Fair Lending Monitorship of Upstart Network's Lending Model

Second Report of the Independent Monitor

PUBLIC

Pursuant to agreement by the NAACP Legal Defense and Educational Fund, the Student Borrower Protection Center, and Upstart Network, Inc.

November 10, 2021

Relman Colfax PLLC
1225 19th St. N.W., Suite 600
Washington D.C. 20036
(202) 728-1888

Table of Contents

Executive Summary 3
A. Fair Lending Overview 6
B. Overview of Upstart's Model Process 8
C. Scope of the Fair Lending Analysis 10
D. Estimating Protected Class Attributes 11
E. Disparate Impact Analysis 12
1. Disparate Impact Step 1 13
a. Disparate Impact Step 1—Assessing Disparities 13
b. Disparate Impact Step 1—Statistical and Practical Significance 14
c. Disparate Impact Step 1—Results 17
2. Disparate Impact Step 2—Legitimate Business Need 18
3. Disparate Impact Step 3—Identifying Less Discriminatory Alternatives 18
a. Disparate Impact Step 3—Technical Methodology 19
b. Disparate Impact Step 3—Choosing Among Potential Alternatives 20
F. Proxy Review 22
1. Variable Background 22
2. Qualitative Variable Review 23
3. Quantitative Variable Analysis 24
a. Quantitative Variable Overview 24
b. Surrogate Modeling 26

Executive Summary

This is the second report of the independent fair lending Monitor regarding Upstart Network's ("Upstart") lending Model. On April 14, 2021, the Monitor issued its Initial Report, which provides a summary of legal principles and fair lending testing, and a descriptive history of the events leading up to the Monitorship.[1] This Second Report provides further detail regarding the methodology and fair lending tests conducted to date.

Upstart is a lending platform that relies on Artificial Intelligence and Machine Learning ("AI/ML") models that incorporate non-traditional applicant data—including data related to borrowers' higher education—to underwrite and price consumer loans. The NAACP Legal Defense Fund ("LDF") is an organization dedicated to furthering racial justice and the Student Borrower Protection Center ("SBPC") is focused on protecting the rights of student borrowers.

In 2020, LDF and the SBPC raised concerns with Upstart that the use of educational criteria can lead to discriminatory lending outcomes, particularly for communities of color. Upstart, LDF, and the SBPC ultimately agreed to appoint Relman Colfax, PLLC, as an independent fair lending Monitor to evaluate and make recommendations regarding certain fair lending implications of Upstart's lending Model specifically related to whether less discriminatory alternatives can be implemented that maintain model accuracy, and to issue a series of reports on its findings and recommendations. This Report does not make any legal conclusions about whether Upstart is in compliance with antidiscrimination law, and this Monitorship does not address other fair lending, fair housing, or civil rights issues related to Upstart—for example, we did not engage in fair lending analyses of marketing, servicing, or other practices.

This Report outlines our methodologies, findings, and progress on testing to date. First, we are conducting a disparate impact and alternatives analysis. This assessment involves ongoing analyses of whether Upstart's Model causes an adverse impact on any protected classes and, if so, whether there are less discriminatory alternative practices that maintain the Model's predictiveness.

- We have identified what we refer to as statistically and practically significant approval disparities for Black applicants as compared to non-Hispanic white applicants. This finding does not, standing alone, demonstrate a fair lending violation. These disparities were measured on an unadjusted basis, *i.e.*, without attempting to control for legitimate creditworthiness criteria. Under our methodology, however, such disparities do trigger an obligation to investigate whether viable less discriminatory alternative models exist.

[1] Relman Colfax, "Initial Report of the Independent Monitor," Fair Lending Monitorship of Upstart Network's Lending Model Pursuant to Agreement by the NAACP Legal Defense and Educational Fund, the Student Borrower Protection Center, and Upstart Network, Inc. (April 14, 2021) ("Monitor's Initial Report"), https://www.relmanlaw.com/media/cases/1088_Upstart%20Initial%20Report%20-%20Final.pdf.

- While we have made significant progress testing and validating potential model alternatives, we are not yet in a position to determine whether a viable alternative model exists and, if so, what changes to Upstart's Model we would recommend. This search is well underway, and any resulting recommendations will be included in future reports.

Second, we are analyzing whether variables in Upstart's Model function as close proxies for protected classes. This proxy question is distinct from our disparate impact assessment. Commonly, if a model contains proxies for protected classes, it is considered a *disparate treatment* issue. Disparate impact and disparate treatment are separate risks. A model can raise disparate impact risks absent the inclusion of proxy-variables. It is also true that proxy-variables may exist in a model even if the model does not cause a disproportionate adverse impact on a protected class. For our proxy-related analyses to date, we prioritized race and national origin because of the concerns giving rise to this Monitorship and the disparities noted above.

- These proxy-related analyses suggest that Upstart's input variables, standing alone, do not appear to be meaningful predictors of race and national origin.
- At the same time, given the AI/ML nature of Upstart's Model, there are inherent limitations to our proxy-related methodologies. For example, we cannot assess whether interaction variables within the AI/ML Model function such that they could be proxies. Because of these limitations, we cannot eliminate the possibility that proxies exist.
- For that reason, we suggest that Upstart continue to weigh this risk and the feasibility of adopting more interpretable model structures alongside the perceived benefits of its AI/ML Model. Those benefits might relate to model performance, as well as the potential that the flexibility of an AI/ML structure may permit improvements on other fairness metrics, such as disparate impact—a possibility we will explore in subsequent reports. Upstart represents that it already performs this type of risk/benefit analysis, which we do not question.

This Report explains the methods we are using to identify whether less discriminatory alternatives exist, and it outlines the conditions under which we would recommend adoption of a less discriminatory alternative model. As explained below, the complexities of AI/ML models require more advanced methodologies for fair lending testing than what might be effective for traditional models. This Report describes the specific methodologies we use in our role as Monitor. At a minimum, in any effective fair lending analysis of a model we would expect some parity between the sophistication of the techniques an entity uses for modeling and those used for fair lending testing those models. Future reports will address our specific disparate impact and alternatives findings, including whether we will make any recommendations to Upstart regarding the adoption of less discriminatory alternative models. We do not make any formal recommendations in this Report.

In our capacity as Monitor, we have engaged Sentrana to serve as a consultant to assist with these analyses. Sentrana is a leading firm in the field of machine learning and artificial intelligence applied to credit risk modeling, pricing, optimization, fraud detection, anti-money laundering, and compliance analytics. We have also engaged Dr. Bernard Siskin, of BLDS, to serve as a statistical consultant. Dr. Siskin is an expert on the use of statistical analyses to measure discrimination in the financial services industry. Unless otherwise noted, this Report uses the term Monitor to refer to the collective contributions of Relman Colfax, Sentrana, and Dr. Siskin.

This Monitorship is intended both to assess Upstart's Model and to contribute to the ongoing dialogue about the growing use of AI and alternative data, including ensuring that such use is consistent with antidiscrimination law and equitable access to credit. Upstart, LDF, the SBPC, and the Monitor continue to share the view that lenders should take steps to avoid the unnecessary perpetuation of discrimination, segregation, and inequity; to date, the testing and analyses conducted pursuant to this Monitorship have been productive.

A. Fair Lending Overview

As discussed in more detail in our Initial Report, antidiscrimination laws such as the Equal Credit Opportunity Act ("ECOA") and the Fair Housing Act ("FHA") prohibit entities in credit markets from discriminating on the basis of certain protected characteristics. ECOA makes it unlawful for a creditor to discriminate against an applicant in "any aspect of a credit transaction" on the basis of protected characteristics such as race, color, religion, national origin, sex, age, or receipt of income from a public assistance program.[2] In addition to ECOA, residential real-estate related loans are also subject to the FHA, which prohibits discrimination based on race, color, national origin, religion, sex, disability, or familial status (meaning the presence in the household of a child under the age of 18).[3] The FHA covers second mortgages, home equity lines of credit, home improvement loans, and refinance loans, in addition to standard purchase loans.

Both ECOA and the FHA prohibit explicit differential treatment or intentional discrimination (known as "disparate treatment"), as well as more subtle forms of discrimination that may occur without any intent to discriminate (known as "disparate impact").[4] Although disparate treatment discrimination often involves animus or the specific intent to harm or disadvantage members of the protected group, such animus is not a required element of a discrimination claim.[5] Under disparate impact, a policy or practice that is neutral on its face but disproportionately disadvantages a protected class in a material way is illegal if it either does not serve a legitimate business interest, or the legitimate interest can be served in some alternative way that results in less disadvantage to the protected class. There are three steps involved in determining whether a policy has an unlawful disparate impact:

> *Step 1*: The first step is to determine whether the policy or practice disproportionately disadvantages a protected class in a material way.
>
> *Step 2*: If the policy or practice does have a disproportionate disadvantage on a protected class, the next step is to determine whether the policy serves a legitimate business need.

