Women's World Banking

CHECK YOUR BIAS!

A FIELD GUIDE FOR LENDERS

Mehrdad Mirpourian
Jonathan Fu
Sonja Kelly

“

Because there’s nothing micro about a billion women.

Mary Ellen Iskenderian

Author, Advocate, President and CEO of Women’s World Banking

Contents

Acknowledgments

This report would not have been possible without our encouraging and patient colleagues across the industry. In particular, we would like to thank Dr. Annette Krauss and Dr. Mrinal Mishra at University of Zurich's Department of Finance and Banking for their instrumental role in ensuring our work was intelligible to a broad audience. Second, we thank Lendingkart Technologies in India, Banco Anonimo in Mexico, and Aflore in Colombia for partnering with us in this project and for providing us with the insights that drove our inquiries. Our colleagues Pallavi Madhok, Gerardo Pedroza, and Harsha Rodriguez deftly translated data science insights into business realities, contributing a great deal to these ideas. Finally, we are grateful for support from data.org, without which we would not have had the freedom and opportunity to pursue this topic.

Introduction

Credit has always been a fundamental part of both formal and informal financial services. In the last three decades, however, the exponential growth in data and computing power has led to new ways of assessing creditworthiness.

Artificial intelligence (AI) and machine learning (ML) have opened up the possibility of scoring new and alternative data sources, either to complement or to replace more traditional lending methodologies. But how do financial services providers (FSPs) ensure these new systems are both efficient and fair? Amidst the backdrop of a rapidly changing credit landscape, this practical field guide walks executives and data scientists alike through recommendations for ensuring that revised and new credit scoring methods are not unintentionally excluding women.[1]

BOX 1

Project Background

Women's World Banking, in partnership with University of Zurich and with support from data.org, set out in 2021 on a two-year technical assistance project with three financial institutions: Lendingkart in India, Banco Anonimo in Mexico, and Aflore in Colombia. With each institution, we conducted a thorough bias audit across either their credit portfolio or a particular credit product line. While our hypothesis was that we would find bias of some sort in nearly any portfolio, we were surprised by our results. All three institutions were "mostly fair," and the biases that emerged came from side inquiries and additional analyses rather than from glaring or obvious gender differences in the most visible parts of the institutions' portfolios.

1 Throughout the paper, we use "men" and "women" to refer to gender because these are the binary terms that most lenders use. Nevertheless, our findings are applicable to all genders, and we recognize that these issues of bias are just as relevant (perhaps even more relevant) for people who identify as non-binary.

This guide combines academic work on bias detection with practical experience analyzing administrative data from real lenders working to increase financial inclusion around the world. The diversity of institutions this report references (Box 1) offered a natural test for generalizability of a core set of easy-to-understand bias detection questions. Although our focus is on detecting gender biases, the same tools and principles can be applied to bias detection for any underrepresented group.

Detecting bias is not a superfluous exercise. For financial institutions, knowing where bias exists can serve as a way of identifying overlooked markets (as is the case with rejected applicants who are highly creditworthy); maximizing the value of current customers (for example, those who are not receiving sufficiently large loans); or proving alignment with regulatory or legal compliance (in demonstrating the likelihood of a credit offer among men versus women, for instance). For customers, an institution attuned to bias detection is more likely to provide equal opportunities for business growth for men- and women-owned businesses. For regulators or policymakers, bias detection processes that ensure fairness contribute to broader economic participation.

This report has three main sections. The first section is a primer on the fundamental concepts of bias and fairness that anyone working in lending should know. In the second section, for the more technical readers, we discuss the statistical foundations of bias audit. The last section offers three examples of bias detection from three different institution types, as well suggestions on potential bias mitigation interventions specific to the institutions' context. This report is relevant for all lenders, even if most of our examples are from institutions using more automated and digital processes.

Part 1: Fundamentals of Bias and Fairness in Lending

Every member of a product team—from senior strategic decision-makers to more junior technical team members—should be able to speak the language of fairness and bias.

This is far more difficult than it sounds. "Is our credit product fair for women-owned businesses?" is a great question, but it is one that will prompt either an imprecise or an invalid response if there is no clear understanding of fairness and bias.

BOX 2

A Diversity of Credit Processes

The credit assessment process is a set of steps within which a lender evaluates the creditworthiness of credit applicants and measures the perceived credit risk these applicants pose to the institution. Credit risk is the probability of a loss resulting from a borrower who fails to meet their contractual credit obligations. We divide credit assessment processes into three main categories: non-digital, digital, and hybrid.

(continued on the next page)

Non-Digital Credit Process

Under a non-digital credit process (a.k.a. traditional), the credit assessment is highly dependent on human judgment. Relative to other types of credit processes, in non-digital processes the credit officers have the maximum level of power for interventions in credit decisions. A credit officer in this context is the person who assesses creditworthiness, supports credit or risk analysts by providing information, makes a credit decision, collects payments on the loan, and/or provides product or financial education to borrowers. Given this potential range of roles, loan officers play a crucial role in a non-digital credit process model. They often have a close connection to their customers and a strong understanding of an applicant's income, businesses, and social status. This credit model and set-up can be very successful in highly relationship-oriented economies and cultures, and under some specific circumstances. For example, this model may be used in a market with a poor credit reporting infrastructure. On the other hand, non-digital models are highly prone to unconscious biases introduced by loan officers.[2]

Digital Credit Process

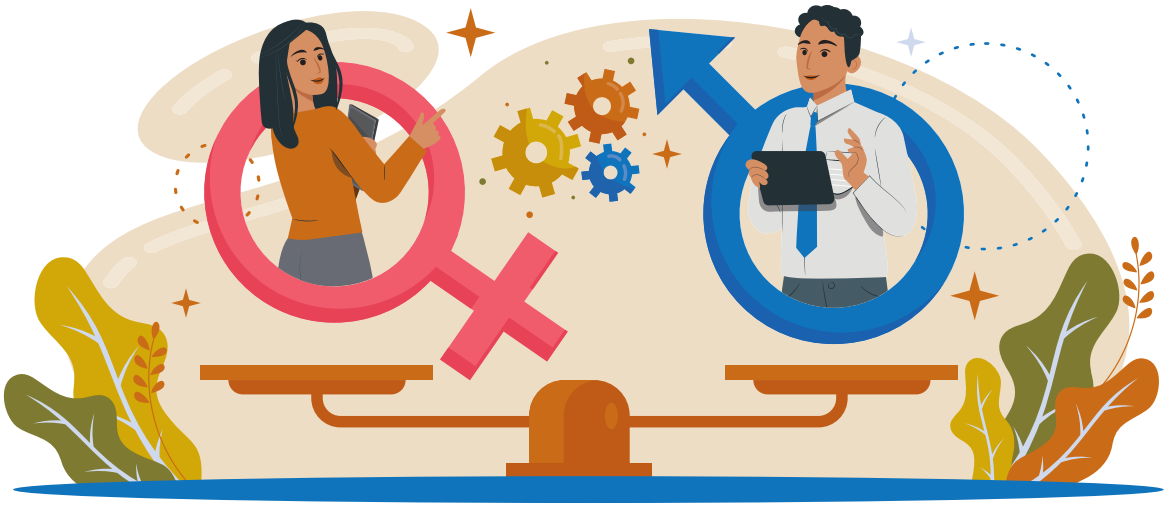
A digital credit process refers to a credit process in which the primary credit assessment activities are digitized. A credit applicant fills out an application form through a personal technology device (phone, computer, tablet, or another device) and submits the application. After submitting the application, the assessment is nearly fully automated and an algorithm or set of predefined coded rules make the credit decision. This category has the least amount of human intervention in credit assessment, and a decision can happen in seconds.

Hybrid Credit Process

Hybrid processes are a combination of both non-digital and digital processes. In this category, credit officers usually collect credit applicants' information, and an algorithm conducts the credit assessment based on the provided information. The credit decision-making might be based on the output of the algorithm, or the algorithm output might be given to a credit analyst who then makes the final credit decision.

2 Women's World Banking (2020). Creating a better banking experience for women-led micro, small, and medium enterprises in Kenya.

