Fair Lending Monitorship of Upstart Network's Lending Model

Third 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.

September 16, 2022

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

Table of Contents

Executive Summary 3

A. Fair Lending Overview (Recap) 6

B. Disparate Impact Analysis 7

1. Disparate Impact Step 1 7

a. Disparate Impact Step 1—Assessing Disparities 7

b. Disparate Impact Step 1—Results 8

2. Disparate Impact Step 2—Legitimate Business Need 9

3. Disparate Impact Step 3—Identifying Less Discriminatory Alternatives 9

a. Adequate Performance—Overview 12

b. Adequate Performance—Legal Standard 13

c. Adequate Performance—Internal Compliance Standards 15

d. Adequate Performance—Upstart's Metrics for Assessing Model Performance 16

e. Adequate Performance—Establishing a Viable Range 17

f. Adequate Performance—Out of Time Analysis 20

g. Translating Model Performance to Business Metrics 23

h. Less Discriminatory Alternative Results 24

i. Less Discriminatory Alternative Recommendation 29

C. Proxy Review 30

a. Proxy Review—Overview 30

b. Proxy Review—Upstart's Policy for Limiting Potential Proxies for Age 31

c. Proxy Review—Methodology Recap 32

d. Proxy Review—Detailed Results 33

e. Proxy Review—Recommendations 36

Executive Summary

This is the Third Report of the independent fair lending Monitor regarding Upstart Network's ("Upstart") lending Model. In April 2021, we issued an 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] On November 10, 2021, we issued a public Second Report, providing further detail regarding the methodology and fair lending tests conducted to date.[2] This Third Report explains application of those tests to a recent version of Upstart's Model, including identifying what would likely have been a viable less discriminatory alternative model. Before the analyses were completed, Upstart updated its Model. Accordingly, instead of recommending adoption of that specific less discriminatory alternative model, we recommend that Upstart apply the methodologies described in this Report to its existing algorithms and resulting Model or to upcoming model updates, and to future model updates. If a viable less discriminatory alternative model is identified, we recommend that Upstart adopt that alternative. This Report separately addresses proxy risks related to sex and age and does not find quantitative evidence that variables in Upstart's model are functioning as close proxies for those characteristics. At the same time, it does provide recommendations to mitigate future potential age-proxy risks specifically related to non-traditional variables. A Confidential version of this Report has been provided to the parties. This Public version of the Report removes confidential proprietary or commercially sensitive information, as appropriate.

Upstart is a lending platform that relies on Artificial Intelligence and Machine Learning ("AI/ML") algorithms, and predictive models generated from those algorithms, that incorporate non-traditional applicant data—including data related to borrowers' higher education—to underwrite and price consumer loans. In 2020, the NAACP Legal Defense Fund ("LDF") and the Student Borrower Protection Center ("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. Unless otherwise noted, this Report uses the term Monitor to refer to the collective work of Relman Colfax, Sentrana Inc., and Dr. Bernard Siskin of BLDS.

[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.

[2] Relman Colfax, "Second Report of the Independent Monitor (Public)," 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. (Nov. 10, 2021) ("Monitor's Second Report"), https://www.relmanlaw.com/media/cases/1180_PUBLIC%20Upstart%20Monitorship_2nd%20Report_FINAL.pdf.

This Third Report first analyzes whether Upstart's Model causes an adverse impact on any protected classes and, if so, whether there are less discriminatory alternative practices that comparably maintain the Model's performance. We identify what we refer to as statistically and practically significant approval disparities for Black applicants as compared to non-Hispanic white applicants. We see these disparities in Upstart-reported data for Q1 of 2022 and on a set of data for Q3 of 2021. Such disparities do not, standing alone, demonstrate a fair lending violation, but they do trigger an obligation under our methodology to investigate whether viable less discriminatory alternative models exist. After conducting this analysis, we identify what would likely have been a viable alternative model that would cause fewer disparities for Black applicants. However, Upstart updated its Model before we completed our analyses.

Rather than recommending that Upstart adopt an outdated model, we recommend that, within a reasonable amount of time, Upstart apply this methodology to its AI/ML algorithms and resulting Model or upcoming Model updates, as well as future updates to the Model. We recommend that, during the course of this Monitorship, Upstart report the results of its application of these methodologies to us as it applies them.

In describing this recommendation, this Third Report details our approach for assessing whether the performance of a potential alternative model would be comparable to the performance of the existing Model (the "Baseline" Model) in terms of reasonably meeting Upstart's legitimate business needs. Under the disparate impact doctrine, a practice could be illegal if a business need can reasonably be achieved as well by an alternative with less impact. As with any model, there is a range of uncertainty regarding how Upstart's Model will perform. We believe there is a likelihood that a court would find that an alternative model that is likely to perform within the Baseline's performance range would achieve Upstart's business interests as well as the Baseline Model. A court could take a different approach, for example by concluding that an alternative is not viable unless it is statistically significantly equal to or better than the Baseline Model using Upstart's chosen accuracy metric. But, largely due to uncertainty not captured by that metric, a court could also reasonably conclude that a model that falls within the likely performance range of the Baseline Model is indistinguishable from the Baseline Model for purposes of achieving Upstart's legitimate business purposes. Accordingly, to account for that likelihood, we presume a less discriminatory alternative model will reasonably achieve Upstart's legitimate business interests if the alternative's performance falls within the likely performance range of the Baseline Model.

Separate from that disparate impact and alternatives analysis, we also analyzed whether variables in Upstart's Model function as close proxies for protected classes. If a model contains proxies for protected classes, it is commonly 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.

In our Second Report, we found that Upstart's input variables, standing alone, do not appear to be meaningful predictors of race and national origin. In this Third Report we find that Upstart's input variables, standing alone, also do not appear to have a high likelihood of functioning as proxies for sex or age. Although we do find evidence that individual input variables in Upstart's Model have a high likelihood of being able to *predict* whether a borrower is age ≥ 62, we do not find evidence that the predictive value of these attributes is solely or largely due to that correlation with age, and therefore they are not likely functioning as proxies for age in the Model. After we shared these observations, Upstart formalized a Policy for Limiting Potential Proxies for Age ("Age Proxy Policy"), which Upstart represents codifies its practice for truncating variables to reduce their ability to predict age. We recommend that Upstart strengthen this Age Proxy Policy by applying more truncation to non-traditional variables that have a high likelihood of predicting whether a borrower is age ≥ 62.

We note that the fair lending analyses described here were unusually complex, and in some ways inherently limited, because of the nature of Upstart's Model and business structure, including use of a common Model for multiple bank and investment partners with different pricing and approval structures and applicant populations. And as noted in our Second Report, given the AI/ML nature of Upstart's Model, there are inherent limitations to our proxy-related methodologies, including an inability to assess whether interaction variables within the AI/ML Model function such that they could be proxies. Despite these limitations, this Report reflects our conclusions and recommendations based on the knowledge and information available to us. This Report does not make any legal conclusions about whether Upstart is or has been 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.

A. Fair Lending Overview (Recap)

As discussed in more detail in our Initial and Second 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, such as race, color, religion, national origin, sex, age, disability, marital status, familial status, or receipt of income from a public assistance program.[3]

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] There are three steps involved in determining whether a policy or practice—here, a model—has an unlawful disparate impact:

(1) Does the model cause a disproportionate adverse impact on a protected class?
(2) Does the model serve a legitimate business need?
(3) If the model causes a disproportionate adverse impact on a protected class and serves a legitimate business need, does a less discriminatory alternative exist that continues to serve the legitimate business need?[5]

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 model. This prohibition also applies to a variable that functions in a model as a close proxy for a protected class.[6]

[3] 15 U.S.C. § 1691(a) (ECOA); 12 C.F.R. § 1002.2(z) (Reg. B); 42 U.S.C. §§ 3604, 3605 (FHA). For a more complete description of these requirements, *see* Monitor's Second Report, *supra* note 2, at 6-8.
[4] *See* Monitor's Second Report, *supra* note 2, at 6.
[5] *See, e.g.,* Monitor's Second Report, *supra* note 2, at 6-7.
[6] *See, e.g.,* Monitor's Initial Report, *supra* note 1, at 8; *see* Monitor's Second Report, *supra* note 2, at 7.

B. Disparate Impact Analysis

1. Disparate Impact Step 1

a. *Disparate Impact Step 1—Assessing Disparities*

In the first step of the disparate impact analysis, we review predicted outcomes of the Upstart Model 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 of applicants.[7] We assess disparities without attempting to control for legitimate creditworthiness criteria because even if criteria are presumptively legitimate, less discriminatory alternatives may exist.[8]

We use two common metrics for assessing disparities at Step 1 of the disparate impact analysis: the adverse impact ratio ("AIR") and standardized mean difference ("SMD").[9]

- 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. AIR is appropriate for models generating approval/denial decisions.

- 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.[10] 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.

Under our assessment, we proceed to Steps 2 and 3 of the disparate impact analysis only if the APR or approval/denial disparities for that class under the Model are both statistically and practically significant.[11] In short:

- 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.[12]

[7] For a description of how protected class attributes were estimated, *see* Monitor's Second Report, *supra* note 2, at 11-12.

[8] *See, e.g.,* Monitor's Second Report, *supra* note 2, at 16-17; Monitor's Initial Report, *supra* note 1, at 10.

[9] Monitor's Initial Report, *supra* note 1, at 9-10; Monitor's Second Report, *supra* note 2, at 13-14.

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

[11] For more discussion of statistical and practical significance, including the rationales for our chosen metrics, *see* Monitor's Second Report, *supra* note 2, at 14-17.

[12] For testing the statistical significance of the difference in scores or continuous outcomes we use the Student's t-test. We consider a disparity 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. For our AIR calculations, we used the Z-test, which is also called the 2-

- 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. Practical significance standards also help ensure that less discriminatory alternatives are not ruled out because of small effects on other protected classes.[13] 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).[14]

b. *Disparate Impact Step 1—Results*

In our Second Report, we calculated disparity metrics on a set of loan applicants that were assessed by Upstart's platform during the first quarter of 2021.[15] 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 under our methodology.

- Asian/Pacific Islander applicants, Hispanic applicants, and applicants 62 years old or older all experienced favorable AIRs.

In this Third Report, Upstart reports adverse approval/denial AIR disparities at the final stage of the loan process for Black, Hispanic, and female applicants. As with our prior findings, only the disparities for Black applicants are below 90% and therefore practically significant by the metrics used for this Report.[16] Because we continue to see practically significant disparities for Black applicants through Q1 2022, the implications for our analysis are the same as identified in our Second Report: the disparities trigger an obligation under our methodology to investigate whether viable less discriminatory alternatives exist.

