eightfold.ai

Responsible AI at Eightfold

Authors
Yovahn Hoole, Manav Mehra, Batuhan Akcay, Sanjeet Hajarnis, Varun Kacholia

Table of contents

Responsible AI at Eightfold . . . 3
AI Principles . . . 3
Fairness . . . 4
Data . . . 4
Training, Evaluation and Model Selection . . . 4
Active Measurement And Monitoring . . . 4
Product Safeguards . . . 5
Right Products & Analytics . . . 6
Candidate Masking . . . 6
Diversity Dashboard . . . 7
Personalized Recommendations . . . 8
Right Data & Features . . . 9
Masked Features . . . 9
Feature Distributions . . . 9
Right Algorithms & Training . . . 10
Model Selection And Training . . . 10
Model Evaluation . . . 10
Parity-based Metrics . . . 11
Confusion Matrix-based Metrics . . . 11
Right Governance & Monitoring . . . 13
Adverse Impact Analysis . . . 13
Background . . . 13
Methodology . . . 14
Approach 1 . . . 14
Approach 2 . . . 19
Approach 3 . . . 19
Perturbation Testing . . . 19
Methodology . . . 20
External Audits . . . 21
Active Monitoring . . . 21
References . . . 22

Responsible AI at Eightfold

Eightfold's Talent Intelligence Platform empowers enterprises to acquire and retain diverse talent, and provides the foundation for public agencies to reemploy and upskill citizens. As the pioneer and leader in talent intelligence, our mission is to enable the right career for everyone.

Eightfold's AI delivers relevant recommendations at scale to predict the next role in an individual's career. Our models understand more than one million unique roles and one million skills across many languages.

With Eightfold technology, candidates can instantly match to the jobs that fit their skills and potential, see why each job is a match, and apply in a matter of seconds. Recruiters and hiring managers get instant ranked lists of candidates who match their requirements, and can engage them through our platform up to the point of making an offer.

Employees can explore future career paths with detailed understanding of the skills and experiences for their next step in their career, and find the projects, courses, mentors, and gigs that can help deliver these skills and experiences. Organization leaders can oversee their talent strategies, find successors for roles, compare scenarios, and determine the upskilling and reskilling plans for their future needs.

Governments and social service organizations can deploy our platform to match individuals with job opportunities at scale in support of reemployment and community building initiatives.

AI Principles

At Eightfold, we are committed to a responsible and ethical development and use of artificial intelligence. As a company, we understand that AI has the potential to significantly impact many aspects of our lives, and we build AI solutions to benefit society while respecting the rights and dignity of our users.

Our team of experts works closely with stakeholders, our committee of representatives from various departments, our AI Ethics Council, and external consultants to design and deploy our AI systems in a responsible and ethical manner. At the core of every design, we prioritize the following principles:

- **Fairness:** Design and Use AI systems that are just and mitigate bias. This includes mitigating discrimination based on factors such as race, gender, age, or other protected characteristics.
- **Transparency:** We believe how AI systems work and how decisions are made should be understandable and explainable.
- **Safety and Reliability:** We strive to design and develop stringent safety measures that our AI has to pass before it rolls out as our product. We believe that it's our responsibility to provide solutions that add value to our society.
- **Active Monitoring And Response:** We believe that any AI system needs to have continuous active monitoring to check that the system behaves as expected. Deviations are treated and responded to on a priority basis.

In this blog post, we'll be taking a deep dive into our thoughts on fairness in the interests of transparency and to hopefully serve as reference to other companies looking to mitigate biases in their AI systems. We should note that this is an active field of research and as things evolve, we will re-visit and update our approaches.

Fairness

Artificial Intelligence can revolutionize employment processes in countless ways. As the industry evolves and increasingly relies on AI systems, it's important to consider the potential of AI to perpetuate social injustices or biases. Fairness is a particularly important issue in the HR recruiting space as biases in AI systems can perpetuate and even amplify existing inequalities in society if left unchecked.

AI fairness refers to the idea that AI systems should not discriminate against groups of people based on characteristics such as race, gender, age, etc.. There are a lot of variables that go into developing an AI system and gaps in oversight can lead to an unfair model. When it comes to applying AI technology to employment practices, we believe the principle of fairness applies at all stages of the development and application of AI technology. It's important for AI developers and users to be aware of the potential for bias in AI systems and take steps to identify and mitigate these issues. The most common pitfalls we've seen can largely be placed into the following buckets:

Data

One possible source of bias is the data used to train models within an AI system. The data used by AI models should be representative across protected categories, and industries. Features should be representative of the population and should not favor any one group. The feature engineering process should be thoroughly vetted. For example, in the case of HR systems, we feel that the model should only need to learn the qualifications of successful individuals rather than their identity.

Training, Evaluation and Model Selection

Alternatively, the choice of model and the training process used can themselves lead to biased outcomes. The models and algorithms used should go through a rigorous and thorough evaluation framework where they are tested for performance across the measurable protected categories. It is crucial that checks and balances are in place during model training to check against learning decisions based on protected categories. At Eightfold, we build models that strive to mitigate amplifying the classic stereotypical patterns in data and human behavior. For example, a model used to recommend candidates for a Software Engineering position should not perform better for one gender than the other.

Active Measurement And Monitoring

In addition to the above, as bias can occur across multiple hiring stages and in a myriad of ways, there is no single test that can test for bias. A robust methodology for measuring bias and monitoring models for biased outcomes is a key component involved in mitigating AI bias.

Product Safeguards

Finally, without any safeguards in the product, even when AI is developed appropriately, outcomes may reflect bias due to human error. While reviewing lists of candidates, for example, people making employment decisions may intentionally or accidentally introduce personal biases into the hiring process. They may favor certain last names or social activities identified in the candidate profiles that reflect historical trends of hiring. Additionally, having detailed monitoring and analytics helps track potentially biased outcomes of human and AI decisions.

By being aware of the potential for bias and consistently aligning our designs with our AI principles, we believe that responsible approaches to AI at Eightfold will help revolutionize employment processes in a fair and equitable way. In the following sections, we'll cover how our principles help us avoid these pitfalls in the development of Eightfold's Talent Intelligence Platform.

These principles through processes are illustrated in Figure 1 below.