Gender Biases in Credit Processes



What is bias, and how is it connected to fairness in a credit assessment? This is the fundamental question at the core of fair lending. In this report, when we talk about bias we are referring to unfair bias or discrimination. Under a discriminatory process, some prioritized groups receive a systematic advantage (being offered credit, for example), and other groups are placed at a systematic disadvantage (being denied credit, for example). Biases can be based on race, color, religion or creed, national origin or ancestry, sex (including gender, pregnancy, sexual orientation, and gender identity), age, physical or mental disability, veteran status, genetic information, citizenship, or other distinguishing factors. Among all these biases, we focus exclusively on gender bias.

Gender-based credit bias happens when a credit process creates results that are systemically prejudiced against certain people, for reasons related to gender. Algorithms created and run by machines can be biased, just as humans can be. Bias stems from a variety of sources. Loan officers can exhibit bias as they assess creditworthiness through interviews and observation. Algorithmic bias sometimes results from conscious or unconscious prejudices introduced by the individuals who create the algorithms—for example data scientists, coders, developers, or others. If they tell an algorithm to pay particular attention to a highly biased variable, the resulting decision might be biased. Other times, data itself can bias the algorithm, such as when a data set represents a sample that is 90 percent men and 10 percent women, and the unbalanced data is used to train the algorithm. There are still other instances in which the construction of an algorithm might prioritize a set of goals that do not include fairness—for example, a highly efficient algorithm may systematically discriminate against a particular group (women, for example) that, on average, is considered less creditworthy than another group.

Examining gender biases in credit-scoring algorithms is an interdisciplinary topic that falls somewhere between data science and finance. The data science community has put most of its efforts into developing techniques for detecting and mitigating bias without emphasizing the impact of these techniques on the accuracy of the unbiased algorithms or on the potential for financial losses. It is the task of financial institutions to determine which biases, when scaled, pose systemic and unnecessary risks to a large population. If an algorithmic approach pushes women deeper into poverty by discriminating on gender at a large scale, it is a systemic and unnecessary risk to a financial system. If a credit process uses income as a proxy for repayment in an environment in which women earn less than men do, it is a justifiable business practice that creates precision in lending and decreases the likelihood of default. Determining what is discrimination and what is good business is an institution-by-institution and market-by-market decision (Kelly & Mirpourian, 2021). This report assumes that in nearly every institution there are avenues for improving fairness that are also good for business. We focus our advice and examples on these areas for improvement.

Assessing and Pursuing Fairness in Credit Processes

Fairness is an intricate and multidimensional concept, and its definition depends on both context and culture. It is impossible to give one specific definition of fairness that applies to all organizations' use cases. For the sake of considering gender-based bias in lending, it is most important for a financial institution to discuss and adopt a definition (or definitions) of fairness and to examine how they balance fairness with efficiency in their credit operations.

What do we mean when we state that fairness is multi-dimensional? Institutions pursue fairness through credit offers (loan approval), credit scores, loan terms (loan amount, interest rate, and collateral), loan maturity, and reasons for rejection. To offer an example of what this might look like in practice, we introduce what is referred to as a "confusion matrix." In a credit assessment process, if someone is predicted to repay the loan and then actually repays it, the result is a "true positive." If that person is predicted to repay the loan but does not repay it, the result is a "false positive," and so on. If women are more likely than men to fall into a "false negative" (creditworthy but denied a loan) than a "true positive" (creditworthy and extended a loan), the decision-making process might be unfair.

Of course, there are other ways of defining and measuring unfairness. While we don't go into the dozens of definitions of fairness here, they are readily available in our recent paper, "Algorithmic Bias, Financial Inclusion, and Gender," in which we summarize Verma and Rubin (2018). What is more important is that an institution develops a definition of fairness that everyone understands. By referring to its own definition of fairness, an institution can benchmark its progress and assess the relative risks of bias.

In looking at a pool of credit applicants, we can separate out two groups for which to assess fairness—those offered credit and those rejected for credit:

For Candidates Offered Credit

In Table 1, candidates offered credit would be concentrated in the first row. Gender fairness might mean that gender does not contribute to building a credit score, or does not correlate with it. Fairness might also mean that men and women receive similar loan amounts, and it ensures that gender does not predict interest rate or collateral requirements.

For Candidates Rejected for Credit

In Table 1, candidates rejected for credit would be in the second row. An institution might be very gender balanced within the candidate pool that is offered credit, but among the rejected candidates there could be a high proportion of women with high credit scores who are rejected. We spend more time on understanding this phenomenon in Part II of this paper, which is focused on "reject inference" techniques.

While we focus our own fairness and bias detection models on these two groups, there are other biases which institutions can assess and address: for instance, bias within the acquisition channel at the top of the data funnel, and bias in renewal processes. Gender bias within the acquisition channel could mean that men have much higher representation among loan applicants. If men account for the majority of the data, the algorithm may over-weight indicators on which men tend to perform strongly (access to and use of a smartphone device, for example). For renewal applicants, we can apply the same bias assessment process used for those candidates who are offered credit, in order to assess whether renewal processes introduce new biases when they prioritize history with the lender.

Table 1. Confusion Matrix Applied to Credit Scoring

	ACTUALLY CREDITWORTHY	ACTUALLY NOT CREDITWORTHY
Predicted to be creditworthy	True Positive (TP)	False Positive (FP)
Predicted to not be creditworthy	False Negative (FN)	True Negative (TN)

BOX 3

Assessing Bias in Non-Digital, Digital, and Hybrid Models

What might be the origins of gender bias, and how do these biases present themselves in different types of institutions? Data bias, unconscious bias by loan officers or algorithm creators, and algorithmic bias might interact with one another, making it difficult to untangle the web of biases and solve bias problems.

When an algorithm is a primary decision-maker in fully digital institutions, the FSP can observe whether some form of algorithmic bias exists. Historic data, unbalanced data sets, design of the algorithm, incorrect interpretation of the algorithm's output, or programming bugs can lead to biased credit decisions. Luckily, there are a wide range of techniques for detecting and mitigating algorithmic biases.

Bias detection and mitigation in hybrid institutions can be challenging. In hybrid models, the algorithm, the credit officers, and the interaction among them all can create gender biases. Bias detection and mitigation in hybrid processes requires careful and close attention to the inputs and outputs of each step in the credit process. The auditor might assess whether the bias appears to be driven more from the credit officer interaction channel (taste-based bias) vs. the algorithmic channel (statistical discrimination). If such a pattern can be found, then the existence of both channels may provide the FSP with realistic opportunities to test whether shifting processes from one channel towards the other will result in the reduction of overall bias.

Loan officers are not immune to unconscious bias, and their credit decisions might be biased and unfair as a result. This type of bias can appear in both non-digital and hybrid credit processes. Given the level of unconscious bias of a loan officer and his/her input in credit decisions, the amount of gender bias in those decisions may vary accordingly. If the model is purely non-digital, we can apply data science to detect gender biases, such as the techniques we propose in Part II. There are a range of solutions to bias in non-digital institutions. For example, loan officer training on gender bias may create a better environment for fairness. Another solution is to limit the influence of the subjective assessment of a loan officer vis-a-vis other more objective data sources. Transforming a non-digital credit process into a more data-driven or hybrid approaches can also mitigate subjective taste-based bias (Vidal & Barbon, 2019).

Part II:
How to Audit for Credit Bias

This section offers two broad approaches lenders can use to audit for credit bias. We assume that prior to assessing bias, an institution has initiated conversations about which dimensions of fairness it considers important, and has established an internal team responsible for bias mitigation.

The first steps in the mainstream approach to mitigation efforts for bias are to: 1) select a normative measure (or measures) for testing for fairness, 2) estimate whether a lack of fairness exists between cohorts of interest based on those measure(s), and then 3) determine whether the economic magnitude of the difference is sizeable enough to warrant mitigation efforts. This typically takes into consideration any protected status classes that an organization or user wants to check for any signs of inequity, bias, or discrimination — in our case, we focus on gender.