Two notes on these observations: First, in our Second Report, we conducted an in-depth analysis of disparities for each model stage based on data provided by Upstart for Q1 2021. Here, we repeat that analysis for Q3 2021 and our observations are consistent with our observations for Q1 2021: approval/denial AIRs for Black applicants were practically significant both with and

standard deviation test ("2-SD test"), because a difference is considered statistically significant if it is more than two standard deviations above zero. Monitor's Second Report, *supra* note 2, at 14.

[13] Monitor's Second Report, *supra* note 2, at 16-17.

[14] An AIR less than 90% can be roughly thought of as a more conservative version of the "four-fifths," or 80%, rule of thumb, developed by the Equal Employment Opportunity Commission. For a discussion of why a more conservative version is appropriate in this context, see Monitor's Second Report, *supra* note 2, at 15.

[15] Monitor's Second Report, *supra* note 2, at 17-18.

[16] We do not assess disparities with respect to American Indian/Alaskan Native, multiracial categories, or membership in other protected class groups because estimates for these groups are not sufficiently reliable. *See* Monitor's Second Report, *supra* note 2, at 11.

without considering Upstart's overlays and adjustments to its AI/ML Model. We also see practically significant disparities for Black applicants looking only at model score SMDs for the AI/ML Model (*i.e.*, model risk scores before they are converted to APRs or approval/denial decisions). 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. In order to prioritize other analyses, we did not replicate this independent model-stage disparity analysis for Q1 2022.

Second, these findings do not, standing alone, demonstrate a fair lending violation, but they do trigger an investigation into whether less discriminatory alternatives exist. In our experience with other credit models, it is not unusual to find statistically and practically significant disparities for at least one protected class at this stage of the analysis.

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."[17] 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 default risk (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 asserts a specific accuracy value in its prediction of these cash-flows across a population of applicants. As noted in our Second Report, meaningfully accurate 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.[18] Assessing whether Upstart could satisfy its burden of showing its Model is necessary to achieve a legitimate interest is beyond the scope of our analysis, and we proceed on the assumption that it could.

3. Disparate Impact Step 3—Identifying Less Discriminatory Alternatives

Because statistically and practically significant disparities were identified for Black applicants, our analysis turns to the third step in the traditional disparate impact framework:

[17] *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.
[18] *See* Monitor's Second Report, *supra* note 2, at 18.

whether less discriminatory alternatives exist.[19] The technical methodology we use in this Monitorship for identifying the existence of less discriminatory alternative models is discussed in our Second Report.[20] In short, we first explore a large number of variable combinations for Upstart's Model using mathematical optimization search techniques to identify combinations that yield reductions in disparate impact, while reasonably preserving the performance of Upstart's Model. Those variable combinations would be maintained in the Model, while other variables would be excluded. Second, we optimally set the configuration parameters associated with Upstart's Model. This process is referred to as Hyperparameter Tuning.[21] Hyperparameter Tuning and retraining a model can be done for each variable combination, or it can be done using all of Upstart's existing variables. In optimizing for less discriminatory alternatives, we focused on disparities in predicted default risk, rather than prepayment risk.[22] For each new trained model, we compute the model performance as well as the disparity metrics consistent with the disparity methodologies described above. A potential alternative model, therefore, includes a combination of predictor variables from the Baseline Model (which could include all variables from the Baseline Model), and tailored hyperparameters that might be different than the hyperparameters of the Baseline Model.

Our Second Report also explains conditions under which we would and would not recommend potential alternative models.[23] In short, the process of identifying potentially viable less discriminatory alternatives is conducted within the following constraints ("Alternative Model Constraints"):

[19] In litigation, it would be the plaintiff's burden to make this showing. We have focused our search to date on whether a less discriminatory alternative exists for Upstart's core AI/ML Model used to predict default and prepayment probabilities for each borrower. Those outputs are eventually translated into APRs and approval/denial decisions. Other stages in the process may be the focus of later reports, if warranted.

[20] *See* Monitor's Second Report, *supra* note 2, at 18-20. Other methodologies for identifying the existence of less discriminatory alternative models exist. We adopt this technique because it emulates characteristics of longstanding methods used on traditional models. *See* Monitor's Second Report, *supra* note 2, at 19. Upstart represents that it is conducting its own research utilizing different ML-powered approaches to identifying less discriminatory alternatives. As discussed below, we recommend that Upstart apply the methodology used in this Report or a different methodology Upstart develops, so long as we and Upstart agree it is comparable or more effective in its ability to identify less discriminatory alternatives.

[21] 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 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.

[22] Optimizing for both may warrant further exploration in future refinements. We also note an inherent limitation: The training data only includes loans that have already been previously approved by Upstart and accepted by applicants. The training data therefore does not show how any applicants that were denied loans by Upstart would have performed. Previous mispredictions or mischaracterizations of the risk of an applicant where the risk was a false-positive (meaning, the prior model predicted high risk of default but the applicant was truly a low-risk) will undermine current models (both the baseline and the alternative) because the false-positive risk rate of the models is not reflected in the performance of the model.

[23] *See* Monitor's Second Report, *supra* note 2, at 20-22.

1. First, we will not recommend adopting a potential alternative model if its performance is meaningfully worse than the performance of the Baseline Model. We discuss this condition in detail below.

2. Second, we will consider reasonable model risk management criteria in assessing whether to recommend an alternative model. For example, the alternative model should also satisfy reasonable validation metrics.

3. Third, we would not recommend an alternative model that introduces *new* statistically and practically significant disparities for other protected classes that were not present in the Baseline Model (for example, new statistically and practically significant disparities based on sex or age, assuming no such disparities existed for those classes in the Baseline Model).

4. Fourth, we would not recommend an alternative model that would exacerbate to a statistically significant extent *existing* statistically and practically significant disparities for other protected classes from the Baseline Model. In other words, if disparities in the Baseline Model are practically significant (*e.g.*, < 90% AIR), the alternative model should not worsen those disparities in a statistically significant manner.[24]

5. 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."

6. It is possible that multiple alternative models could satisfy the above constraints. In such situations, we may need to apply more case-specific criteria.

Finally, we test to ensure that improvements in disparities are statistically significant.[25] If disparities in improvements are statistically significant, we would not reject a less discriminatory alternative simply because improvements in disparities also do not meet some measure of *practical* significance. In contrast, we do apply a practical significance test in assessing whether disparities identified at Step 1 of the disparate impact analysis warrant proceeding to Steps 2 and 3 of the disparate impact analysis because there is a split among courts regarding whether a plaintiff must demonstrate practical significance to establish a prima facie case of disparate

[24] For a discussion of our approach for measuring the statistical significance between two disparities, see *infra* note 25.

[25] We assess the statistical significance of differences in disparities using a statistical method referred to as "Bootstrapping." Under this method, inferences about a population from sample data are made by resampling with replacement from the sample data and making an inference about the population from the resampled data. For example, we may not know the true error in a sample statistic against its population value. In bootstrap-resamples, the samples are the "population," and this population value is therefore known; accordingly, the quality of inference of the sample from resampled data is measurable. For more information on "Bootstrapping," *see* Wikipedia, the Free Encyclopedia, "Bootstrapping (statistics)," https://en.wikipedia.org/w/index.php?title=Bootstrapping_(statistics)&oldid=1104051630.

impact in litigation,[26] and 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.[27] However, we are unaware of agencies or courts applying such a requirement in assessing whether a less discriminatory alternative should be adopted, and such a measure appears inconsistent with agency and judicial formulations of the disparate impact standard, discussed below.

a. *Adequate Performance—Overview*

The first Alternative Model Constraint requires the most discussion—what standard should be used for determining whether the performance of a potential alternative model is acceptable? There is no universally applicable answer to this question. Our methodology here is informed by agency and judicial articulations of the disparate impact legal standard, our experience with internal compliance policies adopted by financial institutions, and our assessment of legal risk based on arguments specific to Upstart's circumstances.

In short, as discussed below, adoption of a less discriminatory alternative is necessary if that alternative can reasonably achieve an entity's legitimate business interests as well as the existing practice. We believe there is a significant likelihood that a court would find that a less discriminatory alternative model could serve Upstart's legitimate business needs as well as the Baseline Model if there is a reasonable probability that the performance of that alternative will fall within the likely performance range of the Baseline Model. Accordingly, we identify that range for the Baseline Model and presume an alternative model within that range would reasonably achieve Upstart's business interests as well as the Baseline. In other words, under this reasoning, it cannot be said that an alternative model within this range would result in a performance loss.

A court could take a different approach. It could, for example, hold that an alternative model need not be adopted unless an accuracy measure of the proposed alternative—here, Upstart's chosen accuracy measure—is equal to or exceeds the accuracy of Upstart's Baseline Model in a statistically significant manner. Our methodology, however, is designed to account for what we perceive to be a likelihood of a court concluding that a model need not meet that test to be viable because an alternative within the range we propose above has a reasonable probability of performing as well as the Baseline Model.

We also do not rule out the possibility that a regulatory agency or court might conclude that a different less discriminatory alternative, including one with greater performance metric

[26] *Compare, e.g., Jones v. City of Bos.*, 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 Dist.*, 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).

[27] *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.").

deterioration, would be required in some circumstances. In our experience some financial institutions, as a matter of internal compliance, would consider alternative models to be viable despite what are likely larger potential drops in model performance metrics than what we recommend here.[28] For example, an institution may not be able to argue plausibly that a drop in performance is unacceptable if evidence suggests the entity does not consider equivalent drops meaningful in other circumstances. The acceptability of performance differences may vary across institution types and business models.

b. *Adequate Performance—Legal Standard*

Neither agencies nor courts have delineated concrete thresholds for determining whether a less discriminatory alternative practice must be adopted because it sufficiently achieves a legitimate business need. That said, agency descriptions of disparate impact under ECOA and the FHA provide that adoption of a less discriminatory alternative is necessary if that alternative would reasonably serve the business interest as well as the practice at issue. Failure to adopt a comparable alternative that has less of a disparate impact could be illegal. For example, the Official Interpretations of Regulation B state that a practice could be illegal if a business need can "reasonably be achieved as well by means that are less disparate in their impact."[29] Similarly, agency guidance documents explain that policies with a disparate impact may be illegal "if an alternative policy or practice could serve the same purpose with less discriminatory effect."[30] Put another way, a violation may exist if an "alternative that is approximately equally effective is available that would cause less severe adverse impact."[31] Illegal disparate impact may occur in a credit scoring system, for example, when a factor contributes to adverse disparate impact and "improves the model's ability to predict risk, but only marginally so,"[32] or if "the business necessity can be achieved by substituting a comparably predictive variable that will allow the credit scoring system to continue to be validated, but also operate with a less discriminatory result."[33]

We are unaware of case law addressing the sufficiency of less discriminatory alternatives in circumstances directly comparable to the testing in this Monitorship. That said, the agency articulations above are consistent with common judicial formulations in disparate impact case law generally. Courts use a range of phrases to describe whether less discriminatory alternatives

[28] *See infra* notes 40-41 and accompanying text.

