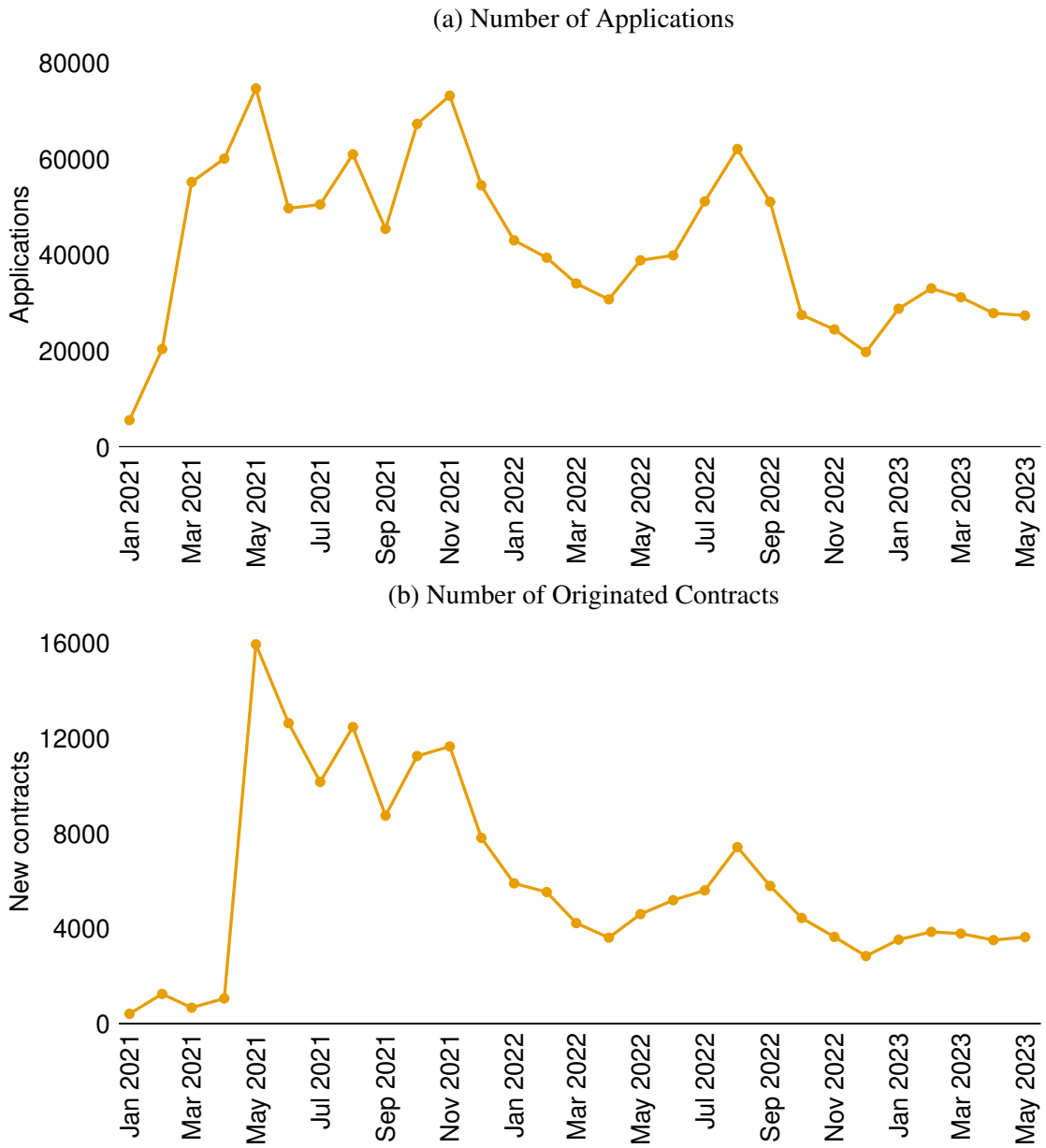
This table shows means for the sample that we use in our machine learning modeling, which consists of applicants with no credit history and no prior credit card who were approved by RappiCard, had at least twelve months with the card by the end of our data period in May 2024, and made at least one transaction during their first 12 months with the card. The mean is calculated separately by group defined by model disagreement between the default and profitability models: accepted by both models, accepted by default model only, accepted by profitability model only, and rejected by both models. The results use $N = 146{,}036$ users, randomly split into training and testing data; summary statistics are shown for the testing sample of $N = 29{,}208$. Panel A shows summary statistics for a small subset of the features used by our machine learning model. For non-binary variables, upper-tail winsorization was performed at the 99.9th percentile consistent with the way features were winsorized for modeling. Census tract for each user is inferred based on login activity on the delivery app. The marginality (SES) index is a summary measure of economic vulnerability at the census track level, which takes values between 0 (high marginality) and 1 (low marginality). Panel B shows summary statistics on the terms and use of the credit cards. Interest rate and credit limit are measured at origination. Statement balance, minimum payment, and repayment are averages over the first 12 monthly statements (measured in MXN per month). Delinquency is measured as at least 60 days delinquent at any point over the first 12 months since origination. Panel C shows summary statistics for the average monthly revenues obtained and costs incurred by the lender for each card, where borrower-level monthly averages are computed over the first 12 months since card origination. For panels B and C, non-binary variables are winsorized at the 1st and 99th percentiles, with the exception of "costs from fraudulent transactions," as positive values occur for this variable for less than 1% of users; borrower-level profits are computed using non-winsorized revenue and cost variables. Std. dev. = standard deviation; perc. = percentile; iOS = Apple device operating system; MXN = Mexican pesos; SES = socioeconomic status; BNPL = buy now pay later.