Broadly speaking, approaches to "fairness" measurement can be separated into those that promote a) group-level fairness and those that move towards b) individual-level fairness. In this section, we briefly describe these high-level approaches and then provide concrete examples of different analyses lenders can use to test for whether group or individual fairness is being met across a variety of outcomes of interest, e.g., approval rates, loan terms, true and false negative rates, and so forth.

Auditing for Group-Level Fairness

Group-based fairness measurements rest on the logic of treating different groups equally. They essentially aggregate the measurement values for each pair of groups and compare the unconditional outcomes of the classification algorithm for those groups. The group that an individual belongs to is indicated by its sensitive attribute—e.g., in our case, the sample split and comparison is between women vs. men applicants. Over time, many different approaches have been suggested, most of which use metrics based on the binary classification confusion matrix to define fairness.

In our case, we provide a concrete example of how this can be applied in the case of assessing credit underwriting processes for bias through the application of positive and negative balance tests.

Auditing for Gender Bias Using Two-Sample T-Test

Positive and negative balance tests are among the most fundamental bias audit tests, and are used to estimate the significance of the difference between two numbers. A positive (or negative) balance test on credit scores, for example, aims to test whether the approved (or rejected) male and female credit applicants exhibit statistically significant differences in their scores. A positive balance test would apply to the applicants who are approved, while a negative balance test would apply to the applicants who are rejected. To run a positive balance test, the auditor compares the average model score of approved men vs. approved women using a two-sample t-test. The negative balance test uses the same construct among rejected applicants.

This method can be applied to most of the variables we discuss in this section.

Auditing for Individual-Level Fairness

Individual-based fairness measurements rest on the intuition that similar individuals should be treated as similarly as possible. The analyses falling under the individual-based approach move towards trying to demonstrate fairness (or the lack thereof) by identifying as similar observations between comparisons group as possible and demonstrating that they are receiving different treatment despite this. In practice, this is commonly done through highly conditional analyses, including those that use quasi-experimental (or even experimental) approaches.

In our case, we provide several concrete examples of how this can be applied in the case of assessing credit underwriting processes for bias through the application of linear regression, regression discontinuity, and matching methods.

Auditing for Gender Bias Using Linear Regression

Linear regression techniques are an approach for modelling the relationship between a dependent variable and one or more explanatory variables. They allow for setting up conditional analyses that take into account other relevant characteristics of the applicants (e.g., demographics, risk factors, requested loan characteristics, etc.), apart from gender, that may explain differential outcomes. For purposes of bias auditing, they thus move closer towards measuring whether a lender is achieving "individual-level fairness", which is grounded on an underlying intuition that similar individuals should be treated as similarly as possible, if one were to not consider protected status class.

One type of regression is called least squares. For example, to understand if there is a strong relationship between the "female" variable and credit scores, relative to male applicants, an auditor can use least squares regression and regress credit score on gender. As mentioned previously, to make this more precise, data scientists can add in a range of relevant control variables to test for whether the relationship between gender and outcome variables of interest hold after other relevant risk factors are accounted for. The assessment is unconditional if it does not account for control variables, and conditional if it accounts for control variables. A least squares regression shows the magnitude and statistical significance of each variable.

There are other types of regression in which the dependent variables are not all linear (such as whether an applicant was approved for credit, or whether a loan is non-performing). Some of these include logistic regression or probit regression. Again, regression techniques are valid for many of the variables we discuss here.

Auditing for Gender Bias Using Regression Discontinuity

Regression discontinuity design (RDD) is a quasi-experimental design that is typically used to determine the causal effects of interventions. The approach compares outcomes for observations where there are (arbitrary) cutoffs or thresholds above or below which an intervention is assigned. The intuition is that observations lying closely on either side of the thresholds are likely to have much more similar characteristics, which increases the validity of the comparison. In our case, we co-opt the method for purposes of bias assessment by testing whether women and men applicants appear to show different discontinuities (e.g., in terms of approval rates or loan terms) around key credit score or risk thresholds of the lender. Many lenders use some form of credit score—whether external or in-house—to segment borrowers into different risk categories, which have bearing on loan approval and/or loan terms to be provided. For some lenders, we find that these can often be "fuzzy" thresholds (e.g., they provide minimum requirements but do not always equate to automated loan approval) and/or that approval or provided loan terms take into consideration additional human decisions (e.g., in the case of hybrid approaches). In such cases, RDD can provide an intuitive method for assessing bias across a range of outcomes.

Broadly speaking, the approach can be separated into two steps. First, the auditor should identify any cutoff or risk thresholds that are used by the lender for decisions on approval or loan terms. For example, credit scores may be categorized into high, medium, or low risk. This categorization is a subjective exercise that is dependent on the actual process of a given lender. The second step is to audit for gender bias in each category. The bias audit can leverage a range of statistical methods, but the one we suggest here is least squares regression, using an interaction between gender and the cutoff threshold. The goal is to assess whether men and women applicants who are close to the cutoff points—and who would be assumed to be fairly similar to one another—have a different likelihood of approval or other loan terms, such as loan amount, interest rate, or loan tenure.

Table 2 provides a sample template for how an auditor could use RDD to assess the presence of bias by comparing approval rates for men and women who are right above and below a hypothetical minimum approval threshold. In the example, the RDD model is re-run using varying windows around the minimum approval threshold (e.g., +/- 15, 30, 45, and 60 points) to test for robustness of the results. The RDD can also be run conditionally, through the addition of other relevant controls.

Based on the model output reflected in Table 2, the auditor can compare general differences in outcomes between women and men and any further discontinuities for women around the lender's cutoff thresholds. They can also test how sensitive or robust the results are to the choice of windows for selecting the sample and to the presence or absence of control variables.

Auditing for Gender Bias Using Simple Augmentation

Simple augmentation is one of many reject inference techniques. Reject inference models are a group of statistical techniques that aim to impute relevant labels (e.g., default risk) for the observations lacking counterfactual outcomes, such as individuals who were rejected for credit. There are two broad approaches in conducting reject inference: statistics-based reject inference models and machine-learning-based reject inference models.[3] The most common statistics models are based on extrapolation and augmentation. On the other hand, machine-learning models may be based on neural networks, genetic algorithms, and semi-supervised learning. The reject inference techniques we have explored for assessing bias and its magnitude are two statistics-based methods: simple augmentation and coarsened exact matching (which we describe in the next section). We use data on approved applicants to impute predicted values on individuals who were denied credit. This technique allows us to explore likely differences in risk across gender for the rejected applicants, along with their respective rates of true and false negatives. The outcome of a reject inference model may suggest that women could have had higher approval rates, if their rates of false negatives are significantly larger than for otherwise similar men.

3 This is a new area of research. For more information, see Li et al. (2017); Anderson (2019); Banasik & Crook (2005); Bücker, van Kampen, & Krämer (2013).

Table 2. Auditing for Bias Using RDD

	+/- 60 WINDOW		+/- 45 WINDOW		+/- 30 WINDOW		+/- 15 WINDOW	
DV = 1 if Applicant received loan offer	b	p value	b	p value	b	p value	b	p value
Above minimum approval threshold = 1 Female = 1 Above minimum approval threshold = 1 X Female = 1								
Control variables								
Observations								

DV: Dependent variable; b: Regression beta coefficients

To use simple augmentation for auditing selection bias, we first run a logistic regression restricted to approved applicants, where the response variable is the predicted probability of whether the applicant's loan was non-performing, and the independent variables include relevant applicant and requested loan characteristics. After running the logistic regression model on the approved applicants, we store the model coefficients and use them to predict the probability of non-performing loans for rejected applicants with the same characteristics, had they been granted a loan. Our suggested approach is based on using simple augmentation or a hard-cutoff technique, a method which assigns rejected applicants with scores below and above a probability threshold to the "bad" or "good" class, respectively. For example, the credit rating industry often assumes that rejects have a "bad rate" of 75%. This is a subjective evaluation that can be adjusted based on the needs and risk tolerance of the lender. In practice, we adopted a very low risk tolerance and set the cutoff value for the "bad rate" at 0.1. That is, if the predicted probability of a non-performing loan is below 10% for a given rejected individual, then we consider it a good class and a false negative. Conversely, if the predicted probability of Non-performing loan is above 10% for a given rejected individual, then we consider it a bad class and a true negative. We then combine the accepted and rejected applicants into a single data set and do basic descriptive statistics on them.

