Explore whether there exists potential bias in renumeration and hiring processes in Black Saber Software

Focusing on three aspects: salary, promotion and pipeline hiring process including three phases.

Report prepared for Black Saber Software by Lin xin

2021-04-21

Contents

Executive summary

Background & Aim:

There has raised concerns about potential bias in renumeration and hiring processes in Black Saber Software company. However, the Officer does not think potential bias matters and would like to invite consultants in Lin xin company to conduct data analysis so that he could report results to the board. Consultants in Lin xin companny will conduct objective analysis on current employees' data about salary plus promotion and hiring data for new grad program provided by Black Saber Software. Overall, this study is aimed to investigate whether there exists potential bias in the salary distribution, promotion and hiring process. Specifically, this study firstly explores the association that gender, productivity, employees' role may have with salary and the size of significant association. This study secondly explores the association that gender, productivity, salary and whether leadership is appropriate for current level may have with promotion. This study thridly explores which significant factors contribute to the final decision in phase1, phase2 and phase3 of hiring process.

Key findings:

- Current employees' role has positive association with salary. Referring to estimated salary for nine roles in table1, after ranking the nine roles by their corresponding received salary on average from highest to lowest, it would be: Vice president, Director, Manager, Senior III, Senior II, Senior I, Junior II, Junior I and Entry-level.

Table 1: Estimated salary amount for nine roles

Role	Received salary amount($)
Director	121592.3302
Entry-level	30784.674
Junior I	36265.5784
Junior II	38644.5145
Manager	71305.0234
Senior I	44390.0161
Senior II	49777.6211
Senior III	55281.2272
Vice president	151572.7082

- Gender has significant association with salary. Holding the employees' current role same, women tend to have salary $2243.7209 less than men in any of one financial quarter on average. There is a bias in salary distribution in Black Saber Software considering the unequal salary distribution between men and women.
- In Black Saber Software company, gender,salary and whether employees leadership exceeds expectations in terms of their current level have significant and positive association with whether being promoted. Productivity has negative association with whether being promoted.
- The probability for women to be promoted is around 0.00297 times the probability for men to be promoted. There is a bias in the promotion process considering the gender issue.
- In terms of phase1 of hiring process, GPAs, the evaluation of applicants' extracurriculars and work experience have positive significant association with probability of passing phase1. There is estimated 17.9% increase in probability of passing phase1 of hiring process for each 0.1 additional increase in applicants' GPAs.
- In terms of phase2 of hiring process, all related skills including technical skills, writing skills and speaking skills and leadership presence have significant and positive association with probability of applicants passing phase2.
- In terms of phase3 of hiring process, the rating of job fit by the first interviewer and the rating of job fit by the second interviewer have significant positive association with probability of being finally hired by Black Saber Software company.
- There is no potential bias in the three phases of pipeline hiring process due to gender issue.

Limitations:

- In term of analysis toward phase1 of hiring process, since consultants in Lin xin company cannot know detailed evaluation steps about work experience and thus cannot know whether bias exists when Black Saber Software evaluating applicants' work experience by company reputation and other related information(e.g. Black Saber Software may value work experience in the company with larger proportion of men higher). The consultants cannot detect potential bias in the evaluation of work experience in the phase1 of hiring process due to limited access to internal information in Black Saber Software.
- This study does not consider the relationship that interactions of two or three factors may have with salary, promotion or hiring process respectively. For example, employees with high speaking skills may create stronger association between leadership and whether passing phase2 of hiring process.

Technical report

Introduction

Recently there has raised some concerns about potential bias in the hiring process, promotion standard and salary distribution in Black Saber Software company. However, the Officer Gideon Blake does not think the significant existence of bias. He would like to invite the consultants in Lin xin company to conduct the analysis on the current employees' data about salary plus promotion and hiring data for their new grad program so that he could report the results to the board. Overall, this study is aimed to investigate which factors are significantly associated with salary amount, promotion and hiring process respectively and if there is a bias in the hiring process, promotion standard and salary distribution in Black Saber Software.

In the following, I am going to detect whether productivity, gender and employees' current role are associated with salary amount. If the significant association exist, to what extent, above specified factors are associated with salary amount. Furthermore, I am going to explore the significant relationship that gender, productivity, salary and whether employees' leadership is appropriate for their current level may have with promotion respectively. Last but not least, I am going to explore which factors would be significantly associated with final decision in phase1, phase2 and phase3 of hiring process respectively and their size of association. Finally, combining with above analysis in three sections, I will conclude if there is a potential bias in the salary, promotion and hiring process in Black Saber Software company and the extent of each bias.

Research questions

1. Are productivity, gender and employees' current position associated with salary amount respectively? If the significant association exists, to what extent the above specified factors are associated with salary amount?
2. Are gender, salary, productivity and whether employees' leadership is appropriate for current level significantly associated with promotion respectively and to what extent?
3. What factors contribute significantly to the final decision(pass/fail) in the phase1, phase 2 and phase3 of pipeline hiring process respectively?

Informative title for section addressing the first research question:

Explore the relationship that gender, productivity and current employees' role have with the salary respectively and their size if association significantly exists.

Methods:

The data used to address this research question is the current employee's data provided by Black Saber Software. It contains related information about 6906 current employees, including their ids, current role, gender, productivity, team, salary in different financial quarters and so on. In terms of current role, the company provides nine roles for current employees, which are Vice president, Director, Manager, Senior III, Senior II, Senior I, Junior II, Junior I and Entry-level ranking from highest to lowest. In the data, there are multiple records of salary for the same employee in the different financial quarters. To make preparations for drawing the box plot and fitting related model, I firstly do certain data manipulation and transform salary with character type into salary with numeric type. After making this preparation, I can regard salary as continuous variable in general, and this is a key factor to decide which appropriate model should be used latter.