[29] Regulation B Official Staff Commentary, 12 C.F.R. pt. 1002, Supp. I, .6(a)-2.

[30] HUD, DOJ, OCC, OTS, Fed. Rsrv. Bd., FDIC, FHFB, FTC, NCUA, Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18266, 18269 (Apr. 15, 1994) ("Joint Policy Statement on Discrimination in Lending"); OCC, FDIC, Federal Reserve Board, OTS, NCUA, Interagency Fair Lending Examination Procedures at iv (Aug. 2009), https://www.ffiec.gov/pdf/fairlend.pdf ("FFIEC Interagency Procedures"); Federal Housing Finance Agency, AB 2021-04 at 8 (Dec. 20, 2021) ("FHFA Bulletin"), https://www.fhfa.gov/SupervisionRegulation/AdvisoryBulletins/AdvisoryBulletinDocuments/AB%202021-04%20Enterprise%20Fair%20Lending%20and%20Fair%20Housing%20Compliance.pdf.

[31] FFIEC Interagency Procedures, Appendix, *supra* note 30, at 27.

[32] FHFA Bulletin, *supra* note 30, at 8.

[33] OCC Bulletin 1997-24, "Credit Scoring Models: Examination Guidance," Appendix at 11 (May 20, 1997), https://www.occ.treas.gov/news-issuances/bulletins/1997/bulletin-1997-24.html. EEOC guidance in the employment context uses similar language, explaining that when two or more procedures are available that serve a legitimate interest and "which are substantially equally valid for a given purpose," the one with lesser adverse impact should be chosen. 29 C.F.R. § 1607.3(B).

would adequately achieve legitimate business needs, such as whether alternatives are "viable" or whether they "serve" or "advance" a legitimate need.[34] Some courts explain that an alternative practice must be "equally effective" as the practice at issue in achieving the legitimate business goals—a phrase first used to interpret a since-amended version of Title VII in employment law.[35] There is some debate about the meaning of this language and whether it appropriately describes the disparate impact standard. In 2013, for example, the Department of Housing and Urban Development ("HUD") declined to adopt this language in its Rule implementing the disparate impact standard under the FHA. Instead, HUD's Rule states liability may exist if the interests supporting the challenged practice "could be served by another practice that has a less discriminatory effect."[36] In declining to adopt the "equally effective" language, HUD explained that it does not agree that *Wards Cove* (the genesis of this language) governs FHA claims, and that HUD's chosen articulation of the standard is consistent with "the [Agencies'] Joint Policy Statement [on Discrimination in Lending], with Congress's codification of the disparate impact standard in the employment context, and with judicial interpretations of the Fair Housing Act."[37] In the FHA context, some courts have relied on HUD's Rule to reject application of the "equally effective" language.[38] Other courts that have used the phrase "equally effective" acknowledge it may mean something other than identical.[39] Courts sometimes vacillate between descriptions of alternatives as being "equally effective" and as "viable alternative practice[s] that would serve

[34] *See, e.g., Tex. Dep't of Hous. & Cmty. Aff.s v. Inclusive Cmtys. Project*, 576 U.S. 519, 533 (2015) (explaining in FHA case that an alternative must have "less disparate impact and serve[] the [entity's] legitimate needs" (quoting *Ricci v. DeStefano*, 557 U.S. 557, 578 (2009)); *Mt. Holly Gardens Citizens in Action, Inc. v. Twp. of Mt. Holly*, 658 F.3d 375, 382 (3d Cir. 2011) (stating in FHA case that plaintiffs "must demonstrate that there is a less discriminatory way to advance the defendant's legitimate interest"); *Darst-Webbe Tenant Ass'n Bd. v. St. Louis Hous. Auth.*, 417 F.3d 898, 906 (8th Cir. 2005) ("[T]he plaintiffs must offer a viable alternative that satisfies the Housing Authority's legitimate policy objectives while reducing the revitalization plan's discriminatory impact."); *Allen v. City of Chi.*, 351 F.3d 306, 313 (7th Cir. 2003) ("To prevail, the officers therefore must demonstrate that an increased percentage of merit-based promotions would be of substantially equal validity as merit-based promotions."); *Newark Branch, NAACP v. Town of Harrison, N.J.*, 940 F.2d 792, 798 (3d Cir. 1991) (stating plaintiffs may prevail "where they are able to suggest a viable alternative to the challenged practice"); *Huntington Branch, NAACP v. Town of Huntington*, 844 F.2d 926, 937-39 (2d Cir. 1988) (analyzing whether the defendant's goal "can be achieved by less discriminatory means"), *aff'd*, 488 U.S. 15 (1988) (per curiam), *superseded by regulation on other grounds, as stated in Mhany Mgmt., Inc. v. Cnty. of Nassau*, 819 F.3d 581, 618 (2d Cir. 2016).

[35] *See, e.g., Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989); *see also Sw. Fair Hous. Council, Inc. v. Maricopa Domestic Water Improvement Dist.*, 17 4th 950, 970 (9th Cir. 2021) (quoting *Wards Cove*, 490 U.S. at 661); *Hardie v. Nat'l Collegiate Athletic Ass'n*, 876 F.3d 312, 315 (9th Cir. 2017).

[36] 24 C.F.R. § 100.500(c)(3) (2013) (emphasis added). This 2013 Rule language is operative. In 2020, HUD published a rule that would have made certain amendments to the 2013 Rule, but those 2020 changes were preliminarily enjoined and never took effect. In 2021, HUD proposed a new rule recodifying the 2013 Rule, which remains in effect due to the preliminary injunction. *See* HUD Proposed Rule, Reinstatement of HUD's Discriminatory Effects Standard, 86 Fed. Reg. 33590 (June 25, 2021).

[37] HUD Final Rule, Implementation of the Fair Hosing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460, 11473 (Feb. 15, 2013).

[38] *See Mhany Mgmt., Inc. v. Cnty. of Nassau*, No. 05-cv-2301 (ADS) (ARL), 2017 WL 4174787, at *8 (E.D.N.Y. Sept. 19, 2017) ("HUD's interpretation could not be clearer that a plaintiff's burden . . . is not to show that the less discriminatory practice would be equally effective, but merely that it must serve a defendant's legitimate interests.").

[39] *See, e.g., MacPherson v. Univ. of Montevallo*, 922 F.2d 766, 773 & 773 n.2 (11th Cir. 1991) (interpreting *Wards Cove* to mean a plaintiff must show, "at the very least," that an alternative practice is "economically feasible" and declining to decide "whether an alternative practice that is economically feasible but is still more expensive than the employer's current practice can be 'equally effective' within the meaning of *Wards Cove*"); *Cureton v. Nat'l Collegiate Athletic Ass'n*, 37 F. Supp. 2d 687, 713 (E.D. Pa.) ("'[E]qually effective' means equivalent, comparable, or commensurate, rather than identical"), *rev'd on other grounds*, 198 F.3d 107 (3d Cir. 1999).

[the entity's] needs," suggesting that differences in terminology may not translate to different substantive applications.[40]

Regardless of the specific judicial formulation, as discussed below, we believe there is a likelihood of a court finding that a less discriminatory alternative model could serve Upstart's legitimate business needs as well as its Baseline Model if there is a reasonable probability that the performance of the alternative will fall within the likely performance range of the Baseline Model, and align our recommended approach with that likelihood.

c. *Adequate Performance—Internal Compliance Standards*

Practices differ across financial institutions conducting internal fair lending testing in terms of whether they will consider a less discriminatory alternative to be viable, in the sense that it reasonably serves their business needs. Some institutions 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.[41] These institutions have made the 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 general model training, development, and risk 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.[42] In other words, these institutions have determined that if performance deterioration is not significant enough to warrant rebuilding a model then it is unlikely to be significant as a real-world matter and so is not considered significant enough to warrant rejecting a less discriminatory alternative model.

In the next sections, we discuss Upstart's Model performance metrics and the basis for our recommended approach here.

[40] *Freyd v. Univ. of Oregon*, 990 F.3d 1211, 1227 (9th Cir. 2021).

[41] See Monitor's Second Report, *supra* note 2, at 20-21. 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.

[42] 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. The concept of out-of-time validation is discussed more in Section B.3.f.

d. *Adequate Performance—Upstart's Metrics for Assessing Model Performance*

Our recommended approach is informed by the fact that there is inherent uncertainty in how any model will perform across datasets.[43] Measuring accuracy on a model training dataset does not guarantee that the model will perform at the same accuracy when used to evaluate different populations—for example, actual applicants once the model is deployed. Accordingly, modelers use techniques to try to ensure that models are accurate and that they will be reasonably valid across different populations and future time periods. The fact that there is inherent and measurable uncertainty around how a model will perform suggests that variations in model performance across alternative models may not be meaningful and alternative models within those bounds of uncertainty are likely to be similarly effective at meeting business needs.

Upstart relies on a primary performance metric that reflects both default risk and prepayment risk; for consistency with Upstart's practices, we rely on the same metric for our analyses, which we refer to as the Error Metric.[44] A smaller Error Metric value is indicative of a more accurate model, but is not indicative of the uncertainty of the model (*i.e.*, how certain is this accuracy when the model will be used in the future where conditions drift).

To help ensure the validity of its Model, Upstart trains and validates its Model using an accepted modeling process called "k-fold cross-validation," which ultimately helps Upstart calculate an Error Metric for the Model. K-fold cross-validation means Upstart trains its Model on a subset of a borrower dataset (*i.e.*, the training set) and then evaluates the Model's performance on the rest of the data in the dataset that are not used in training (*i.e.*, the validation set). Upstart randomly divides the dataset into k-number of equal-sized subsamples (*i.e*., folds). The letter "k" in "k-fold cross-validation" signifies the number of "folds" used for validation. For 5 folds, for example, one would fit the model on a training set comprised of 4 of those folds, and then measure the model's performance by making predictions on the single left-out fold (the validation set). That process would be repeated 5 times, so that each of the 5 folds is left out and used for validation once.