Tables 3 and 4 show the outputs of this approach. To provide an example of this approach, we filled out these two tables with dummy numbers. In Table 3, we can observe that among approved applicants, the actual non-performing loan rates for the total sample compared with men and with women are 16%, 19%, and 15%, respectively. The predicted non-performing loan rates for the rejected applicants in the total sample compared with the men and women applicants are estimated at 88%, 93%, and 83%, respectively. That is, these are the imputed non-performing loan rates for these cohorts based on the reject inference model, in a hypothetical scenario in which those applicants had been provided loans. On the one hand, as can be expected, the rejected cohorts are indeed predicted to be much riskier. On the other hand, the reject inference models also suggest that a sizeable share of the rejected applicants were likely to have been creditworthy (i.e., they are "false negatives" based on the original model). Moreover, this is particularly the case for women rejected applicants, who are nearly 10 percentage points more likely to be predicted as a "false negative". These latter two points are more explicitly shown in Table 4.

Table 3. Reject Inference on Non-Performing Loans

REJECT INFERENCE - NON PERFORMANCE ISSUES	ALL		MALE		FEMALE		MALE - FEMALE	
	mean	sd	mean	sd	mean	sd	b	p_val
Non-performing loan = 1 (Actual outcomes for approved)	0.16	0.40	0.19	0.40	0.15	0.50	0.04***	0.00
Non-performing loan = 1 (Imputed for rejected applicants)	0.88	0.40	0.93	0.30	0.83	0.40	0.10***	0.00
Non-performing loan = 1 (Combined for actual outcomes for approved + imputed rejected applicants)	0.62	0.49	0.69	0.46	0.54	0.50	0.14***	0.00

Table 4. Reject Inference on True Negatives and False Negatives

REJECT INFERENCE - TRUE NEGATIVES VS. FALSE NEGATIVES	ALL		MALE		FEMALE		MALE - FEMALE	
	mean	sd	mean	sd	mean	sd	b	p_val
True Negative rate	0.86	0.31	0.91	0.20	0.81	0.40	0.10 ***	0.00
False Negative rate	0.11	0.33	0.08	0.26	0.17	0.38	-0.10 ***	0.00

Auditing for Gender Bias Using Coarsened Exact Matching

Matching methods are a non-parametric approach that involve taking observational data, and matching people who have similar characteristics but different treatments, as a way of conducting causal inference or other estimations. The intuitiveness of the approach, in additional to a few other advantageous statistical properties, make it a useful method for reject inference as well.

There are many potential matching strategies to choose from, and each has its relative merits.[4] For our purposes, we apply a recently developed method called coarsened exact matching.[5] In our application of this technique, we start by identifying the approved and rejected applicants in the sample (i.e. we tag these as separate cohorts, but keep the sample intact). We then apply the coarsened exact matching algorithm, which divides the full sample into meaningful strata based on relevant variables of interest. The user must select these variables, taking into consideration the ones that are likely to be strong predictors of loan performance issues and also have data that is generally available (i.e., tends to be complete for all applicants). In practice, any categorical variables are subjected to exact matching and any continuous variables are split into bins.[6] For each stratum, the coarsened exact matching process then identifies any matches between the approved and rejected cohorts.

For our purposes, we use only strata that have matched approved and rejected applications. All non-matched observations/strata are ignored. For each stratum, where matches are found, we use the observed loan performance of the matched

4 Giving an extensive review of matching strategies is beyond the scope of this guide. However, Stuart (2010) provides a useful reference and synthesis of the related literature.
5 For further details, see: https://gking.harvard.edu/cem.
6 The user can either denote specific binning rules for each variable or can automate this based on commonly applied decision rules for binning, e.g., Sturge's rule, Scott's rule, Freedman-Diaconis's rule, etc.

approved applications in a given stratum to impute predicted loan performance among the matched rejected applications. For example, we construct a variable, which we call "any NPL imputed" and set it to 1 for a given matched rejected application if the matched approved application in their strata had an NPL and set it to 0 if they did not. In the simplest scenario, if there is only one matched approved/disbursed application in the stratum, then its loan performance is used to impute the matched rejected applications' performance. In the more common scenario where there are multiple approved/disbursed loans in the stratum, we use the mean loan performance. We set a loosely risk-averse threshold: If the actual probability of having an NPL is greater than or equal to 50% among the approved applications in the stratum, then we set "any NPL imputed" equal to 1 for the matched rejected applications and 0 otherwise. (Note that the threshold can be adjusted based on the risk tolerance of the lender.) Under this setup, we define:

a. "False negatives" – rejected applicants where "any NPL imputed" = 0
b. "True negatives" – rejected applicants where "any NPL imputed" = 1

We then combine the accepted and rejected applications into a single data set and do basic descriptive statistics on them. The outputs of this model would be similar to what we showed in Table 3 and Table 4.

Women's World Banking Gender Bias Scorecard for Lenders

There are many ways to measure and track fairness in lending. This scorecard offers some of the more common fairness indicators used by Women's World Banking in its diagnostic process with lenders. Paying attention to these indicators, particularly over time, will offer institutions an evidence base so they can identify areas in which they excel and areas for improvement. Since not all credit processes are the same, the list of fairness-related questions that an institution will ask might vary, and we hope institutions will interpret and adapt these questions to their needs. The six dimensions of fairness in Women's World Banking's Gender Bias Scorecard are: 1) credit score, 2) approval rate, 3) loan amount, 4) interest rate, 5) collateral size, and 6) characteristics of rejected candidates. The complete descriptions of these dimensions and how to measure them are reflected below. To use this scorecard you will need individual-level data on past loan applicants including credit score, decision, loan terms, and any relevant control variables.

HOW TO CHECK YOUR CREDIT PROCESS BIAS

1 ESTABLISH A CLEAR DEFINITION OF FAIRNESS

For example, women applicants have an equal probability relative to men applicants with similar risk characteristics of being approved for credit, being rejected for credit, and receiving the same credit terms.

2 ESTABLISH RELEVANT CONTROL VARIABLES

Using a list of the most influential variables in your approval process, choose 3-4 variables which should be indicative of creditworthiness regardless of gender.

a. *For example, income may be a strategically relevant variable that, regardless of gender of the applicant, maintains high explanatory power in creditworthiness. This is a valid control variable.*

b. *GPS location or number of Facebook friends may be low priority variables that are highly subject to gender bias but with low strategic importance to the approval process. These are invalid control variables.*

The right mix of control variables will vary by institution and business model.

3 EVALUATE FOR BIAS

Using gender-disaggregated data, answer the questions on the next page to assess bias. The word "average" denotes the arithmetic mean. For more advanced analysis, separate applicants into subcategories by risk quantile, or apply the scorecard to separate steps in the lending process, for example, automated systems compared against loan officer assessments.

4 ASSESS THE MAGNITUDE OF BIAS

Starting with the questions outlined here, make an assessment of the magnitude—and resulting financial and consumer impacts—of each bias dimension on your portfolio. How much does this bias cost your company? How many people does this bias exclude or disadvantage unnecessarily?