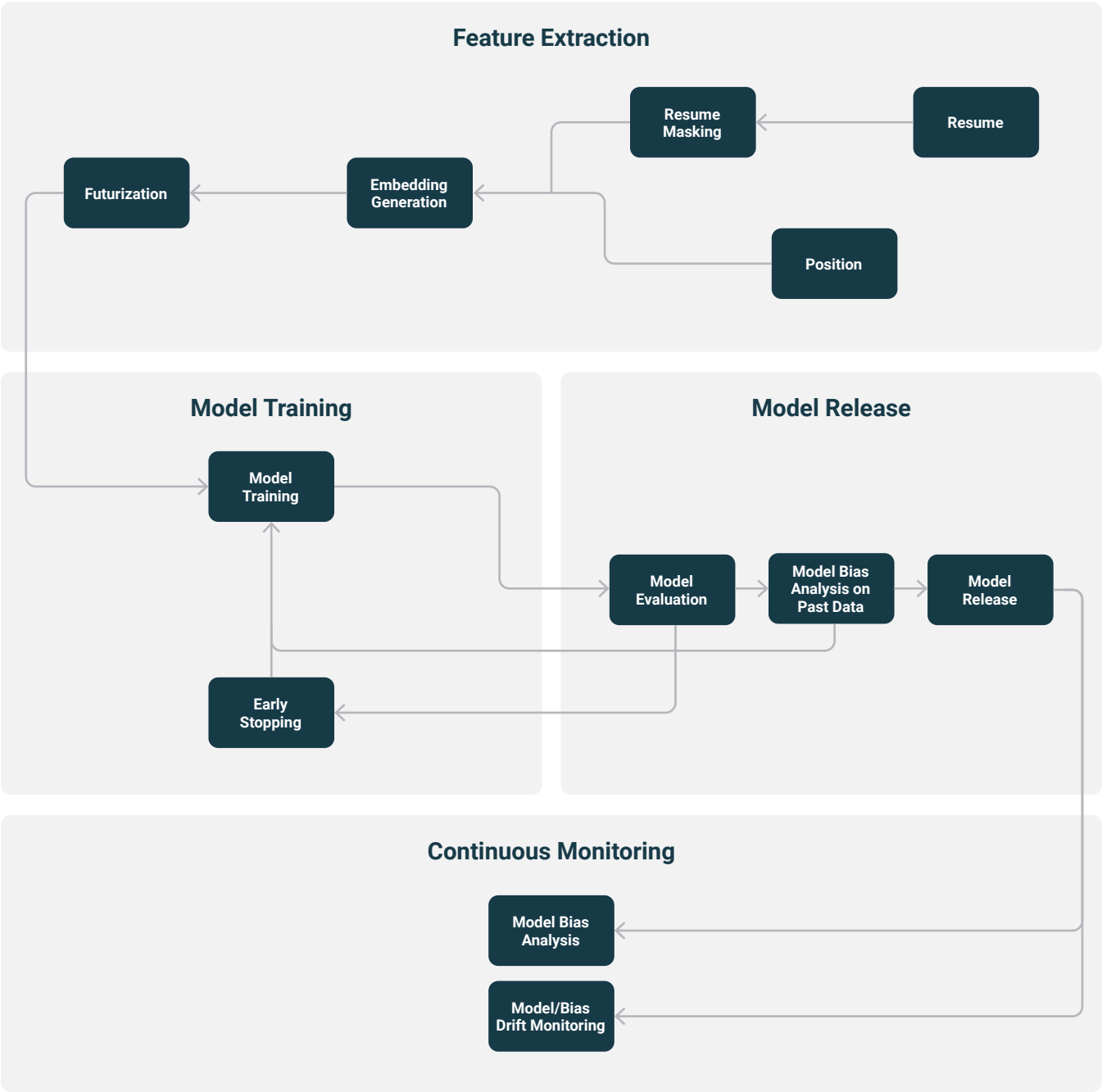


Figure 1: *Flow diagram depicting model development and monitoring pipelines.*

Right Products & Analytics

Candidate Masking

Employment processes are a sensitive area in which societal biases, conscious and unconscious, can foster unfair stereotypes and preferences in selecting candidates. Recruitment processes that are not effective in ensuring fairness can lead to biased hiring outcomes that sustain and amplify imbalances found in society.

Frequently, sensitive, protected information about candidates are identifiable throughout the recruitment stages. Resumes might convey sensitive information about the candidate (such as race, gender, age etc.) that can unintentionally impact the way recruiters make decisions. Additionally, similarity bias in historically imbalanced industries can result in recruiters unintentionally penalizing the disadvantaged groups.

To mitigate these risks, a core functionality of Eightfold's Talent Intelligence Platform is candidate masking. Candidate masking is the process through which we strip sensitive information from a resume that does not affect a prospective employer's ability to evaluate the candidate's skills and suitability for a particular job posting. Particularly when displaying a candidate to the recruiter, the following information can be masked:

Information capable of being masked during Candidate Masking		
Gender, Gender Pronouns and Titles	Disability	Religion
Ethnicity/Race	Marital Status	Sexual Orientation
Name	Email addresses	URLs
Images containing pictures of the applicant		

Table 1: *Types of information that can be masked from resumes during candidate masking.*

Customers may configure to additionally mask the information in Table 2.

Information capable of being masked during Candidate Masking (Optional)		
Phone Numbers	Addresses	Date of Birth
Graduation year	School	

Table 2: *Additional types of information masked from resumes during candidate masking (optionally)*

Masking is performed using algorithms that intelligently capture and mask the aforementioned aspects of the profile. We measure and monitor the performance of these algorithms to ensure that they perform effectively across a variety of types of resumes.