I am going to state the statistical method used to address the research question in general. To check whether gender, productivity and current employees' role are significantly associated with salary, I am going to fit an appropriate model(eg.linear mixed model) and conduct hypothesis/ significance testing for gender, productivity and employees' role respectively. Furthermore, if the association does exist, to estimate the size of this significant association, I will further obtain the corresponding estimate by checking the summary of fitted model. In the process of choosing optimal model, I also apply likelihood ratio test to compare nested models to finally choose the optimal model used for further analysis. Through conducting likelihood ratio test, if the p-value is significant, then we prefer the relatively more complex model. Otherwise, we prefer the simpler model. In the following, I am going to elaborate details about application of each statistical method.

To decide which type of model is used to address this question, I am going to decide it relying on both model assumption and context of the data. As aforementioned, after data manipulation, the salary can be regarded as continuous variable. Besides, there are multiple records for each person in different financial quarter and each person's salary in different financial quarter are related. That is to say, the data are correlated and it violates the independence assumption. Hence, considering the correlated data and continuous response "salary", I am going to apply a linear mixed model to conduct further analysis and address the research question.

Based on research question, I am interested in the relationship that productivity and employees' role have with my response variable salary respectively. Hence, I fit the productivity, gender and

employees' role as fixed effects in my model. As aforementioned, there are multiple records of salary for the same employee in different financial quarters. Hence, I fit the employees' ids as random effects in my model. Considering other possible random effects, I notice that basic salary (base pay) for employees in different teams may be different and basic salary for employees in the same team might be correlated. So I intend to check whether adding random intercept for team would improve my current model and explain data better. According to results from likelihood ratio test, adding random intercept for team does explain better. Thus, we prefer the relatively complex model including the random intercept for team. I also notice that salary amount across all employees are correlated in the same financial quarter under the impact of profits gained by company in that financial quarter. So I intend to check whether adding random intercept for financial quarter would improve my current model. Through conducting likelihood ratio test, I conclude that adding random intercept for financial quarter does not explain data better. Thus, we prefer the simpler model without the random intercept for financial quarter.

Based on above analysis, I finally model salary with linear mixed model, fitting gender, productivity and employees' current role as fixed effects and fitting employees' ids and team as random effects.

After fitting above specified model, to conclusively decide which factor is significantly associated with salary, I intend to conduct hypothesis testing relying on 95% confidence intervals. If 95% confidence interval for corresponding factor of interest does not contain hypothesized value (in this case, it should be zero), then the factor would be statistically significant and we have evidence to reject the corresponding null hypothesis that there is no relationship between factor of interest and salary. Otherwise, the corresponding factor of interest is not significant and we have no evidence to reject the null hypothesis. If the significant association does exist, then I can estimate the extent of association by checking estimate of corresponding regression coefficient.

Results:

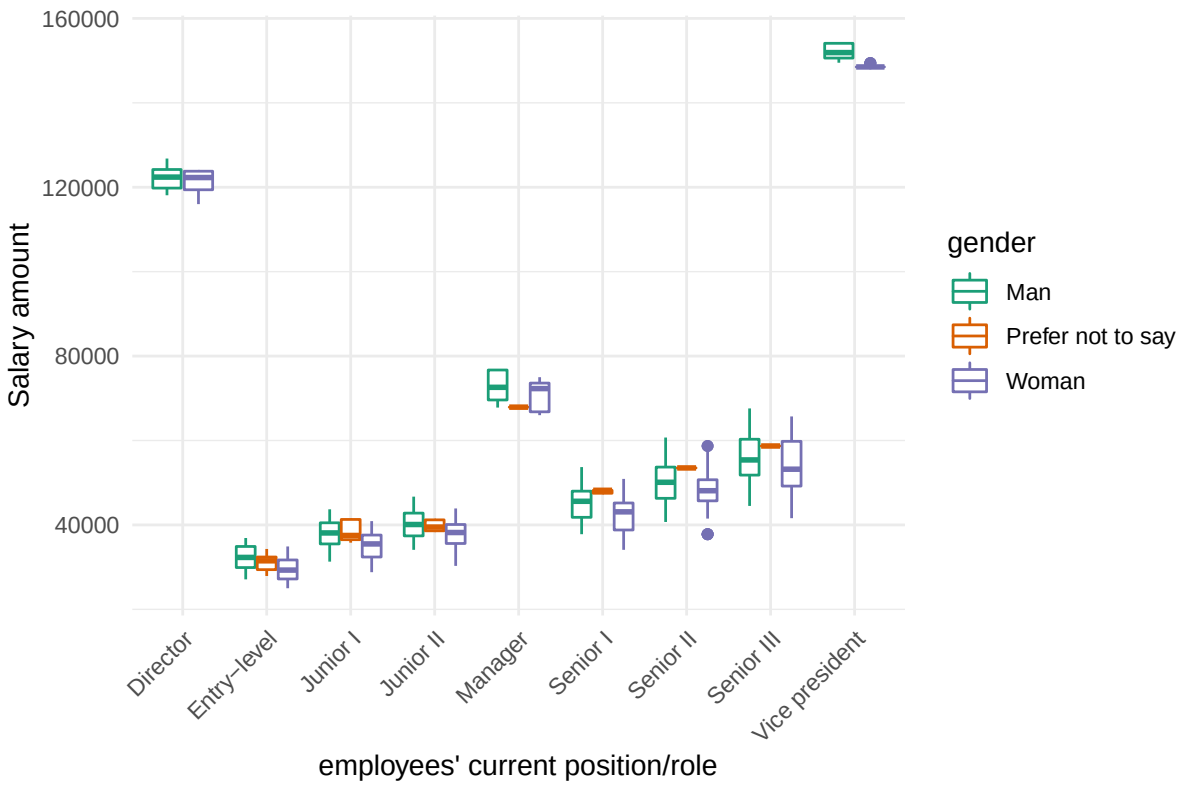

Figure1:Salary amount by employees' role and gender

We can rely on data visualization to potentially check the possible association between factors of interest and salary. Seen from Figure1: Salary amount by employees' role and gender, in general, the salary differs by employees' current role. The employees as a role of vice president tend to have highest salary and employees as a role of entry-level tend to have lowest salary. Overall, I detect the possible relationship between employees' current role and salary.

Furthermore, for each same employee's role, the salary differs by gender. Potentially mean salary and median salary for men are higher than women of same role. Overall, I detect the possible relationship between salary and gender of same employees' role potentially.

Table 2: Association between factors of interest and salary

Factor of interest	Estimate	95% Confidence interval
Women	-2243.7209	(-2787.0602, -1706.9797)
productivity	-1.2302	(-3.4224, 0.9603)
Entry-level	-90807.6562	(-91099.1063, -90515.6891)
Junior I	-85326.7518	(-85605.8365, -85047.2748)

Factor of interest	Estimate	95% Confidence interval
Junior II	-82947.8157	(-83217.3935, -82677.9170)
Manager	-50287.3068	(-50519.9336, -50054.6236)
Senior I	-77202.3141	(-77464.6227, -76939.8290)
Senior II	-71814.7091	(-72069.6670, -71559.5946)
Senior III	-66311.1030	(-66560.6610, -66061.4005)
Vice president	29980.3780	(29608.2630, 30352.4219)

Before conducting hypothesis/significance testing, let me state the three null hypothesis in terms of research question: 1. There is no relationship between productivity and salary. That is to say, regression coefficient before productivity is zero. 2. There is no relationship between gender and salary. That is to say, regression coefficient before gender is zero. 3. There is no relationship between employees' current role and salary. That is to say, regression coefficient before current employees' role is zero.

Seen from Table2: Association between factors of interest and salary, the 95% confidence interval for productivity contains hypothesized value = 0. Hence, the productivity is not statistically significant. We have no evidence to reject the null hypothesis that there is no relationship between productivity and salary.

The 95% confidence interval for gender to be women does not contain hypothesized value = 0, thus the gender to be women is statistically significant. We have strong evidence to reject the null hypothesis that there is no relationship between gender and salary. It indicates that there is significant difference in salary between women and men of same role. The estimate of gender to be woman is -2243.7209. That explains women would have around $2243.7209 less compared with men on average of same role in Black Saber Software.

The 95% confidence intervals for current employees' role at all levels do not contain hypothesized value = 0. Hence, the current employees' role is statistically significant. We have strong evidence to reject the null hypothesis that there is no relationship between current employee's role and their salary. It indicates that there is significant difference in salary between role Entry-level, Junior I&II, manager,Senior I&II&III and vice president.

The estimate for Entry level is -90807.6562. It indicates that current employees at entry level would have 90807.6562 dollars less than the employees' at the role of director of same sex on average.

The estimate for Junior I is -85326.7518. It indicates that current employees at Junior I would

have 85326.7518 dollars less than the employees' at the role of director of same sex on average.

The estimate for Junior II is -82947.8517. It indicates that current employees at Junior II would have 82947.8517 dollars less than the employees' at the role of director of same sex on average.

The estimate for manager is -50287.3068. It indicates that current employees at manager would have 50287.3068 dollars less than the employees' at the role of director of same sex on average.

The estimate for Senior I is -77202.3141. It indicates that current employees at Senior I would have 77202.3141 dollars less than the employees' at the role of director of same sex on average.

The estimate for Senior II is -71814.7091. It indicates that current employees at Senior II would have 71814.7091 dollars less than the employees' at the role of director of same sex on average.

The estimate for Senior III is -66311.1030 . It indicates that current employees at Senior III would have 66311.1030 dollars less than the employees' at the role of director of same sex on average.

The estimate for vice president is 29980.3780. It indicates that current employees at vice president would have 29980.3780 dollars more than the employees' at the role of director of same sex on average.

Based on above analysis, I conclude that gender and employees' role are significantly associated with salary whereas productivity is not associated with salary. Women tend to have less salary compared with men of same role in Black Saber Software. The employees as the role of vice president have highest salary and the employees as the role of entry-level have lowest salary.

Informative title for section addressing the second research question:

Explore the relationship that productivity, gender, whether leadership is appropriate for current level and salary may have with the promotion respectively and their size of association if association significantly exists.

Methods:

The data used to address this research question is the current employees' data provided by Black Saber Software. It includes related information about 6906 current employees, including their ids, current role, gender, productivity, whether their leadership is appropriate for current level, salary and so on. In the data, there are multiple records of salary for the same employee in the different financial quarters. In terms of current role, the company provides nine roles for current employees ranking from highest to lowest, which are vice president, director, Manager, Senior III,Senior II,Senior I,Junior II,Junior I, entry-level. If the employee changes from relatively lower role to relatively higher role, then this employee is regarded as being promoted. To make preparations

for fitting model later, I do certain data manipulation and create a dummy variable to represent whether the current employee is promoted. This dummy variable takes 1 when the employee is promoted and takes 0 when the employee is not promoted. This dummy variable should be the response variable in latter fitted model to address the research question about promotion.