Women's World Banking Gender Bias Scorecard for Lenders

	QUESTION	ANSWER *Your institution's responses go here!*	EXAMPLE *Sample responses from the United Bank of Banking*
1	Do men and women borrowers have the same average credit score to indicate creditworthiness while controlling for relevant variables? Yes/no?		*Yes*
	If not, which gender has a higher credit score?		*n/a*
	What is the magnitude of this gap?		*n/a*
2	Do men and women applicants have the same likelihood of receiving a credit offer, controlling for relevant variables? Yes/no?		*No*
	If not, which gender has a higher likelihood of receiving a credit offer?		*Men*
	What is the magnitude of this gap?		*Men are 65% likely, women are 55% likely.*
3	Do men and women who are extended credit offers receive the same average loan amount while controlling for relevant variables? Yes/No		*No*
	If not, which gender on average has a higher loan amount?		*Men*
	What is the magnitude of this gap?		*Men's average is $3500 and women's average is $2900.*
4	Do men and women who are extended credit offers receive the same average interest rate while controlling for relevant variables? Yes/No		*Yes*
	If not, which gender has a lower interest rate?		*n/a - we use a fixed interest rate.*
	What is the magnitude of this gap?		*n/a*
5	Do men and women who are extended credit offers have the same average collateral requirement while controlling for relevant variables? Yes/No		*No*
	If not, which gender has a lower collateral requirement?		*Men*
	What is the magnitude of this gap?		*Men's collateral requirement is $400 or equivalent, women's is $500 or equivalent.*
6	Do men and women rejected applicants have the same average credit score while controlling for relevant variables? Yes/No		*No*
	If not, which gender has a lower credit score?		*Men*
	What is the magnitude of this gap?		*Rejected women received a 350 on average, rejected men receive 300.*

HOW TO SCORE

Count the number of "yes" answers to the headline questions above and list this answer here.

RESULTING SCORE	SAMPLE SCORE
	2/6

WHAT YOUR SCORE MEANS

1-2	Highly biased on multiple dimensions
3-4	Moderately biased on multiple dimensions
5-6	Little to no bias on multiple dimensions

Check your bias!

Additional Diagnostic Assessments

To supplement the previously described analyses and scorecard, institutions can also employ a wide range of additional diagnostic assessments, as described below. These can give them a more comprehensive view of potential (gender) bias across their full operations. For example:

APPLICATION CHANNEL

How do the above dimensions of fairness differ depending on the application channel?

REJECTION REASONS

Do reasons for rejection differ by gender?

FIRST-TIME VERSUS REPEAT APPLICANTS

How do the above dimensions of fairness differ for first-time compared with repeat applicants?

SEGMENTATION ANALYSIS

If there are highly sensitive intersectionalities with gender, how do the above dimensions of fairness vary between men and women? E.g., Socio-economic strata, formal or informal employment, etc.

CREDIT APPROVAL RATE OVER TIME

How has the approval rate of men and women applicants changed over time compared with the balance of men and women applicants? Is the portfolio trending towards fairness?

CREDIT DENIAL RATE OVER TIME

How has the denial rate of men and women applicants changed over time compared with the balance of men and women applicants? Is the portfolio trending towards fairness?

EX-POST CREDIT PERFORMANCE

How does credit performance differ by gender, and what does this indicate about the accuracy of the model for men and for women?

PORTFOLIO GENDER BALANCE

What percentage of approved loans belongs to men and what percentage belongs to women applicants? How is the gender balance in the portfolio being used to predict the creditworthiness of future applicants?

GAP BETWEEN REQUESTED AND EXTENDED CREDIT

What is the difference between the approved credit line and requested credit amount for each gender? Do these differ?

GENDER BALANCE AT LEAD ACQUISITION STAGE

Are there any lead sources for which significant gender imbalances exist, and what is the magnitude of this difference?

Taking all of the above steps would help an auditor to acquire useful information on credit biases in a portfolio.

Part III:
Bias Detection and Mitigation Examples in India, Mexico, and Colombia

Having compiled this range of tools and methods for bias detection, Women's World Banking and its partner University of Zurich collaborated with three institutions that represent diverse approaches in their use of digitized or hybrid processes.

Lendingkart is primarily digital, Banco Anonimo employs a hybrid approach, and Aflore also uses a hybrid process that leverages more person-to-person engagement relative to the others.[7]

With each institution, our process was straightforward:

1. Understand the steps, both human and data-driven, in the lending process.
2. Conduct a bias assessment.
3. Estimate the magnitude of identified bias(es), from a statistical, ethical, and financial perspective.
4. Recommend bias mitigation strategies.
5. Work with the institution to implement these mitigation strategies where possible.

Important to the understanding of bias is a calculation of the risk it poses to the institution. Risks to reputation—along with regulatory, ethical, and missed market opportunity risks—are all factors that contribute to a financial institution's decision-making process in considering whether to invest in bias mitigation. While we do not go into the risk calculation and decision-making process each institution undertook, we note this as a valid step any financial services provider must take. Not all biases pose a high risk to a financial institution. Similarly, bias mitigation is a highly contextualized and specific process. For some institutions, mitigating bias may mean balancing data. For others, bias mitigation may involve algorithmic revisions. For still others, mitigating bias could mean training staff or loan officers on gender sensitivity in data collection. Rather than offer a set of standard bias mitigation techniques, we offer real-life examples based on our own recommendations.

7 "Banco Anonimo" is a pseudonym for a large bank in Mexico we partnered with on this project. The institution prefers to remain anonymous.

India: Lendingkart

Lendingkart Technologies Private Limited is a fintech startup providing working capital to SMEs in India. The company has developed credit assessment tools that leverage big and alternative data analysis. Legally, Lendingkart is a non-deposit-taking non-bank financial company (NBFC). The company is transforming small-business lending by making credit access for SMEs easier. Unlike banks and other NBFCs, Lendingkart does not focus on applicants' past financial statements or income tax returns to evaluate their creditworthiness. Instead, its algorithm relies on thousands of data points ranging from cash flows to business growth.

Lendingkart's credit assessment process is nearly all digital (Figure 1). The company has developed its own credit assessment procedure using a machine learning algorithm. The majority of Lendingkart's applicants are men, largely because of the gendered nature of the SME landscape in India. Lendingkart has intentionally sought to increase the share of women in its lending process, in part through the partnership with Women's World Banking.

Lendingkart's credit process is fully automated, with nearly no human engagement. An evaluation found the company's credit process to meet traditional definitions of fairness on nearly all metrics (Table 5). An unconditional investigation of the credit portfolio detected a statistically significant—but not substantively large—approval bias against women.[7] This bias was entirely due to policy parameters and disappeared when conditioned on policy. For example, Lendingkart calls its applicants to verify business ownership claims. A larger proportion of women compared to men show false claims of business ownership, such as a woman applying on behalf of a male relative or household member who may have already accessed credit. Beyond bias in the approval rate, we could not detect any gender bias in other areas of Lendingkart's portfolio.