---

[2] 15 U.S.C. § 1691(a); 12 C.F.R. § 1002.2(z). In addition to the listed protected classes, ECOA protects against discrimination based on the good faith exercise of any right under the Consumer Credit Protection Act. 15 U.S.C. § 1691(a)(3). Although sexual orientation and gender identity are not listed as separate classes under ECOA, they are protected. *See* CFPB Interpretive Rule, Equal Credit Opportunity (Regulation B): Discrimination on the Bases of Sexual Orientation and Gender Identity, 86 Fed. Reg. 14363 (Mar. 16, 2021).

[3] 42 U.S.C. § 3605.

[4] *See, e.g., Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Communities Project*, 576 U.S. 519, 539 (2015) (holding that disparate impact claims are cognizable under the FHA); 12 C.F.R. § 1002.6(a) (ECOA regulatory codification of disparate impact); Official Staff Commentary, 12 C.F.R. pt. 1002, Supp. I, 6(a)-2 (explaining that Congress intended to apply the "effects test" to credit discrimination); HUD, DOJ, OCC, OTS, Fed. Rsrv. Bd., FDIC, FHFB, FTC, NCUA, Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18266 (Apr. 15, 1994) ("Joint Policy Statement on Lending Discrimination"); *Barrett v. H&R Block, Inc.*, 652 F. Supp. 2d 104, 108 (D. Mass. 2009) (collecting cases holding that disparate impact is cognizable under ECOA).

[5] *See, e.g., Goodman v. Lukens Steel Co.*, 482 U.S. 656, 668–69 (1987) (explaining that animus is not required for intentional discrimination under Title VII and Section 1981), *superseded on other grounds by statute*, 28 U.S.C. § 1658(a); *Curto v. A Country Place Condominium Assoc.*, 921 F.3d 405, 410 (3d Cir. 2019) (similar for FHA claim).

In the credit context, for example, identifying applicants likely to repay the loan for which they have applied is a legitimate business need. If the policy or practice does not serve a legitimate business need, it is illegal, and the inquiry ends.

Step 3: If the policy or practice does serve a legitimate need, the third and final step is to determine whether there is a reasonable alternative policy or practice that would serve the same end while reducing the disproportionate impact on protected class members. If there is a less discriminatory alternative that satisfies the legitimate business justification, it is unlawful to use the original policy or practice rather than adopting the alternative.

With limited explicit exceptions, it is also a violation of the ECOA and FHA prohibitions against overt, intentional discrimination (*i.e.*, disparate treatment) to use a protected class as a variable in a credit scoring or pricing model.[6] This is equally true for a variable that functions as a close proxy for a protected class.[7]

For years, lenders have been aware that these principles, including disparate impact, apply to their lending- and housing-related activities, and that federal regulatory and enforcement agencies may apply disparate impact analyses in their examinations and investigations under both the FHA and ECOA.[8] Accordingly, many lenders routinely test their models for fair lending risks and make corresponding changes if necessary. While the agencies charged with implementing these laws and regulating financial institutions have not mandated precise methodologies for fair lending testing credit models, many lenders have well-established systems for doing so.

The analyses of Upstart's Model conducted pursuant to this Monitorship are consistent with commonly-used methodologies, and are designed to align with traditional principles relied on in antidiscrimination jurisprudence.[9] At the same time, the complexities of AI/ML models require more advanced methodologies for fair lending testing than what might be effective for traditional models. For example, because of nonlinearities and cross-variable effects that occur in AI/ML models, variables in isolation may be innocuous, but in combination may drive unnecessary disparate impact.[10] Similarly, effectively identifying less discriminatory alternative

---

[6] *See, e.g.,* Monitor's Initial Report, *supra* note 1, at 8; 12 C.F.R. pt. 1002, Supp. I, ¶ 1002.2(p)-4 ("Besides age, no other prohibited basis may be used as a variable."); FFIEC, "Interagency Fair Lending Examination Procedures" at 8 (Aug. 2009) (explaining that "overt discrimination" includes using "variables in a credit scoring system that constitute a basis or factor prohibited by Regulation B or, for residential loan scoring systems, the FHAct"), https://www.ffiec.gov/pdf/fairlend.pdf.

[7] *See, e.g.,* Monitor's Initial Report, *supra* note 1, at 8.

[8] *See id.* at 7.

[9] *See id.* at 7–12.

[10] This observation is provided as an example; we have not at this time made a determination that this occurs in Upstart's Model. "Nonlinearities" refers to a property of AI/ML models in which a variable can have an effect that is not directly linear but can be complex or arbitrary (a linear relationship, in contrast, is one where there is a direct relationship between an independent variable and a dependent variable, such that there would be a straight line if plotted on a graph). The nonlinear impact of a variable can be exponentially large or very small; the precise relationship between a variable and its impact is typically not known *a priori*, although it is deduced by an AI/ML model during training. "Cross-variable effects" refers to the tendency of AI/ML models to form arbitrary

models often requires the use of more sophisticated methods than are adequate for traditional models, such as optimized searches of various permutations of variable combinations. Those searches, for example, can help identify collections of variables that could be modified or dropped entirely from a model without resulting in meaningful losses in model performance. The methods used here are described in more detail below.

B. Overview of Upstart's Model Process

This section provides an overview of Upstart's application and model process as background for describing the specific testing conducted.

To begin an application on the Upstart platform, a consumer applies for a loan through a consumer-facing website. The applicant is asked for information such as the purpose of the loan, the requested loan amount, and identifying information such as name, birthdate, and address, sufficient for a soft inquiry into the applicant's credit report.

Bank Partner Criteria: Once a loan inquiry is submitted and Upstart has received an applicant's credit report information, Upstart applies minimum eligibility requirements imposed by its bank and financial institution partners.[11] These requirements vary by partner, but include criteria such as whether the applicant's identity is verifiable, and whether the applicant is of a minimum age sufficient to contract, is a permanent U.S. resident, meets a minimum credit score, and has had any bankruptcies within some number of months. An applicant must be eligible under the requirements of at least one bank partner to qualify for a loan; if they are not, they are not offered a loan.

Upstart Model: If an applicant would be eligible for a loan from at least one bank partner under those bank partner criteria, they are assessed according to a process that we refer to as "Upstart's Model." At a high level, that process proceeds as follows:

- Stage 1—AI/ML model output: First, Upstart utilizes its AI/ML Model, which predicts default and prepayment probabilities (*i.e.*, risk) for each borrower. This AI/ML Model relies on a combination of credit bureau data, applicant-provided information, and other information captured at the time of application. We call the output of the AI/ML Model "Stage 1."[12]

combinations of so-called "interaction" variables during model training. Typically, these combinations act as new variables and have effects that are distinct from any of their progenitor variables considered individually.

[11] Initially, all Upstart loans were originated by Cross River Bank, with the underlying loans being sold to third party institutional investors or made available for investment to accredited investors. Now, Upstart partners with a number of other banks and financial institutions to provide underwriting and backend services for applications coming through Upstart's referral network or via bank partners' own web portals. *See* Monitor's Initial Report, *supra* note 1, at 18.

[12] Upstart also makes other adjustments to its Model to account for things like economic cycles.

- Stage 2—loan pricing: Second, the outcomes of Stage 1 are fed into a loan pricing engine that generates a borrower-specific Annual Percentage Rate ("APR") that would be offered to each applicant.

- Stage 3—offer generation: The outcome of Stage 2 is then fed into a loan approval engine, which makes final approval and denial determinations based on the recommended APR. If the APR from Stage 2 is within the boundaries specified by a lender, the borrower is presented with an offer of credit. If the APR is above the maximum APR specified by every eligible lender, the applicant is declined.