Upstart calculates the Error Metric for each of those cross-validations. Inevitably, the Error Metric for each individual cross-validation will vary: the Model will perform better on

[43] *See generally, e.g.,* Utakarsh Sarawgi, et al., Uncertainty-Aware Boosted Ensembling in Multi-Modal Settings (Apr. 2021), https://arxiv.org/pdf/2104.10715.pdf; Andrey Malinin, et al., Uncertainty in Gradient Boosting via Ensembles, ICLR 2021, https://openreview.net/pdf?id=1Jv6b0Zq3qi; Martin Krzywinski & Naomi Altman, Importance of Being Uncertain, Nature Methods, Vol. 10, No. 9 p. 809 (Sept. 2013), https://www.nature.com/articles/nmeth.2613.pdf; Zoubin Ghahramani, Probabilistic Machine Learning and Artificial Intelligence, Nature 521:452-459 (May 2015), https://www.repository.cam.ac.uk/bitstream/handle/1810/248538/Ghahramani%202015%20Nature.pdf; Daniel Huynh, Bayesian deep learning with Fastai, Toward Data Science (Nov. 29, 2019), https://towardsdatascience.com/bayesian-deep-learning-with-fastai-how-not-to-be-uncertain-about-your-uncertainty-6a99d1aa686e.

[44] We do not describe the Error Metric in more detail to avoid disclosing proprietary or commercially sensitive information.

some validation datasets than on others. Accordingly, Upstart calculates a single *average* Error Metric representing all of the cross-validations.

Upstart then treats potential model updates as a decision between two competing models: the current Baseline in production and a new candidate model. Upstart compares a version of the Baseline Model retrained on the same newer training data used for the candidate model. Upstart performs cross validation to assess the performance of a candidate model using the same cross-validation procedures described above for the Baseline Model, with both models being trained and evaluated on an identical set of folds. Upstart then evaluates several (at least 15) paired differences of model performance. Finally, it statistically tests if the difference is above zero, meaning the candidate model statistically outperforms the Baseline model. This is Upstart's process for determining whether a candidate model performs statistically significantly better than the Baseline model. If it does, Upstart will generally proceed with the model update.

e. *Adequate Performance—Establishing a Viable Range*

How do we assess what level of change in the Error Metric would still result in a viable model? As noted, our recommended approach is based on the uncertainty associated with Upstart's Model. At a high level: on any single validation dataset, the probability is infinitesimal that the performance of Upstart's Model will be exactly equal to the average Error Metric; it may be higher or lower. But we can say with quantifiable probability that the performance on that new dataset will fall within a range that we call the "Uncertainty Interval."[45] If the Error Metric of an alternative model (when applied to the same validation dataset) is reasonably likely to yield an Error Metric that will fall within that Uncertainty Interval, there is a strong argument it could reasonably serve the same purpose as the Baseline Model.

Put another way, we believe there is a likelihood that a court would find that a less discriminatory alternative model could serve Upstart's legitimate business needs if there is a reasonable probability that the performance of that alternative will fall within the likely range of the Baseline Model. In deciding to adopt that Baseline Model, Upstart implicitly determined that performance within that range was acceptable for meeting its business purpose. It would be difficult to argue that an alternative model could not serve the same business purpose if there is a reasonable likelihood that the alternative model will perform at a level that was considered acceptable for the Baseline Model. Under this reasoning, it cannot be said that an alternative model within this range would result in a performance loss.

To explain the standard in more detail, recall that the *average* Error Metric is a product of the Error Metrics across several validation "folds." Although the average Error Metric across folds is a reasonable shorthand for understanding model performance, exclusive reliance on the

[45] In other words, based on the Error Metric for each of the folds, we can compute the probability that if we randomly selected a new fold, the model's Error Metric on that fold would fall in some range with 95% probability. Technically, this is a probability range, not a confidence band. But they are computed exactly the same and we refer to this probability range as an "Uncertainty Interval."

average Error Metric does not fully capture the uncertainty inherent in the Model. To better understand the uncertainty in the measure of Upstart's Model performance, we performed a statistical analysis of the uncertainty in the average Error Metric of 5 folds of Upstart's Model. Namely, we determined the standard deviation for the Error Metric and the standard error of the average Error Metric determined from cross-validation. In statistics, the standard deviation is a measure of the amount of variation in a set of values and is based on the square of the distance of each value from the mean value. For a normal distribution, about 68% of values fall within one standard deviation of the mean and 95% of values fall within two standard deviations of the mean.[46] The standard error is the standard deviation of the average Error Metric.

At the time of our analyses, Upstart reported to us an average Error Metric. We determined that the standard error of the average Error Metric is 10 points and hence there is a 95% Uncertainty Interval band of + or –20 Error Metric points from that average. In other words, if we validated the model using a different set of 5 folds validation data, it would be unlikely that the average Error Metric would be exactly the reported average Error Metric. But there would be a 95% chance that the average Error Metric value would fall within + or –20 points of that average. And there would be a 68% chance that the value would fall within the narrower range of + or –10 points of that average. These ranges are useful because they give a sense of what range of Error Metric is likely to satisfy Upstart's legitimate business interests. That is, assuming that the applicant pool in production is similar to the training and validation datasets, Upstart would have a 68% chance that the Error Metric will be somewhere within + or –10 points of the average Error Metric (and a corresponding 32% chance that the Error Metric will be outside that range). Upstart would have a higher probability (95%) that the Error Metric will be somewhere within + or –20 points of the average Error Metric (and a 5% likelihood that the Error Metric will be outside that range).[47]

Presumably, before it deploys a model Upstart understands there is some uncertainty regarding how that model will perform and implicitly makes a business decision that performance within that range would reasonably achieve its legitimate business interests. We believe there is a likelihood that a court would find that Upstart's legitimate business interests could "reasonably be achieved as well"[48] by a model whose expected range of predicted Error Metric values generally falls within that expected range of the Baseline Model.[49] An argument

[46] *See* Krzywinski & Altman, Importance of Being Uncertain, *supra* note 43, at 809.

[47] Our approach for measuring uncertainty is designed to be relatively straightforward and practical to apply. More sophisticated methods exist and could warrant further exploration.

[48] Regulation B Official Staff Commentary, 12 C.F.R. pt. 1002, Supp. I, .6(a)-2.

[49] FFIEC Interagency Procedures, Appendix, *supra* note 30, at 27 (emphasis added). This approach of adopting a less discriminatory alternative based on an acknowledgment of inherent uncertainty in validation is analogous to the practice of "banding" test scores in the employment context, where scores falling within certain "bands" are treated as identical because granular score differences are considered insignificant. *See Chi. Firefighters Local 2 v. City of Chicago*, 249 F.3d 649, 656 (7th Cir. 2001) ("[Banding is] a universal and normally an unquestioned method of simplifying scoring by eliminating meaningless gradations."); *Officers for Justice v. Civil Serv. Comm'n*, 979 F.2d 721, 723-24, 728 (9th Cir. 1992) ("Differences between scores within the band are considered to be statistically insignificant due to measurement error inherent in scoring the examination. . . . The district court did not clearly err in finding that banding, as proposed by the City, is more valid, or at least 'substantially equally valid' to rank order

could be made for relying on a 95% Uncertainty Interval (here, +/–20 points) and choosing a less discriminatory alternative within that range. We rely on the narrower 68% Uncertainty Interval (here, +/–10 points) because the counterarguments that an alternative model within that range would not be viable are weaker. Uncertainty exists in the average performance estimate of the candidate alternative model, just like the Baseline Model, and there would necessarily be a higher likelihood that the alternative model would not perform within the range of the Baseline using a 95% Uncertainty Interval, suggesting a narrower range is likely more realistic.

As noted, when we performed a statistical analysis of the uncertainty in the average Error Metric of Upstart's Model, the Uncertainty Interval was 10 Error Metric points. Upstart represents that when it runs its paired difference approach it uses at least 15 folds. Applying our methodology to a 15-fold sample of performance estimates, we would calculate an Uncertainty Interval closer to 5 Error Metric points.[50] As discussed below, the difference between 5 and 10 Error Metric points ultimately does not affect the choice between potential less discriminatory alternative models in this Report. More broadly, we rely on these numbers to guide the analyses in this Report, but do not prescribe a numerical threshold for analyses going forward. The key principle is that when Upstart adopts a model update because of likely expected performance, it implicitly concludes that the corresponding expected range of performance is acceptable, and we recommend that it consider alternatives to be viable if they are also likely to perform within that range (measured, at a minimum, by the 68% Uncertainty Interval).

Accordingly, we would recommend an alternative model that has a less discriminatory effect if its average Error Metric falls within the 68% Uncertainty Interval (equal to the standard error) of the Baseline cross-validation, and the other Alternative Model Constraints noted above are satisfied (*e.g.*, the alternative model does not introduce new practically significant disparities, etc.).

A court might take a narrower view of disparate impact and hold that if an alternative model is not equal to or statistically significantly better than the Baseline Model using Upstart's paired cross-validation methodology for assessing models, then it should not be considered a viable alternative model. In other words, a less discriminatory alternative would be considered viable only if it is statistically significantly equal to or better than the Baseline Model using Upstart's own methodology for assessing model updates. Arguments that such a model would not be viable would be very weak because the model would be equivalent by Upstart's own

promotions."). These Title VII cases are informative because they address scenarios where, because of uncertainty in evaluation scores, quantified differences in evaluation scores are considered insignificant because they represent meaningless gradations—a scenario similar to the proposition that slight differences in a measurement of model accuracy, while quantifiable, may not be meaningful.

[50] This 5 point Uncertainty Interval is likely overly-narrow because each seed does not produce a set of 5 fold Error Metrics that are independent of the fold results from another seed.

metrics, and if it were better Upstart would presumably adopt such a candidate model regardless of whether that model decreased disparities.[51]

However, we believe there is a likelihood that a court would conclude that an alternative model would meet Upstart's legitimate business needs as well as the Baseline even if it did not equal or exceed the Baseline Model on that metric, because that metric does not account for whether the magnitude of the difference is likely to matter in real-world performance. Even if there is a statistically significant difference in Upstart's accuracy measure, that difference might be so small as to not be meaningful in a real-world sense. In our view, certain tests we conducted lend support to that position.[52] The fact that Upstart deploys its Model with an even greater range of uncertainty than we would apply to consider an alternative model viable further supports that position.[53] The Uncertainty Interval methodology asks: is there a reasonable probability that the performance of an alternative model will fall within the likely performance range of the Baseline Model? Because performance within that range presumably meets Upstart's legitimate business needs, we believe there is a likelihood that a court would conclude that an alternative model with a reasonable probability of performing within that range would reasonably achieve Upstart's legitimate business needs as well. Acknowledging that courts may differ in their approaches, we adopt this methodology because of this perceived likelihood.

f. Adequate Performance—Out of Time Analysis

To assess the proposition that relatively small differences in the Error Metric on development data may not reliably translate to meaningful differences in performance once a model is deployed, we conducted an analysis of models on "out of time" ("OOT") data. In this experiment, we observed that, even though a less discriminatory alternative model had worse average performance than a baseline in development, the performance of the models on OOT data was very similar, and on some partitions of OOT validation data the alternative model actually performed better than the baseline. This analysis further suggests that an alternative model that performs within this relatively narrow range is likely to reasonably achieve a legitimate business end.