I am going to state the statistical method used to address the research question in general. To check whether gender, productivity, salary and whether employees' leadership is appropriate for current level are associated with promotion, I am going to fit an appropriate model(eg.Generalized linear mixed model) and conduct hypothesis/significance testing for gender, productivity, salary and whether employee's leadership is appropriate for current level respectively. Furthermore, if the association does exist, to estimate the size of this significant association, I will further obtain the corresponding estimate by checking the summary output of fitted model. In the following, I am going to elaborate details about application of each statistical method.

To decide which model is used to address the research question about promotion, I am going to decide it relying on both model assumption and context of the data. As aforementioned, the created dummy variable(1 or 0) that represents whether the employee is promoted would be the response variable. Besides, there are multiple records for each employee in different financial quarters. Considering the bernoulli response and correlated data, I am going to apply a generalized linear mixed model to conduct data analysis.

Based on research question, I am interested in the relationship that gender, productivity,whether leadership is appropriate for current level and salary may have with the promotion respectively. Hence, I fit the gender, salary, productivity, whether leadership is appropriate for current level as fixed effects in the model. As aforementioned, there are multiple records for the same employee in the data. Hence, I fit the employees' ids as random effects.

Based on above analysis, finally I model whether being promoted with generalized linear mixed model, using productivity, gender, salary and whether leadership is appropriate for current level as fixed effects and employees' ids as random effects.

Results:

Table 3: Association between factors of interest and promotion

Factor of interest	Estimate	P-value
Women	-17.3309	0.0000
Productivity	-0.0168	0.0000
leadership: Exceeds expectations	3.6902	0.0638
leadership: Needs improvement	-0.9245	0.4136

Factor of interest	Estimate	P-value
Salary	0.0001	0.0000

Before conducting hypothesis/significance testing, let me state the four null hypothesis in terms of research question: 1. There is no relationship between gender and whether being promoted. 2. There is no relationship between productivity and whether being promoted. 3. There is no relationship between salary and whether being promoted. 4.There is no relationship between whether leadership is appropriate for current level and whether being promoted.

Seen from Table3: Association between factors of interest and promotion, the p-value for predictor gender to be women is 0.0000 after appropriate rounding, which is smaller than the chosen significance level = 0.05. We have evidence to reject the null hypothesis that there is no relationship between gender and whether being promoted. The estimate for predictor to be women is -17.3309. It indicates the the odds of being promoted for women is exp(-17.3309) = 0.00297 times the odds of being promoted for men.

The p-value for predictor productivity is 0.0000 after appropriate rounding, which is smaller than the chosen significance level = 0.05. We have strong evidence to reject the null hypothesis that there is no relationship between productivity and whether being promoted. The estimate for predictor productivity is -0.0168. It indicates there is estimated 1- exp(-0.0168) = 1- 0.9833 = 0.0167 = 1.67% decrease in odds of being promoted for each additional one score increase in productivity.

The p-value for predictor salary is 0.0000 after appropriate rounding, which is smaller than the chosen significance level = 0.05. We have evidence to reject the null hypothesis that there is no relationship between salary and whether being promoted. The estimate for predictor salary is 0.0001. It indicates there is estimated exp(0.0001) - 1 = 0.01% increase in the odds of being promoted for each one additonal dollar increase in salary.

The p-value for the employees' leadership exceeding expectations than current level is 0.0638, which is smaller then significance level = 0.1. We have evidence to reject the null hypothesis that there is no relationship between employees' leadership exceeding expectations than current level and whether being promoted. The estimate for this predictor is 3.6902. It indicates that the odds of being promoted for employees whose leadership exceeds expectations is exp(3.6902) = 40.0528 times the odds of being promoted for employees whose leadership is appropriate for current level.

Informative title for section addressing the third research question:

Explore which significant factors contribute to the final decision in phase1, phase2 and phase3 of pipeline hiring process respectively.

Conduct modeling analysis on phase1:

Methods:

The data used to address which significant factors contribute to the final decision in phase1 of hiring process is the phase1 of hiring data for the new grad program provided by Black Saber Software. The phase1 of hiring data includes related information about 613 new applicants, including their id, gpa, work experience, extracurricular activities, gender, the team they are applying for and whether they provide CV as well as cover letter. To make preparations for fitting an appropriate model and address the question later, I do certain data manipulation and create a dummy variable that represents whether the applicant passes phase1. It takes 1 if the applicant passes phase1 and it takes 0 if the applicant does not passes phase1. This dummy variable is intended to be a response variable in my fitted model later. Since the applicants' id in the data in each record is unique and thus each record is independent, we do not consider any possible random effects.

I am going to state the statistical method used to address the research question for phase1 in general.To explore which factors contribute significantly to the final decision in the phase1 of hiring process, I am going to fit an appropriate model (eg.Generalized linear model/Logistic model) and conduct hypothesis/significance testing for related factors such as gpa shown in phase1 of hiring data. Furthermore, if the association does exist, to estimate the size of this significant association, I further obtain the corresponding estimate by checking summary of fitted model.In the following, I am going to elaborate details about application of each statistical method.