Figure 1. Lendingkart's Credit Assessment Process

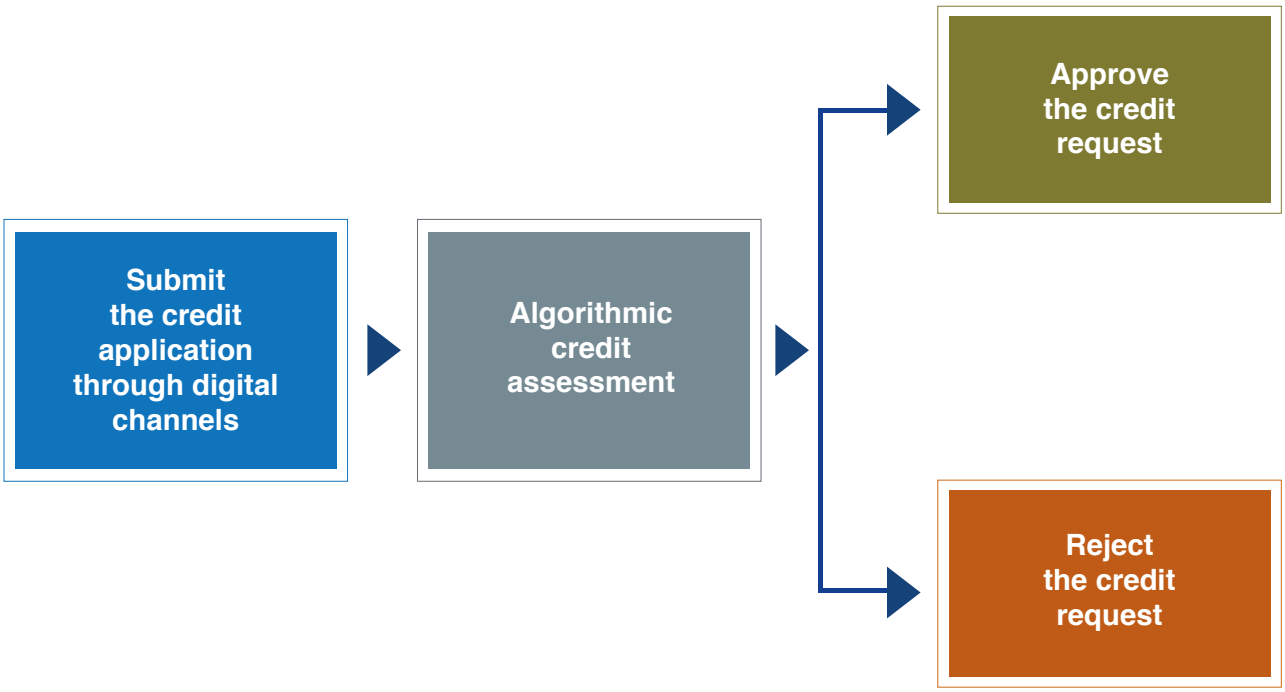


Table 5: Lendingkart's Scorecard Results

NO	QUESTION	ANSWER
1.	Do men and women applicants have the same average credit score to indicate creditworthiness, while controlling for relevant variables? Yes/No?	Yes
1.1	If not, which gender has a higher credit score?	n/a
1.2	What is the magnitude of this gap?	n/a
2	Do men and women applicants have the same likelihood of receiving a credit offer, while controlling for relevant variables? Yes/No?	Yes
2.1	If not, which gender has a higher likelihood of receiving a credit offer?	n/a
2.2	What is the magnitude of this gap?	n/a
3	Do men and women who are extended credit offers receive the same average loan amount, while controlling for relevant variables? Yes/No	Yes
3.1	If not, which gender on average has a higher loan amount?	n/a
3.2	What is the magnitude of this gap?	n/a
4	Do men and women who are extended credit offers receive the same average interest rate, while controlling for relevant variables? Yes/No	Yes
4.1	If not, which gender has a lower interest rate?	n/a
4.2	How much is the magnitude of this gap in percentages?	n/a
5	Do men and women who are extended credit offers have the same average collateral requirement, while controlling for relevant variables? Yes/No	Yes
5.1	If not, which gender has a lower collateral requirement?	n/a
5.2	How much is the magnitude of this gap in percentages?	n/a
6	Do men and women rejected applicants have the same average credit score, while controlling for relevant variables? Yes/No	Yes
6.1	If not, which gender has a lower credit score?	n/a
6.2	What is the magnitude of this gap?	n/a

8 Statistically significant with 99% confidence.

After a thorough review of Lendingkart's overall credit assessment process, a comprehensive data audit, and an estimation of the relative risk of the identified bias, our joint conclusion was that the company's greatest concern was not approval of women applicants but rather gender imbalance in its clientele. This gender imbalance stems from issues with lead acquisition at the top of the data funnel. Having significantly more men than women applying creates the risk that "representation bias" — a well-documented concern in machine learning — could creep into Lendingkart's models over time. The team, composed of a range of departments and employees with varying levels of seniority, developed a broad array of strategies aimed at increasing acquisition and retention of women applicants across the lender's operations, and decreasing the considerable gender imbalance in the company's clientele. As part of this effort, together we developed and launched a pilot that employs a gender-differentiated product and marketing strategy to target and increase take-up among women leads.

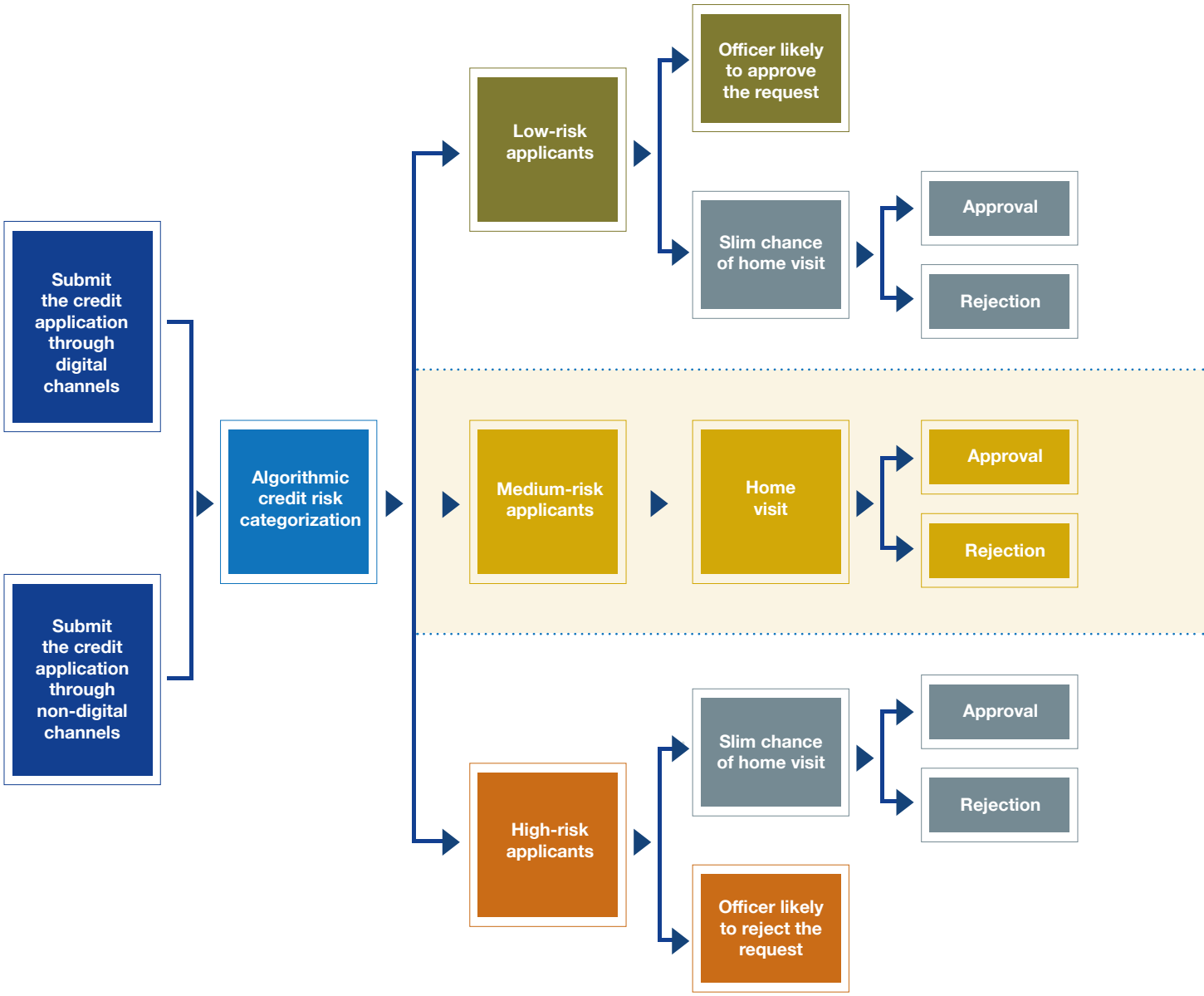
Mexico: Banco Anonimo

Banco Anonimo is a large bank operating in Mexico with a physical branch infrastructure and online presence. The company's products include consumer credit for goods, personal loans, small business loans, credit cards, mortgages, and payroll systems. In Mexico, Banco Anonimo has thousands of bank branches and has millions of outstanding loans. We partnered primarily with Banco Anonimo's small-business loan team focused on the Mexican market.

Banco Anonimo's credit assessment process is a hybrid model. The company accepts credit applications through both digital and non-digital (i.e. physical branches) channels. Regardless of the application channel, applications go through a similar assessment process. Banco Anonimo uses an algorithm for categorizing its customers into one of the three risk buckets: low risk, medium risk, and high risk. Applicants categorized as low risk usually receive credit approval without the necessity of going through any further investigation. Medium-risk applicants often receive home visits or video calls to add information on their sources of income and upcoming expenses. With the risk categorization and these additional pieces of information, a loan officer makes the final credit decision. Applicants with high credit risk are likely to be rejected, with a portion being offered the opportunity to go through an interview process, after which they sometimes receive a loan. These high-risk applicants often need a guarantor or collateral, provided they get loan disbursement approval.