As discussed in the Initial Report, a perceived risk central to this Monitorship is the concern that the use of certain information related to higher education may contribute to discriminatory outcomes that disproportionately affect communities of color, including students who attend minority-serving institutions, such as Historically Black Colleges and Universities ("HBCUs") and Hispanic Serving Institutions ("HSIs").[13] Upstart's Stage 1 AI/ML Model considers information related to higher education. However, in response to conversations with LDF and the SBPC, as well as a congressional inquiry, Upstart made certain changes to how its Model utilizes educational data. Most notably, it eliminated the use of average incoming SAT and ACT scores to group education institutions in its Model. Instead, Upstart's Model groups schools based on median post-graduation income.[14] Upstart also established a "normalization" process for Minority Serving Institutions ("MSIs")—which Upstart defines as schools where 80% or more of the student body are members of the same minority racial demographic group.[15] Under that process, Upstart normalized MSIs as a group to have equal graduate incomes to non-MSIs by calculating and using the distance, as a percentage, between a school's graduate incomes and its respective school group average (*i.e.*, MSIs, non-MSIs). According to Upstart, this process results in MSIs and non-MSIs being on average equal. Put another way, above average MSIs (in terms of graduate income) are treated above average overall by as much as they are above the MSI average. Any decisioning by Upstart's Model is then performed on this normalized information.[16]

---

[13] *See* Monitor's Initial Report, *supra* note 1, at 17.

[14] *See id.* at 23.

[15] *See id.* at 24. Upstart's definition is not the same as federal definitions of "minority-serving institution." Seven categories of MSIs are defined in federal law. *See* 20 U.S.C. § 1067q(a). One is HBCUs (referred to as "part B institutions"), which are historically Black colleges or universities established prior to 1964, whose "principal mission was, and is, the education of Black Americans" and that are properly accredited. 20 U.S.C. § 1061(2). Another category is Hispanic-serving institutions, which are eligible institutions with enrollments that are at least 25% Hispanic students. 20 U.S.C. § 1101a(a)(5). Other categories include: Tribal College or University, Alaska Native- or Native Hawaiian-serving institution, Predominantly Black Institution, Asian American and Native American Pacific Islander-serving institution, and Native American-serving nontribal institution. 20 U.S.C. § 1067q(a). According to Upstart, it used its MSI definition because the relatively low thresholds for certain categories under the statutory definitions would have resulted in the inclusion of too many institutions being normalized; for example, many of the University of California schools would qualify as HSIs and/or AANPIs. Upstart represents that over 400 schools qualify as MSIs under its definition.

[16] As noted in our Initial Report, Upstart voluntarily adopted these changes; Upstart reports that none of its internal fair lending tests—which are reported to the CFPB—have identified unlawful discrimination against any protected class, including any racial group.

The fair lending analyses performed as part of this Monitorship, including the quantitative proxy analysis discussed below, are of Upstart's Model following adoption of these changes. However, we did not attempt to test the fair lending-related effects of these changes—in other words, we do not know whether our fair lending test results to date would have been different prior to these changes.

C. Scope of the Fair Lending Analysis

The focus of this Monitorship is on Upstart's Model, including whether it causes an adverse impact on any protected class and, if so, whether viable less discriminatory alternatives exist that maintain the Model's predictiveness. Accordingly, we focused on fair lending testing Upstart's Model. We did not engage in other fair lending analysis, for example of marketing, servicing, or other practices.

Moreover, the data used for our fair lending analyses was limited to the pool of applicants that, if approved under Upstart's Model, would have been eligible for a loan from at least one bank partner. In other words, applicants that would not be eligible for a loan under the criteria of at least one bank partner were excluded from the fair lending analyses under the rationale that those applicants were not assessed or excluded by the Upstart Model that forms the basis of this Monitorship.[17]

Although not part of the fair lending testing conducted for this Report, we did consider the effects that bank partner criteria generally might have on the demographic characteristics of Upstart's applicant pool by comparing the protected class characteristics of all applicants for the first quarter of 2021 to the protected class characteristics of just those applicants that would have been eligible for a loan from at least one bank partner. These comparisons show very similar percentages by protected classes for both pools of applicants. We did not assess and do not comment on these bank partner criteria, other than to note that they did not appear to disproportionately adversely affect members of any protected class to a significant degree for the first quarter of 2021.

[17] This methodological decision is based on the scope of this Monitorship. We do not intend to make any suggestions regarding the potential scope of liability, legal responsibilities, or the like.

D. Estimating Protected Class Attributes

Any statistical fair lending analysis requires an awareness of consumers' likely protected class status.[18] Outside the mortgage context, creditors usually do not have information about the race, national origin, or sex of consumers, and therefore estimation techniques are required. Here, Upstart provided flags identifying protected class status of consumers in the training dataset. These flags are the same flags that Upstart uses for its own fair lending assessments, and for the fair lending assessments provided to the CFPB as part of Upstart's No Action Letter Model Risk Assessment Plan.[19]

Each consumer in the dataset is determined to be 62 or older, or less than age 62, based on the date of birth provided at the date of application. ECOA generally prohibits discrimination on the basis of "age," and because a central purpose of the law is to protect older persons, this provision is often applied to prohibit discrimination against older consumers.[20] Age 62 is chosen as a cut off because Regulation B defines the term "elderly" to include all persons age 62 or older.[21] Sex is estimated based on first name, using Social Security Administration probability tables and an 80% cutoff for classification.

Race and national origin are estimated using a proportional method of Bayesian Improved Surname Geocoding ("BISG"), a method that relies on a combination of surnames and geography.[22] First, this method estimates the probabilities that an individual is a member of each of six race and national origin categories—Hispanic, African American, Asian American/Pacific Islander, American Indian/Alaskan Native, non-Hispanic white, and multiracial. Second, using the proportional approach, an individual's data are associated with multiple race/national origin categories, in proportion to the probability an individual belongs to each category. For example, an individual whose BISG probabilities were 50% for African American, 50% for non-Hispanic white, and 0% for all other categories would be included in the analysis as both African American and non-Hispanic white, but with only a 50% weight in each case. This proportional method is in contrast to a classification method, where an individual is categorized as belonging to a particular group if the BISG probability they belong to that group meets some minimum threshold, such as 80%. Finally, disparate impact testing is not performed with respect to the American Indian/Alaskan Native or multiracial categories because BISG estimates for these groups are not sufficiently reliable.[23]

---

[18] Monitor's Initial Report, *supra* note 1, at 7–8.
[19] *See* CFPB, Letter Response to 2020 NAL Request at 1 (Nov. 30, 2020), https://files.consumerfinance.gov/f/documents/cfpb_upstart-network-inc_no-action-letter_2020-11.pdf.
[20] *See* NCLC, "Credit Discrimination," § 3.4.2 ("Age") ("In practice, however, the ECOA mainly prohibits discrimination against older consumers.").
[21] *See* 12 C.F.R. § 1002.2(o).
[22] *See* CFPB, "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity" (Summer 2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.
[23] *See* Marc. N. Elliot, et al., "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Servs. and Outcomes Rsch. Methodology*, 9:69–83 (Apr. 10, 2009).

Taken together, our testing comparisons are broken down as follows:

Table 1: Control and Test Groups

Protected Class	Control Group	Test Groups
Age	<62	≥62
Gender	Male	Female
Race	Non-Hispanic White	African American, Asian/Pacific Islander, Hispanic

E. Disparate Impact Analysis

A primary focus of this Monitorship is testing for disparate impact risk, following the three disparate impact steps:

(1) Does the Model cause an adverse impact on a protected class?
(2) Does the Model serve a legitimate business need?
(3) If the Model causes an adverse impact on a protected class, does a less discriminatory alternative exist that continues to serve the legitimate business need?

A disparate impact review typically starts after a model has been developed that has been optimized for performance and designed consistent with applicable model risk management principles, such as ensuring appropriate use, accuracy, robustness, and the like.[24] This is referred to as the "Baseline Model." Here, the Parties agreed that the Monitor would begin its fair lending analysis on an accurate facsimile of Upstart's Baseline Model for Stage 1. The facsimile model was designed by Sentrana to mimic as closely as possible the Upstart Stage 1 AI/ML Model. We refer to this as the "Facsimile Baseline Model." Using the Facsimile Baseline Model protects Upstart's proprietary information during the Monitorship. To have confidence in potential alternative models, the Facsimile Baseline Model must faithfully reflect the Stage 1 AI/ML Model. Accordingly, we performed continuous quality control tests to ensure the Facsimile Baseline Model is robust. The remaining stages of the Upstart Model process are assessed via a software interface provided by Upstart that generates outputs based on the inputs provided by the Stage 1 predictions. These results are validated in Upstart's own environment to confirm their accuracy.