[51] This assumption may depend on the improvement being at least 1 point of the Error Metric. Upstart notes that in addition to its statistical test on accuracy it typically imposes a 1-unit Error Metric threshold on the average performance gain of a candidate model to justify the costs of a model update.

[52] *See* the Out of Time analysis described in Section B.3.f and the default-risk Out of Sample analysis described in Section B.3.h.

[53] Conceptually it is useful to think of three forms of uncertainty: First, it can be difficult to determine the average accuracy correctly. This is sometimes called uncertainty arising from latent noise. Second, uncertainty arises from pure chance—once one determines an average score, any deployment of the model on a new dataset is unlikely to land on that average, even if the dataset is similar to the data the model was trained on. This is sometimes called idiosyncratic trends. The Uncertainty Interval is generally designed to account for these first two types of uncertainty, and, as noted, we rely on a narrower 68% Uncertainty Interval instead of a 95% Uncertainty Interval. Third, uncertainty arises as the model is deployed on new populations and under new socioeconomic conditions. This is sometimes called concept drift and is what our Out of Time analysis discussed in the next section explores. Similar to the body of work addressing the explainability and interpretability in AI-systems, there is a growing body of work recognizing the importance of evaluating uncertainty of AI-based systems in a trustworthy manner. *See, e.g.,* sources cited *supra* note 43.

As background, models can be validated on “in-time” out-of-sample (“OOS”) data or OOT data. In-time OOS data simply means that part of the data used to develop the model are randomly “held out” for validation: the model is trained on one subset (say, 70%) and then tested on the held out OOS subset (the remaining, 30%) to confirm the model is accurate. In contrast, OOT data contain data from an entirely different time period than what was used for model development. OOT validation shows model robustness over different populations and may help give a sense of how the model will perform in production, as populations and socioeconomic factors shift over time. Recall that the Uncertainty Interval discussion above assumed that the model in production would operate on an applicant pool that is very similar to the development pool; that assumption often does not hold because the macroeconomic circumstances surrounding even the same demographic pool of applicants drifts over time, which causes the performance on OOT validation data to vary from performance observed on in-time OOS data.

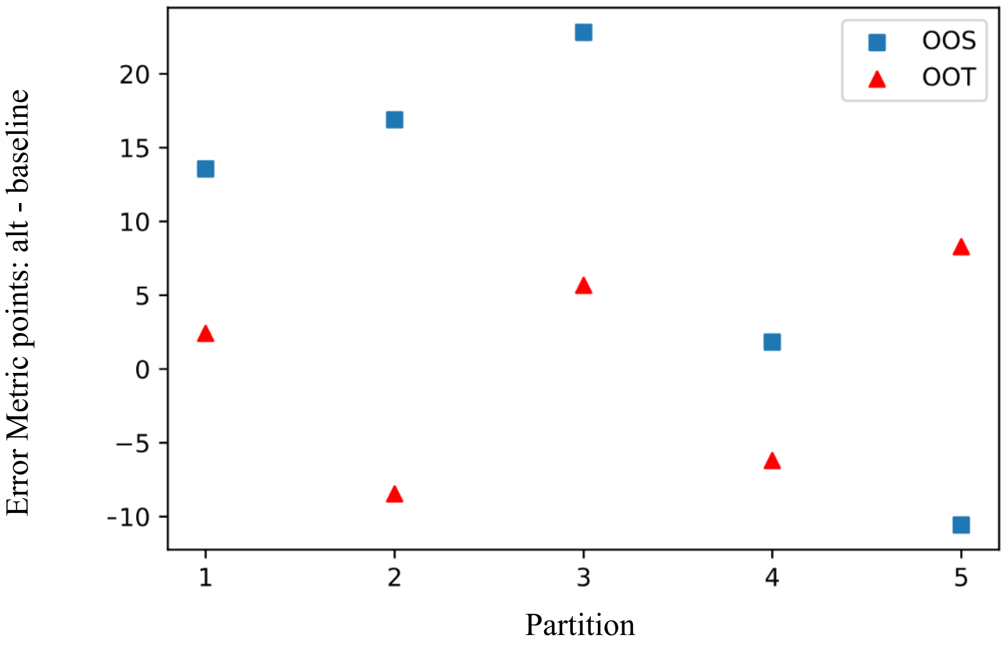
For our analysis, we trained a baseline model on approximately four years of Upstart borrower data from pre-December 2018, using all of Upstart’s variables at the time of analysis. We then trained an alternative model with improved disparity metrics for Black borrowers. This less discriminatory model was trained on the same borrowers as the baseline model. We assessed the performance of both models on OOS and OOT data. The OOS results are based on an in-time set of borrowers that received 3- and 5-year loans in 2014-2018, and the OOT results are based on an out-of-time set of borrowers that received 3- and 5-year loans in Q1-2019.[54]

On the OOS set, the alternative model had an Error Metric score that was **8.9** points worse than the baseline model. However, on the OOT dataset, the alternative model had an Error Metric score that was only **0.4** points worse than the baseline. In other words, the OOT performance of the alternative model was very similar to the OOT performance of the baseline model.

We also observed that the baseline model did not consistently outperform the alternative model when looking at pair differences on partitions. On the chart below, the X-axis represents 5-partitions of validation. The Y-axis represents the difference in Error Metric score between the alternative and the baseline. A difference below 0 signifies that the alternative outperformed the baseline on that partition. As shown below, the alternative model outperformed the baseline model on 2 out of 5 OOT partitions, despite slightly worse average Error Metric (+0.4 points).

[54] This analysis differs from a true OOT analysis because information from outside the training set was used to create the alternative model: we used the entire borrower dataset through Q1 2021 to identify features to remove, and then retrained the alternative model on data prior to December 2018. Because of that characteristic, the analysis might instead be analogized to an “in-sample goodness-of-fit” (ISGF) analysis, comparable to statistical model quality assessments performed on linear models, such as statistical model quality assessment metrics like R^2 and p-values. These metrics are not computed on any hold-out sample data, but on the entire dataset, including the data used to train the model. These ISGF metrics provide another means for comparing the relative fitness of two or more models that are applied to the same data. We refer to our analysis as an OOT analysis for simplicity, even though it lies somewhere between traditional OOT and ISGF methods.

Partition-wise Error Metric differences



This analysis has some inherent limitations. First, we trained a single XGBoost model using Upstart's variables, but could not conduct this OOT experiment using Upstart's full Model because that Model is trained on Upstart's entire borrower population (not just the pre-2018 data we used to train the "baseline" model to allow for this OOT experiment). Second, information from outside the training set had to be used to create the alternative model. Here, we used the entire Upstart borrower dataset through Q1 2021 to identify features to remove, and then retrained the alternative model on data prior to December 2018, meaning information from the evaluation sets (both OOT and in-time OOS) necessarily "leaked" into the alternative model's training data. This is a technical limitation, but whether it impacts the results, or the direction of impact, is not clear. Correcting this issue would introduce a new problem because it would require identification of features for removal using only the pre-December 2018 dataset, which would severely limit the data available for the test. Third, to conduct this test we necessarily assess OOT alternative model performance on Upstart's existing loan portfolio, but loans are included in that set because their perceived risk at the time of origination was below an acceptable threshold. Relatedly, we do not know the loan performance of applicants that were rejected; the baseline and any alternative models will not capture rejected applicants that would not have defaulted. However, adjusting for this limitation is not practical for this analysis because it is very difficult to realistically simulate how loans not originated on Upstart's platform would perform.

With those limitations in mind, we interpret this OOT analysis with some caution. At the same time, in our view it supports the proposition that two models within this relatively narrow range of performance difference are likely to perform similarly, and that a less discriminatory alternative model within this range could reasonably serve the same legitimate goals as a baseline model. Here, despite an in-time 8.9 point difference in performance, the two models were nearly identical in their OOT performance, and the baseline did not consistently outperform the alternative.

g. *Translating Model Performance to Business Metrics*

The prior sections establish a presumed range of comparable performance based on a likelihood that a court would find that a less discriminatory alternative model within this range could serve Upstart's legitimate business needs. As we assessed potential alternative models, we also separately conducted analyses to estimate the expected change in business metrics from adopting an alternative model assuming that there was *no* uncertainty in how the models would behave. Assessing how model performance differences could affect business metrics even assuming no uncertainty could illuminate what performance differences might mean as a practical matter, assuming comparable performance differences persist. It can also inform decisions about alternative models that fall outside the Uncertainty Interval: even if there is a higher likelihood that such alternative models will underperform as compared to the Baseline, would that degree of underperformance be meaningful as a business matter? We treat these analyses as secondary to our uncertainty analyses above because they suffer from a serious limitation: they assume that the Baseline model represents the ground-truth, which we know not to be true.

That said, to assess these possible changes on business metrics, we look at changes in:

(1) Expected net present value (NPV) profit of a loan pool: This metric can be thought of as the expected profit from the performance of a pool of loans. The direct impact from loan underperformance (either because of default or prepayment), measured by NPV Profit, falls largely on Upstart's lender and investor partners. These institutions stand to lose profit if a pool of loans underperforms. Conversely, if a pool of loans overperforms, lender and investor partners benefit. Loan underperformance will also indirectly impact Upstart because of relationship harms with partners. Not meeting investor expectations will create negative business impacts to Upstart, and a sustained period of meaningful loan underperformance would likely cause lenders and investors to invest their capital elsewhere.

(2) Fees to Upstart: Upstart receives revenue in the form of fees from lender and investor partners for each applicant that accepts an offered loan. Changes in these fees are a direct cost or benefit to Upstart. These fees vary but Upstart provided a per acceptance fee estimate for use in our analyses.[55]

[55] We do not include that per acceptance amount to avoid disclosing proprietary or commercially sensitive information.

Of course, there may be countervailing regulatory, reputational, and other benefits to partner institutions or Upstart from adopting a less discriminatory model, which we do not attempt to quantify.

h. *Less Discriminatory Alternative Results*

We identified a number of potential alternative models for consideration. We focused our analysis on one core component of Upstart's AI/ML Model, which is a primary driver of the results of Upstart's Model.[56] Improving disparities caused by this core component of the Model should improve downstream disparities in pricing and approval/denial rates. But those downstream disparities are also affected by other elements of the process and will vary depending on circumstances such as adjustments Upstart makes to model outcomes to account for things like economic cycles, pricing strategies and approval cut-offs (which vary by lending and institutional partner), and applicant populations (which also vary by partner). The results below are useful to illustrate the directional and likely relative effects of alternative models, but they are not concrete representations of these effects, which will vary across circumstances.[57]

Many iterations of training produced numerous potential alternative models, most of which we ruled out because their Error Metric increases were outside even a 95% Uncertainty Interval or they did not improve AIRs as compared to the Baseline or other alternative models.