To decide which model is used to address the question, I am going to decide it relying on both model assumption and context of the data. As aforementioned, the dummy variable that represents whether applicants pass phase1 is considered as a response variable. Furthermore, there is no correlated data and the assumption of independence satisfies. Hence, consider the bernoulli response and satisfied independence asssumption, I am going to apply one category of generalized linear model called logistic model to conduct further analysis.

Based on research question for phase1, I initially intend to fit all related factors as predictors in this logistic model, including the team applied for, gender, gpa, work experience, extracurricular activities, whether to provide CV and whether to provide cover letter. However, I notice that the proportion of applicants providing CV in the phase1 of hiring process is around 90%. Since

there are almost all applicants providing CV in this phase1, the relationship between whether to provide CV and final decision in phase1 of hiring process does not arise my interest. I wil exclude this predictor from my model and just fit rest aforementioned factors in my final model.

Based on above analysis, I finally model whether passing the phase1 of hiring process with logistic model, using the team applied for, gpa, work experience, extracurricular activities, whether to provide cover letter as variables.

After fitting above specified model, to conclusively decide which factor is significantly associated with whether passing the phase1, I intend to conduct hypothesis testing relying on p-values. If p-value for corresponding factor of interest is smaller than the chosen significance level = 0.05, then the factor would be statistically significant and we have evidence to reject the corresponding null hypothesis that there is no relationship between factor of interest and whether passing phase1. Otherwise, the corresponding factor of interest is not significant and we have no evidence to reject the null hypothesis. If the significant association does exist, then I can estimate the size of association by checking estimate of corresponding regression coefficient.

Results:

Table 4: Association between factors of interest and final decision in phase1

Factor of interest	Estimate	p-value
Team applied for	-0.1734	0.5978
Cover letter	23.249	0.9806
GPA	1.6501	0.0000
Extracurriculars	1.4632	0.0001
work experience	2.2141	0.0000
women	-0.2616	0.4241

Seen from Table4: Accosiation between factors of interest and final decision in phase1, the p-value for the predictor team that applicants apply for is 0.5978, which is larger than the chosen significance level = 0.05. The predictor team that applicants apply for is not significant. I have no evidence to reject the null hypothesis that there is no difference in final decision in phase1 of hiring process between applying for data and applying for software.

The p-value for the predictor cover letter is 0.9806, which is larger than the chosen significance level = 0.05. The predictor cover letter is not significant. I have no evidence to reject the hypothesis that there is no difference in final decision in phase1 between providing cover letter

and not providing cover letter.

The p-value for the predictor gpa is 0.0000 after appropriate rounding , which is smaller than the chosen significance level = 0.05. The predictor gpa is statistically significant. I have strong evidence to reject the null hypothesis that there is no relationship between gpa and final result decision in phase1 of hiring process. The estimate for predictor gpa is 1.6501 . It indicates that there is an estimated exp(0.1*1.6501) -1 = 17.9% increase in the odds of passing the phase1 of hiring process for each additional 0.1 increase in GPA.

The p-value for the predictor Extracurriculars is 0.0001, which is smaller than the chosen significance level = 0.05. The predictor extrecurricular activities is significant. I have evidence to reject the null hypothesis that there is no relationship between the relevance or skills of building extracurriculars and final decision in phase1. The estimate for extracurricular is 1.4632. It explains there is 1.4632 increase in the log odds of passing phase1 for each one score increase in the evaluation of extracurriculars on average after controlling other variables.

The p-value for the predictor work experience is 0.0000 after appropriate rounding , which is smaller than the chosen significance level = 0.05. The predictor work experience is statistically significant. I have strong evidence to reject the null hypothesis that there is no relationship between the evaluation of work experience and final decision in phase1 of hiring process. The estimate for predictor working experience is 2.2141. It explains there is 2.2141 increase in the log odds of passing phase1 for each one score increase in the evaluation of work experience on average after controlling other variables.

The p-value for the predictor gender to be women is 0.1164, which is larger than the chosen signficance level = 0.05. The predictor gender to be women is not signficant. I have no evidence to reject the null hypothesis that there is no difference in final decision in phase1 between women and men after controlling other variables.

Overall, I conclude that extracurriculars, GPA and work experience contribute significantly to the final decision in phase1 of hiring process. However, gender, whether to provide cover letter and the team applied for do not have association with final decision in phase1 of hiring process.

Conduct modeling analysis on phase2:

Methods:

The data used to address which significant factors contribute to the final decision in phase2 of hiring process is the phase2 of hiring data for new grad program provided by Black Saber Software. The phase2 of hiring data contains related information about 300 applicants who pass phase1, including the evaluation of their techinical skills, writing skills, leadership presence and speaking skills. To make preparations for fitting an appropriate model and address the question later, I do certain data manipulation and create a dummy variable that represents whether

applicants pass phase2.It takes 1 if applicants pass phase2 and takes 0 if applicants don't pass phase2. Since the applicants' id in the data is unique and thus each record is independent, I do not consider random effects.

I am going to state the statistical method used to address the research question for phase2 in general. To explore which factors contribute significantly to the final decision in phase2 of hiring process, I am going to fit an appropriate model (e.g.Logistic model) and conduct hypothesis/significance testing for related factors. Furthermore, if the association does exist, to estimate the size of this association, I will obtain estimate by checking summary of fitted model. In the following , I am going to elaborate details about application of above specified statistical methods.

To decide which model is used to address the question, I am going to decide it relying on both model assumption and context of the data. The dummy variable that represents whether applicants pass phase2 and enter into phase3 should be considered as responser variable. As aforementioned, the data is independent. Considering bernoulli repsonse and independent data, I am going to apply logistic model to conduct further analysis.

