Potential Gender Biases in the Workplace

Investigating the Hiring Process, Promotions, and Salaries in Black Saber Software

Report prepared for Black Saber Software by The Hive

April 21, 2021

Contents

Executive summary

Guidelines for the executive summary:

- *No more than two pages*
- *Language is appropriate for a non-technical audience*
- *Bullet points are used where appropriate*
- *A small number of key visualizations and/or tables are included*
- *All three research questions are addressed*

Technical report

Introduction

Background

A critical area of concern in today's workplace is gender bias, and Black Saber Software's culture is no exception. It is critical that a company shows that their workplace practices are not only unbiased, but also that they embrace diversity. As a result, we have been hired as an external, third-party consultancy to review Black Saber Software's hiring and promotion processes, as well as their employee salaries in order to determine whether or not the company is biased in their practices. We conclude that there is (in)sufficient evidence to suggest that Black Saber Software (is/is not) biased in their hiring and remuneration processes and that further action to solve this potential problem (is/is not) required. (Our suggestions follow if there is bias)

Research questions

Use bullet points to to describe the research questions you are going to address. Write in full sentences. * Does the hiring algorithm favour a certain gender? Do the humans conducting the interviews have bias towards a certain gender? * Promotion * Salary

Methods

Hiring Process

Black Saber's current new graduate hiring process proceeds in 3 stages, the first two of which are assessed by an artificial intelligence algorithm. It is not until the third and final round that a human becomes involved in the recruitment process. At the beginning of the process, each applicant is assigned a unique ID number that follows them throughout the process to help anonymize the data, as well as keep track of their progress. The applicants specify their gender (male, female, prefer not to say), and the team they wish to apply to (data or software). They then have the option of uploading a cover letter, resume, their GPA (scale from 0.0 to 4.0), extracurricular activities, and work experience. In phase 1, the algorithm rates each applicant's level of extracurriculars and work experience (0, 1, or 2; 2 being the best). These in conjunction with their GPA and the presence of a cover letter and resume are used to decide which applicant moves onto phase 2. Phase 2 consists of a technical task, writing sample, and re-recorded video.

The algorithm uses these materials to rate each applicant's technical skills (0-100), writing skills (0-100), speaking skills (1-10), and leadership presence (1-10). These scores determine who moves

Table 1: Gender Count for Phase 1

Gender	Quantity
Man	291
Woman	311
Prefer not to say	11

onto the final phase, the only one that has human involvement on the company's side. Phase 3 is an interview with 2 interviewers, who each score the applicant on how fit they are for the job on a scale from 0 to 100. We will use this information to investigate whether or not there is gender bias in the rating system both in the algorithm, but also in the interviewers.

Phase 1

To wrangle the data, we added a column that specified whether or not an applicant moved onto the next round (denoted 0 and 1 for no and yes, respectively). This was done by first checking if there were any missing values in the dataset, which there were not. Then, we (fully) joined the phase 1 and phase 2 datasets, and noting which applicants had a value for "technical skills" in a new "next round" column. This is because technical skills were rated in phase 2; thus, if an applicant did not have a technical skill rating, then they did not make it to the next round. This could be done because it was confirmed prior that there were no other missing values in the dataset. We marked a 0 for applicants who did not have a technical skill rating and 1 for those who did. Then, we kept only the columns that were rated in phase 1, along with the new "next round" column.

We then compare the number of applicants that identify as either male, female, or preferred not to specify (Table 1). We see that the applicant pool seems to be fairly even between 311 women and 291 men; with a smaller portion of 11 applicants who preferred not to specify. The even distribution between women and men is a good basis to investigate whether or not there is gender bias in the AI algorithm that determines who moves onto the next round. We investigate this effect with generalized linear models and generalized linear mixed models, with the response variable being a binary response of whether or not the applicant moved onto the next round. Since the algorithm considers each factor (ie. existence of cover letter, resume, level of GPA, work experience, and extracurriculars), these will be the fixed effects in the base generalized linear model.

Next, we created models with a fixed effect for gender. This second model is the same as the base linear model, but with an additional fixed effect of gender. The third model is a generalized linear mixed model, and adds an additional random effect for the team that the applicant

Table 2: Log Likelihood Test for Random Effect of Team in Phase 1

# Df	Log Likelihod	Df	Chi-squared	P-value
8	-17.45086	NA	NA	NA
9	-17.45086	1	1e-07	0.9997396

Table 3: Log Likelihood Test for Fixed Effect of Gender in Phase 1

# Df	Log Likelihod	Df	Chi-squared	P-value
6	-17.90805	NA	NA	NA
8	-17.45086	2	0.9143913	0.6330565

applied for. We want to see if this potential bias exists in one, or both of the teams.