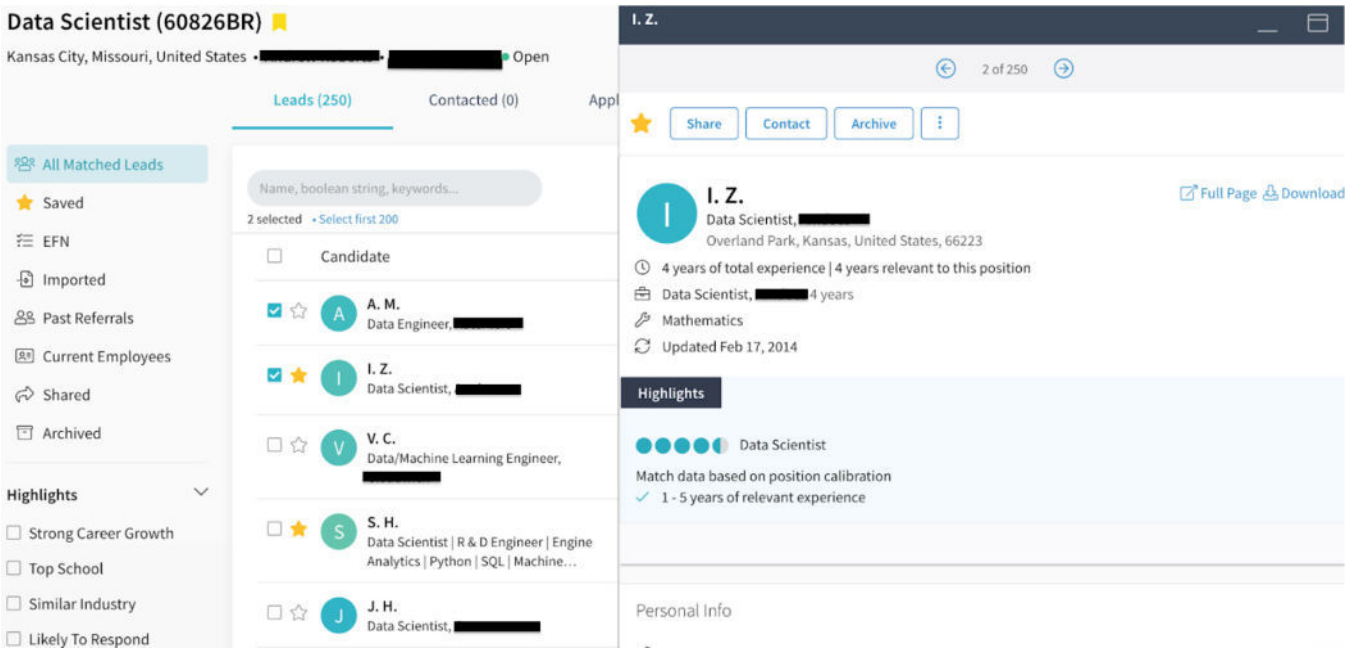


Figure 2: *Candidate masking in the products. Candidates are displayed in a way that only displays information relevant to selection as configured by the customer.*

Through candidate masking, Eightfold's Talent Intelligence Platform minimizes a recruiter's ability to identify and make decisions based on stereotypes. We provide a candidate profile to the recruiter for evaluating job fitness based on fair and objective factors related to the job description. Candidate masking helps empower recruiters to focus on the skills and experiences of candidates that are relevant for the job in order to make strong hiring decisions for the success of their company.

Diversity Dashboard

Beyond candidate masking, the product allows customers to track, review and analyze human users' aggregate usage statistics through each stage of the hiring process via the Diversity Dashboard.

The Diversity Dashboard enables employers to analyze differences in pipelines across various breakdowns.

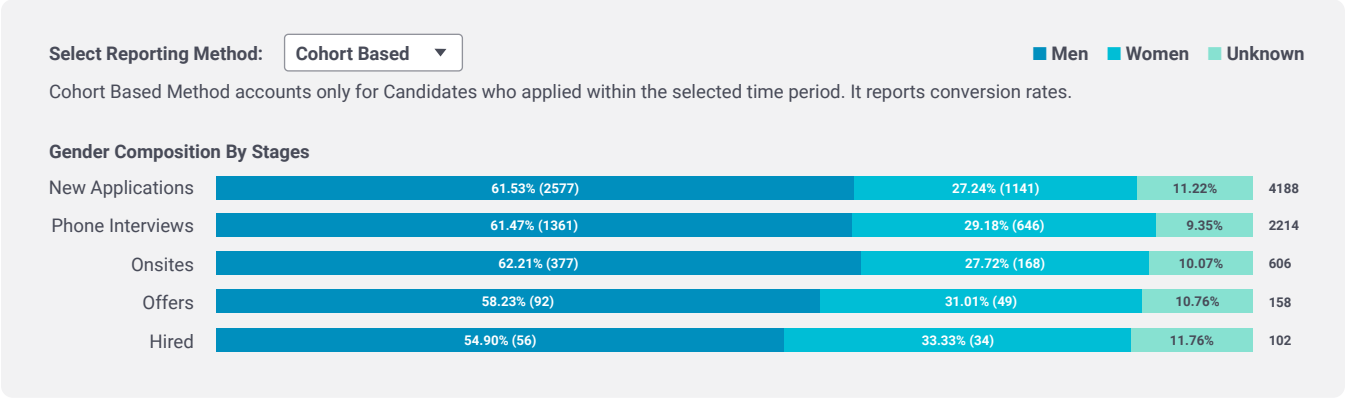
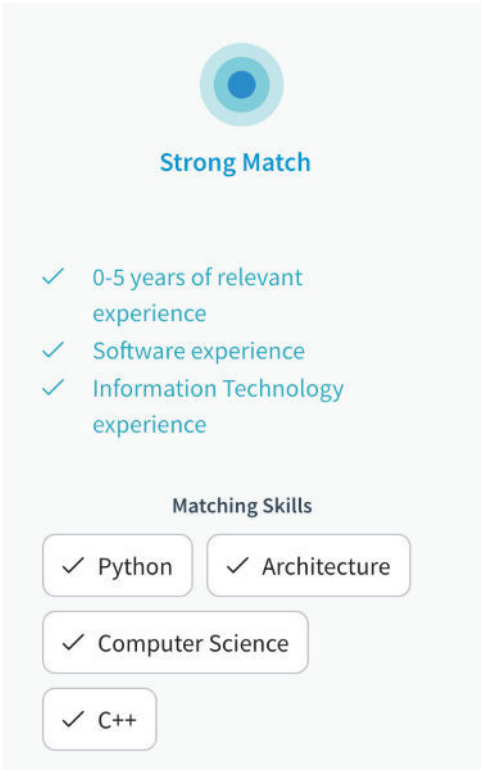


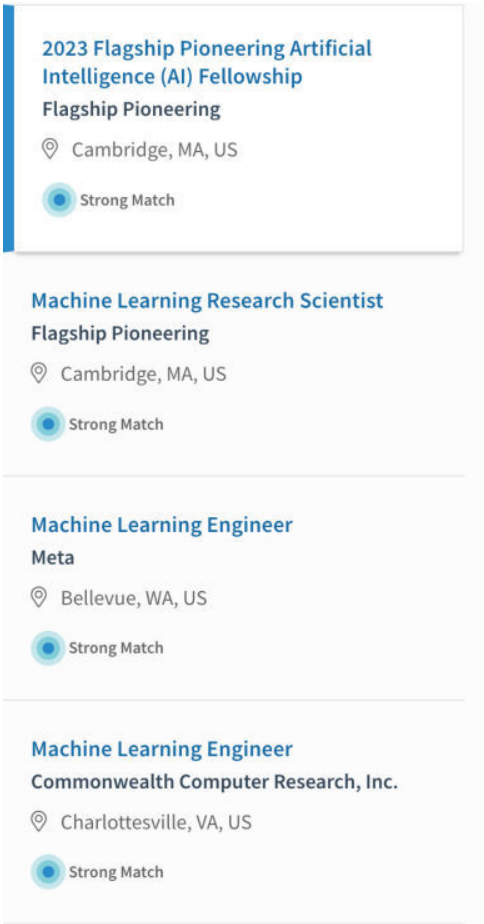
Figure 3: *Diversity Dashboard displaying Gender composition at each stage of the hiring process.*



Personalized Recommendations

On the other side of the hiring process are the individuals who are searching for jobs. Studies [REF] have shown that women are statistically less likely to apply to jobs they are qualified for than men. These differences can have a significant impact on underrepresentation of protected groups in certain industries. To combat such disparities, we provide potential applicants with an objective analysis of their fit with a job to encourage qualified applicants to apply.

