FinTech Lending to Borrowers with No Credit History

Laura Chioda Paul Gertler Sean Higgins Paolina Medina*

October 9, 2024

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. We also show that segmenting our machine learning model by gender can improve credit allocation fairness without a substantive effect on the model's predictive performance.

*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 José Antonio Murillo, Luz Téllez, and Roberto Arriaga 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, and Jora Li 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.

So far, when lending to clients with no credit history, our FinTech partner has not relied on a machine learning algorithm; instead, they use 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, Gombović, 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, as repayment data from a credit card would be deemed sufficient data by the credit bureau to generate a traditional credit score.

[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. 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 solely on publicly available

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 will implement our model to make lending decisions for borrowers with no traditional credit score in the coming months.

Our baseline model achieves an area under the receiver operating characteristic curve (AUC) of 0.752. 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 then test how important each group of features is by estimating the AUC of a model which excludes them and comparing it to the model that includes all features. We find that the digital footprints data—which include the exact same set of variables as in Berg, Burg, Gombović, 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 have the highest marginal contributions to the model's predictive power. The digital footprints data contribute 0.044 to the AUC and the digital transactions data contribute 0.028.[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 at least 27 transactions through the delivery app—has an AUC of 0.777 while that for the lowest quintile—who have at most 2 transactions through the delivery app—has an AUC of 0.707.

Next, we show that the fairness and equity of algorithmic decisions—which have become increasingly important points of discussion and regulation for the use of machine learning models to predict creditworthiness (Bartlett, Morse, Stanton, and Wallace, 2022; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2022)—can be addressed by adopting gender-segmented models without meaningful losses in predictive accuracy (as proxied by AUCs) nor a deterioration in the portfolio default rates. We train gender-segmented models on women-only and men-only samples and allow all aspects of the algorithm (e.g., feature selection, feature importance, and hyperparameter tuning) to vary by gender. We show that this approach would permit a lender to identify low-risk female

data sets aggregated at the local level merged with the location where the applicant lives; see CRIF (2018). In addition, the other data sources include a score based on cell phone records, sold to FinTech companies by local providers, and socioeconomic characteristics at the census tract level.

[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 0.016 to the AUC. The mobile phone-based scores contribute 0.009. The socioeconomic characteristics at the census tract level contribute 0.001.

borrowers, who are approved by the gender-segmented model but rejected by a pooled model, as the pooled model is not able to fully capture how different behaviors may differentially predict creditworthiness for men and women.[4]

Specifically, we find that 12.3% of women who would be rejected by a standard pooled machine learning model would instead be approved by the gender-segmented model. In contrast, only 4.0% of women who would be approved by the pooled model would be rejected by the gender-segmented model. In absolute terms, the number of women approved by the gender-segmented model but rejected by the pooled model is 2.6 times larger than the number of women approved by the pooled model but rejected by the gender-segmented model. This is achieved without cost to the predictive power of the model: our gender-segmented model has an AUC of 0.750 compared to the pooled model's AUC of 0.752. Furthermore, it is achieved without a deterioration in the portfolio default rate: the overall portfolio default rates of both models are very similar, at 10%. Finally, it is achieved without a significant change to the allocation of credit to men, with 2.8% of male applicants approved by the pooled model but rejected by the gender-segmented model and 2.6% of male applicants approved by the gender-segmented model but rejected by the pooled model. Thus, the FinTech lender could increase access to credit for a subset of women deemed sufficiently low-risk by the gender-segmented model without a substantive change to the performance of their portfolio.

From a regulatory perspective, not only are there no restrictions in Mexico on the use of gender variables in credit scoring, but as of 2021, by law, banks must differentiate reserve requirements by borrower gender (SEGOB, 2021). However, in some regulatory environments like that of the US, protected features such as gender and race cannot be explicitly leveraged in credit scoring models. Our findings add to the evidence that regulating algorithms to be gender-"blind" could be exacerbating the inequities that originally motivated the regulations (Dwork et al., 2012; Kearns and Roth, 2019).

We make two 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 Ratnadiwakara, 2024) or borrowers with a thin credit file that is nevertheless sufficient for the credit bureau to generate

[4]Our approach aligns with the framework in Kleinberg, Ludwig, Mullainathan, and Rambachan (2018) for thinking about algorithmic fairness. In the context of US college admissions, they also find no meaningful trade-off empirically between efficiency and fairness considerations once they grant their algorithm access to race variables; however, they do not implement a race-segmented model. In the next sections, we detail how these two approaches differ.

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 0.752.

Our second contribution is to show that machine learning models using alternative data to predict creditworthiness of applicants with no credit history can achieve fairness objectives without a substantive effect on the predictive performance of the models. Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022) show that, holding input data fixed, the gains from more sophisticated machine learning models are not evenly distributed and accrue more to White borrowers than to Black and Latinx borrowers in the US. Bartlett, Morse, Stanton, and Wallace (2022) find that FinTech lenders charge Black and Latinx borrowers more than otherwise observationally equivalent White borrowers to the same extent that traditional lenders do. We consider a regulatory environment where these variables *can* be used in lenders' default prediction models, as recommended by the algorithmic fairness literature (Kleinberg, Ludwig, Mullainathan, and Rambachan, 2018; Kearns and Roth, 2019).

Although creditworthiness models typically pool data from men and women and either omit gender entirely due to discrimination concerns (Mester, 1997), or include gender without fully capturing the ways in which gender interacts with other variables (e.g., Johnston and Morduch, 2008), we show that a gender-segmented model identifies a subset (12.3%) of women who would be rejected by a pooled machine learning model to approve for credit, while rejecting a only smaller subset (4.0%) of women who would be approved by the pooled model. The gender-segmented model accomplishes this by forcing the algorithm to consider how different features in the alternative data *differentially* predict creditworthiness for men and women, and also by allowing other aspects of the model such as feature selection and hyperparameter tuning to differ across genders. Because the overall predictive accuracy of the gender-segmented models is similar to that of the pooled model, we conclude that our FinTech partner can increase the credit allocation fairness of its model without a substantive cost in terms of predictive accuracy or default rates.