Since the second model is nested in the third, we can first compare these last two models with a log likelihood test to see if there is a significant difference between the model that includes the teams, and the one that does not (Table 2). Since the p-value value is large and close to 1 (0.9997), we conclude that there is not a statistically significant difference that the algorithm is biased towards a certain team. Thus, we can use the simpler model to compare to our base linear model that does not include gender as a fixed effect.

Since the base model is nested within the second linear model with a fixed effect for gender, the log likelihood test can be used again to compare them (Table 3). We find a large p-value again (0.6331) that shows that there is not a significant difference between the two models. This shows that the algorithm is not significantly biased towards gender in the first round, otherwise the random intercepts between these models would be different and the p-value value would be very small (<0.05).

Phase 2

We wrangled the data in phase 2 similarly to how we did in phase 1. We fully joined the phase 2 and phase 3 datasets, and added a column denoting which applicants had an interviewer rating, which is how we determined who moved onto the third round. We denoted "moved forward to the next round" a 1, and "did not move forward" as 0. Again, if the applicant did not have an interviewer rating, it meant that they did not move forward in the process since we confirmed in the beginning that there were no missing values prior to the wrangling process.

We can see that in Table 4, the split between men and women is still fairly even, and there is a similar ratio of people in each gender category removed by the algorithm. This makes sense from

Table 4: Gender Count for Phase 2

Gender	Quantity
Man	145
Woman	152
Prefer not to say	3

Table 5: Log Likelihood Test for Random Effect of Team in Phase 2

# Df	Log Likelihod	Df	Chi-squared	P-value
7	-34.62372	NA	NA	NA
8	-34.43333	1	0.3807755	0.5371885

our last statement that the algorithm as not biased in phase 1.

Then, we want to investigate whether or not there exists gender bias in this phase. Since the algorithm necessarily considers technical skills, writing skills, leadership presence, and speaking skills, these will be the factors in our base generalized linear model.

Next, we created a second and third model with a fixed effect for gender. The second model is the baseline model, with an additional fixed effect for gender. The third model is the same as the second model, but with a random effect for the team each applicant applied for. First, let's compare the second and third model to see if the algorithm is more biased for one team than the other (Table 5). We test this with a log-likelihood test, since the second model is nested within the third. We see that the p-value is 0.5372, which is insignificant at the 5% level, signifying that there is not a significant difference between the model with the random effect for team and the one without it. Therefore, we can move forward with our comparison using the similar model without the random effect.

Next, we compare this simpler model with a random effect for gender with its nested generalized linear model without this effect using a log-likelihood test (Table 6). We see that the p-value is 0.4099, meaning it is insignificant at the 5% level. Thus, we can see that there is not a statistically significant difference between the model accounting for gender and the one that doesn't. Thus, we can say that there is insufficient evidence to suggest that the algorithm is biased in Phase 2 of the hiring process.

Moreover, if we look at the 95% confidence interval for the coefficients in the model with a fixed

Table 6: Log Likelihood Test for Fixed Effect of Gender in Phase 2

# Df	Log Likelihod	Df	Chi-squared	P-value
5	-35.51555	NA	NA	NA
7	-34.62372	2	1.783654	0.4099062

Table 7: 95% Confidence Interval for GLM with Gender Effect

	Estimate	2.5%	97.5%
Baseline	-20.7734073	-29.3292613	-14.4336940
Technical Skills	0.0810649	0.0456063	0.1266256
Writing Skills	0.0922199	0.0499704	0.1446250
Leadership Presence	0.8959324	0.5447564	1.3598746
Speaking Skills	0.7155576	0.4156908	1.0830039
Woman	-0.5665788	-2.0554497	0.8225913
Prefer not to Say	-16.1919849	NA	243.4255353

effect for gender (Table 7), we can see that 0 lies within the interval for women, and that the estimate for those that prefer not to specify is negative, and the upper bound for the estimate is positive, so we know that 0 also lies within this interval. Therefore, we can say that there is insufficient evidence that suggests gender is associated with moving onto the interview round from phase 2.

Phase 3 and Final Hires

In phase 3, we wrangled it similarly to Phases 1 and 2 and created a new binary column, but instead of a new column indicating if the applicant moved onto the next round, it indicated whether or not the person was hired. First, we had to create two intermediate datasets. Let's call them intermediate dataset 1 and 2 respectively. By right joining the phase 2 and phase 3 datasets respectively, we were able to see the variables critical to our analysis (such as gender and team applied for) from phase 2 on only the applicants that made it to phase 3. Then, we remove the columns we don't need to analyse phase 3: the scores the algorithm gave each applicant in phase 2. Thus, we get a dataset with each applicant's ID, gender, team, and both interviewer ratings - intermediate dataset 1.