Figure 4: *Card displayed to potential applicants depicting a "strong match" between a candidate and position.*



Additionally, by providing a personalized list of open jobs ranked by compatibility, we further help mitigate self selection biases among potential applicants. This allows each job opportunity to reach the strongest potential applicants, and increase the likelihood of a diverse applicant pool for recruitment.

Figure 5: *Ranked list of open positions supplied to individuals searching for jobs.*

Right Data & Features

Masked Features

Data sanitization for model training is an ongoing and evolving process. At Eightfold, we are committed to investing in and improving our approach. Some examples of our approach are discussed below.

For our models, the input data is stripped of identity information such as names, contact information and address information, which is not relevant to the qualification of the candidate with respect to the job requirements. This helps minimize the ability of a model to explicitly incorporate information pertaining to protected categories like race, gender, age etc. into its decision making processes. With the myriad possible variations of resumes however, masking is a non-trivial problem to solve. To handle such lapses in resume masking and to mitigate the impact of more nuanced indicators, we see resume masking as only one part of a strong defense against biased data.

Feature Distributions

As a part of the feature engineering process, It is important to have a theoretical understanding of what the feature represents and what it will help the model to learn and as a part of that feature engineering process, we establish a hypothesis around the expected distribution of the feature values, which may reveal the classification power of the feature. For example, a feature that has a constant value for a range, then it may not be a useful feature for a classifier. As part of our vetting process of features we seek to ensure that the feature computations hold true with the initial hypothesis.

Feature	10p	30p	50p	70p	90p	min	max	mean	stddev	% nonzero
	-0.001	0.11	0.287	0.45	0.644	-0.712	1.0	0.29213478	0.26668702	51.64158104
	-0.002	0.206	0.439	0.632	0.801	-0.575	1.0	0.40985002	0.31352084	96.01332752
	-0.003	0.118	0.273	0.419	0.614	-0.526	1.0	0.27998945	0.25003912	52.78457251
	-0.0085	0.324	0.6085	0.81	0.932	-0.583	1.0	0.52418678	0.36079574	56.31166362
	-0.01225	0.16075	0.359	0.5365	0.7325	-0.557	1.0	0.3491731	0.29036782	54.12209397
	-0.014	0.0	0.0	0.207	0.573	-0.673	1.0	0.14432393	0.26344408	28.22064675
	-0.018	0.1585	0.3665	0.553	0.746	-0.557	1.0	0.35480977	0.29749635	53.98139189
	-0.022	0.216	0.444	0.632	0.822	-0.66	1.0	0.40909378	0.32453249	54.80652612
	-0.031	0.0	0.215	0.418	0.65	-0.69	1.0	0.24695503	0.28542278	36.98113979
	-0.031	0.0	0.215	0.418	0.65	-0.69	1.0	0.24695503	0.28542278	36.98113979
	-0.031	0.0	0.215	0.418	0.65	-0.69	1.0	0.24695503	0.28542278	36.98113979
	-0.031	0.0	0.215	0.418	0.65	-0.69	1.0	0.24695503	0.28542278	36.98113979
	-0.06	0.207	0.44	0.625	0.78	-0.675	0.989	0.39565019	0.31581496	95.39481753
	-0.081	0.067	0.296	0.506	0.73	-0.655	1.0	0.30073577	0.31222277	42.77258047
	-0.081	0.067	0.296	0.506	0.73	-0.655	1.0	0.30073577	0.31222277	42.77258047

Figure 6: *Feature Distributions before training the models.*

Right Algorithms & Training

Model Selection And Training

Eightfold's web-based platform utilizes an underlying mathematical model built using machine learning techniques. The underlying model predicts the match between a candidate profile and a job position, and displays candidates for a given job position in a rank-list manner as a use case, and supports other use cases such as candidates viewing jobs on career pages. The model operates on a candidate-position pair. It is not a stand-alone score for a candidate. Rather, it is the match of the candidate to job requirements as specified by the calibration of the job position.

At Eightfold, we believe in employing the use of algorithms that support a high degree of explainability. The mindful choice of the algorithms becomes another guard rail towards fairness in the decision making process powered through our AI. Explainable algorithms allow users of the product to understand the reasoning behind algorithms' scoring.

In addition o the right algorithms, data and features the training process for these models/algorithms act as an added layer of protection against unwanted bias from creeping in during inference. During training, we divide all the data that we have into train and test sets. This distribution is done keeping in mind that there is no data leakage between the two sets, to ensure that the model is tested on samples that it has not seen during training.

We also incorporate early stopping based on classification results made by the model across input datasets of different groups and protected categories. Additionally, keeping up with the recent literature we're exploring different ways of integrating anti-bias and fairness efforts as part of the loss function on which the model optimizes.

Model Evaluation

Post training, we conduct rigorous evaluations on the model so that it meets our standards before it's launched to our customers. On passing the initial check, a myriad of metrics are computed to evaluate the model performance and fairness to compare it with the previous iteration of the model. We highlight below certain metrics we use to measure fairness. In addition to the standard accuracy metrics like AUC, Precision, Recall and F1 we compute metrics specifically to measure bias.

Cluster ID	Cluster Title	AUC %	TPR/Recall	FPR	F1	FNR	Precision
		94.77985777939958	0.8973	0.145	0.8788	0.1027	0.861
		95.06376612539265	0.9595	0.2109	0.8964	0.0405	0.8411
		91.34768989291014	0.8891	0.2113	0.8574	0.1109	0.8279
		94.97101418145813	0.9198	0.1644	0.8685	0.0802	0.8226
		91.31696435444331	0.9043	0.2574	0.8883	0.0957	0.8728

Figure 9: *Metrics evaluated by Job Title*

Language	AUC %	TPR/Recall	FPR	F1	FNR	Precision
	94.77985777939958	0.8973	0.145	0.8788	0.1027	0.861
	94.8022994189819	0.898	0.145	0.8784	0.102	0.8597
	94.75717411859937	0.9203	0.1685	0.9122	0.0797	0.9043

Figure 9: *Metrics evaluated by Language*

Fairness criteria on which these metrics are based upon can be of two types,

1. **Group Fairness:** These metrics essentially compare the outcome of a classification algorithm for two or more groups that are defined on the basis of a protected category.
2. **Individual Fairness:** In this set of metrics we ensure that the outcome of a classification algorithm is the same for two similar inputs. Two inputs are considered to be similar on the basis of a predetermined threshold on a similarity (distance) metric.

We are including a summary of certain commonly used metrics. This is an evolving field, and we continuously aim to improve, evaluate and update our methodology periodically.

Parity-based Metrics

These metrics only take into account predicted positive rates. The metrics computed here are discussed as follows,

We first start the discussion on the metric of **Demographic Parity**, This metric examines fairness as an equal probability of being classified as a positive. Ideally, each group should have the same probability of being labeled as a positive outcome.