9 There was an initial gap between men's and women's likelihood of receiving a credit offer. We investigated the rejection reason codes that explain this differential. We found that, after removing those applications where rejection is based on non-negotiable and legitimate policy rules, this gap no longer remained.

Figure 2. Banco Anonimo's Credit Assessment Process



A gender bias audit found some statistically significant differences between men and women on a few metrics. These statistically significant differences, while indicating some unfairness in the process, did not have a high economic impact and were also not always biased against women. For example, rates of loan approvals proved to be higher for women than for men with otherwise similar characteristics. Moreover, loan amounts for women were significantly larger than for men in statistical terms, albeit the economic magnitude was not large — around 120 pesos or about US $6. Looking at rejected applicants, however, revealed statistically significant and substantively important gender bias against rejected women applicants. This bias was more profound among women applicants who had received home visits and had gone through extra credit assessment verifications, indicating some additional unconscious bias. See Table 6 for details.

Table 6: Banco Anonimo's Scorecard Results

NO	QUESTION	ANSWER
1	Do men and women applicants have the same average credit score to indicate creditworthiness, while controlling for relevant variables? Yes/No?	No
1.1	If not, which gender has a higher credit score?	Women
1.2	What is the magnitude of this gap?	7.2*** (Approved women have a 561; Approved men have a 554. The score ranges from 468 to 665, with a standard deviation of ~25).
2	Do men and women applicants have the same likelihood of receiving a credit offer, while controlling for relevant variables? Yes/No?	No
2.1	If not, which gender has a higher likelihood of receiving a credit offer?	Women
2.2	What is the magnitude of this gap?	~1.5 percentage points***
3	Do men and women who are extended credit offers receive the same average loan amount, while controlling for relevant variables? Yes/No	No
3.1	If not, which gender on average has a higher loan amount?	Women
3.2	What is the magnitude of this gap?	~120 Mexican pesos ($6 USD)** Statistically significant at the 5% level, but not economically significant.
4	Do men and women who are extended credit offers receive the same average interest rate, while controlling for relevant variables? Yes/No	Yes
4.1	If not, which gender has a lower interest rate?	n/a
4.2	How much is the magnitude of this gap in percentages?	n/a
5	Do men and women who are extended credit offers have the same average collateral requirement, while controlling for relevant variables? Yes/No	Yes
5.1	If not, which gender has a lower collateral requirement?	n/a
5.2	How much is the magnitude of this gap in percentages?	n/a
6	Do men and women rejected applicants have the same average credit score, while controlling for relevant variables? Yes/No	No
6.1	If not, which gender has a lower credit score?	Men
6.2	What is the magnitude of this gap?	6.5*** (Rejected women have a 545.35; Rejected men have a 538.87. The score ranges from 468 to 665, with a standard deviation of ~25).

Note: *As in many presentations of inferential statistics, asterisks denote the level of confidence in the statistical significance of the finding. For example, *** means p<0.001 or 99.9% confident, ** means p<0.01 or 99% confident, * means p<0.05 or 95% confident.*

In other words, our audit (particularly the reject inference analysis) revealed that women applicants have a significantly higher likelihood of being "false negatives," i.e., a disproportionate share are being rejected despite having a higher propensity to repay than comparable male applicants. Moreover, we observed that the partner's credit assessment process exhibits some signs of subjective biases at stages where credit officers are heavily engaged in decision-making.

As such, we recommended two different pilots aimed at mitigating both problems, and we engaged in discussions with the partner to integrate them into their processes. The first was an experiment to reduce the disproportionately high "false negative" rate for women applicants and extend more loan access to such borrowers. Specifically, we would utilize "reject inference" techniques to identify applicants who were marginally rejected, but likely to have had no repayment issues if they had been granted a loan. A subset of such marginally rejected applicants would be randomly selected to either be approved (overturning the original decision) or to remain rejected. We would use the experiment to gauge how accurately the methods work in selecting safe marginally rejected applicants, and then we would consider a broader roll-out of the process.

The second recommendation was an experiment to test whether greater provision of hard-coded information could mitigate the high potential for subjective "taste-based" discrimination in the partner's credit assessment process, which is currently reliant on credit officers to make the final decision for riskier borrowers. The partner's status quo operations use credit scoring models to sort applicants into different assessment channels, but they do not currently provide the actual scores to credit officers when making assessments. As such, we planned an intervention in which some credit officers would be (randomly selected to be) "primed" with the model scores instead of continuing with the original process. While both of these recommendations were superseded by other organizational priorities, they may be relevant to other institutions with similar biases.

Colombia: Aflore

Aflore is a fintech company based in Colombia, serving low-income customers who operate primarily in the informal sector. Aflore provides its customers with loans and insurance products, and maintains a tech-enabled direct sales channel to bring simple financial products into the household via people the community trusts. Building on social networks, Aflore has more than 14,000 informal advisors engaging directly with its customers.

Aflore uses a hybrid credit model and is heavily dependent on its network of informal advisors. These informal advisors are not employed by Aflore; however, they advocate for Aflore in their communities and bring new credit applicants to the top of the application channel. The informal advisors receive a commission for recommending new applicants, and they consider the opportunity to serve the community as an important motivation, aside from the financial gains. The advisors enjoy seeing their community thrive and helping people get out of the informal credit markets. Aflore has a digital credit application process through which informal advisors help prospective borrowers apply. Before the Covid-19 outbreak, Aflore had been using its own credit-scoring algorithm. However, during the pandemic, Aflore found that the accuracy of its model had deteriorated. Therefore, Aflore decided to retire this model and instead started to rely on credit scores from formal credit-rating agencies. Only two thirds of applicants who apply for credit at Aflore have a credit history with credit bureau agencies. Aflore passes these applicants through a set of screening checks. If the applicants pass the initial screening, a loan officer will look at the credit score, income, estimated expenses, and other financial and socioeconomic indicators before making a decision. However, assessing the creditworthiness of the remaining one third of applicants who have no credit history (also known as thin-file applicants) can be more challenging. This involves a more subjective assessment by loan officers, who conduct interviews with applicants and often require a physical or virtual visit at the applicant's home or business.

Looking into Aflore's credit portfolio from January 2016 - June 2022, we did not detect much substantive gender bias against approved women applicants. However, we learned that there is a significant bias against rejected women applicants (Table 7).