We focus on two potential alternative models. What we refer to as "Model 2" results in the largest AIR improvement for Black applicants within the Uncertainty Interval range of +10 Error Metric points (Model 2 falls within +4.8 points).[58] The performance of Model 2 is also within the narrower +5 Error Metric points mentioned above. Under Model 2, AIRs for Black and Hispanic applicants improve and AIRs for other groups remain above parity (and well above practical significance, as defined here).[59]

What we call "Model 4" would result in more significant AIR improvements for Black and Hispanic applicants than Model 2, and, in this analysis, it would bring the AIRs for all groups to levels that we would consider to be practically insignificant (*i.e.*, ≥ 90%).[60] But its performance would be outside the 68% Uncertainty Interval range relied upon above (Model 4 falls within +15.3 Error Metric points) and so we would not recommend it. That said, we do not

[56] The results in this section were calculated on 100,000 random applicants from Q1 2021 using a pricing engine provided by Upstart. A sample was used to allow for efficient calculations; the sample size was chosen to ensure representativeness.

[57] Upstart uses different pricing strategies in different circumstances. To assess disparities, we rely on a pricing strategy provided by Upstart at the outset of our analyses, which allows us to compare estimated differences between models. Use of a different pricing strategy would result in different absolute numbers, although relative differences may be similar because the choice of pricing strategy affects both the Baseline and any potential alternatives. As noted in note 62, *infra*, a different pricing strategy is used to assess business impacts.

[58] Many specific testing results in this section are not included to avoid disclosing proprietary or commercially sensitive information.

[59] These improvements in AIRs are statistically significant. *See supra* note 25.

[60] These improvements in AIRs are also statistically significant.

rule Model 4 out entirely because it is within the 95% Uncertainty Interval and is useful for comparison.

Turning from approval/denial AIR disparities to pricing, APR disparities as measured by SMD would also improve for Black borrowers under both Models 2 and 4. For all groups, APR disparities remain well under practical significance thresholds (*i.e.*, 0.30) in this analysis.[61]

We also observe that the mean APR rates for Black borrowers would also improve to some extent under either alternative, and both alternatives would lead to increased approval rates for Black applicants.

We also assessed potential business metric impacts of these alternatives, assuming no uncertainty in the models' performance—in other words, assuming that the Baseline Model accurately represents the ground-truth of what will happen with borrowers. The expected *direct* economic impact to Upstart of either alternative Model 2 or Model 4 would be positive because both models result in an increase in applicants that accept offered loans. In this analysis, these acceptances would increase for Model 2 as compared to the Baseline Model by 3.7%. These acceptances would increase for Model 4 by 10.5% as compared to the Baseline Model. Accordingly, under either scenario, Upstart would receive a direct economic benefit via increased revenue from per applicant fees from partners.

There could also be potential direct effects on lender and investor partners, and corresponding potential indirect effects on Upstart. Assuming no uncertainty, Upstart partners might expect about a 3.6% decrease in expected NPV Profit under Model 2, as compared to the Baseline. They might expect about a 12.2% decrease under Model 4.[62] A court might consider those differences as evidence that potential alternative models would not meet Upstart's business needs as well as the Baseline Model.[63] As with other analysis in this report, this one is inherently limited. For example, and principally, it relies on the assumption that the Baseline Model is the ground truth for accuracy—an assumption we know not to be true. As discussed above, there is a reasonable probability that Alternative Model 2 would perform similarly to the Baseline Model. This analysis is also not able to account for performance of applicants that Upstart's Model

[61] These comparisons are done on the set of applicants who are approved by both the alternative and the Baseline models.

[62] These metrics are assessed on 567K applicants from Q1 2021, with results normalized by 100K to reflect the same scale as the 100K applicants used in the disparity analyses above. A number of assumptions and constraints are applied in this analysis: First, these results are based on a different pricing strategy than is used for the disparity results above because Upstart argues the pricing strategy here is more realistic for measuring business impacts. Second, the cashflows used (meaning the amount the partner receives) are those after subtracting fees that Upstart charges to the investors. Third, we use a discount rate of 1.5%, which loosely approximates the Federal discount rate.

[63] Potential business impacts can also be represented through an Internal Rate of Return (IRR), which is a metric used to estimate the profitability of a potential investment. Upstart loan investors set a target IRR for the loans they fund. An internal rate of return is the annual rate of growth than an investment is expected to generate. When Upstart prices loans, it does so with the intent of meeting the investor's target rate of return. Model 2 would result in an IRR that is 0.45 percentage points less than the Baseline, and Model 4 would result in an IRR that is 1.27 percentage points less than the Baseline.

denied, which could offset these results, especially because our analyses suggest that Upstart's Model overpredicts default (and in fact, Alternative Models 2 and 4 would also likely overpredict default).[64] In other words, this analysis does not capture the potential revenue from applicants the models would decline but that would not default. Accordingly, a court might be skeptical of the assumption that the Baseline model should be considered the ground truth and, in turn, interpret these results cautiously.

Finally, in assessing these alternatives, we consider whether they might inadvertently harm consumers because these expansions of credit might be to borrowers more likely to default. Exploring that potential concern further supports that Model 2, in particular, would likely be a viable alternative. Recall that Upstart's Model predicts both default and prepayment risks. Potential concerns about applicants' ability to repay loans relate only to prediction of default. Accordingly, we explore how these two alternative models would perform as compared to the Baseline Model with respect only to predicting borrower default on an out-of-sample dataset of Upstart's borrower data from 2013-2021.[65] To do so, we look at the "F1 scores" of the three models. An F1 score considers both the "precision" and the "recall" of a model: "precision" measures the percentage of correctly predicted defaults within all of the *predicted* defaults, whereas "recall" measures the percentage of correctly predicted defaults within all of the *actual* defaults.[66]

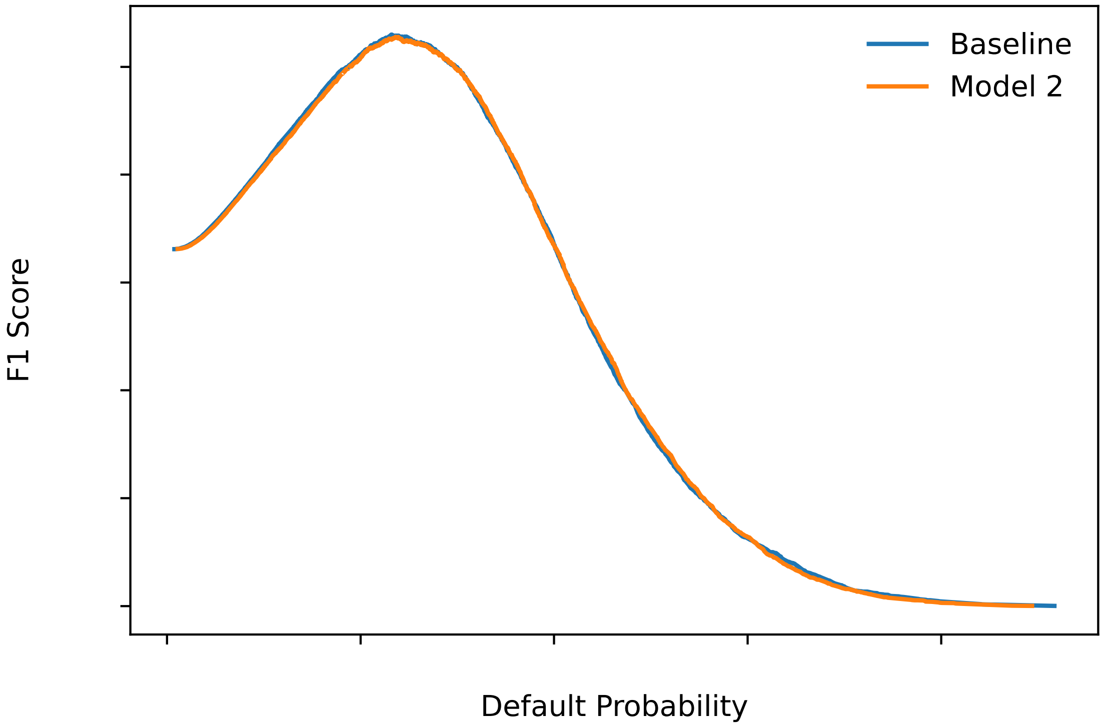
As shown in the figure below, for all default risk ranges Model 2 has the same performance in identifying defaulters as the Baseline Model, indicating that Model 2 and the Baseline Model are essentially identical in their ability to predict default.[67]

[64] We conducted an in-sample goodness-of-fit analysis of borrower default prediction errors on borrowers from Q1-Q4 2017 and Q1 2018 for the Baseline Model, as well as Alternative Models 2 and 4. We found that all three models overpredicted defaults, with Model 4 overpredicting the least.

[65] Here, default includes delinquencies.