Formula,

$$DP = \frac{P(Y_{pred}=1 | s_i)}{P(Y_{pred}=1)} \text{ for all } s_i$$

Secondly, we discuss the metric of **Impact Ratio**, similar to parity, except the fact that the ratio is calculated between unprivileged and privileged groups. A model is said to have failed Impact Ratio if it's value falls below 0.8 or above 1.25

Formula,

$$\frac{P(Y_{pred}=1 \mid s_i=1)}{P(Y_{pred}=1 \mid s_i=0)}$$

Indicates Impact Ratio for that protected feature s_i

Another Interpretation discusses Impact Ratio as a ratio of positivity rate of a lesser-represented group to the positivity rate of a more represented group.

These metrics may not take into account potential qualitative differences between the predictions of the groups. More metrics can help in that regard. Algorithmic fairness is an important and heavily researched area. To get a more holistic perspective on fairness, in addition to above, we examine the algorithm's prediction quality using the following metrics.

Confusion Matrix-based Metrics

These metrics take into account True Negative Rate (TNR), True Positive Rate (TPR), False Negative Rate (FNR), and False Positive Rate (FPR). The advantage of these metrics is that they take into account the underlying qualitative differences between groups that are otherwise not included in the parity-based metrics.

The first metric that we would like to talk about here is the **Equality of Opportunity**. It's defined as the probability of a person in a positive class assigned to a positive outcome of the model's classification. The goal of it is to have very close ratios for all the members of a protected category (such as female and male). In the formula below, s means a particular group; Y is the ground-truth label, and Y_{pred} is the predicted outcome.

$$P(Y_{pred} = 1|s_i = 1, Y = 1) = P(Y_{pred} = 1| s_i = 0, Y = 1)$$

And the ratio can be calculated for a particular group s using the following formula,

$$\text{EOP} = \frac{P(Y_{pred}=1|S=s_i, Y=1)}{P(Y_{pred=1}| Y=1)}$$

Another metric would be **Equalized Odds** which is the probability of a person in the positive class being correctly assigned a positive outcome of the model's classification, and the probability of a person in the negative class being incorrectly assigned a positive outcome of the model's classification. The goal is to have very close ratios for all the members of a protected category .

$$P(Y_{pred} = 1|s_i = 0, Y = y) = P(Y_{pred} = 1|s_i = 1, Y = y) \; where \; y \in \{0, 1\}$$

Finally we also track whether the ratio of false negatives and false positives is close for all categories in a protected group. In literature, this is called Treatment Equality.

Right Governance & Monitoring

In recent years, the EEOC (Equal Employment Opportunity Commission) launched its “Artificial Intelligence and Algorithmic Fairness Initiative” [REF], additionally New York City’s AEDT (Automated Employment Decision Tools) Law precludes employers from using AEDTs that have not completed an independent bias audit within the past year [REF]. As a result, employers and regulators alike have grown increasingly cognizant of the risks of using AI based tools in connection with employment decisions.

A key aspect in mitigating these risks is a robust and transparent methodology for measuring AI bias in selection processes. Particularly, we seek a solution that effectively bridges the gap between model evaluation frameworks in place today and the decades of research in employment law and adverse impact analysis. Model evaluation frameworks focus on a machine learning model’s ability to understand and generalize patterns within a dataset. In the context of algorithmic fairness, these frameworks help answer the question:

“Is the performance of the model employed by the recruitment tool dependent on subgroup membership?”

In cases where the underlying data is biased however, even a model that performs equally well across subgroups can result in unequal outcomes. To this end, adverse impact analysis broadly covers the analysis of disparities in employment outcomes. As a result, adverse impact analysis helps answer the question:“Does the use of the recruitment tool in question result in disparate outcomes across subgroups?”

Both model evaluation frameworks and adverse impact analysis provide unique insights into algorithmic fairness and are part of Eightfold’s measurement of bias analysis as applied to real-world data.

Adverse Impact Analysis

Background

Existing methodologies of adverse impact analysis have been historically prevalent for evaluating and analyzing adverse impact in case of human decisions. Given the scale of data at which AI operates, some of the assumptions behind these methodologies do not apply causing incongruent behavior. We, however, can take inspiration from these to develop tests that are applicable at different scales of data.

A core component of adverse impact analysis examines selection rate differences among subgroups. It is intended to assess disparities in selection processes. Even unbiased selection processes, when evaluated on a finite sample, may result in selection rate differences due to sampling error [12]. Significance testing is the process through which selection rate differences that are potentially indicative of discrimination are distinguished from those that occur simply due to chance.

In statistical significance testing, a null hypothesis about the total population is tested against a sample of the population. In the context of adverse impact analysis, the null hypothesis is that there is no substantial difference in selection rates between two subgroups [12]. Under a set of assumptions, a statistical significance test validates the null hypothesis against applicant flow data by determining the probability of observing the selection rates seen in the sample when the null hypothesis is true. When this probability is below a certain threshold, the selection rate differences are deemed statistically significant. When this probability is greater than the threshold, the differences in selection rates are not significant enough to reject the null hypothesis. As non-significant differences can also result from insufficient data due to small sample sizes, a failure to reject the null hypothesis does not necessarily imply an impartial selection process.

Tests of statistical significance have their share of limitations. Type I and Type II error rates express the probabilities of a test resulting in a false positive and false negative respectively. Statistical power is the complement of the Type II error rate and denotes the probability that a test will correctly reject the null hypothesis when a substantial difference is present. In an ideal world, both Type I and Type II error rates would be low, however, reducing one type of error often results in increasing the other. In the development of a testing framework, we seek a balance between the two.

Additionally, when large sample sizes are used in statistical significance testing, even small, practically insignificant differences can be statistically significant. To alleviate such concerns practical significance testing is used. Practical significance tests offer domain specific heuristics that are used to determine whether a difference has meaningful impact in the real world [12]. At large sample sizes where statistical tests are practically unreliable, practical significance tests are a useful complement. However, practical significance tests may be unreliable in small sample sizes.