[5] An exception is Björkegren and Grissen (2020), where the random forest models using mobile phone data for the 15% of their sample with no credit history have an AUC of 0.719, compared to an AUC of 0.708 for the model combining mobile phone data and credit bureau data for the 85% of their sample with credit histories.

2 Institutional Context

2.1 Financial Inclusion

Only 37% of Mexicans have bank accounts and 32% have made or received digital payments, both significantly below the equivalent rates for countries with similar levels of development. Moreover, there is an 8 percentage point (p.p.) gender gap in the probability of having a bank account, which is significantly higher than that of other countries in Latin America and of other OECD countries (World Bank, 2021).

In the context of credit markets, a national survey (INEGI, 2021) finds that 31% of Mexicans have at least one credit account in formal institutions and women are 2 p.p. less likely than men to have such accounts. However, the same survey also reveals differences in the types of credit to which they have access, which may translate into gender gaps in the intensive margin of access to credit. For instance, women are 4 p.p. less likely to own credit cards issued by financial institutions (compared to 12% of men who do), and 2 p.p. more likely to have store credit cards (compared to 19% of men who do). Notably, store credit cards are typically associated with higher interest rates than credit cards issued by financial institutions (CONDUSEF, 2016). Women are also significantly less likely to have collateralized credit and more likely to have microcredit.

To promote the financial inclusion of women, in 2021, the National Banking and Securities Commission (CNBV) issued a new regulation that explicitly differentiates reserve requirements for banks, when issuing credit for women and men (SEGOB, 2021). Reserve requirements are set by law, as a function of an estimated probability of default for each loan. This amendment includes a downward adjustment factor in creating reserves for loans granted to women. For non-revolving consumer loans, the factor reduces the probability of default by 4% for personal loans, durable goods loans, and automobile loans, and by 2% for payroll loans. For most housing mortgages, the factor reduces the default probability by 3%. This regulation was based on international evidence that points to women defaulting less than men, but also having less access to credit than men (D'Espallier, Guérin, and Mersland, 2011; Global Banking Alliance, 2017). For example, in Chile, loan requests submitted by women were 18% less likely to be approved compared to otherwise equivalent loan requests submitted by men, despite women repaying at higher rates (Montoya, Parrado, Solis, and Undurraga, 2022).[6]

[6]Related evidence from Turkey found that loan officers were 26% more likely to require a guarantor for identical loans submitted by female applicants compared to male applicants (Brock and De Haas, 2023).

2.2 FinTech Lending

Fostering a dynamic FinTech environment has also 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 2023, Mexico is one of the largest FinTech markets in Latin America with 650 FinTech start-ups (CNBV, 2019; 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), which are one of the most common ways for new borrowers to access formal credit. In Mexico, for instance, credit cards were the first loan type for 74% of all formal sector borrowers (Castellanos et al., 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 control 56.5% of the cards issued by traditional financial institutions and the top five largest banks control 87.0% of them. However, during 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.[7]

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 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, 24.2% 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.[8]

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

[8] 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/.

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.[9] 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 686,277 individuals who applied for a credit card between November 2020 and November 2022, and performance data for 136,062 credit cards originated from these applications. This sample is a random sample of all individuals who applied for a credit card during this period and were flagged by the Mexican credit bureau as having null or insufficient credit history to have a traditional credit score. Of the sample of 136,062 approved applications, 52,334 of those approved for credit cards were women and 83,728 were men.

Individuals in the sample are the subset of those with no credit history (or with too limited of a credit history to have a traditional credit score issued by the credit bureau) who were approved for a loan by our FinTech partner using their ad hoc decision rules for applicants with no credit score, i.e., 136,062 applicants. To not understate default rates associated with inactive or recent card holders, we impose two additional restrictions on the analysis sample: card holders (i) must have completed at least one transaction using their credit card and (ii) must have held the card for at least 120 days after their first transaction. This leaves us with an analysis sample of 123,042 approved applicants, of which 46,928 are women and 76,114 are men.

3.2 Data Sources

For each applicant, irrespective of their application's outcome (approved or rejected), 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, Gombović, and Puri (2020).[10]
>
> *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,

[9] Burlando, Kuhn, and Prina (2023) study the effects of a digital lender in Mexico imposing a waiting period.

[10] The variable "email error" used in Berg, Burg, Gombović, 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.

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., Higgins, n.d.), 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 which 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). We confirm in the data that none of the borrowers in our sample had a credit card prior to applying for a card from our FinTech partner, suggesting that repayment data on a credit card would be deemed sufficient for the credit bureau to generate a credit score. 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.

Mobile phone-based proprietary scores, based on cell phone records. These scores are sold to FinTech companies by independent local providers.

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.

Our data also feature monthly information on outstanding balances and number of days in default for individuals who received a credit card. We define the target variable for the machine learning models as overdue for more than 60 days, which we refer to throughout the paper as "default."

3.3 Summary Statistics

Table 2 shows descriptive statistics for our target population. The applicants in our sample are relatively young: the average user age in our sample is 24.9. Younger people are more likely to lack formal credit histories, more likely to use smartphones and delivery apps, and also more likely to consider a FinTech lender as a potential source of credit. Less than half of the sample (37%) uses an Apple product, which is an important predictor of creditworthiness (Berg, Burg, Gombović, and Puri, 2020). There is not a lot of variation in the no-hit score, which has a mean of 638.9, a standard deviation of 20.7, and an interquartile range of 631 to 648; 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.[11] 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 23.7 with a standard deviation of 57.7 and interquartile range of 3 to 22, the average percent of orders paid in cash is 48% with an interquartile range of 14% to 81%, and the median amount per order is 298 Mexican pesos with a standard deviation of 333 pesos. The majority (80%) of purchases are orders from food establishments, while 5% are from supermarkets and 3% are from pharmacies.