[24] Monitor's Initial Report, *supra* note 1, at 9. "Robustness" in the AI/ML context generally means the stability of accuracy of a model when it is deployed for use. AI/ML models are trained on a specific subset of data and are then used to make predictions on new data in the real world. A robust model, therefore, is one where the training process

1. Disparate Impact Step 1

a. *Disparate Impact Step 1—Assessing Disparities*

In the first step of the disparate impact analysis, predicted outcomes of the Upstart Model are reviewed to assess whether the Model is likely to cause any material adverse impacts on any protected class. At a high level, this is done by assessing whether each tested protected class disproportionately ends up with negative outcomes as compared to a control class.

As noted in the Initial Report, two common metrics for assessing disparities at Step 1 of the disparate impact analysis are the adverse impact ratio ("AIR") and standardized mean difference ("SMD").[25] The AIR "is equal to the ratio of the proportion of the protected class that receives a favorable outcome and the proportion of the control class that receives a favorable outcome."[26] AIR is commonly used in various antidiscrimination scenarios such as financial services and employment, and is appropriate for models generating approval/denial decisions. In contrast, SMD is often used to assess disparities in model outcomes in two situations. The first is when the decision being made is not binary, but rather is a choice from a numerical range, such as an interest rate or a credit line assignment (as compared to a discrete decision, like approval/denial). The second is when the decision is based on the model output in combination with other factors.[27] The SMD is equal to the difference between the average protected class outcome and the average control class outcome, divided by a measure of the standard deviation of the outcome across the overall population.[28]

Here, the disparate impact Step 1 disparity analysis of Upstart's Model was performed as follows:

1. First, the SMDs at Stage 1 were calculated. Recall that at Stage 1, Upstart's AI/ML Model predicts the default and prepayment probabilities for each borrower.

2. Second, we fed the Stage 1 model outputs into Stages 2 and 3. Recall that at Stage 2, Upstart calculates APRs for each applicant, and at Stage 3 applicants are either approved or rejected for a loan. Therefore, this step in the analysis lets us measure whether any Stage 1 AI/ML Model disparities would translate into APR and

---

ensures the model will be accurate not just on training data but on other datasets when deployed, for example on different populations, time periods, or economic conditions.

[25] *Id.* at 9–10.

[26] Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt, "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing," at 5 (2020), https://www.bldsllc.com/publications/20200229_A_Responsible_Machine_Learning_Workflow.pdf; *see also* 29 C.F.R. § 1607.3 (describing adverse impact test for assessing employee selection procedures under Title VII); 29 C.F.R. § 1607.16 (defining "adverse impact" as a "substantially different rate of selection in hiring, promotion or other employment decision which works to the disadvantage of" a protected class).

[27] Monitor's Initial Report, *supra* note 1, at 9–10.

[28] *Id*.

approval/denial disparities for applicants. We use SMD to measure APR disparities and AIR to measure approval/denial disparities.[29]

We calculate these metrics on a set of loan applicants that were assessed by Upstart's platform during the first quarter of 2021. For each applicant, we standardized loan amount and loan term in order to eliminate any effects that variations in amount and term might have on disparate impact. We set the loan amount at the median requested loan amount for all applicants. We set the loan term at the most commonly requested loan term. As noted, Upstart partners with various banks that have differing APR and loan limits. We conducted our analysis of APR and approval/denial disparities by choosing as a representative bank an institution that does a significant portion of the loan volume on Upstart's platform; that institution also has the most permissive maximum APR and loan amount upper and lower limits.

b. *Disparate Impact Step 1—Statistical and Practical Significance*

Any adverse APR or approval/denial disparities for each protected class are assessed for whether they are statistically and practically significant. Courts assessing disparate impact claims under ECOA, the FHA, and Title VII of the Civil Rights Act of 1964 usually evaluate the strength of a plaintiff's statistical evidence by assessing the statistical, and, in many instances, the practical significance of the adverse impact. Under our assessment, we proceed to Steps 2 and 3 of the disparate impact analysis—identifying a legitimate business justification and searching for the existence of less discriminatory alternatives for any protected class—only if the APR or approval/denial disparities for that class under the Model are both statistically and practically significant.

Statistical significance is a standard used to determine whether a disparity is likely explained by chance instead of a specific facially neutral practice or policy.[30] Here, a disparity must be statistically significant to be considered meaningful. For testing the statistical significance of the difference in scores or continuous outcomes we used the Student's *t*-test. We consider a disparity here to be statistically significant if it has a *p*-value level of less than or equal to 0.05, which is a commonly used significance level.[31] For our AIR calculations, we used the Z-test, which is also called the 2-standard deviation test ("2-SD test"), because a difference is considered statistically significant if it is more than two standard deviations above zero.[32]

---

[29] Our analysis also lets us isolate the effects of model adjustments Upstart conducts to account for things like economic cycles.

[30] *See, e.g., Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994 (1988). In addressing statistical significance standards, the Supreme Court has observed that "[a]s a general rule for . . . large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the [result] was random would be suspect to a social scientist." *Casteneda v. Partida*, 430 U.S. 482, 496 n.17 (1977).

[31] *See, e.g., Stephanie Glen*, "T Test (Student's T-Test): Definition and Examples," https://www.statisticshowto.com/probability-and-statistics/t-test/.

[32] *See* David Morgan, "Statistical Significance Standards for Basic Adverse Impact Analysis," DCI Consulting White Paper (July 2010), https://adverse-impact.com/wp-content/uploads/2015/02/Statistical-Significance-Testing-for-Adverse-Impact-Measurement.pdf; Scott B. Morris, Russell Lobsenz, "Significance Tests and Confidence

In contrast to statistical significance, practical significance is a measure of whether the magnitude of the effect being studied is sufficiently important substantively for a court, regulator, or entity to be seriously concerned, as a real-world matter. There is a split among courts regarding whether a plaintiff must demonstrate practical significance to establish a prima facie case of disparate impact in litigation.[33] However, institutions commonly employ practical significance thresholds in their internal analyses of automated models to determine whether disparities are meaningful enough to warrant investigating whether less discriminatory alternatives exist.[34]

For the analysis of Upstart's Model, we consider an APR disparity to be practically significantly adverse if it has an SMD greater than 0.30 (where a higher SMD means greater disparities), and we consider an approval/denial disparity to be practically significantly adverse if it has an AIR less than 0.90 (where a lower AIR means greater disparities).[35] In our experience, these thresholds are commonly used by many financial institutions in their internal fair lending analyses. An AIR less than 90% can be roughly thought of as a more conservative version of the "four-fifths" rule of thumb, developed by the Equal Employment Opportunity Commission.[36] In other words, a 90% threshold would be triggered more frequently than the more forgiving 80% "four-fifths" threshold sometimes used in employment. In addition to prompting more frequent searches for less discriminatory models, a more conservative practical significance threshold (*i.e.*, 90% AIR) is sensible when dealing with a fully automated model where the model is relatively easy to validate and the effects of model inputs on the model outcome can be defined and adjusted with some precision. For example, a credit modeler can reasonably predict the effects that changing model inputs will have on a discrete model outcome, like predicting the likelihood of default at month X. The modeler can then build and measure the impact of various alternatives and understand those alternatives' effects on disparate impact and performance. In contrast, in a hiring or employment selection test—most of which are not algorithmic or model based—understanding the impact or validity of an alternative can be much more difficult. Hence,

---

Intervals for the Adverse Impact Ratio,"
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.218.7150&rep=rep1&type=pdf.

[33] *Compare, e.g., Jones v. City of Boston*, 752 F.3d 38, 53 (1st Cir. 2014) (holding that a plaintiff need not demonstrate practical significance to establish a prima facie case of disparate impact), *with Southwest Fair Hous. Council v. Maricopa Domestic Water Improvement District*, 9 F.4th 1177, 1190 n.10 (9th Cir. 2021) ("'Significance' in the context of disparate-impact claims is not limited to statistical significance; 'practical significance,' which examines whether minor statistical disparities have any discriminatory effect in practice, also plays a role."); *Waisome v. Port Auth. of N.Y. & N.J.*, 948 F.2d 1370, 1376 (2d Cir. 1991) (finding no disparate impact where impact was of "limited magnitude," despite being statistically significant).

[34] *Cf. Jones*, 752 F.3d at 52 ("Notwithstanding these limitations, [a practical significance standard] may serve important needs in guiding the exercise of agency discretion, or in serving as a helpful rule of thumb for [institutions] not wanting to perform more expansive statistical examinations.").

[35] The 0.30 threshold assumes that higher model scores are less favorable; if lower model scores are less favorable, the sign would be reversed such that SMDs below -0.30 would be considered practically significant.