Based on research question for phase2, I am going to fit all related factors as fixed effects and no random effects. Overall, I finally model whether passing phase2 with logistic model, using techinical skills, speaking skills, writing skills and leadership presence as variables.

After fitting above specified model, to conclusively decide which factors contribute significantly to the final decision in phase2, I conduct hypothesis testing relying on p-values. If p-value for factor of interest is smaller than 5%, then the factor is significant and I have evidence to reject the null hypothesis that there is no relationship between factor of interest and whether passing phase2. Otherwise, I have no evidence to reject the null hypothesis that there is no relationship between factor of inerest and whether passing phase2. If the association is significant, then I will check the corresponding estimate to detect the size of this significant association.

Results:

Table 5: Association between factors of interest and final decision in phase2

Factor of interest	Estimate	P-value
Techinical skills	0.0777	0.0001
Writing skills	0.0877	0.0001
Leadership presence	0.9282	0.0000
Speaking skills	0.6998	0.0000

Seen from Table5: Association between factors of interest and final decision in phase2, the p-value of predictor technical skills is 0.0001, which is smaller than the chosen significance level = 0.05. The predictor technical skills is statistically significant. I have evidence to reject the null hypothesis that there is no relationship between level of mastery of technical skills and final decision phase2 of hiring process. The estimate for predictor technical skills is 0.0777. It means that there is estimated exp(0.0777)-1 = 8.08% increase in the odds of passing phase2 for each additional one score increase in the evaluation of technical skills.

The p-value of predictor writing skills is 0.0001, which is smaller than the chosen significance level = 0.05. The predictor writing skills is statistically significant. I have evidence to reject the null hypothesis that there is no relationship between level of mastery of writing skills and final decision in phase2. The estimate for predictor writing skills is 0.0877. It means that there is estimated exp(0.0877) - 1 = 9.17% increase in the odds of passing phase2 for each additional one score increase in the evaluation of writing skills.

The p-value of predictor leadership presence is 0.0000 after appropriate rounding, which is smaller than the chosen significance level = 0.05. The predictor leadership presence is statistically significant. I have evidence to reject the null hypothesis that there is no relationship between extent of leadership presence and final decision in phase2. The estimate for predictor leadership presence is 0.9282. There is 0.9282 increase in the log odds of passing phase2 for one additional score increase in the evaluation of leadership presence on average.

The p-value of predictor speaking skills is 0.0000 after appropriate rounding, which is smaller than the chosen significance level = 0.05. The predictor speaking skills is statistically significant. I have evidence to reject the null hypothesis that there is no relationship between speaking skills and final decision in phase2 of hiring process. The estimate for predictor speaking skills is 0.6998. There is 0.6998 increase in the log odds of passing phase2 for one addtional score increase in the evaluation of speaking skills.

Based on above analysis, I conclude that technical skills, writing skills, leadership presence and speaking skills are all significantly associated with whether passing phase2 of hiring process.

Conduct visualizing analysis on phase3:

Methods:

The data used to address which significant factors contribute to the final decision in phase3 is the phase3 of hiring data provided by Black Saber Software. The phase3 of hiring data includes related information about 22 applicants who pass phase2, including their rating or job fit by the first and second interviewer. To make preparatons for drawing plots of each rating and final hiring decision, I create a dummy variable that represents whether applicants are finally hired. It takes “hired” if applicants are hired and takes “not hired” if applicants are not hired.

To explore whether final hiring deicison is associated with ratings given by the first and ratings given by the second interviewer respectively, I am going to draw the box plot of final hiring decision by the rating of job fit by the first interviewer and draw the box plot of final hiring decision by the rating of job fit by the second interviewer respectively.

Results:

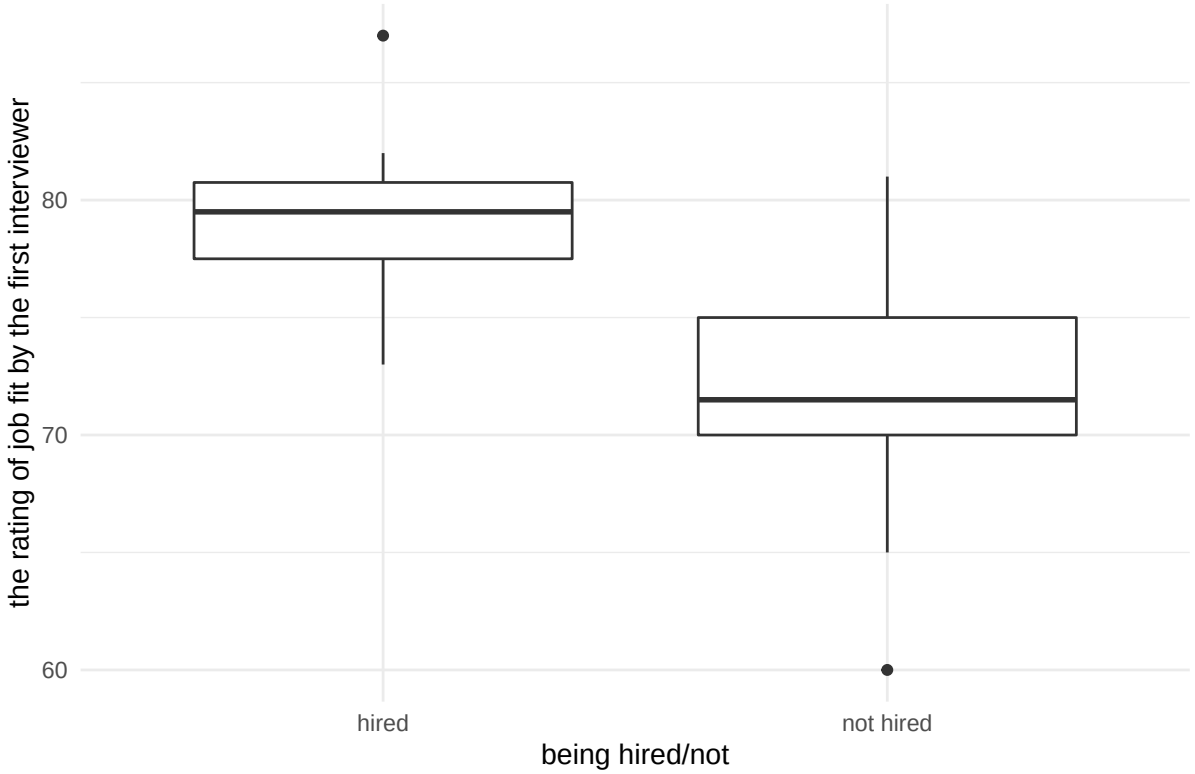

Figure2: Whether being hired by the rating of job fit by the first interviewer

Seen from figure2: Whether being hired by the rating of job fit by the first interviewer, compared with hired group and not hired group, hired group tend to have higher rating of job fit by the first interviewer than not hired group. Hence, I detect there exists the positive relationship between the rating of job fit by the first interviewer and whether being finally hired.

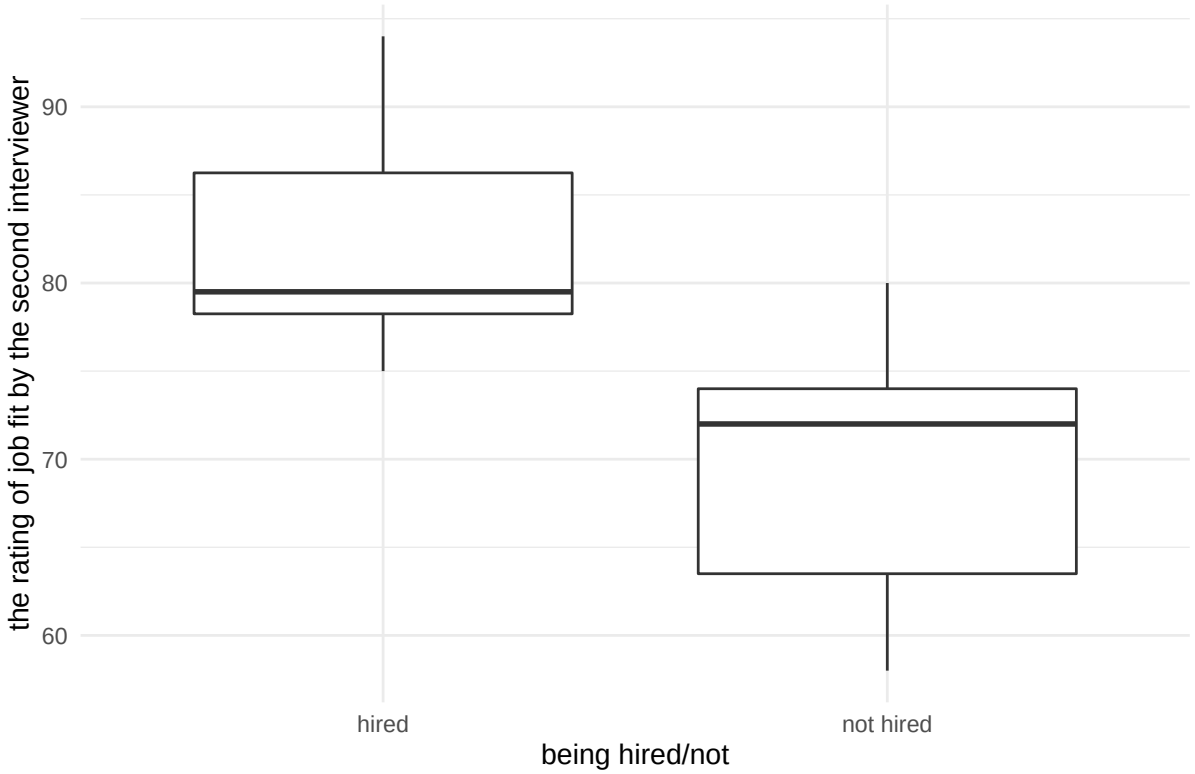

Figure3: Whether being hired by the rating of job fit by the second interviewer

Seen from figure3: Whether being hired by the rating of job fit by the second interviewer, compared with hired group and not hired group, hired group tend to have higher rating of job fit by the second interviewer than not hired group. Hence, I detect there exists the positive relationship between the rating of job fit by the second interviewer and whether being finally hired.

Discussion

To sum up, to investigate potential possible bias in the salary distribution, I model salary with the linear mixed model, fitting productivity, gender as well as current employees' role as fixed effects and employees' ids as well as team as random effects. According to results from hypothesis/significance testing, I get significant factors including gender and current employees' role since their 95% confidence intervals do not contain hypothesized value zero. Women tend to have $2243.7209 less salary than men. Vice president, director, manager, Senior III, Senior II, Senior I, Junior II, Junior I and Entry-level tend to have salary ranking from highest to lowest. Hence, considering the unequal salary distribution between women and men, I conclude there is potential bias in salary distribution in Black Saber Software.