Figure 3. Aflore's Credit Assessment Process

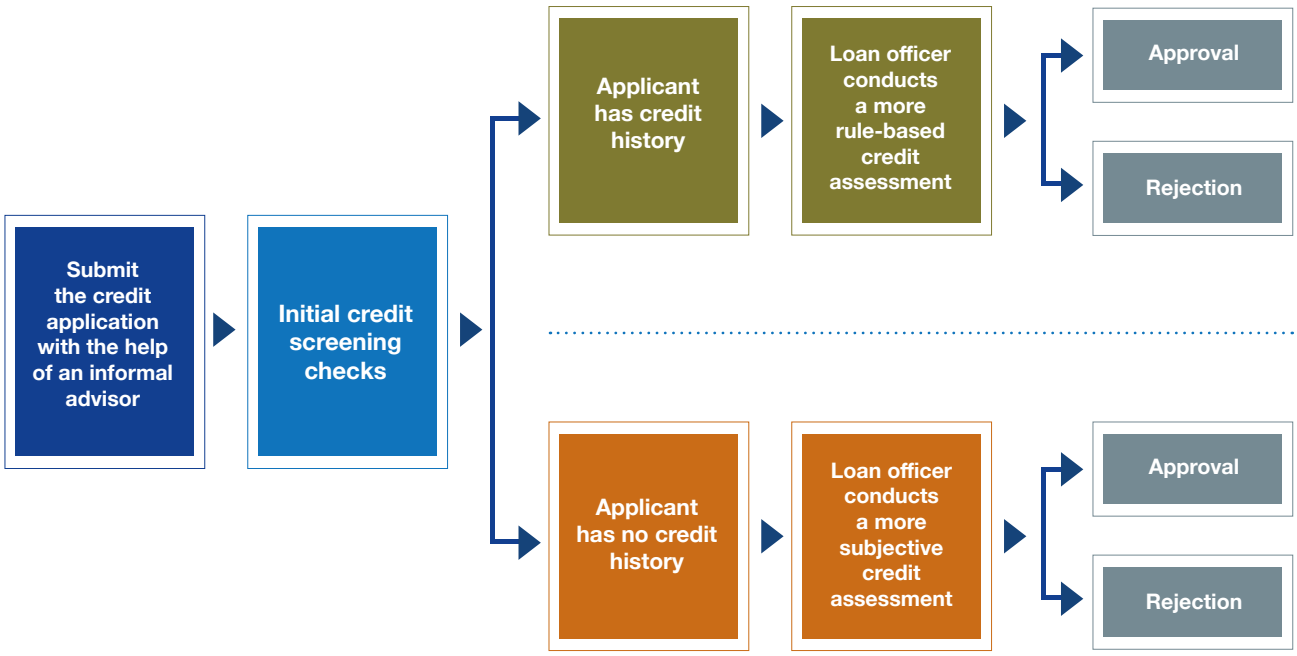


Table 7: Aflore's Scorecard Results

NO	QUESTION	ANSWER
1	Do men and women applicants have the same average credit score to indicate creditworthiness, while controlling for relevant variables? Yes/No?	n/a±
1.1	If not, which gender has a higher credit score?	n/a
1.2	What is the magnitude of this gap?	n/a
2	Do men and women applicants have the same likelihood of receiving a credit offer, while controlling for relevant variables? Yes/No?	No
2.1	If not, which gender has a higher likelihood of receiving a credit offer?	Women
2.2	What is the magnitude of this gap?	0.3-0.8 percentage points
3	Do men and women who are extended credit offers receive the same average loan amount, while controlling for relevant variables? Yes/No	Yes
3.1	If not, which gender on average has a higher loan amount?	n/a
3.2	What is the magnitude of this gap?	n/a
4	Do men and women who are extended credit offers receive the same average interest rate, while controlling for relevant variables? Yes/No	Yes
4.1	If not, which gender has a lower interest rate?	n/a
4.2	How much is the magnitude of this gap in percentages?	n/a
5	Do men and women who are extended credit offers have the same average collateral requirement, while controlling for relevant variables? Yes/No	Yes
5.1	If not, which gender has a lower collateral requirement?	n/a
5.2	How much is the magnitude of this gap in percentages?	n/a
6	Do men and women rejected applicants have the same average credit score, while controlling for relevant variables? Yes/No	n/a±
6.1	If not, which gender has a lower credit score?	n/a
6.2	What is the magnitude of this gap?	n/a

*Note: As in many presentations of inferential statistics, asterisks denote the level of confidence in the statistical significance of the finding. For example, *** means p<0.001 or 99.9% confident, ** means p<0.01 or 99% confident, * means p<0.05 or 95% confident.*

± Aflore does not currently use an in-house credit score for its credit assessment, and only collects a 3rd-party credit score for a subset of its applicants. The reject inference model we used to impute the share of rejected applicants expected to have an non-performing loan is based on other factors. Consequently, this metric is less applicable in their case.

For Aflore, our audit similarly revealed that their women applicants have a significantly higher likelihood of being "false negatives," i.e. a disproportionate share are being rejected despite having a higher propensity to repay than comparable male applicants. Our main bias mitigation strategy also revolves around using reject inference methods to complement their regular process. We would apply reject inference techniques to identify "marginal" applicants rejected by their status quo process but predicted by the reject inference models to be creditworthy. We would experimentally offer loans to a subset of these marginal applicants to test if the approach was accurately identifying "false negatives," before a potential large-scale roll-out. The team developed the algorithm to identify these false negative applicants and is working to make it open-source for any institution to use.

Conclusion: Why Bias Detection Matters

In this paper, we started by emphasizing that fairness is a complex goal, one that cannot be achieved through any singular pathway. Nevertheless, there are some clear first steps that institutions can take to understand where their biases exist and the extent to which these biases get in the way of their goals.

We offered a menu of bias detection tools, both basic and advanced, that financial services providers can use in their journeys towards fairness. We ended with three case examples of financial services providers that care about measuring potential biases in their portfolio and understanding the relative impact of these biases.

Bias in lending—whether algorithmic or human—matters, but not always for the reasons we assume it does. Perhaps most relevant to financial services providers is the business case. If gender biases in a credit assessment process are a result of inefficiencies, they cost a financial institution money. In the portfolio of every institution with which we partnered, a proportion of rejected applicants would have been likely to have paid their loans if they had been extended credit. Financial services providers can decrease their non-performing loans by paying attention to bias in their portfolios.

A second reason to care about bias is the perception in the market. Highly biased institutions may be seen to be unfair at best, and discriminatory at worst. Perception is important to institutions because it affects the loyalty of their existing customers, their ability to acquire new customers, and their access to investment capital.

Further, as we see in many markets across the world, there is likely to be increased regulation of innovation in lending. Europe's new regulation of algorithms and related data-protection measures, alongside the U.S. government's request for information about machine-learning-based models, indicate that regulations surrounding fairness are

imminent for the financial sector. The most forward-thinking institutions will take the opportunity to "future-proof" their processes to ensure they can assess and improve their ability to be fair.

Finally, bias detection and mitigation matters for financial inclusion. New data and new technology combined create the conditions for faster, more efficient financial services. It is the responsibility of all of us to ensure that these services, rather than creating new lines of exclusion, will present new opportunities to increase women's financial inclusion.

References

Anderson, B. (2019). Using Bayesian networks to perform reject inference. *Expert Systems with Applications*, 137. Retrieved from https://doi.org/10.1016/j.eswa.2019.07.011

Banasik, J., & Crook, J. (2005). Credit scoring, augmentation and lean models. *Journal of the Operational Research Society*, 56(9), 1072-1081. Retrieved from https://doi.org/10.1057/palgrave.jors.2602017

Bücker, M., Van Kampen, M., & Krämer, W. (2013). Reject inference in consumer credit scoring with nonignorable missing data. *Journal of Banking and Finance*, 37(3), 1040-1045. Retrieved from https://EconPapers.repec.org/RePEc:eee:jbfina:v:37:y:2013:i:3:p:1040-1045

Kelly, S., & Mirpourian, M. (2021). *Algorithmic bias, financial inclusion, and gender: A primer on opening up new credit to women in emerging economies*. Women's World Banking. Retrieved from https://www.womensworldbanking.org/wp-content/uploads/2021/02/2021_Algorithmic_Bias_Report.pdf

Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 75(15), 105-114. Retrieved from https://doi.org/10.1016/j.eswa.2017.01.011

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Shen, F., Zhao, X., & Kou, G. (2020). Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory. *Decision Support Systems*, 137. https://doi.org/10.1016/j.dss.2020.113366

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1), 1.

Tian, Y., Yong, Z., & Luo, J. (2018). A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines. *Applied Soft Computing Journal*, 73, 96–105.

Verma, S., & Rubin, J. (2018, May 29). *Fairness definitions explained* [Paper presentation, pp. 1-7]. IEEE/ACM International Workshop on Software Fairness (Fairware), Gothenburg, Sweden. https://fairware.cs.umass.edu/

Vidal, M. F., & Barbon, F. (2019). *Credit scoring in financial inclusion*. Retrieved from https://www.cgap.org/research publication/credit-scoring-financial-inclusion

Women's World Banking. (2020). *Creating a better banking experience for women-led micro, small, and medium enterprises in Kenya*. Retrieved from https://www.womensworldbanking.org/wp-content/uploads/2020/11/MSME_Report_2020.pdf

Women's World Banking

@womensworldsbnkg

www.womensworldbanking.org