[36] For a discussion of the EEOC "four-fifths" threshold, see *Jones*, 752 F.3d at 49–53. Even in employment cases, the four-fifths rule of thumb is not strictly or uniformly applied. *See, e.g., United States v. City of New York*, 637 F. Supp. 2d 77, 97-99 (E.D.N.Y. 2009) (granting summary judgment in favor of plaintiffs' prima facie disparate impact showing where some disparities would not have triggered the four-fifths rule but statistical significance was high and court found other practical significance metrics persuasive).

in the latter context, it can be difficult to have confidence that an alternative policy or practice is valid or will improve disparities, unless the disparities are more pronounced.

Use of a practical significance threshold in analyses aimed at identifying less discriminatory alternatives to automated models is not just a way to prioritize internal assessments or regulatory exam resources. It also has the important benefit of expanding the universe of less discriminatory alternatives that may be viable. This is true because alternative models might affect protected classes differently, and deciding whether an alternative is viable requires a framework for guiding those decisions. Section E.3.b, *infra*, presents the framework used for the analysis of Upstart's Model. But to illustrate this narrower point about practical significance, consider a stylized hypothetical:

> Imagine an alternative model is identified that maintains acceptable performance levels and improves statistically and practically significant disparities for Black applicants (assume AIRs for Black applicants as compared to non-Hispanic white applicants would improve from 68% to 88% under the alternative model). However, that alternative model would create worse disparities for female applicants as compared to the baseline model (assume AIRs for female applicants as compared to male applicants would decrease from 99% to 98%). Imagine disparities for all other groups remain the same.

Table 2: Hypothetical AIR Scenarios

	Baseline Model	Alternative Model
Black Applicants	68%	88%
Female Applicants	99%	98%

If no practical significance threshold is used, then this alternative model may not be acceptable because the disparities for female applicants are worse than those of the baseline model (assuming statistical significance). Without using a practical significance standard, Black applicants—who experience serious disparities under the baseline model—would not get the benefit of a significant improvement in disparities because of increased disparities experienced by female applicants, even though neither the original disparities nor the disparity increase for females is particularly large. In contrast, if a practical significance threshold of 90% AIR is applied, then the alternative model might be acceptable because it has not introduced any new statistically *and* practically significant disparities for other protected groups. Because the 98% AIR for females is well above the 90% practical significance threshold, the alternative model would be acceptable.

Importantly, identifying even statistically and practically significant disparities at disparate impact Step 1, standing alone, does *not* demonstrate the existence of a fair lending violation. In part, that is because for the disparate impact analysis, we assess disparities without attempting to control for legitimate creditworthiness criteria.[37] While such controls are helpful

[37] Monitor's Initial Report, *supra* note 1, at 10.

for certain fair lending analyses—for example, analyses to ascertain whether disparities are attributable to discriminatory decisions in a judgmental process—in a disparate impact review of a fully automated model, controlling for certain criteria at this stage can unnecessarily mask the disparate impact of the model and the need to search for potential less discriminatory alternatives (*e.g.*, even if a risk score is business justified, an alternative risk score may be as effective and have less disparate impact). Traditional and commonly used credit criteria can (and often do) cause disparate impacts; controlling for these criteria can inappropriately and incorrectly assume no less discriminatory alternatives exist. Accordingly, merely identifying even a meaningful disparity at Step 1 does not demonstrate the existence or absence of a violation.

c. *Disparate Impact Step 1—Results*

Applying these methodologies, we found adverse approval/denial AIR disparities at the final stage of the loan process for both Black applicants and female applicants, but only the disparities for Black applicants were below 90% and therefore practically significant by the metrics used for this Report. Asian/Pacific Islander applicants, Hispanic applicants, and applicants 62 years old or older all experienced favorable AIR rates.

With respect to pricing disparities, Black, Hispanic, and female applicants experienced some APR disparities. However, none of these disparities were above 0.30 and therefore none were practically significant by the metrics used for this Report. Asian/Pacific Islander and applicants 62 years old or older experienced favorable APR SMDs.

Finally, looking just at the lifetime default risk SMD for the Stage 1 AI/ML Model alone, Black, Hispanic, and female applicants experienced adverse disparities, but these too fell below the 0.30 practical significance threshold used in this Report.

These results are generally consistent in important ways with certain internal Upstart analyses for the first quarter of 2021. Although conducted under different constraints, by the metrics used in our Report, those results also show practically significant adverse disparities for Black applicants, but not for other groups. Upstart represents that it is actively researching less discriminatory alternatives, including via this Monitorship.

We offer a few observations on these results: First, this finding does not, standing alone, demonstrate a fair lending violation, but it does trigger an investigation into whether less discriminatory alternatives exist. Second, as noted, these disparity measurements were calculated on a dataset representing the pool of applicants that, if approved under Upstart's Model, would have been eligible for a loan from at least one bank partner. Accordingly, these results can generally be attributed to Upstart's Model, rather than bank partner criteria. Third, we conducted this analysis based on a representative bank partner and results might differ for analyses conducted using the data of other bank partners because, for example, the demographics of those datasets differ. Finally, we note that there is likely a relationship between a bank partner's approval threshold and whether any disparities caused by Upstart's Model appear in pricing or in

approval/denial decisions. For example, a bank with a low APR maximum might deny more applicants, resulting in higher approval/denial disparities, whereas a bank with a high APR maximum might approve more applicants, resulting in lower approval/denial disparities but higher pricing disparities.

2. Disparate Impact Step 2—Legitimate Business Need

Under Step 2 of the disparate impact analysis, if there are meaningful disparities adverse to a protected class, the entity should establish a legitimate business need for the model—in other words, showing that the model is "necessary to achieve one or more substantial, legitimate, nondiscriminatory interests."[38] In the credit context, a model or variable is often considered to advance a valid business need if it is predictive of a relevant outcome—for example, a variable that is predictive of loan performance (and its predictive relationship is not simply because it is a proxy for protected class status) or a model that meets a minimum standard of accuracy for predicting default.

Upstart's Model predicts default and pre-payment probabilities, which are combined to compute a cash-flow estimation. Upstart represents that its Model is accurate in making these predictions. Prediction of default and pre-payment probabilities would likely be considered legitimate business interests at Step 2 of the disparate impact analysis, although we note that the scope of what qualifies as a legitimate business interest in the credit context is not settled and some have argued that legitimate interests in this field should be construed narrowly.[39]

3. Disparate Impact Step 3—Identifying Less Discriminatory Alternatives

Because statistically and practically significant disparities were identified for approval/denial decisions for Black applicants, our analysis turns to the third step in the traditional disparate impact framework: whether less discriminatory alternatives exist. Significant work has been done to design and develop a faithful Facsimile Baseline Model and methodology tailored to Upstart's Model. As of the date of this Report, we are testing and validating potential model alternatives, but we are not yet in a position to determine whether a viable alternative model exists, and, if so, what changes to Upstart's Model we would recommend. Instead, we describe below our testing methodology and parameters under which we would recommend an alternative model.

---

[38] *See, e.g., Mhany Mgmt., Inc. v. Cty. Of Nassau*, 819 F.3d 581, 617 (2d Cir. 2016) (quoting 24 C.F.R. 100.500(c)). In litigation, it would be the defendant's obligation to make an evidentiary showing to this effect.

[39] *See, e.g.,* FFIEC Interagency Fair Lending Examination Procedures, Appendix (Aug. 2009) ("There is very little authoritative legal interpretation of [the term business necessity] with regard to lending."); NCLC, "Credit Discrimination," § 4.3.2.5 ("Business Justification") ("With respect to claims under the ECOA, no guidance is given in Regulation B as to what might constitute a legitimate business necessity in credit discrimination cases."); Robert P. Bartlett, *et al*., "Algorithmic Discrimination and Input Accountability Under the Civil Rights Act" at 33 (2020) (arguing that in credit determinations, decision-making outcomes can lawfully vary across protected groups only if decisions are based on a target variable of creditworthiness).

We have focused our search to date on whether a less discriminatory alternative exists at Stage 1 of Upstart's process. As explained above, Stage 1 is Upstart's core AI/ML Model, which Upstart uses to predict default and prepayment probabilities for each borrower. Those outputs are eventually translated into APRs and approval/denial decisions. We operate under the premise that identifying a less discriminatory version of the Stage 1 Model is likely to translate into less discriminatory APRs and approval/denial decisions.[40]

In the case of traditional statistical models, identifying a less discriminatory alternative has often included a process of adding, dropping, or substituting variables in the model, with the goal of identifying variations of the model that maintain similar performance but that have less disparate impact on protected classes.[41] The increased interest in and reliance on AI/ML models has sparked development of more sophisticated methods for identifying less discriminatory alternative models.[42] Those alternative models might also involve excluding variables, although the process for identifying effective candidates for exclusion will be more sophisticated than under traditional methods. At bottom, the foundation for many of these newer methods for identifying less discriminatory alternatives is akin to that of the traditional methods: using an awareness of the likely effects of a model on protected groups to inform a search for and development of protected-class neutral alternative models that achieve similar performance metrics.

a. *Disparate Impact Step 3—Technical Methodology*

Our chosen methodology for this Monitorship involves exploring a large number of variable combinations for Stage 1 using AI/ML techniques, to identify variable combinations that yield the highest reduction in disparate impact, while reasonably preserving the performance of the Baseline Model. Those variable combinations would be maintained in the Model, while other variables would be excluded.