Table A.6: Data Drift Test Statistics

Data Drift Test	Variables	Definition	Reference
Population Stability Index (PSI)	Numerical Categorical Binary	Compares distributions by variables and computing a log-ratio of frequencies between expected (baseline) and observed (updated data) datasets.	(Siddiqi, 2017)
Jensen-Shannon (JS) Divergence	Numerical Categorical Binary	Jensen-Shannon is a symmetric divergence metric that measures the relative entropy or difference in information represented by two probability distributions. It is always finite and interpretable as a measure of similarity. It is derived from the Kullback-Leibler divergence but smoother.	(Lin, 1991)
Chi-squared Test	Binary Categorical	Tests for independence between distributions using a comparison of observed vs. expected category frequencies.	(Pearson, 1900)
Kolmogorov-Smirnov (KS) Test	Numeric	Non-parametric test comparing empirical cumulative distributions from two samples. Sensitive to differences in shape, location, and spread.	(Massey, 1951)

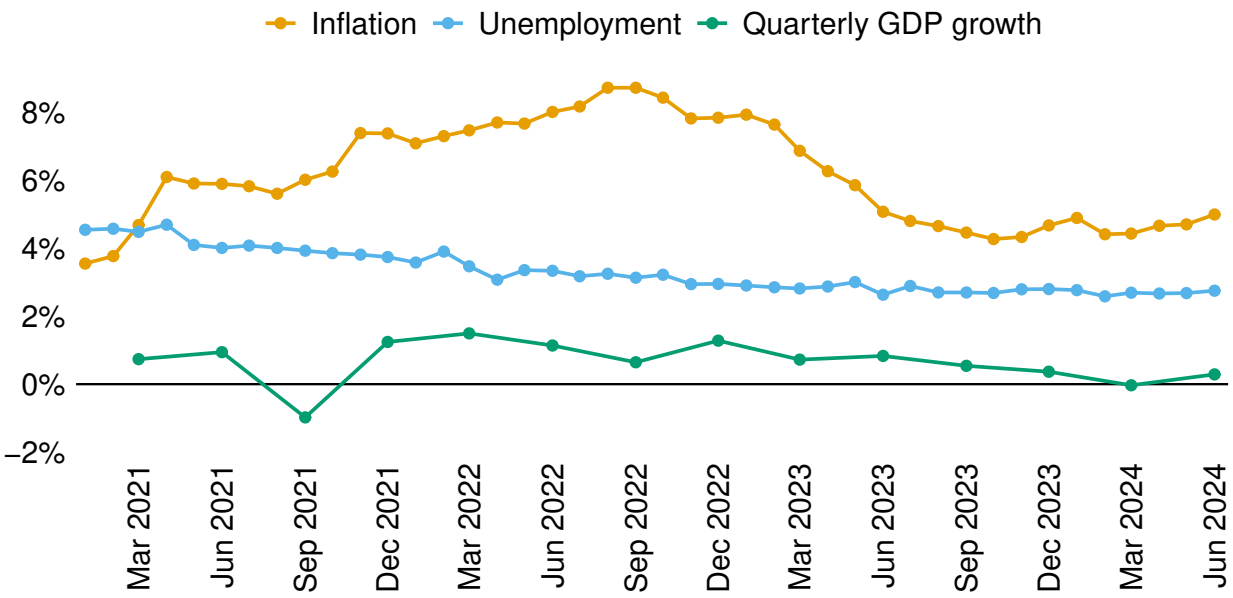
This table describes the data drift test statistics that we use.