4 Machine Learning Methods

4.1 Algorithm Details

We use data on credit card default 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.[12] XGBoost, like other boosting

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

[12] 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,

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).[13] 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 pooled models and gender-segmented 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.1. The Bayesian optimization algorithm was implemented with the aid of 5-fold cross-validation. The evaluation metric in all models is log-loss, which is preferred in scenarios in which we are interested not only in a predicted class (e.g., default vs. no default), but also in the predicted probability of being classified into a given class.

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, 2023; Meursault, Moulton, Santucci, and Schor, 2023). Contributions not using XGBoost opt for random forests (Björkegren and Grissen, 2020; 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, Gombović, and Puri, 2020) and deep neural networks (Sadhwani, Giesecke, and Sirignano, 2021).[14] 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). Stratification guarantees that the incidence of each class (default and no default) is preserved in both sets. In our context and given our interest in comparing

2022).

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

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

pooled and gender-segmented models, we also ensure that the bivariate distribution of gender and class is preserved. 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.

We train pooled models that combine data on both men and women, as well gender-segmented models that first split the data by gender prior to training them. It is worth noting that the pooled models are not gender-blind: that is, they are allowed to access the gender variable. This approach enables the models to capture how gendered behaviors and data patterns differentially predict creditworthiness for men and women, which may not be fully captured by the pooled model.

By training gender-segmented models, we allow all aspects of the XGBoost algorithm to vary. These include *initialization of the base learner* (i.e., a simple prediction for all observations: the log odds of default) as well as the *learning path and aggregation* of weak learners into the ensemble model, some aspects of which are governed by hyperparameters.[15] As such, even when we allow the pooled model to access the gender variable, the learning and aggregation process may look quite different between the gender-segmented and pooled models. Finally, gender segmentation yields gender-specific regularization hyperparameters.

4.2 Model Performance Measures

We use three measures of model performance: AUC, recall, and F1 score. The AUC 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 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. The two additional performance measures that we use, recall and F1 score, do depend on the approval threshold. Recall measures the proportion of actual positive cases 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, where precision measures the proportion of positive predictions that were actually correct, calculated as true positives divided by the sum of true positives and false positives. In this paper, when reporting recall and F1 figures, we assume an approval threshold of a 20% predicted probability of default (i.e., the lender approves anyone with a predicted probability of default at or below 20%), which is consistent with the FinTech lender's target default rate in practice.

[15] Key elements of the iterative learning process include residual errors for each weak learner; the direction in which predictions should be modified to reduce the loss in subsequent learners (gradient descent step); and the learning rate and minimum loss reduction required to make a further partition on a leaf node of the tree.

5 Results

5.1 Pooled Model

In our benchmark model using all of the data sources and features available on the full sample of training data, our out-of-sample AUC estimated on the testing data is 0.752. We refer to this as the pooled model to distinguish it from the model where we segment the data by gender, which we discuss below. Studies in highly data-rich environments such as the US 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, 2023; 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, 2023; Rishabh, 2024). The AUC of our model is at the upper end of those from middle-income settings—even though traditional credit scores are also used as an input in the algorithm in those studies, but not in ours as our sample has no credit bureau score (Table 1).

We next assess the importance of each data source by comparing the AUC of our benchmark model using all of the data sources to that of separate models trained with features from all but one data source (Table 3). The digital footprint user characteristics, which include the same set of features as in Berg, Burg, Gombović, 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 0.709, a reduction in AUC of 0.044 compared to the benchmark model. The transaction-level data from the delivery platform is the second most-important data set, as a model without those data has an AUC of 0.724, a reduction of 0.028. Omitting the no-hit score and limited credit history, mobile phone-based proprietary score, or census tract-level socioeconomic characteristics lead to smaller AUC reductions of 0.016, 0.009, and 0.001, respectively.

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. Indeed, the predictive power of the models is increasing in transaction history: for those in the first quintile with only 2 or fewer transactions the AUC is 0.707, while for those in middle quintile with

6–12 transactions the AUC is 0.742, and for those in the fifth quintile with 27 or more transactions the AUC is 0.777 (Table 4).

5.2 Gender-Segmented Model

We now turn to the gender-segmented models, where we segment the sample by gender prior to estimating the models. Table A.2 shows how the descriptive statistics of our samples vary by gender (as well as by quintile of number of transactions through the delivery app). The men and women in our modeling sample are quite similar on most observable characteristics, with the exceptions that the women are slightly older (25.8 average age compared to 24.3 for men), more likely to use an Apple device (42% for women compared to 34% for men), and have completed slightly more orders through the app (24.9 orders on average compared to 23.0 for men).

Table 5 compares the predictive performance of the pooled and gender-segmented models. When we estimate the gender-segmented models separately for men and women and then calculate the AUC for predictions in the full sample of testing data (both men and women), the gender-segmented model has an AUC of 0.750, which is very close to the AUC of the pooled model of 0.752 (column 1). Similarly, Recall and F1 score are only slightly lower in the gender-segmented model than in the pooled model when calculated for the pooled sample (columns 2 and 3).

When we calculate AUCs of both the pooled model and gender-segmented models separately on men and women, we again find only slight differences. The AUC of the gender-segmented model for predictions on men only is 0.755 compared to an AUC of the pooled model for predictions on men only of 0.757 (column 4). For predictions on women only, the AUC of the gender-segmented model is 0.740 compared to 0.744 for the pooled model (column 7). Table A.3 shows that the relative importance of each data set reported for the pooled model in Table 3 is similar in the gender-segmented models for men only and women only, and Table A.4 shows that the performance of the gender-segmented models is increasing in the "thickness" of the transactions data, as was shown in Table 4 for the pooled model.