We start by experimenting in the Facsimile Baseline Model with various combinations of predictor variables from the universe of available Upstart variables—each combination of variables is referred to as a "combinatorial subspace point." The collection of all possible permutations of predictor variable combinations is referred to as the "combinatorial subspace." We conduct a tailored search through the combinatorial subspace to identify whether there exists a subset of predictor variables drawn from the full set of variables used in the Baseline Model that could decrease adverse impact. To maximize the predictive accuracy of each subset of variable combinations stipulated for each alternative model, we have to optimally set the configuration parameters associated with Upstart's Stage 1 AI/ML Model.[43] This process is

[40] Other stages in the process may be the focus of later reports, if warranted.
[41] *See* Monitor's Initial Report, *supra* note 1, at 11–12.
[42] *See id*.
[43] A model parameter is a configuration internal to the model; it can be thought of as a way to tailor the model to a specific set of data. AI/ML models typically have parameters that are set to optimal values *during* model training. In addition, there are several configuration parameters that must be set *prior* to model training—these configuration

referred to as Hyperparameter Tuning. Accordingly, for each variable combination, we tune the model hyperparameters and retrain a model. For each new trained model, we compute the disparity metrics (consistent with the disparity methodologies described above) and the model performance.[44] A potential alternative model, therefore, includes a combination of predictor variables from the Baseline Model, and tailored hyperparameters that might be different than the hyperparameters of the Baseline Model. Through this process, variables currently used in the Model may be excluded.

After identifying a promising alternative model based on running this process on the Stage 1 Facsimile Baseline Model, the predictor variables and hyperparameters of the alternative Stage 1 model are fed back into a Stage 1 model training interface to conduct the model training again, this time in a non-facsimile environment internal to Upstart. That process validates that the results from the Facsimile Baseline Model environment hold true in Upstart's own environment.

These alternative models from Stage 1 generate predictive outputs for prepayment probability and default probability for each borrower. Those outputs are fed into Upstart's software interface for the remaining stages of its model process to discern how the alternative Stage 1 AI/ML Model would actually perform with respect to mitigating disparities and predicting APR and approval/denial decisions.

b. *Disparate Impact Step 3—Choosing Among Potential Alternatives*

As noted, alternative models might affect different protected classes differently, and deciding whether an alternative is viable requires a framework for guiding those decisions. The process of identifying potentially viable less discriminatory alternatives is conducted within the following constraints:

First, we will not recommend adopting a potential alternative model if its performance is meaningfully worse than the performance of the Baseline Model. This means that model performance metrics must be within some tolerance of the original model. Neither courts nor agencies have delineated concrete thresholds for this determination and internal practices differ across financial institutions. Some institutions, for example, adopt internal thresholds beyond which model performance metrics such as KS or R^2 should not drop—for example, a deterioration in KS of 5% might be deemed unacceptable.[45] These institutions have made the

parameters may govern the model training process or the model architecture. The process of finding optimal values of configuration parameters is called Hyperparameter Tuning, and results in training high quality AI/ML models.

[44] To assess model performance and guard against model drift, we compute the accuracy for each new model against an out-of-sample dataset representing the most recent year of data and compare it with the complete multi-year dataset. We separately compute each disparity metric on the out-of-sample dataset representing the most recent year, using models trained on the entire data set, and compare those results with the same metrics using a model trained only using the most recent year.

[45] KS and R^2 are both statistical metrics of model accuracy, and are provided here as examples because, although not used by Upstart, they are commonly used to measure performance. The KS metric (or Kolmogorov-Smirnov Test) is used when the output of a model is a probability distribution (which is the range of probabilities associated with

decision that alternatives that perform within that threshold are sufficiently effective to advance their legitimate business needs. Lenders might also align such thresholds with criteria they use to evaluate performance deterioration for original model training and development purposes. For example, if modelers consider a model to be acceptable as long as there is no more than a 5% deterioration in KS when comparing model development and out-of-time validation data, then they might also apply a maximum 5% deterioration in KS when assessing the viability of potential alternative models.[46] In other words, if performance deterioration is not significant enough to warrant rebuilding a model, then it is not considered significant enough to warrant rejecting a less discriminatory alternative model.

Upstart uses more than one performance metric to assess its models. At this time, we are not prepared to recommend a specific performance deterioration threshold. We will base any recommendations regarding whether alternatives should be considered viable on our experience in this area, including with other institutions and models, alongside Upstart's modeling practices, any impacts on performance, the commercial reasonableness of implementing any recommendations, and corresponding improvements in disparities.

Second, other model risk management criteria that would normally be used to establish the viability of a model apply to the potential alternative as well. For example, if Upstart's standard model development procedures require that models have similar performance metrics across validation samples, that requirement would also apply to a potential alternative. Accordingly, we will consider reasonable model risk management criteria in assessing whether to recommend an alternative model.

Third, we would not recommend an alternative model that introduces *new* statistically and practically significant disparities that were not present in the original model. Recall that we consider an APR disparity to be practically significantly adverse if it has an SMD greater than 0.30, and we consider an approval/denial disparity practically significantly adverse if it has an AIR less than 90%. Therefore, for example, if the original model showed an AIR of 80% for Black applicants and an AIR of 91% for Hispanic applicants, a potential alternative that dropped the AIR for Hispanic applicants below 90% would not be recommended, regardless of the improvement in AIR for Black applicants.

Fourth, we would not recommend an alternative model that would exacerbate *existing* statistically and practically significant disparities from the original model. For example, if the original model showed an AIR of 85% for Hispanic applicants and 89% for female applicants, a potential alternative that resulted in a statistically significant deterioration in the AIR for female applicants would not be recommended. There is no practical significance requirement for the size of the deterioration in AIR, so long as the change is statistically significant.

---

each possible outcome) and is a measure of how close the predicted distribution is to the expected distribution. The R^2 (or R-squared) metric is used when the output of a model is a numerical quantity (*e.g.*, an APR). It measures the proportion of the variation from the mean of the quantity of interest that can be predicted by the model.

[46] Exceptions might exist, such as a decision to accept a drop in performance beyond this threshold in order to achieve an exceptionally large benefit to an affected class.

Fifth, we would not recommend an alternative model that would improve disparate impact for one protected class but that would introduce meaningful new adverse bias for a different protected class, such as predicting risk meaningfully less accurately for different protected class groups—a form of model bias that is sometimes referred to as "differential validity."

Finally, it is possible that multiple alternative models could satisfy the above criteria. In such situations, we may need to apply more case-specific criteria. For example, we may recommend an alternative model that results in the greatest improvement in disparate impact. It might also be true that there are multiple groups with practically significant disparities, and while all affected groups are benefited by all the alternatives, which group is benefited the most varies.[47] In this case, to resolve a choice between multiple potential alternatives, we may recommend the model that would result in the greatest overall reduction in shortfall among members of affected protected classes.[48]

As noted, we are actively testing and validating potential alternatives and have not yet determined whether Upstart should adopt an alternative model, and if so, what that alternative might be.

F. Proxy Review

1. Variable Background

As noted, it is generally a violation of the ECOA and FHA prohibitions against overt, intentional discrimination (*i.e.*, disparate treatment) to use a protected class or a close proxy for a protected class as a variable in a credit scoring or pricing model.[49] Agencies and courts have not clearly defined what qualifies as a close proxy, but it is often understood to mean a variable whose predictive value in a model is attributable solely or largely to its correlation with a protected characteristic. This proxy analysis is independent of the disparate impact analysis described above: a model can raise disparate impact risks even if it does not contain any protected class or close proxy variables. Similarly, a model that uses protected class or proxy variables would raise disparate treatment risks, even if that model did not cause disparate impacts adverse to a protected class.

[47] For example, imagine that both Black and Hispanic applicants have an AIR of 80%. It could be that potential alternative #1 improves the AIRs to 85% and 82% respectively, whereas potential alternative #2 improves the AIRs to 82% and 85% respectively.