Methodology

Approach 1

A commonly used approach to structure adverse impact analysis is through a 2 by 2 contingency table [12]. The contingency table compares the selection rates of a given process between a focal and comparator group. The focal and comparator groups are two subgroups within a protected category we want to compare. In the context of match scores, a candidate is, for purposes of this approach, considered "selected" if the match score they received is greater than some cut off score T. The simulated selection rates can't be controlled due to the nature of the computation and solely depend on the model's predictions and thresholds set. The comparison of selection rates is as follows (Table 3):

	Selected	Not Selected	Total
Focal Group	NP_{focal}	NF_{focal}	N_{focal}
Comparator Group	$NP_{comparator}$	$NF_{comparator}$	$N_{comparator}$
Total	NP_T	NF_T	N_T

Table 3: *2 by 2 Contingency Table used for adverse impact analysis.*

$NP_{focal}, NP_{comparator}, NP_T$ are the number of applicants from the focal, comparator and overall applicant pools that were selected by the test. Conversely, $NF_{focal}, NF_{comparator}, NF_T$ are the number of applicants from the focal, comparator and overall applicant pools that were not selected by the test. Finally, $N_{focal}, N_{comparator}, N_T$ reflect the total number of applicants in the focal comparator and overall applicants pools. The primary attribute analyzed in adverse impact analysis is the selection rate. The selection rates for the focal group SR_{focal}, the comparator group $SR_{comparator}$ and overall applicant pool SR_T are defined as follows:

$$SR_{focal} = \frac{NP_{focal}}{N_{focal}}, \quad SR_{comp} = \frac{NP_{comp}}{N_{comp}}, \quad SR_T = \frac{NP_T}{N_T}$$

To illustrate the application of Table 3, consider the following scenario: a given position receives 100 applicants. Of these 100 applicants, 15 applicants declared their race as Asian, 25 declared their race as Black, and 60 declared another race or chose not to declare their race/ethnicity. A recruiter then uses a cut off match score of 3.5 to filter out applicants. Of the applicants who declared their race/ethnicity, 7 Asians out of 15 received a match score greater than or equal to 3.5 and were thus "selected." Similarly 14 out of 25 Blacks who declared their gender received a score above 3.5 and were selected. In this scenario, the generated contingency table will be

	Selected @ 3.5	Not Selected @ 3.5	Total
Asian	7	8	15
Black	14	11	25
Total	21	19	40

Table 4: *Sample contingency table depicting the selection process of a particular position.*

The goal of this analysis is to determine whether applying such a cut off score will lead to adverse impact. Contingency tables such as the one above, provide a digestible view of applicant flow across two subgroups of a protected category and also simplify statistical calculations.

As to statistical tests, the first test we will consider is the Z or Two Standard Deviation Test which is calculated as follows [12]:

$$Z = \frac{SR_{focal} - SR_{comp}}{\sqrt{SR_T(1-SR_T)(\frac{1}{N_{focal}}+\frac{1}{N_{comp}})}}$$

This test is used to determine the statistical significance of selection rate differences. When the absolute value of the test statistic is greater than 1.96 (i.e., Z <-1.96 or Z > 1.96), the test indicates a statistically significant difference between the two selection rates. At an intuitive level, the test assumes that, under the null hypothesis, the differences in selection rates are normally distributed with a mean centered at 0, and a standard deviation estimated from the contingency table as:

$$\sqrt{SR_T(1 - SR_T)(\frac{1}{N_{focal}} + \frac{1}{N_{comp}})}$$

The estimation of the standard deviation from the contingency table, particularly its reliance on sample sizes in the term from the above equation (Eq. 1)

$$\frac{1}{N_{focal}} + \frac{1}{N_{comp}}$$

results in monotonically increasing z-statistic with increasing sample size.. Consider that the overall selection rate is fixed at 30% and that selection rates between the focal and comparator groups vary by 1%. Further assume that $N_{focal} = N_{comp}$ such that the test statistic equation simplifies to:

$$Z = \frac{0.01}{\sqrt{0.3(1-0.3)(\frac{2}{N_{focal}})}}$$

The following plot (Fig. 1) shows the value of the Z statistic with N_{focal} varying from 2 to 50,000 applications.

As can be seen from the above figure, the same difference in selection rates increases in statistical significance as the sample size increases. Practically however, an absolute difference of 1% in selection rates may not be a significant difference regardless of the sample size. Intuitively, as the number of applications increases, the estimated standard deviation decreases. As a result, even small differences in selection rates may be more than 2 standard deviations away from 0. Particularly at the scale of millions of applications, the Z test becomes an unreliable indicator of bias.

In these cases of very large sample sizes, commonly used practical significance tests such as the 4/5ths rule can be more reliable. The 4/5th rule [REF], is a guideline that suggests that the adverse impact ratio can be defined as [12],

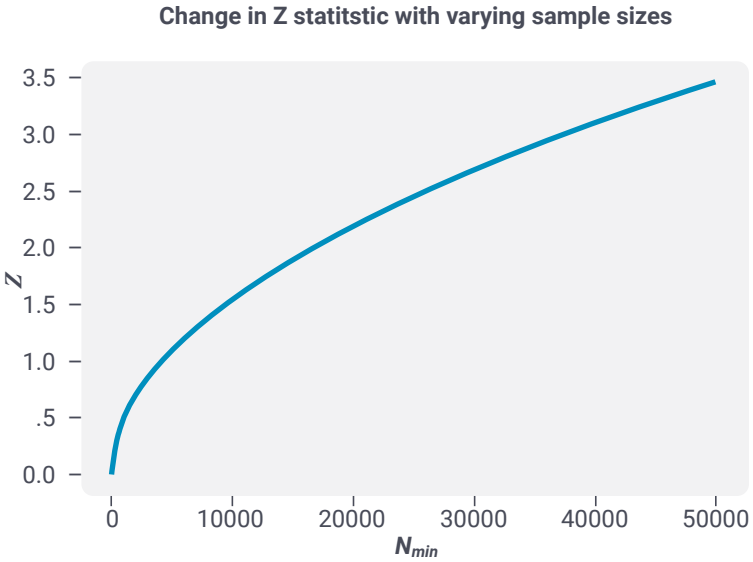


Figure 14: *Change in Z score for a 1% difference in selection rates with increase in sample size.*

$$IR = \frac{SR_{focal}}{SR_{comp}}$$

should be between 0.8 and 1.25. When the IR is below 1, it is an indicator that the comparator group is preferred over the focal group and when IR is above 1 it is an indicator that the focal group is preferred over the comparator group. In a perfectly neutral process the ratio would be 1, however the 4/5th rules sets the guideline that slight deviations from 1 will generally not be considered a substantially different rate of selection, while ratios outside of the 0.8 to 1.25 range will generally be considered a substantially different rate of selection. As the 4/5ths rule's notion of significance is independent of sample size, the 4/5ths rule provides practically useful results at large sample sizes. At small sample sizes however, selecting one more applicant from the disadvantaged group instead of the advantaged group can flip the result of the test. Statistical significance makes statistical significance tests robust to such small perturbations at small sample sizes. As practical significance tests do not have such an understanding, additional heuristics such as the "flip flop" rule are applied in practice to make the 4/5ths rule more robust at small sample sizes [12]. The sensitivity of the impact ratio at small sample sizes and the correction applied by the 4/5ts rule can be understood through the following example. Consider a position with the following contingency table (Table 5):