Figure A.1: Number of Applications and Originated Contracts over Time



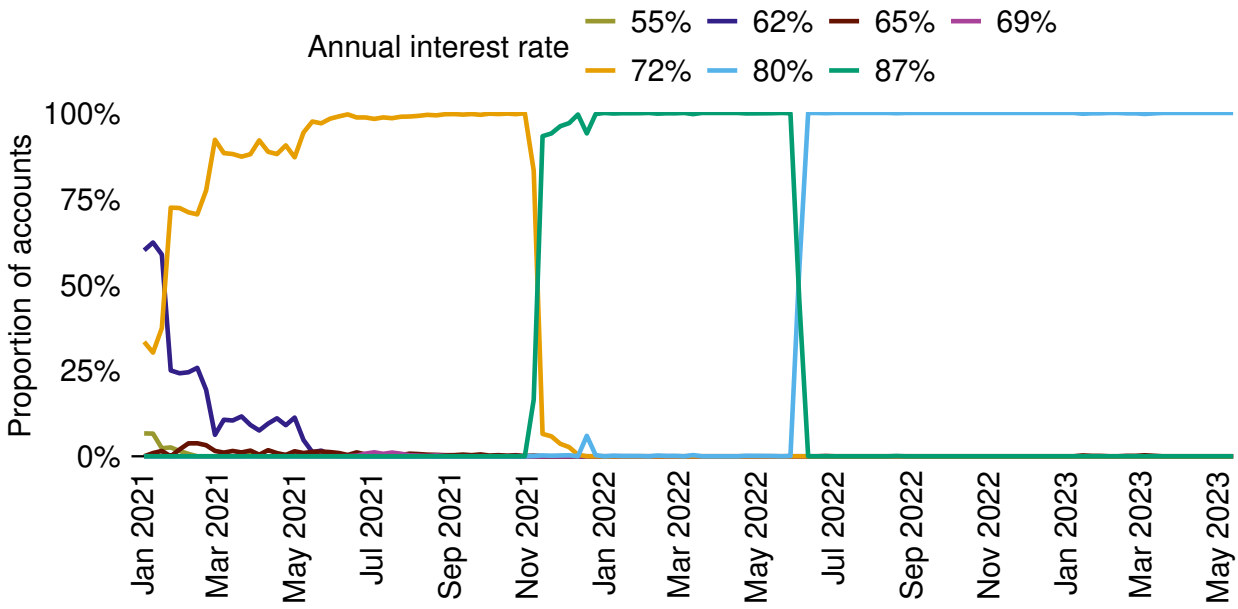
This figure shows the number of applications by month and the number of originated contracts by month from January 2021 through May 2023.

Figure A.2: Macroeconomic indicators for Mexico over period of analysis



This figure shows annual inflation, unemployment and quarterly GDP growth rates between January 2021 and June 2024. Annual inflation is published by Banco de México and is the difference in the consumer price index (CPI) on a given month, and the CPI of the same month one year earlier. Unemployment rate is expressed as a fraction of labor force, it is seasonally adjusted and published by Mexico's National Institute of Statistics (INEGI). Quarterly GDP growth is published by INEGI, it is expressed in real terms (GDP is measured in prices of 2018) and seasonally adjusted.

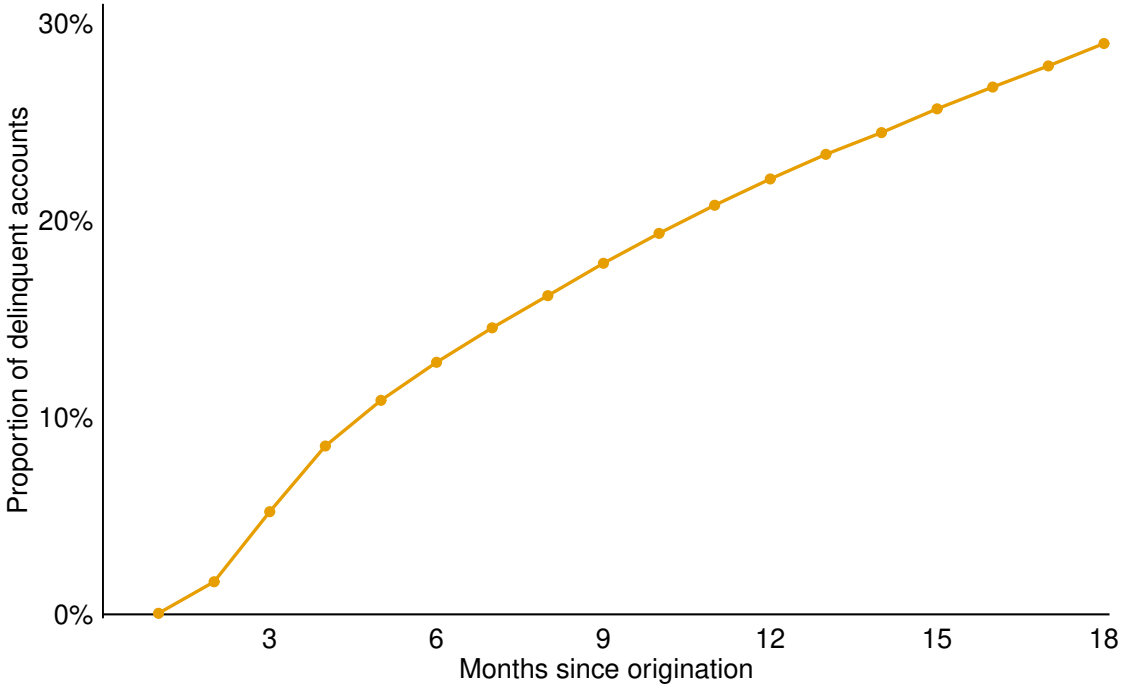
Figure A.3: Distribution of Interest Rates Assigned at Origination, Over Time