[48] The shortfall is defined as the difference between the number of protected class members who actually received a favorable outcome and number who would have received the favorable outcome, if they were to receive such outcomes at the same rate as the control group.

[49] *See, e.g.,* Monitor's Initial Report, *supra* note 1, at 8.

Upstart provided a list of base variables available for use in its Model.[50] We did not identify any protected classes included explicitly in the variable list provided.[51] The vast majority of the variables provided are variations of criteria commonly used in credit determinations. Examples include typical credit file variables related to things like spend, payment, balance activities, delinquencies on credit accounts, loan terms, and third-party credit scores.

The variable list provided by Upstart does include a group of what Upstart describes as "non-credit variables." Among Upstart's non-credit variables are education-related variables, including:

- Aggregated income data of graduates of an applicant's school;
- Aggregated graduate income data for an applicant's area of study;
- Graduation year of an applicant's most recent degree;
- The highest level of education attained by an applicant;
- An indicator of whether an applicant is attending a coding bootcamp;
- An indicator of whether an applicant is attending a coding bootcamp whose quality has been reviewed by Upstart.

Although all of these variables are available for use in Upstart's Model, as of February 2021, the Upstart Model does not include either: (1) aggregated graduate income data for an applicant's area of study; or (2) an indicator of whether an applicant is attending a coding bootcamp.

As discussed above, perceived risks central to this Monitorship concern the use of information related to higher education and whether such information may contribute to discriminatory outcomes that disproportionately affect communities of color.

2. Qualitative Variable Review

Many financial institutions conduct a qualitative variable review in which they flag for further scrutiny, or may simply remove from a model, variables that they consider to be high risk because the variables may be perceived to be close proxies for protected classes, they may raise reputational risk, or for other reasons. Some institutions flag individual variables during a qualitative review because, in part, they do not quantitatively review proxy risks, or they only quantitatively review select variables for such risks. Such variables might be considered particularly problematic if they significantly contribute to disparate impact adverse to a protected class.

[50] These variables can then be engineered in various ways to create manually generated variables for use in the Model (for example, from monthly loan balance data, one could manually engineer standardized mean monthly balance).

[51] We did not assess whether any of these variables might risk violating state antidiscrimination laws applicable to credit.

As noted, the variable list provided by Upstart includes a group of what Upstart describes as "non-credit variables." Upstart represents that these variables are statistically related to model performance, but the relationship between some of these variables and an individual applicant's creditworthiness is not as direct as the credit-related variables noted above and in some cases is not necessarily intuitive.

However, aside from determining that Upstart's variable list does not include protected class statuses as attributes, we do not make any qualitative determinations about Upstart's variables—meaning, we do not identify any variables as raising or not raising potential risks based on a qualitative assessment alone. Instead, as described below, we use quantitative methods to attempt to assess proxy risks across all variables.

3. Quantitative Variable Analysis

a. *Quantitative Variable Overview*

A predictor variable used within a credit risk model might be considered a proxy for a protected class if the variable is strongly correlated with a protected class label or is a strong predictor of protected class, and all or a significant part of the variable's contribution to model performance derives from its correlation with the protected class characteristic.[52] We designed analyses to assess both considerations, although, as discussed, the methodologies have inherent limitations.

First, we used a statistical technique called "Surrogate Modeling" to assess whether the individual input variables that are fed into Upstart's AI/ML Model may be significant predictors of race and national origin.[53] A surrogate model is a model whose predictions closely approximate those of a given baseline model on specific datasets but that has some desirable properties, such as greater interpretability. Surrogate models can be used as approximations of more complex AI/ML models to enable interpretation or explanation of the results of those more complex models. Here, Surrogate Modeling is used to ascertain the degree to which the entire collection of input variables in a model is predictive of protected class labels, and to assess the relative significance of each variable compared to all others regarding its individual significance in predicting protected class.[54]

Second, if through that Surrogate Modeling process we identify variables with meaningfully higher significance in predicting protect class, we would assess what happens when Upstart's Stage 1 AI/ML Model is trained on a population of only white applicants with

[52] *See* Monitor's Initial Report, *supra* note 1, at 8.
[53] We prioritized race and national origin for this stage of our review, in part because of the concerns giving rise to this Monitorship and because the most meaningful disparities observed were related to race. We may assess proxy risks for gender or age at a later stage.
[54] For this stage, these quantitative proxy analyses were done without the Monitor's full awareness of the actual names of each predictor variable; generic replacement titles were used at this time.

and without these variables included, and what happens when Upstart's Stage 1 AI/ML Model is trained on a population of only non-white applicants with and without these variables included. If this group of variables does not contribute to performance in either the all-white model or the all-non-white model, it might be an indication that the variables' contributions to performance in Upstart's AI/ML Model are attributable to their correlation with protected class characteristics.

Finally, we would be particularly concerned with potential proxy variables if they also drive adverse disparate impacts. Accordingly, we would track identified variables in future analyses by assessing whether they would be excluded in potential recommended less discriminatory alternative models.

Before describing these methodologies, it is important to note that the Surrogate Modeling approach has inherent limitations and it cannot conclusively demonstrate that a model does or does not contain proxies for protected class. Importantly, in certain models, namely nonlinear and nonparametric models that stem from AI/ML, input variables may combine inside of the model and interact with one another to produce temporary internal variables sometimes called "interaction variables."[55] These interaction variables that are automatically created within a model might correlate with or predict protected class labels in ways that could be considered proxies, even if the individual input variables do not. Surrogate Modeling does not illuminate whether any interaction variables that are automatically created within a model are predictive of protected class labels.

In other words, although the Surrogate Modeling approach can reveal whether input variables in the model are functioning as proxies relative to each other, and it can provide insights into whether the entire collection of variables are materially predictive of protected class labels, it cannot demonstrate—or rule out—that interaction variables generated inside of an AI/ML model are functioning in ways such that they might be considered proxies.

Because of these limitations, we cannot conclusively eliminate the possibility that proxies exist. For that reason, we suggest that Upstart continue to weigh this risk and the feasibility of adopting more interpretable model structures against the perceived benefits of its AI/ML Model. Those potential benefits might include improved model performance, as well as the potential that the flexibility of an AI/ML structure may permit improvements on other fairness metrics, such as disparate impact—a possibility we will explore more in future reports. Upstart represents that it already performs this type of risk/benefit analysis, which we do not question.

[55] The term "nonlinear" is described above at footnote 10. The term "nonparametric" means a model that does not have a fixed set of parameters that are computed during the model training process. (Model parameters are described above at footnote 43.) Instead, the size of their parameter set is unbounded and (generally) grows with the amount of training data. Parametric models, in contrast, have a fixed set of parameters. A nonparametric structure allows a model to generate new variables automatically, including by combining several input variables. Because of this feature, nonparametric models are almost always nonlinear, since the effect of an input variable may be disproportionately attenuated or amplified with respect to its magnitude.

b. *Surrogate Modeling*

Our Surrogate Modeling methodology for proxy detection uses the BISG-estimated class label of individuals in the historical borrower dataset along with all the variables used in Upstart's Baseline Model. For this stage of our analysis, we prioritized looking at proxies for race—Hispanic, African American, Asian American/Pacific Islander, American Indian/Alaskan Native, and multiracial.

We trained several proxy assessment surrogate models to assess the proxy status of all input variables used by Upstart. The goal of each surrogate model is to predict the protected class of each borrower in the dataset. These protected class predictions from the surrogate models can be compared to the actual BISG-estimated class labels of the borrowers to yield a "goodness-of-fit" metric for each surrogate model, as measured by an F1 score.[56] If a surrogate model has a strong goodness-of-fit, then input predictor variables or some portion thereof that are fed into the credit-risk model may be functioning as proxies for protected class labels.

The surrogate models we created are listed in Table 3 and described below.

Table 3: Surrogate Model Results

	Model Type	Percentage of Protected Class Predictor Variables Removed	Max F1 Score
1	Null Model	100.0%	0.434
2	Ridge Logistic Regression Surrogate	0.0%	0.495
3	RandomForest Surrogate	0.0%	0.480
4	Principal Components Regression Surrogate	69.2%	0.434

We first created a Null model (shown in row 1 above), which attempts to predict the protected class labels of borrowers in the dataset without including *any* of Upstart's input variables. We use this Null model as a baseline from which we can assess the performance of other surrogate models that *do* include Upstart's input variables. If a surrogate model that uses all of Upstart's input variables can predict the protected class label of a borrower with a significantly greater F1 score than the Null model, that would be indicative of the possible presence of protected class proxy variables among Upstart's input variables.