We conclude that using gender-segmented models does not lead to a meaningful change in predictive performance of the model (proxied by AUC). Next we turn our attention to the gender-segmented models' implications for the allocation of credit. Figure 1 shows the predicted probability of default for each observation in our testing data under both the gender-segmented models (y-axis) and the pooled model (x-axis). While the predictions of the two models are highly correlated (as evidenced by the mass of points near the 45-degree line), many individuals have substantially different predicted probabilities of default in the two models (as evidence by the points far from the 45-degree line). In addition, we can note that women are more likely to receive different predicted probabilities of default in the two models than men: the female observations in the figure tend to be farther from the 45-degree line. This provides initial evidence that moving from a tra-

ditional pooled model to gender-segmented ones will lead to a larger reallocation of credit among women than men.

To characterize the credit allocation of each model, an approval threshold must be selected. We use a threshold of a 20% predicted probability of default, i.e., we assume the lender approves anyone with a predicted probability of default at or below 20%. This threshold is consistent with our FinTech partner's target default rate. The vertical line in Figure 1 shows this approval threshold for the pooled model and the horizontal line shows this approval threshold for the gender-segmented models. These lines divide the figure into four quadrants, and Table 6 reports the percent of observations in each quadrant, separately for men and women. The lower-left quadrant includes those approved by both models as their predicted probability of default is below 20% in both models; this includes 52.0% of women and 52.0% of men. The upper-right quadrant includes those rejected by both models as their predicted probability of default is above 20% in both models; this includes 40.2% of women and 42.6% of men. The upper-left quadrant includes those *approved* by the pooled model but *rejected* by the gender-segmented models, which includes 2.2% of women and 2.8% of men. Finally, the lower-right quadrant includes those *rejected* by the pooled model but *approved* by the gender-segmented models, i.e., those who benefit from gender-segmenting the machine learning model used to assess risk. This quadrant includes 5.7% of women and 2.6% of men.

In absolute terms, the number of women who would be be approved by the gender-segmented model but rejected by the pooled model is substantially larger (2.6 times larger) than the number of women who would be rejected by the gender-segmented model but approved by the pooled model. Table 6 also reports the percent of women rejected by the gender-segmented model who would be approved by the pooled model, which is 12.3%, and the percent of women rejected by the gender-segmented model who would be approved by the pooled model, which is only 4.0%. The corresponding figures for men are much more similar, at 5.1% and 5.8%, respectively.

Thus, shifting from a traditional pooled machine learning model, which nevertheless has access to the gender variable, to a gender-segmented model increases the allocation of credit to women and improves the equity and fairness of the algorithm's lending decisions. This is because segmenting the models by gender enables the models to capture how gendered behaviors and data patterns can differentially predict creditworthiness for men and women, which may not be fully captured by the pooled model.

This increase in the equity and fairness of credit allocation achieved by the gender-segmented models is achieved without a meaningful change in portfolio default rates. Panel C of Table 6 reports the overall portfolio default rates of both models, which are quite similar at 9.8% for the pooled model and 10.2% for the gender-segmented models.

6 Conclusion

Traditional financial institutions such as banks typically do not lend to borrowers with no 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., 2023). 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 studies 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.

Furthermore, gender gaps in access to credit have persisted despite the automation of creditworthiness evaluations and the entry of many FinTech lenders into the market (IFC, 2024). We argue that this is at least partly due to the way in which credit scoring models are trained and deployed. Even in regulatory environments that allow the credit scoring models to observe and use gender—as recommended in the algorithmic fairness literature, but not allowed in the US—FinTech lenders estimate pooled models that do not fully capture how gendered behaviors and data patterns can differentially predict the creditworthiness of men and women. Intuitively, modeling

default behavior for men and women separately not only allows gender-specific decision trees, but also allows gender-specific learning paths and aggregation into the ensemble learner, which may not be replicated by a pooled model accessing the gender variable.

We show that segmenting the machine learning model by gender—that is, splitting the modeling sample into separate samples of women and men before training the models, and then training separate models on each sample—can improve the equity and fairness of credit allocation without meaningfully impacting the predictive accuracy of the model or the overall default rates of the portfolio. In particular, the gender-segmented model identifies a substantial subset of women who are low-risk, but who would be rejected by a model that pools data on men and women and thus does not fully take into account how various behaviors may differentially predict creditworthiness for men and women.

A limitation of our data is that we only observe repayment and default outcomes for applicants who were approved for credit by our FinTech partner according to a predefined set of rules for applicants with no conventional credit score. That is, there is a substantial portion of applicants whose default behavior is not observed and who may significantly differ from our current sample. In order to address this selection bias, a lender would need to either lend to all applicants initially to obtain data on repayment for both populations and include both in the application scoring model, or infer performance of the rejected applicants (i.e., reject inference). In our case, the sample selection arises from the FinTech's ad hoc approval rules. These could have created meaningful differences in characteristics and behaviors between the approved and rejected samples and potentially dampened gender differences.

We note that this limitation affects most FinTech lenders' models, as well as other studies that use machine learning models to predict creditworthiness, regardless of whether they use traditional or alternative data. Selection bias will be present unless the algorithm is trained on a sample of credit card holders that is representative of the pool of potential applicants. However, in practice models are trained using repayment data only from borrowers who received credit organically, based on the risk appetite of the lenders. One solution to assess this bias, which our lender plans to implement in the future after adopting the models presented in this paper, 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 would enable lenders to identify whether the model they are using to allocate credit is biased in the sense that it fails to capture how behavior and data patterns may differentially predict creditworthiness among the populations that would be approved or rejected by the model.

Our findings suggest two potential policy interventions. First, governments or trade associations could require FinTech lenders to disclose whether algorithms use conventional credit scores

as an input, and what percent of borrowers receiving loans from the FinTech have formal credit histories or conventional credit scores generated by the credit bureau at the time of loan application. This would give consumers a sense of the extent to which FinTech companies are expanding access to credit to those who were previously financially excluded, as is often reported in the media but may not be true in practice.