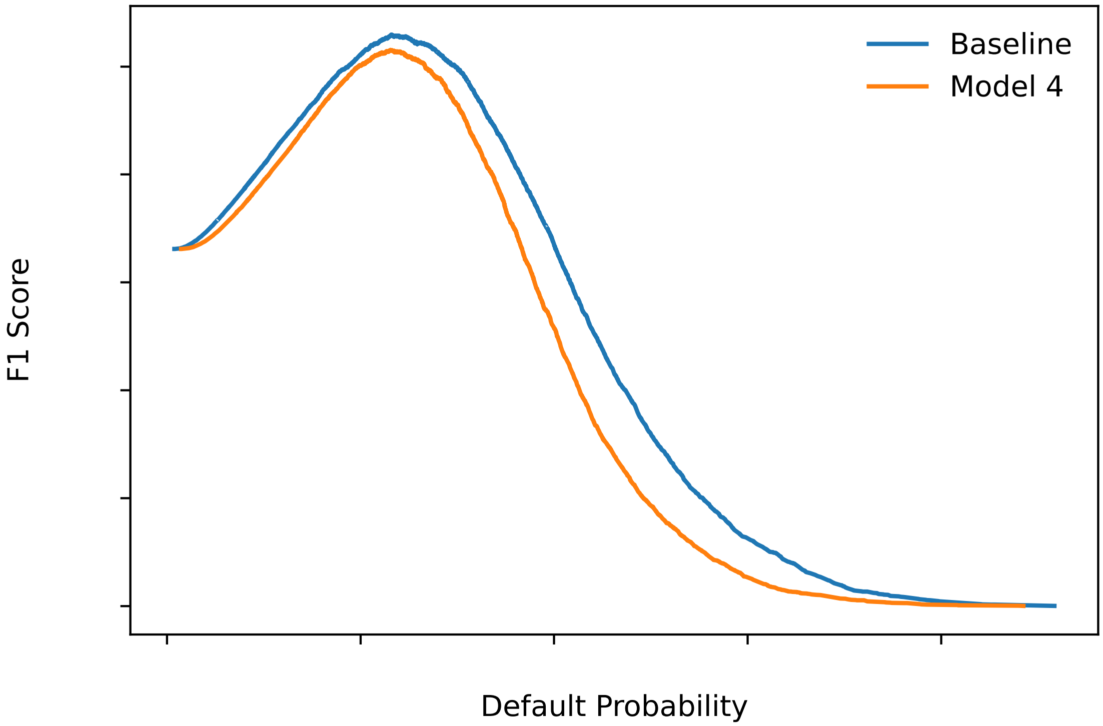
[66] To be more precise, the F1 score is a commonly-used statistical measure of a model's accuracy, calculated by taking the harmonic mean of the "precision" and "recall" of a model. The "precision" is the fraction of true positives divided by the number of false positives plus true positives. Here, precision means the number of defaulters correctly identified by the model divided by the number of defaulters correctly identified by the model plus the number of non-defaulters identified as defaulters. The "recall" is the number of true positives divided by the number of true positives plus false positives. Here, recall means the number of defaulters correctly identified by the model divided by the number of defaulters correctly identified by the model plus the number of defaulters incorrectly labeled as non-defaulters. *See* The F1 score, Towards Data Science (Aug. 31, 2021), https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

[67] Actual F1 scores and default probabilities are not included to avoid disclosing proprietary or sensitive information.

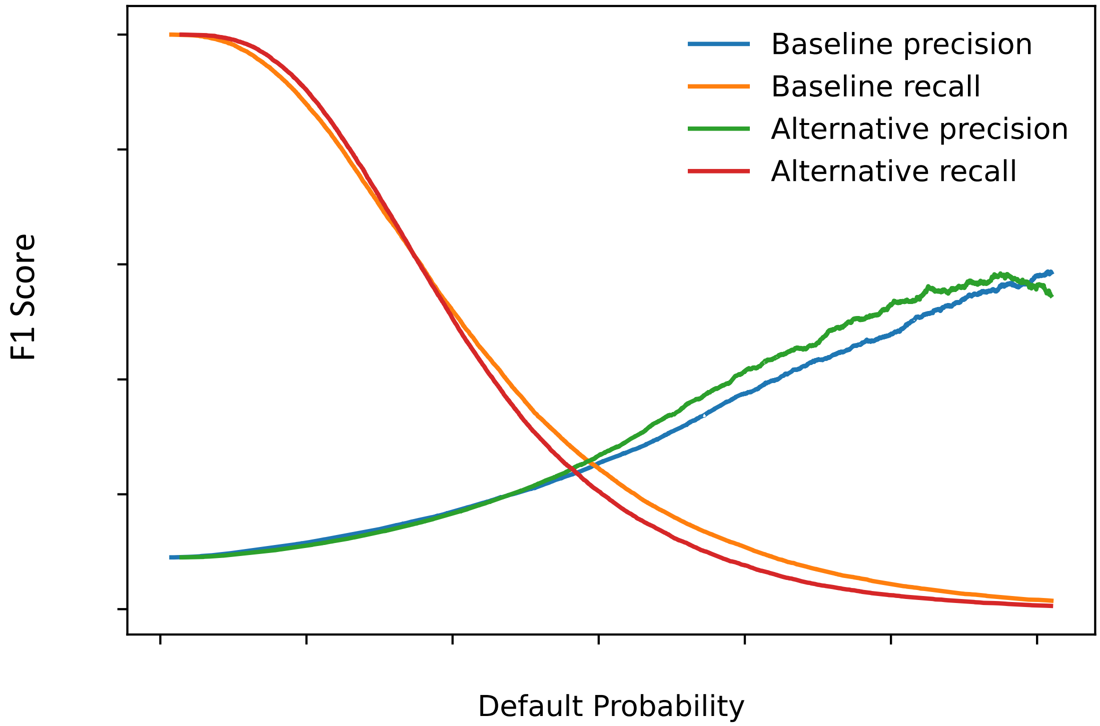


In other words, when just considering the ability to predict likelihood of default, Model 2 is essentially indistinguishable from the Baseline Model. We consider this to be additional support for a likelihood that a court would consider Model 2 to be viable, and further supports the proposition that differences within the 68% Uncertainty Interval of the Error Metric are unlikely to be meaningful.

This analysis for Model 4 is more complicated. Model 4 has lower F-1 scores than the Baseline Model, especially for higher default risk populations, as shown in the figure below.



Remember that the F1 score considers both Precision and Recall. As shown in the figure below, Model 4 has a lower F1 score because it has worse recall than the Baseline Model. However, it has better precision than the Baseline Model. In other words, Model 4 would label slightly fewer defaulters as non-defaulters (*i.e.*, recall) but would do a better job avoiding labeling people that will pay as likely defaulters (*i.e.*, precision).



Given that Model 4 falls outside of the 68% Uncertainty Interval, that potential tradeoff—adopting an alternative model that would improve disparate impact and that would decrease the number of denials to applicants that will pay, but that would extend credit to more likely defaulters—could further weigh against adopting Model 4.

In sum, we believe there is a likelihood that a court would find that less discriminatory alternative Model 2 would meet Upstart's legitimate business interests. Model 4 would further improve model disparity metrics for Black applicants. However, we would not recommend it because its performance falls outside the 68% Uncertainty Interval of +10 Error Metric points.

i. *Less Discriminatory Alternative Recommendation*

Based on our analyses, we would likely have recommended that Upstart adopt alternative Model 2, discussed above. Upstart could have considered adopting Model 4, although there would be reasonable counterarguments to doing so. However, as noted, Upstart updated its model prior to completion of these analyses. Accordingly, rather than recommending adoption of a specific model alternative at this time, we recommend that as Upstart continues to retrain and update its Model, it does so using the fair lending methodologies discussed in these Monitorship Reports.

- Those methodologies include identifying whether statistically and practically significant disparities exist for protected classes, ensuring the existence of a legitimate business need, and searching for less discriminatory alternative models.

- We apply a technical methodology for identifying alternative models described above, which includes Hyperparameter Tuning and exploring variable combinations using mathematical optimization search techniques to identify combinations that yield reductions in disparate impact.[68] Other technical methodologies for identifying less discriminatory alternative models exist, and Upstart represents that it has conducted its own research utilizing different approaches. We recommend that Upstart apply the methodology for searching for alternatives we use in this Report, or a methodology Upstart develops, so long as we and Upstart agree that the different methodology is comparable or more effective in its ability to identify less discriminatory alternatives.
- Regardless of the technical methodology used to search for alternatives, we recommend application of the Alternative Model Constraints described in this Report, including that Upstart adopt alternative models that improve any identified statistically and practically significant disparities as long as the Alternative Model's performance falls within a 68% Uncertainty Interval of the Baseline Model, understanding that that Uncertainty Interval might vary depending on the Baseline Model.
- Finally, we recommend that, during the course of this Monitorship, Upstart report to the Monitor the results of its application of these methodologies and constraints as it applies them. Upstart's process should be repeatable and verifiable by the Monitor.

Understanding there may be compliance and operational obstacles to immediate adoption, we recommend that Upstart implement these steps within a reasonable amount of time following issuance of this Report.

C. Proxy Review

a. *Proxy Review—Overview*

In addition to our disparate impact analysis, we also conducted further testing for proxy variable risks. It is generally a violation of the ECOA and FHA prohibitions against overt, intentional discrimination (*i.e.*, disparate treatment) to use protected class status or a close proxy for protected class status as a variable in a credit scoring or pricing model.[69] A close proxy 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.

In our Second Report, we concluded that, 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 origin.

[68] *See supra* notes 19-20 and accompanying test.

[69] *See, e.g.,* Monitor's Initial Report, *supra* note 1, at 8; Monitor's Second Report, *supra* note 2, at 22-23.

In this Third Report, we conduct similar analyses for sex and age, and find the following:

1. Sex: 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 sex.

2. Age: We also do not find sufficient evidence to conclude that input variables have a high likelihood of functioning as proxies for age. Although we do find evidence that individual input variables in Upstart's Model have a high likelihood of being able to *predict* whether a borrower is age ≥ 62, we do not find evidence that the predictive value of these attributes is solely or largely due to that correlation with age.

3. Recommendations: After we shared these observations, Upstart formalized a Policy for Limiting Potential Proxies for Age ("Age Proxy Policy"), which Upstart represents codifies its existing practice of truncating variables to mitigate age correlations. We recommend that Upstart strengthen this Age Proxy Policy by further truncating certain non-traditional variables that have a high likelihood of being able to predict whether a borrower is age ≥ 62, as described below.

b. *Proxy Review—Upstart's Policy for Limiting Potential Proxies for Age*

It is not uncommon for individual input variables in credit models to have a high likelihood of being able to predict age because many commonly-used credit factors relate to characteristics like length of time. Accordingly, lenders often take measures to mitigate age-risk proxies.[70] For example, a lender might "top-code" at 120 months a variable that relates to the number of months a consumer has had an open trade line, meaning all time-period values greater or equal to 120 months (*i.e.*, ≥ 10 years) would be treated the same. These types of techniques can help mitigate risks that these variables' contributions in a model may be related to differentiating between older and younger consumers.

After sharing preliminary results in this Report with Upstart, Upstart documented in an Age Proxy Policy a methodology it represents it has been using for truncating variables to mitigate the risk that variables may be functioning as close proxies for age. Under Upstart's now-formalized Policy, an analysis is conducted to estimate the relationship between a variable and status as an applicant age ≥ 62, and then Upstart truncates variables at the maximum value of the variable such that a set percentage (not included in this Public Report) are younger than 62 years

[70] Correlations with age are sometimes considered differently than correlations with other protected characteristics because ECOA and Regulation B treat age differently than other protected characteristics (age is not a protected class under the FHA). For example, ECOA does not prohibit direct consideration of the age of an elderly applicant in any system of evaluating creditworthiness when age is used in favor of an elderly applicant, and it does not prohibit direct consideration of age in an empirically derived, demonstrably and statistically sound credit scoring system as long as the age of an elderly applicant is not assigned a negative factor or value. *See* 15 U.S.C. § 1691(b)(3)-(4); 12 C.F.R. § 1002.6(b)(2)(ii)-(iv). In other words, unlike protected characteristics like race, national origin, or sex, *direct* consideration of age is not prohibited if such consideration is favorable or not negative for an elderly applicant. Therefore, even if a variable is *solely* functioning as a direct proxy for age, use of that variable may not be illegal if it is favorable for elderly applicants.

of age. Values less than the threshold stay the same, values more than the threshold are set equal to the threshold value.[71] In a sense, Upstart's Age Proxy Policy is similar to the "top-coding" policies described above.