Then, by fully joining intermediate dataset 1 with the IDs of the final hires, we were able to get a dataset with all the variables from intermediate dataset 1 but for only the applicants that got hired. Then, by adding a column of all 1's, indicating that these applicants were hired, we get intermediate dataset 2. By fully joining the two intermediate datasets, and changing the missing values in the hired column (for those that were not hired) to 0's, we get a dataset that can be used for analysis. For the purpose of this analysis specifically, however, we know that both the interview scores are taken into consideration when hiring an applicant. We noticed that the 10 applicants with the highest average score were hired, so we consolidated the two scores into one column for the average of the two scores to run our analysis.

We see that there was only 1 woman hired for each team (Table 8). From this we can say two things: (1) The number of women hired is fairly even, so we don't need to create and test a

Table 8: Gender Breakdown of Final Hires

	Man	Woman
Not Hired	7	5
Hired	8	2

Table 9: Log Likelihood Test for Fixed Effect of Gender in Final Hires

# Df	Log Likelihod	Df	Chi-squared	P-value
2	-1.386294	NA	NA	NA
3	0.000000	1	2.772589	0.095891

model with a random effect for team; especially since the sample size is so small, it will be hard to make good statistical inferences even with the model. (2) Even though more women applied and made it through the unbiased algorithm twice, only 2 out of the 20 final hired applicants were women. We should investigate this and see if the actual people conducting interviewers are significantly biased or not.

For our analysis, we used two generalized linear models - one with a fixed effect for gender, and one without. We use a log-likelihood test to test whether there is a difference in models when we account for gender (Table 9). We see that the p-value is 0.096, which is significant at only the 10% level. Since we are considering significance only at a 5% level, we can say that this test did not show significant evidence for gender bias in the interviewers; however, this result is still important to note. The significance at a 10% level should not be ignored, especially since the phases in the hiring process done by the AI algorithm did not show any significant results. This may suggest that the people involved in the interview process may have a slight bias in gender, even though we technically did not find significant results at the 5% level.

Promotions

Salary

at first glance the salaries per seniority level per gender is generally even, though women seem to be less than men… let's investigate if this is significant or not

Discussion

In this section you will summarize your findings across all the research questions and discuss the strengths and limitations of your work. It doesn't have to be long, but keep in mind that often people will just skim the intro and the discussion of a document like this, so make sure it is *useful as a semi-standalone section (doesn't have to be completely standalone like the executive summary).*

In Black Saber's hiring process, we investigate whether the algorithm that grades each applicant is biased towards a certain gender. By running a log-likelihood comparison of two nested generalized linear models (one with a fixed effect for gender, one without) on each of the two AI-rated phases. We found that in both phases, there was insufficient evidence to show that the algorithm was biased towards any gender. Moreover, we considered whether or not the algorithm was more biased towards a certain gender within each team; however, we did not find a significant difference when we added a random effect for team in either phase.
The final, human-involved interview round of Black Saber's hiring process may be something of concern. By running a log-likelihood comparison of two nested generalized linear models (one with a fixed effect for gender, one without) on the average interview rating given by Black Saber employees, we find a slight bias towards males. Despite a larger number of women than men that initially applied to Black Saber's new grad program, and more women moving forward through the phases ranked by the algorithm, only 1 woman on each team was hired. We found bias at the 10% significance level, which we recommend further investigating - we will discuss the data required for this in our limitations section.

Strengths and limitations

something about systemic barriers women face in the professional world also small sample size in last round where we found some bias so it might not be that improtant We recognize we may not have all the right answers - there may be additional insights that are also helpful. We are not declaring that our ideas are the best and should be necessarily followed - these are merely our honest suggestions from the analyses we have run.

10% signficance in interviews- we recommend some HR stuff. . . . . we should see what exactly they're rating them on , and whos interviewing the people to properly investigate whether or not theres bias in the hiring process.

Consultant information

Consultant profiles

Yian Wang. Yian is a junior data analyst at The Hive. She specializes in reproducible visualization and making actionable insights. Yian earned her Bachelor of Science, double majoring in Statistics and Economics, and minoring in Mathematics from the University of Toronto in 2021.

Claire Hsiung. Claire is a junior financial analyst at The Hive. She specializes in interpretable visualizations and Big Data. Claire earned her Bachelor of Science, double majoring in Statistics and Economics from the University of Toronto in 2021.

Code of ethical conduct

Not only is The Hive passionate about making actionable insights, but we also practice ethical statistics. Our main values lie in, but are not limited to:

- Confidentiality of client information and respecting their rights
- Using appropriate methods and interpreting them correctly and completely
- Reporting results impartially even if they may pose harm to the parties involved, so as to encourage action against our insights so as to not fall into trouble in the future
- Only using methods we have sufficient knowledge in to use in order to prevent any misinterpretation and summary of the data
- Only using data that is provided to us directly by Black Saber Software, and not to scrape other data that we do not have permission to access
- Declaring our relationship with the clients (ie. financial or other interests) to maintain transparency regarding the influence it may have on the outcomes