To investigate possible bias in the promotion, I model whether being promoted with the generalized linear mixed model, fitting productivity, gender, salary as well as whether their leadership is appropriate for current level as fixed effects and fitting employees' ids as random effects. According to results from hypothesis/significance testing, I get significant factors including gender, salary, productivity and whether employees' leadership exceeds expectations since their p-values are smaller than significance level. Women tend to be harder to be promoted than men. Employees with higher salary tend to be easier to be promoted whereas employees with higher productivity tend to be harder to be promoted. Last not but least, employees whose leadership exceeds expectation than their current level have higher probability to be promoted. Hence, considering the unequal salary distribution between women and men, I conclude there is a potential bias in promotion in Black Saber Software.

In terms of possible bias in the three phases of pipeline hiring process, to investigate the possible bias in the phase1, I model whether passing phase1 with logistic model, using cover letter, the team applied for, GPAs, gender, extracurriculars and work experience as variables. According to results from hypothesis/significance testing, I get significant factors including GPA, relevance or skills of building extracurriculars and work experience since their corresponding p-values are smaller than 5%. GPA, relevance or skills of building extracurriculars and work experience are all positively associated with the possibility of passing phase1 of pipeline hiring process.

To investigate possible bias in the phase2, I model whether passing phase2 with logistic model, using technical skills, writing skills, speaking skills and leadership presence as variables. According to results from significance testing relying on p-values, I obtain that all above specfied skills and leadership presence are significant. They are have significant positive association with whether passing phase2 of pipeline hiring process.

To investigate possible bias in the phase3, I draw the box plot of whether being hired by the rating of job fit by the first interviewer and the box plot of whether being hired by the rating of job fit by the second interviewer respectively. Through visualizing, applicants who get higher rating of job fit by the first interviewer have higher possibility to be hired and applicants who get higher rating of job fit by the second interviewer also have higher possibility to be hired. Hence, considering no gender issue and any other biased factors affecting the final decision in each phase of hiring process, I conclude there is no potential bias in the hiring process.

Overall, I conclude there are potential biases arisen by gender issue in salary distribution and promotion in Black Saber Software. Only pipeline hiring process is fair for all applicants from new grad program potentially.

Strengths and limitations

My study has completed the effective investigation of potential bias and has certain strengths. It has taken account into all possible factors that would affect the salary, promotion and hiring process. It achieves the comprehensiveness given available data. Furthermore, the related modeling for salary, promotion and hiring process considers the correlated data and takes account into any random effects in the appropriate fitted model to address the violation of independence, further giving relatively accurate and correct estimate. Last but not least, my analysis not only explore which factors are significantly associated with salary, promotion and hiring process respectively, but also study the size of this significant association and further explore the extent of those detectable bias, which helps to evaluate potential bias in salary and promotion more accurately and more convincingly.

However, there are some limitations in my study. As aforementioned, during the investigation of possible bias in the salary distribution, the model used for analysis is the linear mixed model and the response variable is salary. As we know, one assumption for linear mixed model is normal distribution of response variable. However, the normal distribution of salary for current employees in Black Saber Software may not be perfectly guaranteed. Thus, it may not give very precise estimate for corresponding size of association and the extent of bias due to gender issue.

Furthermore, in the phase3 of hiring process, since consultants in Lin xin company cannot know the detailed evaluation steps about work experience and thus cannot know whether bias exists when Black Saber Software evaluating applicants' work experience by company reputation and other related information. For example, Black Saber Software may value companies with larger proportion of men higher and give higher score in the evaluation of applicants' work experience.

Last but not least, during the modeling analysis for each research question, I did not consider any possible interactions and this may affect the accuracy of final results. For example, employees with higher speaking skills may create stronger association between leadership presence and whether passing phase2 of hiring process.

Further study should check the diagnostics of fitted model and adjust the model if necessary to ensure that model follows model assumption. It also needs to take account into any possible interactions effects when fitting the model to analyze the possible bias in the salary, promotion and hiring process. This would improve the accuracy of related results such as estimate of extent of bias due to gender issue. Last but not least, besides the detectable bias due to gender issue, further study could investigate potential bias in salary, promotion and hiring process caused by other factors such as race issue or age issue etc.

Consultant information

Consultant profiles

Xinyi Yao. Xinyi is a senior consultant with Lin xin. She specializes in data mining and big data analytics. Xinyi earned her Bachelor of Science, double major in Statistics and Economics, from the University of Toronto in 2022.

Code of ethical conduct

1. The consultants in Lin xin company stick to their responsibilities and carry out every work conscientiously. They will conduct effective data mining and data analysis with the most professional methods and skills. Furthermore, they will never carry their personal subjective thoughts and feelings when working on the data provided by employers or customers.
2. The consultants in Lin xin company guarantee to use real data provided by employers or customers for analysis and ensure 100% authenticity of data analysis results. They guarantee that they will not change data or forge results.
3. The consultants in Lin xin company will keep the personal information and data related to the interest of employers or customers absolutely confidential. They will protect the interests of employers or customers. Without the permission of the employers or relevant laws, they will never disclose the relevant information of employers or customers to the third party.
4. The consultants in Lin xin company will maintain good communication with employers or customers. The consultants in Lin xin company will timely and accurately inform employers or customers of ethical standards of statistical practice and potential conflicts of interest between employers/customers and Lin xin company.