The following section discusses our proxy analyses, with a focus on age and sex proxy risks. Our analyses are conducted on variables after Upstart's Age Proxy Policy was applied; presumably, the age-related associations observed in this Report would likely be more significant absent Upstart's mitigation techniques. But because these age-related associations persist despite application of the Policy, we recommend that Upstart apply more truncation than otherwise called for by its Policy to higher-proxy risk non-traditional variables, which can raise increased fair lending risks because they are less commonly used and the relationships between such variables and an individual applicant's creditworthiness is not always as direct as with traditional variables.

c. *Proxy Review—Methodology Recap*

Our Second Report describes our quantitative methodologies for assessing proxy risks across all variables.[72] To summarize here:

- First, we use a statistical technique called "Surrogate Modeling" to assess whether the individual input variables that are fed into Upstart's Model may be significant predictors of sex or age. 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, we include Upstart's variables in different Surrogate Models to ascertain the degree to which those variables are predictive of sex or age. We find those variables are not predictive of sex, but they are predictive of whether consumers are < 62 or ≥ 62 years of age.

- Second, because we identify variables with meaningfully higher significance in predicting age, we next assess if their contributions to performance in Upstart's Model are separable from that correlation with age. We do so by assessing what happens when Upstart's Model is trained on populations of only: (1) consumers age < 62, with and without these variables included; and (2) consumers age ≥ 62, with and without these variables included. If this group of variables contribute to performance in either the "all < 62 model" or the "all ≥ 62 model," the variables' contributions to performance in Upstart's Model are likely not fully dependent on their correlation with predicting whether consumers are < 62 or ≥ 62 years of age. Here, the group of variables identified contribute to performance in a statistically significant way, suggesting that they are not functioning solely as proxies for whether an applicant is ≥ 62 years of age.

[71] A converse process is applied where larger values of the variable related to lower proportions of applicants age ≥ 62. The above procedure is effective if the relationship between the potential proxy-variable and group membership is approximately monotonic. Monotonic means that the relationship is one-directional (*e.g.*, increasing the value of an input variable will always cause the output variable to increase or will always cause the output to decrease).
[72] *See* Monitor's Second Report, *supra* note 2, at 24-30.

d. *Proxy Review—Detailed Results*

The following describes the results of these methodologies in more detail. We first created Null models for sex and age—shown in row one of Table 2 (Sex) and Table 3 (Age) respectively. These Null models attempt to predict the sex and age labels of borrowers in the dataset without including *any* of Upstart's input variables. We use these Null models as a baseline from which we can assess the performance of other surrogate models that *do* include Upstart's input variables. As Table 2 indicates, a Null model has an F1 score of 0.649 with respect to predicting sex. As Table 3 indicates, a Null model has an F1 score of 0.062 with respect to predicting age. These performance metrics should not be considered good or bad. They simply demonstrate as a baseline that, in the absence of using any of Upstart's input variables, we can achieve an accuracy of 0.649 and 0.062 in predicting the sex and age respectively 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 these protected class labels of a borrower with significantly greater F1 than the Null models' scores, that would be indicative that Upstart's input predictor variables might include protected class proxies.

Table 2: Surrogate Model Results for Sex

	Model Type	**Max F1 Score**
1	Null Model (no Upstart variables)	0.649
2	Ridge Logistic Regression Surrogate (all Upstart variables)	0.748
3	RandomForest Surrogate (all Upstart variables)	0.694

Table 3: Surrogate Model Results for Age

	Model Type	**Max F1 Score**
1	Null Model (no Upstart variables)	0.062
2	Ridge Logistic Regression Surrogate (all Upstart variables)	0.557
3	RandomForest Surrogate (all Upstart variables)	0.538

Rows two and three in Tables 2 and 3 show results from two separate surrogate models: a Ridge Logistic Regression surrogate model and a RandomForest surrogate model.[73] These surrogate models were specified with the entire set of Upstart's input variables. For sex, the Ridge Logistic Regression achieved an F1 score of 0.748 and the RandomForest achieved an F1 score of 0.694. Although there is some improvement, these scores do not represent substantial gains in accuracy for predicting the sex of each borrower over the Null model (0.649). 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 sex.

For age, however, the Ridge Logistic Regression achieved an F1 score of 0.557 and the RandomForest achieved an F1 score of 0.538. These F1 scores do represent substantial gains in accuracy for predicting the age of each borrower over the Null model (0.062). In other words, including all of Upstart's input variables in these surrogate models does provide a practically significant improvement in absolute accuracy for predicting age.

We then assessed whether certain variables are *relatively* more significant in predicting a borrower's sex and age status than other variables. To do so, we used 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.[74] If a variable's Information Value level exceeds 0.3, it would be deemed significant and flagged for further analysis.

After implementing our Information Value test for sex, we found that all individual variables had Information Values that were below 0.3, the threshold we use for relative significance. In other words, even using a relative test, the relative contributions of individual variables to predicting sex were generally flat across variables. Accordingly, we did not proceed with further analyses for sex and conclude it does not appear that individual input variables in Upstart's Model have a high likelihood of functioning as proxies for sex.

However, for age, we found certain features that had Information Values that were above 0.3, suggesting these features have meaningfully more significant predictive power with respect to whether a borrower is greater or equal to age 62 as compared to other features.[75]

[73] For more details on these models, *see* Monitor's Second Report, *supra* note 2, at 27.

[74] *See, e.g.,* Towards Data Science, "Model? Or do you mean Weight 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.

[75] As in our Second Report, we also address collinearity concerns for sex and age by conducting a supplemental Principal Component surrogate test. This test is largely confirmatory of our results so far: we see from this analysis that a relatively small number (approximately 100) of the Principal Components are responsible for classifying between borrowers age < 62 and ≥ 62. We also see that even the most heavily weighted Principal Components are not particularly strong predictors of sex on their own. For a description of the rationale behind this analysis, see Monitor's Second Report, *supra* note 2, at 29-30.

Because our Surrogate Modeling and Information Value analyses identified features with high contributions for differentiating between older and younger borrowers, we next conducted analyses designed to see whether the contributions of these features to performance in Upstart's Model are entirely reliant on their association with age. To do so, we trained Upstart's Model:

(1) on a population of only borrowers age < 62, with and without these features included; and
(2) on a population of only borrowers age ≥ 62, with and without these features included.

If removing these features affects the performance of these "all older" and "all younger" models, it suggests that the features contribute to performance independent of their ability to differentiate older and younger borrowers.

The results of this analysis suggest that these features are not functioning solely as proxies for whether applicants are older or younger borrowers. Excluding these variables from the models trained only on younger borrowers and only on older borrowers affects model performance. Although not large, the effects in at least one of these models are statistically significant, suggesting that these variables retain predictive power independent of their association with age, which would likely not be the case if their performance were solely attributable to their ability to predict age.

Although these analyses suggest these features are not functioning solely as proxies for age, we paid particular attention to "non-credit variables"—meaning variables that are not traditionally used in credit scoring and underwriting—because of the increased risk that such variables can pose. To do so, we focused on the "non-credit" features within the identified features, which include: (1) graduation year of most recent degree; and (2) one other category of non-credit related features.[76]

Our analyses consider graduation year of most recent degree after application of Upstart's Age Proxy Policy. As described below, we recommend applying a more robust version of the Policy with enhanced truncation to mitigate potential associated risk.

Upstart's Age Proxy Policy cannot be applied to the other non-credit related features because they are not categorical and not time related, and therefore cannot be truncated. Therefore, in order to further confirm that these features are not functioning solely as proxies for age in Upstart's model, Upstart provided information comparing the contributions of the features for applicants < age 62 and ≥ age 62.[77] This analysis showed that the contributions of the features

[76] This latter feature is not provided in this Public Report to avoid disclosing proprietary or commercially sensitive information. Under the Parties' agreement establishing this Monitorship, all education-related variables are considered publicly known.

[77] Specifically, Upstart measured the difference in average and median contributions of the features for both groups of applicants (using a sample of 100K applicants). Contributions were measured using SHAP, which is a relatively common approach to explaining the contributions of particular features to a model's prediction. *See* generally Scott Lundberg & Su-In Lee, "A Unified Approach to Interpreting Model Predictions," arXiv:1705.07874 (Nov. 2017).

were nearly identical for both older and younger applicants, which would not be likely if the variables were solely acting as proxies for age in the model.

In sum: although we include specific recommendations below, our analyses suggest that these features are unlikely to be functioning solely as proxies for whether applicants are older or younger. Moreover, because of the unique way that ECOA and Regulation B treat age as a protected characteristic, even if the features were functioning solely as direct proxies for age, further analysis would need to be done to assess any potential violation—for example, whether the variables uniformly favor elderly applicants.

As we note in our Second Report, the Surrogate Modeling approach has inherent limitations and it cannot conclusively demonstrate that a model does or does not contain proxies for protected class.[78] 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." 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 show whether any interaction variables that are automatically created within a model are predictive of protected class labels. Because of these limitations, we cannot conclusively eliminate the possibility that proxies exist.

e. *Proxy Review—Recommendations*

After we shared these observations, Upstart formalized its Age Proxy Policy, which Upstart represents codifies its existing practice. We recommend that Upstart continue to apply this Age Proxy Policy, monitor its effectiveness, and apply more robust truncation it if it does not mitigate proxy risks for higher risk features.

Despite our quantitative findings, we think there is some qualitative risk that non-traditional variables—such as graduation year of most recent degree—could be perceived as raising age-proxy risk, particularly as the length of time between graduation year and application date increases. To be clear, this risk is not based on our quantitative findings; instead it is a reflection of the fact that non-traditional variables are generally viewed as riskier from a qualitative perspective. To mitigate this potential risk, we recommend that Upstart apply a more robust version of its Age Proxy Policy for non-traditional variables that have a high likelihood of predicting age—such as graduation year of most recent degree—by truncating such variables at a shorter time period than they would be truncated under Upstart's existing Policy.[79]

[78] Monitor's Second Report, *supra* note 2, at 25.

[79] The specific level of recommended truncation is not provided in this Public Report to avoid disclosing proprietary or commercially sensitive information.