	Selected @ 3.5	Not Selected @ 3.5	Total
Asian	2	3	5
Black	3	2	5
Total	5	5	10

Table 5: *Contingency table for a position with small sample sizes.*

In this case, assuming Asian being the focal group:

$$SR_{focal} = 2/5 = 0.40 \text{ and } SR_{comp} = 3/5 = 0.60$$

As a result,

$$IR = 0.4/0.6 = 0.667$$

While only one more Black applicant was chosen, the impact ratio in this case falls below the threshold of 0.8 implying that the difference of just 1 selected applicant is practically significant. The flip flop rule can be applied as follows: we observe that Blacks candidates are advantaged in this scenario and perturb the contingency table as follows:

	Selected @ 3.5	Not Selected @ 3.5	Total
Asian	3 (2 + 1)	2 (3 - 1)	5
Black	2 (3 - 1)	3 (2 + 1)	5
Total	5	5	10

Table 6: *Contingency table from Table 5 modified with the flip flop rule.*

The above perturbations simulates the contingency table in the case that 1 more Asian candidate was selected as opposed to a Black candidate. In this case, the impact ratio is:

$$IR = 0.6/0.4 = 1.5$$

The test now indicates that Asians are the advantaged groups, having a higher selection rate unlike before. As the test result has changed dramatically with just a small perturbation in the selection, we consider the result practically insignificant. Another notable behavior of the IR can be observed at extremely low selection rates (<5%). When overall selection rates are low, small differences in selection rates have a much larger impact on the IR than at high selection rates. To understand this point, assume that there are 100 male applicants and 100 female applicants. Of these, 1 male applicant is selected and 2 female applicants are selected. The selection rates for males and females are 1% and 2% respectively and the selection rate with men as the focal group is 0.5 which falls below the threshold of 0.8. Conversely if 4 male applicants and 5 female applicants were selected, the impact ratio is 0.8 which just passes 4/5ths rule. In essence, the same difference of 1 additional selection yields a significant result at low selection rates and an insignificant result at slightly higher selection rates.

While the notion of statistical significance allows statistical significance tests to differentiate statistically significant differences from those that occur simply due to chance, statistical tests tend to be too conservative in flagging statistically significant results when sample sizes are small. In these cases, the test is said to have "low power." As such, one of the assumptions of the Z test is that the large sample assumption holds [12]. Fisher's exact test (FET) is used when the large sample assumption does not hold. In the case of FET, the test assumes that marginal frequencies (i.e., N_{focal}, N_{comp}, NP_T, NF_T and N) are held constant, and the test calculates the "exact" probability of selecting NP_{focal} candidates from the focal group under the null hypothesis. This probability p can be expressed as [REF]:

$$p = \frac{NP_T! * NF_T! * N_{focal}! * N_{comp}!}{NP_{focal}! * NP_{comp}! * NF_{focal}! * NF_{comp}! * N!}$$

As exact tests do not rely on approximating the null distribution, but rather compute the p value directly from the true null distribution, exact tests such as FET are the preferred test when the large sample assumption is not met. At large sample sizes however, calculation of the FET becomes non trivial as the product of large factorials quickly leads to arithmetic overflows.

Overall, each of the three tests described so far – Z test, FET, and IR – has a set of key limitations that prevent practitioners from solely relying on one.

Statistically, Z-Score tends to be less reliable when it comes to small sample sizes. At large sample sizes too, z-score has the tendency to be sensitive to small, practically insignificant differences in selection rates. That leaves us with moderately sized samples where the test becomes more reliable.

Fisher's Exact Test, on the other hand, is effective in both small and moderately sized samples, however as the sample size increases, it gets harder to compute as factorials lead to arithmetic overflow.

Impact ratio is sensitive in small sample sizes or small selection rates. This may happen due to the sampling error that leads to over interpretation of small differences as practically significant. In moderate and large sample sizes, the metric is effective and reliable but may lead to over interpretation of small differences at low selection rates.

Approach 2

Similar to the above approach of simulated selection rate, in this method we have a predetermined threshold computed using the median of the scores present in the dataset of interest. Using this median value as the threshold, we compute the selection ratios for each group within the protected category. The selection ratios are then used to compute the impact ratios using the group with the maximum selection rate as the comparator.

Approach 3

Another approach that we've seen in the literature employed to perform the adverse impact analysis involves taking the ratio of the average scores associated with different groups within a protected category. This approach might be applicable for systems that assign a score to the position-profile pair based on the suitability of the candidate for that role. The idea being that the ratio between the focal-comparator group should be as close to 1 as possible.

	Average Score	Ratio
Group 1	3.032	0.988
Group 2	3.028	1.0

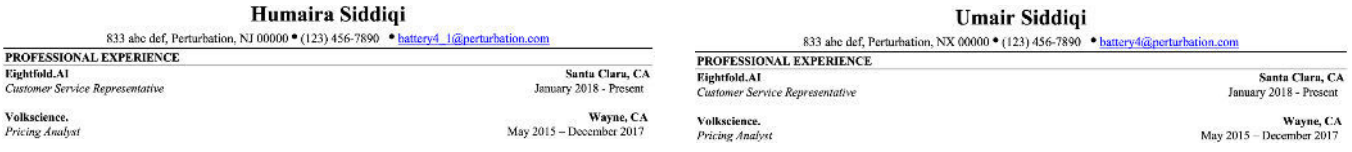
Table 7: *Example Table for Average Score Ratios computation*

For this particular analysis, the ratios closer to 1 are preferable.

Perturbation Testing

Adverse impact analysis evaluates fairness of the match score model for different candidates and provides a global picture of the fairness of the match score model. With perturbation testing, fairness of the match score model is evaluated on a more granular level by evaluating fairness on an individual candidate basis. This is done by using two slightly different resumes to create the candidate data that is used by the match score model. One of the resumes is the original resume of the candidate and the other resume is a slightly modified version of the original resume, in which some text is modified to imply the candidate may belong to a different gender/race subgroup than the gender/race subgroup of the candidate.

An example of a resume modification used in perturbation testing can be seen in the image below, where the text in the resume describing the name of the candidate is replaced to imply candidate may belong to another gender subgroup.



Humaira Siddiqi

833 abc def, Perturbation, NJ 00000 • (123) 456-7890 • battery4_1@perturbation.com

PROFESSIONAL EXPERIENCE

Eightfold.AI — Santa Clara, CA
Customer Service Representative — January 2018 - Present

Volkscience. — Wayne, CA
Pricing Analyst — May 2015 – December 2017

Umair Siddiqi

833 abc def, Perturbation, NX 00000 • (123) 456-7890 • battery4@perturbation.com

PROFESSIONAL EXPERIENCE

Eightfold.AI — Santa Clara, CA
Customer Service Representative — January 2018 - Present

Volkscience. — Wayne, CA
Pricing Analyst — May 2015 – December 2017

Figure 15: *A pair of resumes used in perturbation testing. The resume on the left is the original resume and the resume on the right is depicts the perturbed resume.*

Methodology

To evaluate fairness of the match score model on a more granular level, perturbation tests measure whether the match scores for a position are statistically similar for the candidates using the original resumes and candidates using the modified resumes, given a resume modification, a position, and a list of candidates.

