Navigating Corporate Labyrinth: A 10,000 Company Rating Analysis

Bridget Sheng, Class of 2024
Data Science Major Capstone

Background and Research Questions

Understanding the corporate landscape is important for students to make informed decisions for life after college. Based on the profiles of 10,000 companies, **this capstone project seeks to understand the relevant factors that influence corporate success**, aiming to provide peers with an understanding of what drives company ratings in today's environment.

This study used top 10,000 company or startup profiles from Ambition Box, a website that provides company reviews [1]. By investigating three variables, this project intended to answer the following research questions:

1. How are average salary, total number of interviews conducted by the company, and total number of benefits related to a company's rating?
2. Among average salary, total number of interviews taken, and total number of benefits, what variables are relevant to predict a company's ratings?

Data Processing and Methodology

Overview and Cleaning:

This dataset of company profiles includes character variables (company name, description, etc.) and numeric variables (ratings, total review, average salary, etc.). The dataset was processed in the following steps:

10,000 rows and 10 columns

↓

Selected relevant metrics: ratings, average salary, total interviews taken by the company, total jobs available, total number of benefits

↓

Total jobs available: dropped due to high missingness
Average Salary/total interviews taken/total benefits: dropped small number of missing values

↓

9,708 rows, 4 variables: Ratings;Salary;Interviews;Benefits

Method:

- Based on initial visualization analysis, the original data may not meet linearity assumption. Hence, log transformations were performed on all four variables (Figure 1).
- Simple and first-order linear regression models were used to analyze the dataset.
- 9,708 rows of cleansed data was split: 80% was used for training; 20% was reserved for testing on the best fit model.

Data Modeling: Simple Linear Regression

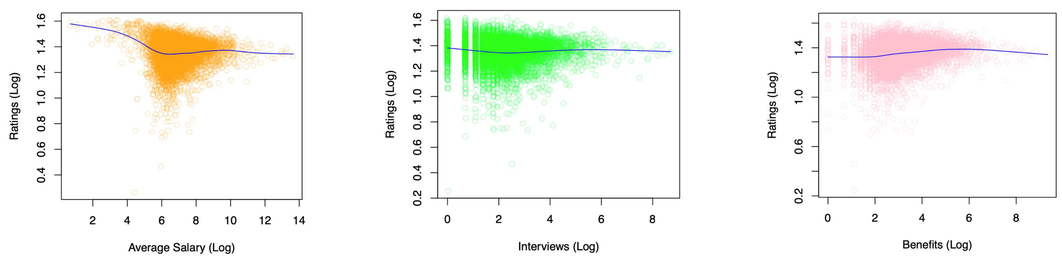

Figure 1. Relationship between each predictor (average salary, total interviews taken, total benefits) and company rating on a log scale

- Simple linear regression models for each variable were run to identify their associations with company ratings.
- All three models generated statistically significant results.
- The model using *Benefits Log* as a predictor had positive coefficient and highest adjusted R^2 of 0.0267 among the three models.
- *Benefits Log* was chosen as the base predictor variable for larger first-order models for better fits.

Model	Intercept	Coefficient	P-Value	R^2	Adjusted R^2	BIC
Salary Log	1.383	-0.00443	6.38e-05	0.00206	0.00193	-13094
Interviews Log	1.358	-0.00249	0.0133	0.000789	0.000661	-13084
Benefits Log	1.305	0.0154	<2e-16	0.0269	0.0267	-13289

Data Modeling: First-Order Regression

•*Interviews Log* and *Salary Log* were fitted with *Benefits Log* for first-order models respectively. A full model was fitted with all three variables.
•Adjusted R^2 increased for first-order models and BIC decreased, indicating better fits with more predictors.

Model	Coefficient	R^2	Adjusted R^2	BIC
Benefits Log +Salary Log	(Intercept) 1.489 Benefits Log 0.0447 Salary Log -0.0400	0.0965	0.0963	-13857
Benefits Log +Interviews Log	(Intercept) 1.300 Benefits Log 0.0363 Interviews Log -0.0273	0.0721	0.0719	-13650
Benefits Log +Interviews Log +Salary Log	(Intercept) 1.450 Benefits Log 0.0478 Salary Log -0.0321 Interviews Log -0.0116	0.102	0.102	-13895

*p-values were statistically significant for all the first-order models

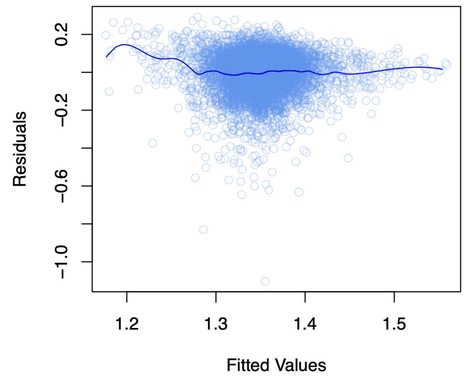

Figure 2. Residuals versus Fitted values for the best fit model. All the variables were log-transformed.

Results

- The model with all three predictors appears to be the best fit one with the highest adjusted R^2 of 0.102 and lowest BIC of -13895.
- The best fit model shows that average salary and total number of interviews conducted have a slightly negative association with company ratings, while total number of benefits has a positive relationship.
- When the best fit model was applied to the testing data, it yielded a mean absolute error (MAE) of 0.277 and a mean squared error (MSE) of 0.129.

Discussions

Best Fit Model: The model with all three predictor variables presents the best fit, indicating that a company's rating is a multifaceted issue.

Unexpected Negative Coefficients Under Best Fit Model: Higher salaries and more interviews correlate with lower company ratings, possibly due to demanding job environments and competitive selection processes. Conversely, a greater number of benefits improves ratings, which highlights the value placed on workplace welfare.

Testing Data: The model's predictions closely match actual ratings with low errors. It performed well on test data and is potentially suitable for real-world applications.

Limitations: 1) The analysis focuses on three variables. Incorporating additional ones, such as aspects of the company that are highly rated for, could offer for a more comprehensive understanding for the corporate landscape. **2)** Despite using log transformation, the residual and fitted value pattern indicates potential model assumption violations (Figure 2). Further data transformation or alternative modeling practices might be needed.

Data Ethics: 1) Privacy and transparency concerns may emerge concerning methods used to gather and share data on company information. **2)** Subjective rating procedures could introduce biases into the datasets, potentially causing skewed outcomes and discrimination.

Conclusion

In conclusion, multiple factors are relevant for company ratings: benefits increase ratings, while higher salaries and more interviews conducted decrease ratings. Therefore, peers should thoroughly assess companies before considering future job opportunities.

Reference:
[1] Kaggle Dataset: https://www.kaggle.com/datasets/vedantkhapekar/top-10000-companies-dataset/data