Second, the algorithmic fairness literature has already suggested that *if combined with appropriate complementary regulations*, equity and fairness can be improved by allowing lenders to use protected variables such as gender and race in their machine learning algorithms (Kleinberg, Ludwig, Mullainathan, and Rambachan, 2018; Kearns and Roth, 2019). Our findings suggest an alternative complementary regulation to those suggested in the algorithmic fairness literature, such as imposing that false negatives are approximately equal across groups (Hardt, Price, and Srebro, 2016). In particular, we show that imposing no additional constraints on the algorithm but instead segmenting the sample by the groups across which inequities exist (e.g., men and women) and estimating separate machine learning models for each group can increase the equity and fairness of credit allocation.

Table 1: Comparison of studies that predict creditworthiness

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
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	0.752
Agarwal, Alok, Ghosh, and Gupta (2023)	India	FinTech loan	63%	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
Berg, Burg, Gombović, 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	0.711
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

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
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
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
Frost et al. (2019)	Argentina	FinTech SME loan	100%	Sales data and internal rating from e-commerce platform, credit score	Logit, XGBoost	0.764
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 for default, tobit for loss rate	0.607
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

Citation	Country	Loan Type	% with Credit Bureau Score	Data	Methods	AUC
Jagtiani and Lemieux (2019)	US	FinTech P2P loan	100%	Personal installment loan-level data from LendingClub's 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
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%	Supermarket's loyalty card data and credit card spending and payment history, sociodemographic data, credit scores	XGBoost	0.679
Meursault, Moulton, Santucci, and Schor (2023)	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%	Payment history data, demographic data, TransUnion credit scores	Random forest, logit	0.70 for bank loans, 0.68 for FinTech loans
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
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; since both are related to the method we use, we report the AUC from the better-performing of these, 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. Berg, Burg, Gombović, and Puri (2020) report in-sample and out-of-sample AUCs; we use their out-of-sample AUCs to be consistent with our study, using the AUC with credit bureau scores, digital footprints, and fixed effects. Björkegren and Grissen (2020) use both RF and logistic regression; we report AUCs from RF as it is closer to the XGBoost method used in our paper. 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 De Cnudde et al. (2019), we report the highest AUC, which is from the ensemble model that uses 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 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 highest AUC, which is from the model with 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 model using all data sources predicting ever-delinquent, which is the highest AUC in the paper. For Meursault, Moulton, Santucci, and Schor (2023), we report the overall XGBoost model, averaged over all years. For Netzer, Lemaire, and Herzenstein (2019) we report AUCs 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; we do not use the model that also incorporates "soft information" because the paper uses loan terms on the loan being applied for as "soft information", but loan terms are a function of predicted default and not available to the lender as an input to the model at the time they are predicting the applicant's default. 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: Modeling Sample: Summary statistics

	Mean	Std. dev.	Median	25th perc.	75th perc.
User age	24.9	8.3	23	20	26
User iOS (Apple) operating system - dummy	0.37				
No-hit score	638.9	20.7	641	631	648
Number of orders on app	23.7	57.7	8	3	22
Proportion orders paid in cash	0.48	0.36	0.47	0.14	0.81
Median amount per order (MXN)	298.3	332.8	247	174	351
Proportion orders at supermarkets	0.05	0.14	0	0	0.03
Proportion orders at pharmacies	0.03	0.10	0	0	0
Proportion orders at food establishments	0.80	0.27	0.93	0.69	1
Marginality (SES) index of census tract	0.96	0.01	0.97	0.96	0.97
Years of schooling among age 15+ in census tract	12.4	1.7	12.4	11.3	13.6
Proportion households own a motor vehicle in census tract	0.64	0.17	0.64	0.52	0.76

This table shows summary statistics for selected variables from various data sources for the sample that we use in our machine learning modeling. Observations are at the user level, and $N = 123{,}042$ users. 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 that takes into account education, housing, public services and income. It takes values between 0 and 1, with 0 representing the highest levels of marginality observed in the cross-section of geographies in a given year, and 1 representing the lowest. Std. dev. = standard deviation; perc. = percentile; iOS = Apple device operating system; MXN = Mexican pesos; SES = socioeconomic status.

Table 3: Marginal contribution of each data source to AUC

Feature set	AUC	Reduction in AUC
All	0.7522	0
All, but digital footprint user characteristics	0.7087	0.0435
All, but transaction-level data from delivery platform	0.7238	0.0284
All, but no-hit score and limited credit history	0.7358	0.0164
All, but mobile phone-based proprietary score	0.7431	0.0091
All, but census tract socioeconomic characteristics	0.7516	0.0006

This table shows the differences in AUCs between a model trained with all features and a separate model trained with features from all but one data source. The results use $N = 123{,}042$ users, split into training data to train the machine learning models and testing data to calculate out-of-sample AUCs. AUC = area under the receiver operating characteristic curve.

Table 4: AUC by quintile of number of transactions through delivery platform

Quintile	Number of transactions	AUC
1	2 or fewer	0.7069
2	2–6	0.7386
3	6–12	0.7419
4	12–27	0.7593
5	27 or more	0.7772

This table shows AUCs 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 AUCs are calculated on the testing data for each quintile. The results use $N = 123{,}042$ users, split into training data to train the machine learning models and testing data to calculate out-of-sample AUCs. AUC = area under the receiver operating characteristic curve.