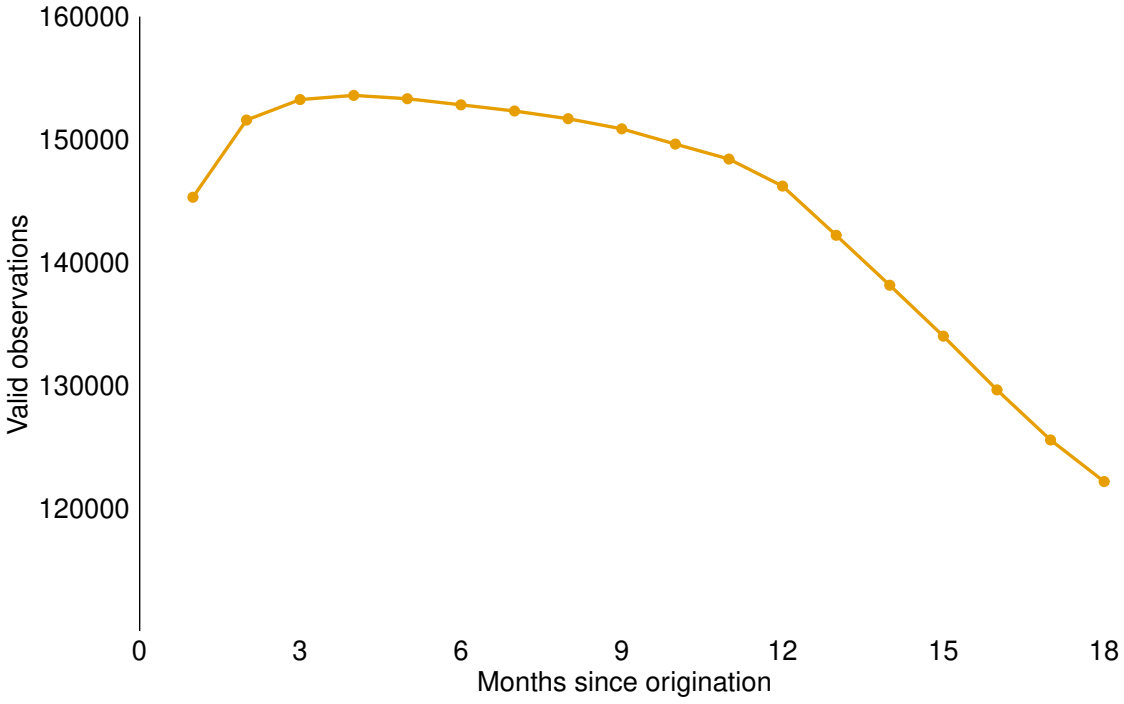
This figure shows the proportion of accounts assigned each annual interest rate by week of origination. $N = 146{,}030$. For 6 accounts, we do not observe the interest rate assigned at origination, and thus these accounts are excluded from the figure.

Figure A.4: Trade-offs in Threshold of Months since Origination over which Default is Measured

(a) Proportion of Delinquent Accounts



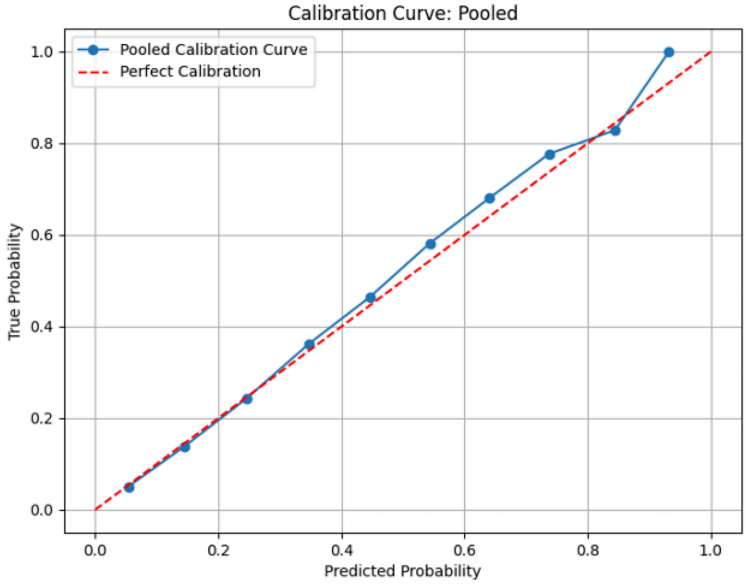
(b) Number of Observations in Modeling Sample



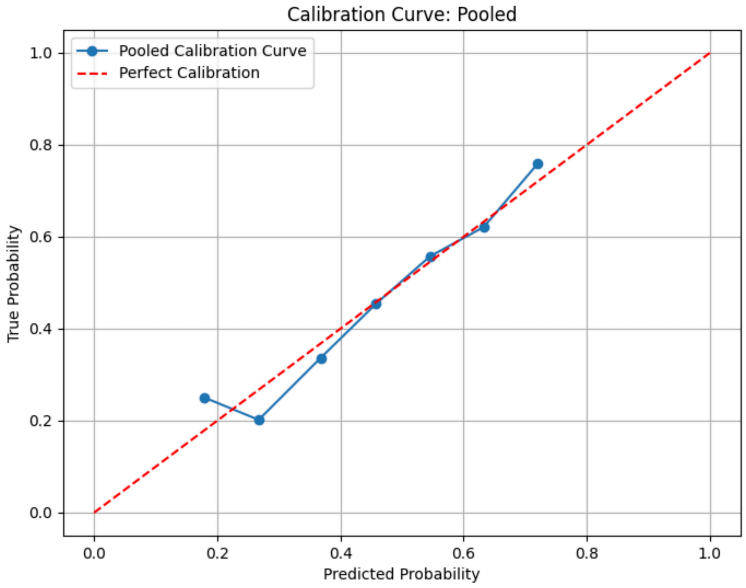
This figure illustrates the trade-offs underlying setting the threshold of number of months since origination over which default is measured. Panel (a) shows the proportion of accounts that are delinquent within x months since origination. Panel (b) shows the number of valid observations available to be included in the modeling sample based on different thresholds over which default is measured.