After computing the match scores for a position both for candidates using the original resumes and candidates using the modified resumes, independent Samples T-Test is used to compute t-score and p-value for the null hypothesis that the match scores using the original resumes and match scores using the modified resumes have identical mean (expected) values. The t-score quantifies the difference between the means and the p-value quantifies the probability of obtaining a t-score with an absolute value at least as large as the one observed if the null hypothesis is true. A higher p-value for this null hypothesis suggests that there is not strong evidence for the difference between the means (t-score) of match scores to be statistically significant. Therefore, a low t-score and a high p-value for perturbation tests suggest that the difference in match scores for candidates using the original resumes and candidates using the modified resumes are low and not statistically significant, and so it can not be claimed that bias is introduced with the resume modifications. On the other hand, a high t-score and a low p-value suggest that the difference in match scores for candidates using the original resumes and candidates using the modified resumes are high and statistically significant, and so it can be claimed that bias is introduced with the resume modifications.

Below are the formulas used to compute t-score and p-value for the Independent Samples T-Test used in the perturbation tests:

$$t = \frac{\overline{x_{original}} - \overline{x_{modified}}}{s_p \sqrt{\frac{1}{n_{original}} + \frac{1}{n_{modified}}}} \qquad s_p^2 = \frac{((n_{original} - 1)\, s_{original}^2) + ((n_{modified} - 1)\, s_{modified}^2)}{n_{original} + n_{modified} - 2}$$

Formula: t-score *Formula: pooled variance*

$x_{original}$: match scores using original resumes

$x_{modified}$: match scores using modified resumes

t : t-score of $x_{original}$ and $x_{modified}$

s_p^2 : pooled variance of $x_{original}$ and $x_{modified}$

$\overline{x_{original}}$: mean of $x_{original}$

$n_{original}$: size of $x_{original}$

$s_{original}$: standard deviation of $x_{original}$

$\overline{x_{modified}}$: mean of $x_{modified}$

$n_{modified}$: size of population $x_{modified}$

$s_{modified}$: standard deviation of $x_{modified}$

$$p_{value} = 2 * cdf_{t,d}(-|t|) \qquad d = n_{original} + n_{modified} - 2$$

Formula: p-value *Formula: degrees of freedom*

p_{value} : p-value of t-score of $x_{original}$ and $x_{modified}$

t : t-score of $x_{original}$ and $x_{modified}$

d : degrees of freedom for the t-student distribution

$cdf_{t,d}$: cumulative distribution function of t-student distribution with d degrees of freedom

External Audits

External bias audits provide objective perspective from industry experts on biases within AI systems. At Eightfold, we employ external bias audits to build trust with stakeholders, customers, and the public, as we demonstrate our commitment to transparency and fairness.

Active Monitoring

As part of our beliefs, we set the groundwork for monitoring the performance of our AI models extensively across different aspects including latency and bias.

The latency and accuracy metrics are plugged-in and visualized with the help of a dashboard monitored by the engineers in the team on a regular cadence to ensure the performance is within acceptable parameters. Alarms are set-up on these metrics that are triggered when these values cross a certain predetermined threshold notifying the engineers of the same and prompting immediate response. For all the job positions that users accessed in production, calculate the median of the probabilities of every profile's match to the position for which it was considered. Our standard is for this graph to have a generally flat trend.

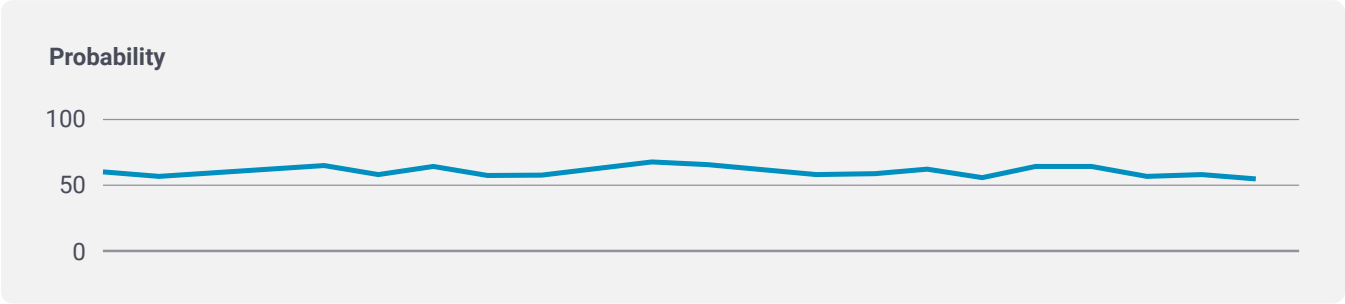


Figure 16: *Probability prediction over time*

We curate continuously growing golden datasets using a human-in-the-loop approach for all our currently active models. Models in production are evaluated on those datasets on a regular basis. In addition to above, we have dashboards that can generate the above discussed bias metrics on varying parameters and granularities to observe and address any anomalies.

References

[1] Wong, A., Hryniowski, A., & Wang, X. Y. (2020). Insights into Fairness through Trust: Multi-scale Trust Quantification for Financial Deep Learning. *arXiv preprint arXiv:2011.01961*.

[2] Golfouse, F. (2020). Partially Aware: Some Challenges Around Uncertainty and Ambiguity in Fairness. *arXiv preprint*

[3] Martinez, N., Bertran, M., Papadaki, A., Rodrigues, M., Sapiro, G. (2020). Pareto Robustness for Fairness Beyond Demographics. *arXiv preprint*

[4] Jones, G. P., Hickey, J. M., Di Stefano, P. G., Dhanjal, C., Stoddart, L. C., & Vasileiou, V. (2020). Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv preprint arXiv:2010.03986*.

[5] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., ... & Chi, E. H. (2020). Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*.

[6] Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

[7] Saha, D., Schumann, C., Mcelfresh, D., Dickerson, J., Mazurek, M., & Tschantz, M. (2020, November). Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning* (pp. 8377-8387). PMLR.

[8] Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2021), 35 pages. DOI:https://doi.org/10.1145/3457607

[10] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259-268).

[11] Adverse Impact Analysis: Understanding Data, Statistics, and Risk. United Kingdom, Taylor & Francis, 2016.

[12] Morris, Scott & Dunleavy, Eric. (2017). Adverse Impact Analysis: Understanding Data, Statistics and Risk.

[13] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems* 29 (2016).

eightfold.ai

Eightfold AI's market-leading Talent Intelligence Platform™ helps organizations retain top performers, upskill and reskill their workforce, recruit talent efficiently, and reach diversity goals. Eightfold's patented deep learning artificial intelligence platform is available in more than 155 countries and 24 languages, enabling cutting-edge enterprises to transform their talent into a competitive advantage.

For more information, visit **www.eightfold.ai**