Table 5: Predictive performance of pooled and gender-segmented models

Model	Full sample			Men only			Women only		
	AUC	Recall	F1	AUC	Recall	F1	AUC	Recall	F1
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooled model	0.7522	0.4781	0.7490	0.7571	0.4894	0.7529	0.7443	0.4597	0.7425
Gender-segmented model	0.7496	0.4770	0.7338	0.7549	0.4875	0.7523	0.7398	0.4588	0.7021

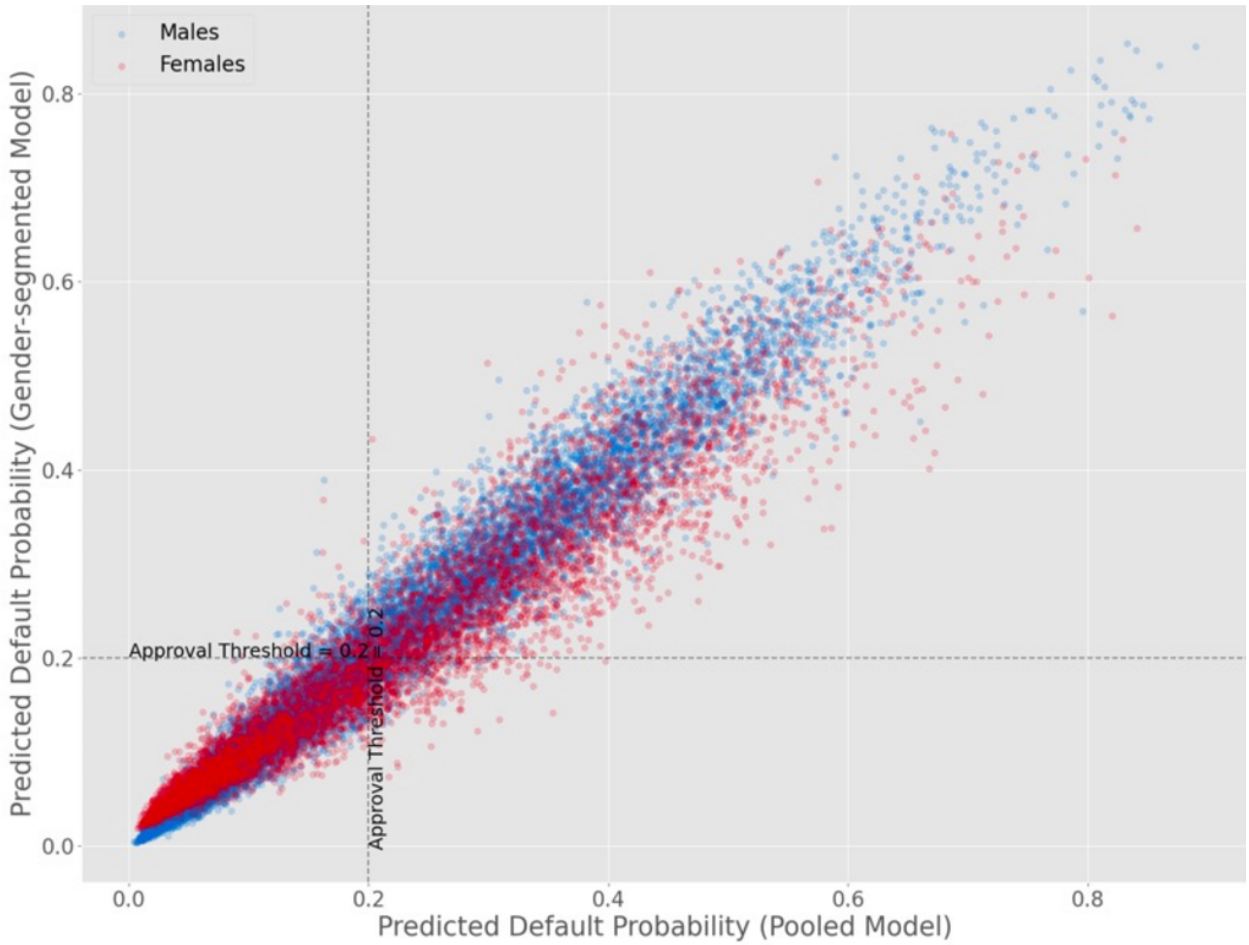
This table shows AUC, recall, and F1 (harmonic mean of precision and recall) for both the pooled and the gender-segmented models, calculated for three samples: the full sample (all observations), men only, and women only. That is, for the gender-segmented model predictions from the men-only and women-only samples, we estimate separate models and use them to make predictions on the segmented testing data. For the gender-segmented predictions on the full sample, we estimate separate models on the segmented training data and use these models to make predictions on the full sample of testing data (including both men and women). For the pooled model predictions on the full sample, we estimate one model on the pooled (men and women) training data and use it to make predictions on the full sample of testing data. For the pooled model predictions on the men-only and women-only samples, we estimate one model on the pooled (men and women) training data and use it to make predictions on the segmented samples of testing data, according to the gender of the applicant. The results use $N = 123{,}042$ users ($N = 76{,}114$ men and $N = 46{,}928$ women), split into training data to train the machine learning models and testing data to calculate out-of-sample measures of model performance. Recall and F1 are calculated with an approval threshold of up to a 20% predicted probability of default; AUC is a threshold-free measure. AUC = area under the receiver operating characteristic curve.

Table 6: Agreements and disagreements between pooled and gender-segmented models

Definition	Proportion
Panel A: Women	
Proportion of all women approved by both models	0.520
Proportion of all women rejected by both models	0.402
Proportion of all women approved by pooled & rejected by gendered	0.022
Proportion of all women rejected by pooled & approved by gendered	0.057
Proportion women approved by pooled model | rejected by gender-segmented model	0.040
Proportion women rejected by pooled model | approved by gender-segmented model	0.123
Panel B: Men	
Proportion of all men approved by both models	0.520
Proportion of all men rejected by both models	0.426
Proportion of all men approved by pooled & rejected by gendered	0.028
Proportion of all men rejected by pooled & approved by gendered	0.026
Proportion men approved by pooled model | rejected by gender-segmented model	0.051
Proportion men rejected by pooled model | approved by gender-segmented model	0.058
Panel C: Default rates of overall portfolio	
Pooled model	0.098
Gender-segmented model	0.102

This table shows agreements and disagreements between the pooled and gender-segmented models, expressed as proportions of either all applicants of that gender (first four rows of Panels A and B), or conditional on being approved or rejected by the gender-segmented model (last two rows of Panels A and B). It also shows the overall portfolio default rates of the two models (Panel C), i.e. the default rate of all applications approved under each model. The results use $N = 123{,}042$ users ($N = 46{,}928$ women and $N = 76{,}114$ men), split into training data to train the machine learning models and testing data to calculate the measures reported in the table out-of-sample. For all of these calculations, we assume that the lender uses a 20% predicted probability of default as its threshold to determine credit allocation in all models, i.e. that the lender approves anyone with up to a 20% predicted probability of default according to the model. The | symbol = conditional on.