Figure A.5: Model Calibration: Default and Profitability Predictions

(a) Default Model Calibration

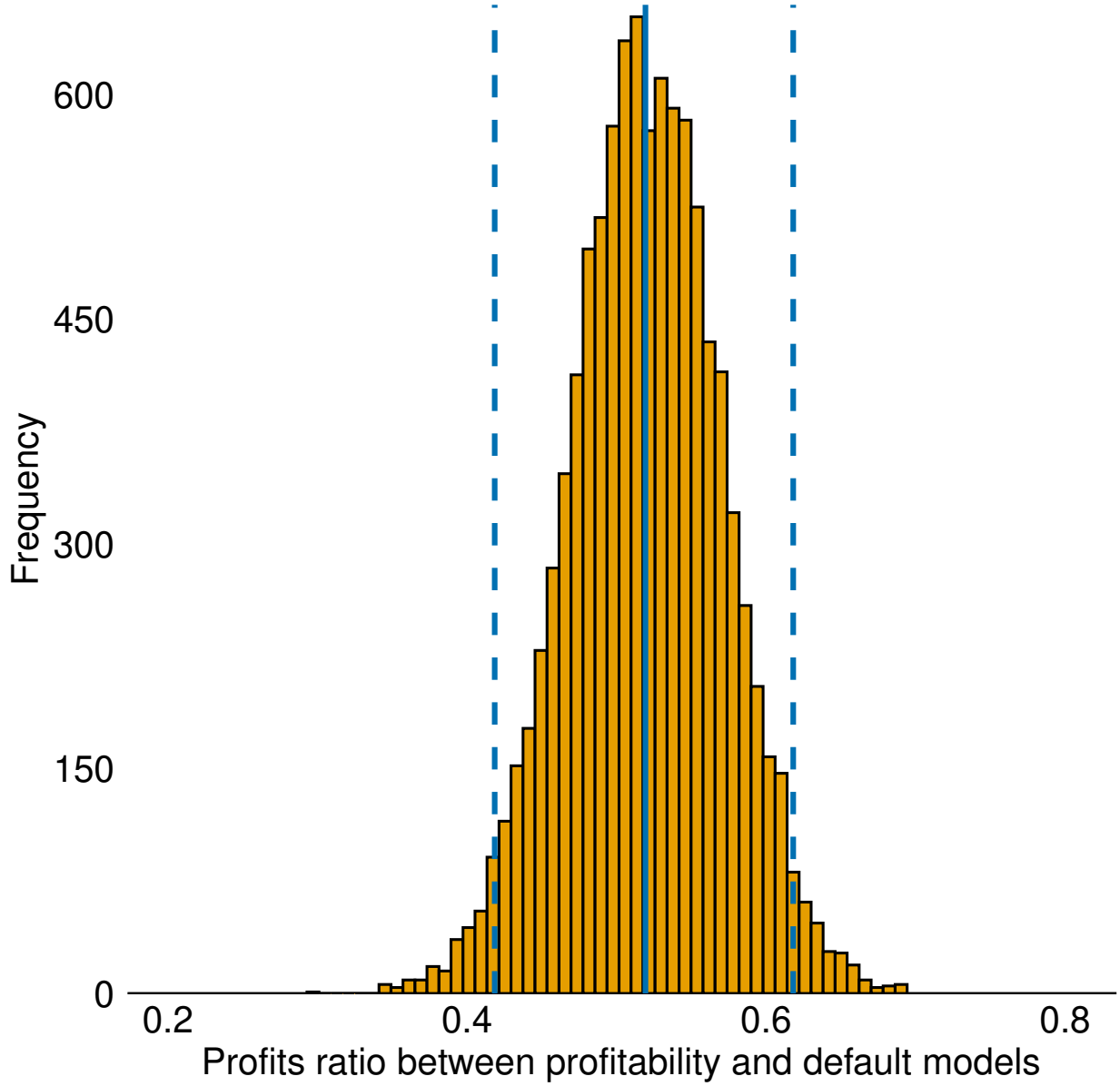


(b) Profitability Model Calibration



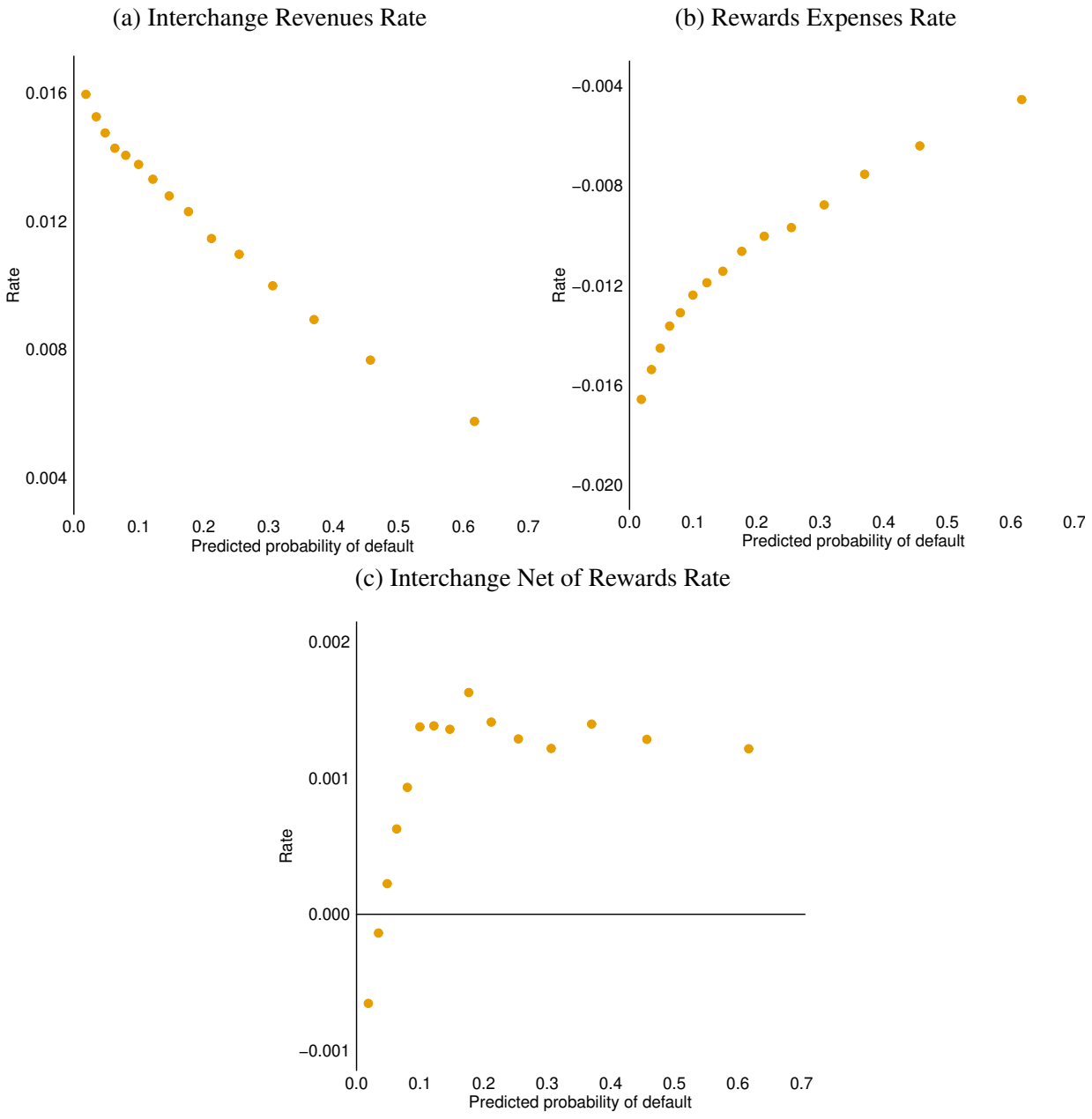
This figure shows calibration curves for the benchmark XGBoost models trained with log-loss optimization. Perfect calibration (45-degree line) represents the ideal scenario where predicted probabilities exactly match the true frequency of the positive label. Panel (a) displays the relationship between the model predicted probabilities of default and observed default rates, with the blue line representing the model calibration curve and the red dashed line indicating perfect calibration. Panel (b) displays the profitability model calibration curve.

Figure A.6: Bootstrap Estimates of Ratio of Profits between Profitability and Default Models



This figure plots estimates from 10,000 bootstrap samples of the ratio of total profits from the profitability model to total profits from the default model, all within the testing sample, using the profit-maximizing threshold for each model. The results use $N = 146{,}036$ users, randomly split into training and testing data. The solid vertical blue line shows the estimated ratio in the full sample, while the dashed vertical lines show the upper and lower bounds of the 95% bootstrapped confidence interval.