[56] The F1 score of a model represents the best combination of the precision (*i.e.*, specificity) and recall (*i.e.*, sensitivity) achievable by the model. It is a measurement of a model's predictive power defined by the combination of true positive, false positive, true negative, and false negative rates.

This Null model is straightforward: it simply uses the ratios of the number of individuals in each protected class group within the dataset to predict the BISG protected class membership of each individual. For example, imagine the historical data shows that 12% of the borrowers are Black. This gives us a naïve probability that for every 100 future applicants, 12 will be Black. If we randomly apply this naïve predictive probability to each applicant and then compare the predicted class label to the BISG-class label, we will find in some cases we have inaccurately predicted the applicant as Black and in other cases we have accurately predicted the applicant as Black. In other words, this basic Null model will result in a number of true positives, false positives, true negatives, and false negatives. These numbers are combined to form the F1 score, which is a measure of the accuracy and precision of the model. An F1 score of 1 suggests the null model is perfectly accurate and precise. An F1 score of 0 suggests that the Null model is completely inaccurate and imprecise.

As Table 3 indicates, the Null model has an F1 score of 0.434. This performance metric should not be considered good or bad. It simply demonstrates as a baseline that, in the absence of using any of Upstart's input variables, we can achieve an accuracy of 0.434 in predicting the protected class label of each borrower simply by rolling a die with probabilities informed by the historical ratio of protected class populations. If a surrogate model that uses Upstart's input variables can predict the protected class label of a borrower with significantly greater F1 than the Null model's 0.434 score, that would be indicative that Upstart's input predictor variables might include protected class proxies.

We then developed three surrogate models designed to predict the protected class labels of borrowers using Upstart's input variables. Row two in Table 3 shows a Ridge Logistic Regression surrogate model. This Ridge Logistic Regression surrogate model is a linear classification model for predicting the probability of a person belonging to a specific race using variables in the Upstart borrower dataset. It is a much simpler and more interpretable model than Upstart's AI/ML Model, but it enables us to see the strength of each input variable's relative predictive power for identifying the protected class label of a borrower.[57] Row three in Table 3 shows another surrogate model—a RandomForest surrogate model. RandomForest models are less interpretable, but are often more powerful than Ridge Logistic Regression models. The RandomForest is a nonparametric surrogate model designed to predict the probability of a person belonging to a specific race using variables in the Upstart borrower dataset. We use this surrogate model because Upstart's actual Model is nonparametric, so this model can serve as a better approximation than the Ridge Logistic Regression, but it still allows for computation of the importance of each input variable in the trained model.

Both of these surrogate models were specified with the entire set of Upstart's input variables and trained on the borrower dataset with BISG protected class labels. The Ridge

[57] It is "interpretable" in the sense that it is a linear model where a weighted sum of predictor variables determines the prediction of the model. The weights are determined during model training and are known, therefore the effects of individual variables on the predictions of this model are precisely known.

Logistic Regression achieved an F1 score of 0.495 and the RandomForest achieved an F1 score of 0.480. These F1 scores do not represent significant gains in accuracy for predicting protected class membership of each borrower over the Null model. In other words, including all of Upstart's input variables in these surrogate models does not provide a practically significant improvement in absolute accuracy for predicting protected class membership. This observation offers evidence that Upstart's individual input variables are not strong predictors of race and national origin in an absolute sense.

Among Upstart's input variables, it still might be the case that certain variables exert more influence on driving the F1-gains of our two surrogate models. Identifying any such variables could help inform our analyses, and provide a set of variables that warrant further analysis or monitoring. Accordingly, we conducted an analysis aimed at discerning the protected class predictive power of each variable in these surrogate models. This assessment can provide insights into the *relative* proxy significance of each variable as compared to all other variables. The more important a variable is in the surrogate model, the more likely it is to be predictive of protected class, relative to the other variables in the surrogate model.

To assess whether certain variables are *relatively* more significant in predicting a borrower's protected class label than others, we first use an analysis called Information Value regression, which is a statistical technique designed to rank variables on the basis of their importance. We use an Information Value threshold of 0.3, which is commonly used in statistical literature as a threshold for relative significance.[58] If a variable's Information Value level exceeds 0.3, it would be deemed significant and flagged for further analysis. Namely, we would train another surrogate Ridge Logistic Regression model that excludes the variables with levels above 0.3. The theory animating that analysis would be that if removing the higher correlation variables resulted in a significant drop in performance of the surrogate model, it might suggest those variables are significant predictors of protected class. Separately, we would assess what happens when Upstart's Stage 1 AI/ML Model is trained on a population of only white applicants with and without those variables included, and what happens when Upstart's Stage 1 AI/ML Model is trained on a population of only non-white applicants with and without those variables included. If that group of variables does not contribute to performance in either the all-white model or the all-non-white model, it might be an indication that the variables' contributions to performance in Upstart's AI/ML Model are attributable to their correlation with protected class characteristics.

However, after implementing our Information Value test, we found that all individual variables had Information Values that were below 0.1—significantly below the 0.3 threshold we use for relative significance. In other words, even using a relative test, the relative contributions of individual variables to predicting race and national origin were generally flat across variables, meaning no variables were identified as having meaningfully more significant predictive power

[58] *See, e.g.,* Towards Data Science, "Model? Or do you mean Weigh of Evidence (WoE) and Information Value (IV)?" (Mar. 9, 2020), https://towardsdatascience.com/model-or-do-you-mean-weight-of-evidence-woe-and-information-value-iv-331499f6fc2.

as compared to other variables. Accordingly, we did not proceed with the further analyses described above.

Finally, the fourth surrogate model was motivated by our observation that variables within Upstart's input variable set have significant collinearity, which means that variations in the values of the variables are highly interrelated and, in many cases, nearly identical. That high level of collinearity can mean that the surrogate model approach of finding relative variables can misleadingly identify one variable as having higher relative significance on proxy prediction relative to the other variables, when in fact collinear "sibling" variables might have just as much significance. To address this concern, we transformed variables into non-collinear representative predictor variables called Principal Components.[59] After identifying the Principal Components, we retained only those that explain more than 95% of the variation in the data in order to remove the collinearity effects. That step resulted in dropping about 54% of the Principal Components.

We then trained another surrogate model to predict protected class labels using the remaining Principal Components (rather than the original variables) as predictors. Within that surrogate model, we identified the Principal Components with the highest coefficients (meaning the strongest at predicting protected class labels). We then attempted to assess what percentage of these remaining Principal Components we would need to remove from a retrained surrogate model such that the performance of that surrogate model would degrade to match that of the Null model. If that number of removed Principal Components is very low, we might surmise that those few removed Principal Components are important predictors of protected class. However, as shown in row four in Table 3 above, the performance of the surrogate model did not equal that of the Null model until we removed another 15% of the Principal Components (which equates to about 200 Principal Components, resulting in 69.2% of the Principal Components removed in total). That figure suggests that even the most heavily weighted Principal Components are not particularly strong predictors of race and national origin on their own.

In sum, the Surrogate Modeling techniques described here allow us to identify whether variables have a higher contribution to predicting protected class labels—a signal that their model contribution is correlated with protected class. If such variables existed, they would receive further scrutiny, including: (1) assessing whether they would contribute to performance in models trained on only white applicants and only minority applicants; and (2) tracking whether they may be included in future recommendations for less discriminatory alternative models. However, based on the methodologies used, it does not appear that individual input variables in Upstart's Model have a high likelihood of functioning as proxies for race or national

[59] The Principal Components Regression Surrogate model is the same as the Ridge Logistic Regression model, with one crucial difference: the variables in Upstart's borrower dataset are condensed to a much smaller set. Each condensed variable is formed by doing a weighted sum of the original variables. The weights and number of condensed variables are chosen such that more than 95% of the original variables' explaining power is captured. For a background on principal component analysis, *see* Wikipedia, "Principal Component Analysis," https://en.wikipedia.org/wiki/Principal_component_analysis.

origin as compared to a Null model baseline or relative to other variables, and so we did not proceed with further proxy analyses with respect to race and national origin at this time.

As noted, these Surrogate Modeling techniques have inherent limitations and do not allow us to conclude definitively that Upstart's Model does or does not contain proxies for protected class. In an AI/ML model, variables interact to create new interaction variables within the model. Our Surrogate Modeling approach does not provide visibility into whether any of those interaction variables within the AI/ML model may be functioning in ways such that they could be considered proxies.