Figure 1: Predicted probabilities of default in pooled and gender-segmented models



This figure shows the predicted probabilities of default for each observation in our out-of-sample testing data under both the pooled and gender-segmented models. Red dots represent women and blue dots represent men. The results use $N = 123{,}042$ users ($N = 46{,}928$ women and $N = 76{,}114$ men), split into training data to train the machine learning models and testing data to calculate out-of-sample predicted default probabilities which are shown in the figure. Fixing the approval threshold to 20% predicted probability of default, the lower-left quadrant shows applicants who would be approved by both the pooled and gender-segmented 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 gender-segmented model but approved by the pooled model, and the lower-right quadrant shows applicants who would be rejected by the pooled model but approved by the gender-segmented model. The mass of women in the lower-right quadrant, i.e., those who would be approved by the gender-segmented model but not by the pooled model, is substantially larger than the mass in the upper-left quadrant, i.e., those who would be approved by the pooled model but rejected by the gender-segmented model.

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."

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace (2022). "Consumer-lending discrimination in the FinTech Era." *Journal of Financial Economics* 143(1), 30–56.

Berg, Tobias, Valentin Burg, Ana Gombović, 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). _eprint: https://doi.org/10.1146/annurev-financial-101521-112042, 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.

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*. EC '22: The 23rd ACM Conference on Economics and Computation. Boulder CO USA: ACM, 559–559.

Breiman, Leo (1996). "Bagging predictors." *Mach Learn* 24(2), 123–140.

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

Brock, J. Michelle and Ralph De Haas (2023). "Discriminatory Lending: Evidence from Bankers in the Lab." *American Economic Journal: Applied Economics* 15(2), 31–68.

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.

Burlando, Alfredo, Michael A Kuhn, and Silvia Prina (2023). "Too Fast, Too Furious? Digital Credit Delivery Speed and Repayment Rates."

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.

Castellanos, Sara G., Diego Jiménez Hernández, Aprajit Mahajan, Eduardo Alcaraz Prous, and Enrique Seira (2023). "Contract Terms, Employment Shocks, and Default in Credit Cards."

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*. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 785–794.

CNBV (2019). "Boletin Trimestral de Inclusion Financiera." URL: https://www.gob.mx/cms/uploads/attachment/file/468607/Boletin_IF_2doT_2019.pdf.

CNBV (2023). "Panorama Anual de Inclusion Financiera." URL: https://www.cnbv.gob.mx/Inclusi%C3%B3n/Anexos%20Inclusin%20Financiera/Panorama_2023.pdf.

CONDUSEF (2016). "¿Sabes cuál es la diferencia entre una tarjeta de crédito departamental y una bancaria?" URL: http://www.gob.mx/condusef/articulos/sabes-cual-es-la-diferencia-entre-una-tarjeta-de-credito-departamental-y-una-bancaria?idiom=es.

CRIF (2018). "CRIF desarrolla nuevo Score No Hit para Buró de Crédito en México." URL: https://www.crif.com.mx/noticias/notas-de-medios/2018/abril/crif-desarrolla-nuevo-score-no-hit-para-bur%C3%B3-de-cr%C3%A9dito-en-m%C3%A9xico/.

D'Espallier, Bert, Isabelle Guérin, and Roy Mersland (2011). "Women and Repayment in Microfinance: A Global Analysis." *World Development* 39(5), 758–772.

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.

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

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012). "Fairness through awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12: Innovations in Theoretical Computer Science. Cambridge Massachusetts: ACM, 214–226.

Finnovista (2023). "Fintech Radar Mexico 2023." URL: https://www.finnovista.com/wp-content/uploads/2023/02/Finnovista_Fintech_Radar_MX_23_ENG.pdf.

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). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13090, 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.

Global Banking Alliance (2017). "The Economics of Banking on Women." Financial Alliance for Women. URL: https://financialallianceforwomen.org/download/the-economics-of-banking-on-women/.

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.

Hardt, Moritz, Eric Price, and Nati Srebro (2016). "Equality of Opportunity in Supervised Learning."

Higgins, Sean (n.d.). "Financial Technology Adoption: Network Externalities of Cashless Payments in Mexico." *American Economic Review* ().

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.

IFC (2024). "Her Fintech Edge: Market Insights for Inclusive Growth."

INEGI (2021). "National Survey of Financial Inclusion (ENIF) 2021." URL: https://en.www.inegi.org.mx/programas/enif/2021/.

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." Rochester, NY.

Johnston, Don and Jonathan Morduch (2008). "The Unbanked: Evidence from Indonesia." *The World Bank Economic Review* 22(3). Publisher: Oxford University Press, 517–537.

Kearns, Michael and Aaron Roth (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Illustrated edition. New York: Oxford University Press. 232 pp.

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.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan (2018). "Algorithmic Fairness." *AEA Papers and Proceedings* 108, 22–27.

Lee, Jung Youn, Joonhyuk Yang, and Eric Anderson (2023). "Using Grocery Data for Credit Decisions." Rochester, NY.

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

Mester, Loretta J (1997). "What's the Point of Credit Scoring?"