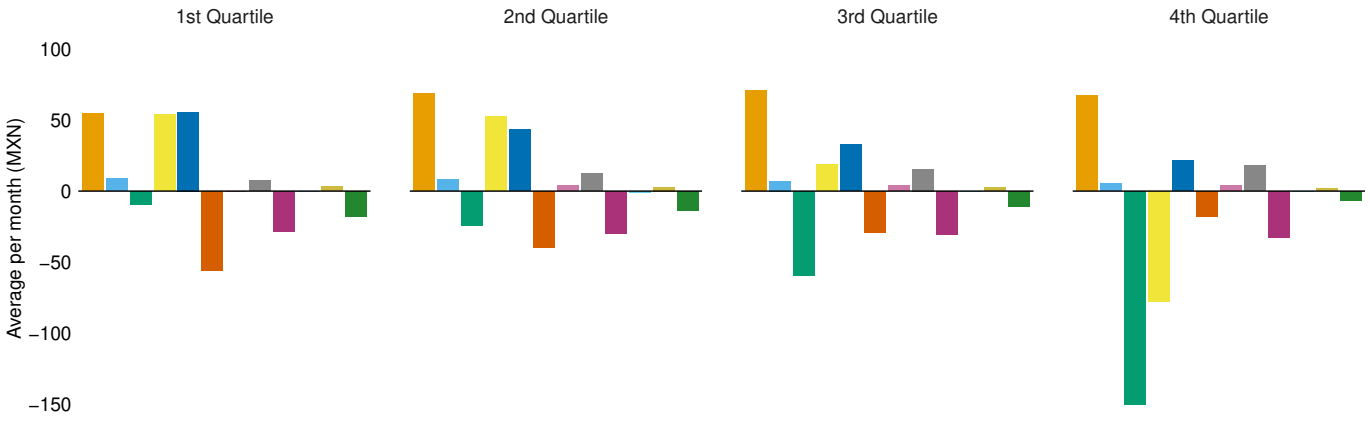
Figure A.7: Interchange Fees and Rewards as a Proportion of Spending



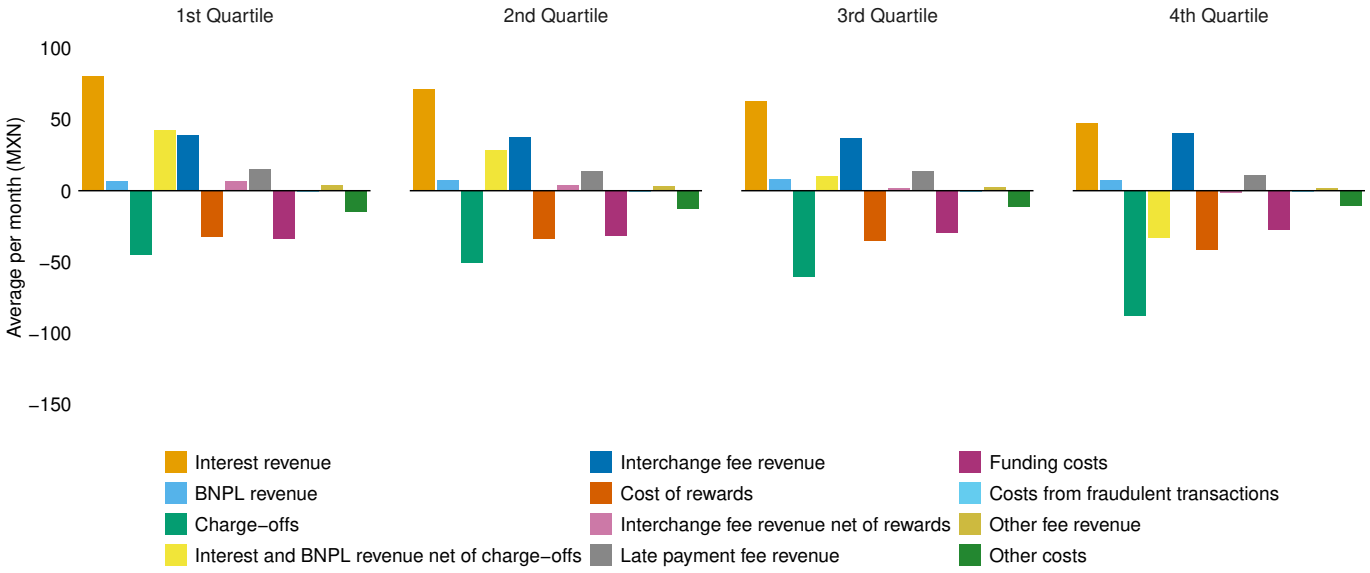
This figure plots a binscatter of interchange revenues, rewards expenses, and interchange revenues net of rewards expenses, all measured as a rate (i.e., as a proportion of spending). The results use $N = 146{,}036$ users, randomly split into training and testing data. All estimates are in the testing sample across 15 bins in the predicted probability of default or of negative profits, with approximately equal number of observations in each bin. We construct each measure at the account level by dividing interchange fees, rewards, or interchange fees net of rewards by spending, all over the first 12 months since card origination. We then calculate spending-weighted averages of the rates within bin. (This is equivalent to summing total interchange fees, rewards, or interchange fees net of rewards within bin and dividing by the sum of spending within the bin.)

Figure A.8: Revenues, Costs, and Profits by Quartile of Predicted Probabilities

(a) By Quartile of Predicted Default



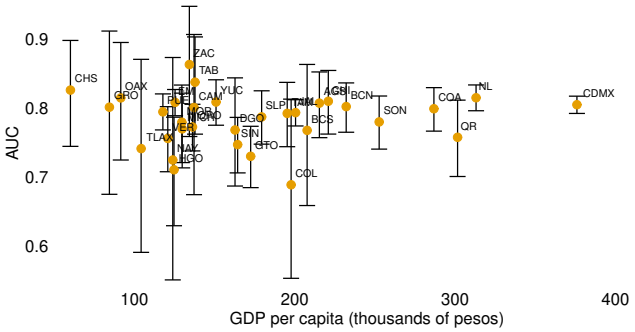
(b) By Quartile of Predicted Probability of Negative Profits



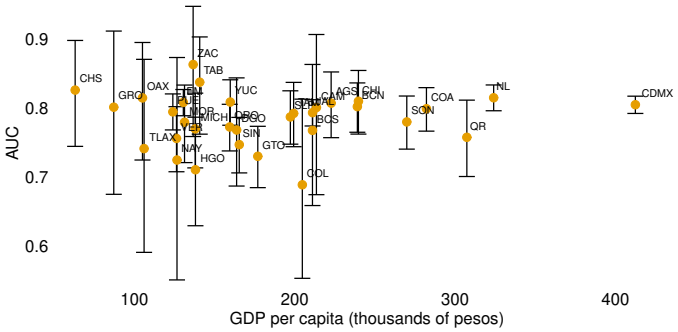
This figure shows the components of revenues, costs, and profits in the testing sample by quartile of predicted probabilities for the default and profitability models. The results use $N = 146{,}036$ users, randomly split into training and testing data. In both panels, variables are winsorized at the 1st and 99th percentiles, with the exception of "costs from fraudulent transactions," as positive values occur for this variable for less than 1% of users.

Figure A.9: State-level AUCs and economic activity (without revenue from oil)

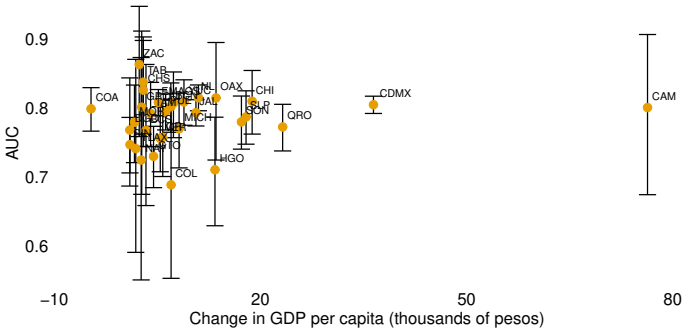
(a) 2021



(b) 2023

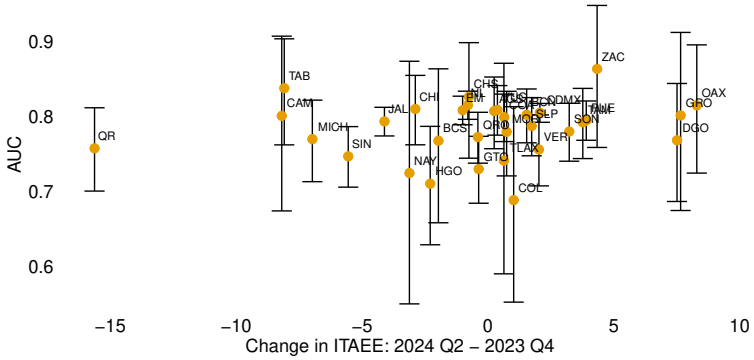


(c) 2023 - 2021



This figure shows scatterplots of state-level AUCs from our benchmark default model (vertical axis) against levels (or changes) in GDP per capita, excluding revenue from oil, for each state (horizontal axis, in thousands of pesos of 2018). Borrowers are assigned to states based on the address provided at the time of credit card application. The results use $N = 146{,}036$ users, randomly split into training data and testing data. AUCs are computed on the testing data considering only observations from borrowers in the corresponding state. Predictions come from the default model of Section 4 (not from separate models by state). GDP without revenue from oil is calculated by substracting the contribution of revenue from oil to state level GDP, as published by Mexico's National Statistical Institute (INEGI). This value is then divided by the population in each state and year, as reported by Mexico's National Population Agency (CONAPO). Panel A uses GDP per capita for 2021. Panel B uses GDP per capital for 2023. Panel C uses the change in GDP per capita between 2023 and 2021. Vertical lines represent 95% confidence intervals for AUCs, obtained with 10,000 bootstrap repetitions. Labels correspond to state names.

Figure A.10: State-level AUCs and economic activity during the first semester of 2024



This figure shows scatterplots of state-level AUCs from our benchmark default model (vertical axis) against the change in the Quarterly Index of Economic Activity (ITAEE) between 2024Q2 and 2023Q4 (horizontal axis). ITAEE is published by Mexico's National Statistical Institute (INEGI) and takes the value of 100 for every state in 2018. Borrowers are assigned to states based on the address provided at the time of credit card application. The results use $N = 146{,}036$ users, randomly split into training data and testing data. AUCs are computed on the testing data considering only observations from borrowers in the corresponding state. Predictions come from the default model of Section 4 (not from separate models by state). Vertical lines represent 95% confidence intervals for AUCs, obtained with 10,000 bootstrap repetitions. Labels correspond to state names.