Meursault, Vitaly, Daniel Moulton, Larry Santucci, and Nathan Schor (2023). "The Time Is Now: Advancing Fairness in Lending Through Machine Learning." Series: Working paper (Federal Reserve Bank of Philadelphia).

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

Montoya, Ana Marıa, Eric Parrado, Alex Solis, and Raimundo Undurraga (2022). "Bad Taste: Gender Discrimination in Consumer Lending."

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.

Of Commerce, Department (2023). "Mexico - Financial Technologies (Fintech) Industry." URL: https://www.trade.gov/country-commercial-guides/mexico-financial-technologies-fintech-industry.

Rishabh, Kumar (2024). "Beyond the Bureau: Interoperable Payment Data for Loan Screening and Monitoring." *SSRN 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. Vol. 9146. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 195–207.

SEGOB (2021). "DOF - Diario Oficial de la Federación." URL: https://www.dof.gob.mx/nota_detalle.php?codigo=5624744&fecha=23/07/2021#gsc.tab=0&gsc.tab=0.

Trecone (2023). "La evolución del sector de delivery en México: un vistazo al pasado, presente y futuro." Trecone. URL: https://trecone.com/la-evolucion-del-sector-de-delivery-en-mexico-un-vistazo-al-pasado-presente-y-futuro/.

World Bank (2021). "Expanding Financial Access for Mexico's Poor and Supporting Economic Sustainability." World Bank. URL: https://www.worldbank.org/en/results/2021/04/09/expanding-financial-access-for-mexico-s-poor-and-supporting-economic-sustainability.

Yin, Jiangning and Nan Li (2022). "Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping." *Ore Geology Reviews* 145, 104916.

Internet Appendix

A Appendix Tables

Table A.1: Search space used in machine learning algorithm

Panel A: XGBoost Classifier	
Evaluation metric	logloss
Tuning	hyperopt, max eval 500
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, 5000,1)
Learning task-specific 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.2: Summary statistics by gender and quintile of number of transactions

	By gender		By quintile of number of transactions through delivery app				
	Men	Women	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
User age	24.3	25.8	25.6	24.7	24.4	24.3	25.1
User iOS (Apple) operating system - dummy	0.34	0.42	0.26	0.32	0.36	0.42	0.53
No-hit score and limited credit history	638.2	640.2	639.1	638.5	638.1	638.3	640.5
Number of orders on app	23.0	24.9	1.1	4.3	9.2	18.6	87.4
Proportion orders paid in cash	0.48	0.48	0.59	0.55	0.48	0.44	0.35
Median amount per order (MXN)	298.9	297.5	347.7	303.2	282.7	274.8	273.9
Proportion spending at supermarkets	0.05	0.06	0.06	0.04	0.05	0.05	0.06
Proportion spending at pharmacies	0.03	0.04	0.03	0.02	0.03	0.05	0.02
Proportion spending at restaurants	0.80	0.81	0.80	0.83	0.81	0.79	0.77
Marginality (SES) index of census tract	0.96	0.96	0.96	0.96	0.96	0.97	0.97
Years of schooling among age15+ in census tract	12.4	12.4	11.8	12.1	12.3	12.6	13.2
Proportion households own a motor vehicle in census tract	0.64	0.64	0.59	0.62	0.64	0.66	0.70

This table shows the mean of selected variables from various data sources for the sample that we use in our machine learning modeling. The mean is calculated separately by gender and by quintile of number of transactions in the delivery app. Observations are at the user level, and $N = 123{,}042$ users ($N = 76{,}114$ men and $N = 46{,}928$ women). 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 that takes into account education, housing, public services and income. It takes values between 0 and 1, with 0 representing the highest levels of marginality observed in the cross-section of geographies in a given year, and 1 representing the lowest. Std. dev. = standard deviation; perc. = percentile; iOS = Apple device operating system; MXN = Mexican pesos; SES = socioeconomic status.

Table A.3: Marginal contribution of each data source to gender-segmented model's AUC

	Men only		Women only	
Feature set	AUC	Reduction in AUC	AUC	Reduction in AUC
All	0.7549	0	0.7398	0
All, but digital footprint user characteristics	0.7086	0.0463	0.7061	0.0337
All, but transaction-level data from delivery platform	0.7283	0.0266	0.7159	0.0239
All, but no-hit score and limited credit history	0.7376	0.0173	0.7269	0.0129
All, but mobile phone-based proprietary score	0.7466	0.0083	0.7347	0.0051
All, but census tract socioeconomic characteristics	0.7545	0.0004	0.7432	–0.0034

This table shows the differences in AUCs between a model trained with all features and a separate model trained with features from all but one data source, for the gender-segmented models. The results use $N = 123{,}042$ users ($N = 76{,}114$ men and $N = 46{,}928$ women), split into training data to train the machine learning models and testing data to calculate out-of-sample AUCs. AUC = area under the receiver operating characteristic curve.

Table A.4: AUC by quintile of number of transactions, gender-segmented model

		AUC	
Quintile	Number of transactions	Men only	Women only
1	2 or fewer	0.7098	0.6898
2	2–6	0.7443	0.7223
3	6–12	0.7361	0.7379
4	12–27	0.7631	0.7335
5	27 or more	0.7731	0.7691

This table shows AUCs for separate models estimated for each quintile of the distribution of number of transactions made through the delivery platform, for the gender-segmented models. Data are split into five quintiles of the full modeling sample for each gender, where quintile cut-offs are based on the full sample (i.e., they do not vary by gender); machine learning models are then trained on the training data for each quintile by gender, and AUCs are calculated on the testing data for each quintile by gender. The results use $N = 123{,}042$ users ($N = 76{,}114$ men and $N = 46{,}928$ women), split into training data to train the machine learning models and testing data to calculate out-of-sample AUCs. AUC = area under the receiver operating characteristic curve.